Numerical Evidence Extraction from Scientific PDFs: A Systematic Review and Empirical Evaluations

Athira Rajendran[1] **Jonne Kotta**[1] **Tariq Yousef**[2]

[1]Tallinn University of Technology [2]University of Southern Denmark

Correspondence: athira@taltech.ee

Abstract

Scientific PDFs remain a major source of quantitative evidence, but numerical values in tables and figures are difficult to extract in a reusable and traceable form. This paper examines this problem through a systematic review of 66 tools and an empirical evaluation of representative systems for table extraction, chart digitisation, full-document processing, and LLM-supported interpretation. The evaluation uses a curated scientific dataset, including marine-domain PDFs relevant to Cumulative Impact Assessment. Results show that table extraction is more mature than chart extraction, but both are sensitive to layout complexity and lose important context such as units, labels, captions, and provenance. Full-document tools preserve structure more effectively, but they do not provide sufficiently reliable numeric extraction on their own. We conclude that numerical evidence extraction is not only an extraction task, but also a context-preservation challenge. Reliable workflows therefore require hybrid pipelines that combine document parsing, specialised extraction, structured normalisation, and LLM-based validation.

1 Introduction

Scientific evidence synthesis increasingly depends on collecting quantitative results from research publications. However, in many domains, important numerical evidence such as effect sizes, concentrations, rates, and thresholds remains embedded in tables and figures within PDF documents. These representations are designed for human interpretation rather than machine processing, making reliable automatic extraction difficult. This challenge involves two interconnected problems. The first is document layout understanding, which concerns detecting tables, figures, and their structural organization within a document. The second is semantic reconstruction, which involves recovering the relationships between extracted numerical values and their associated meaning, including variables, units, captions, and contextual metadata.

This paper presents a systematic review and empirical evaluation of AI-based and software tools for extracting numerical evidence from tables, figures, and charts in scientific PDFs. Beyond extraction accuracy, we focus on whether extracted values remain reusable, traceable, and linked to sufficient contextual information. We review 66 tools spanning table extraction, chart digitisation, document-level processing, OCR pipelines, and LLM-supported interpretation. In addition, we empirically evaluate representative systems on a curated scientific dataset containing tables, figures, and full research PDFs.

The work is motivated by the practical requirements of *Cumulative Impact Assessment* (CIA), where experts manually identify and extract quantitative evidence describing ecological impacts, controls, and propagation across environmental systems. In such workflows, extracted values are only useful if they remain connected to their semantic and methodological context. This includes provenance information, units, variable descriptions, captions, and links to the originating table cells or figure elements. Manual extraction is therefore time-consuming, difficult to scale, and often challenging to reproduce consistently across reviewers.

Recent advances in AI-based document understanding have produced a wide range of tools targeting different stages of the extraction pipeline (Blecher et al., 2023; Livathinos et al., 2025). Some approaches focus on converting entire PDFs into structured machine-readable representations, while others target specific components such as tables or charts. These systems often generate intermediate representations containing text, layout, figures, and positional metadata, enabling partial reconstruction of document context. More recently, large language models (LLMs) and vision-language models have enabled combining extraction with higher-level se-

mantic interpretation, making them increasingly attractive for evidence synthesis workflows (Ji et al., 2023).

Despite this progress, practitioners still face a practical selection problem: which tools reliably extract numerical evidence from real scientific PDFs, and which preserve sufficient context for downstream evidence synthesis tasks? To address this question, we combine a systematic review with an empirical evaluation of representative systems. We organize existing approaches using a unified task-oriented framework, evaluate their extraction capabilities and context preservation behavior, and analyze their suitability for integration into practical workflows.

The main contributions of this paper are threefold: (1) a systematic review and categorization of 66 tools for numerical evidence extraction from scientific PDFs, (2) an empirical evaluation of representative systems across tables, charts, and full-document processing, focusing on both extraction accuracy and context preservation, and (3) the identification of context preservation and provenance linkage as central bottlenecks motivating hybrid extraction pipelines.

2 Problem Definition and Task Decomposition

2.1 Numerical Data Extraction

In this work, numerical data extraction refers to recovering quantitative values from scientific documents together with the contextual information required for their interpretation. The objective is not only to extract numbers, but to recover usable and traceable evidence. This includes the numeric value itself as well as its associated semantic context, such as variable names, units, captions, measurement conditions, and provenance. Figure 3 in the Appendix illustrates this requirement with an example in which extracted impact values from a scientific figure must be linked to species, response variables, units, sample sizes, source figure, and caption information.

This task differs from standard text extraction or Optical Character Recognition (OCR). OCR converts scanned documents into machine-readable text, while PDF text extraction methods recover textual content from digital documents. However, these approaches do not reconstruct the structure of tables or figures, nor do they reliably link numerical values to their semantic meaning. This distinction is particularly important in applications such as CIA, where extracted evidence must remain interpretable, reusable, and traceable to its original document context.

Numerical data extraction from scientific PDFs is therefore not a single task, but a pipeline of interdependent components. **Table extraction** focuses on detecting tables and reconstructing their internal structure, including rows, columns, headers, and cell boundaries. **Chart and figure extraction** aims to recover data from visual representations such as bar charts, line plots, and pie charts by interpreting graphical elements, axes, and legends. **Context reconstruction** links extracted values to their semantic meaning, including variables, units, captions, and relationships between document elements. Finally, **document-level integration** combines outputs from multiple components into a unified representation that preserves structure, reading order, and provenance across tables, figures, and surrounding text. Since these stages are often handled by different tools or models, integration becomes a central challenge in practical extraction pipelines.

2.2 Core Challenges

Despite recent progress, numerical data extraction from scientific PDFs remains difficult. Scientific publications often contain complex layouts, nested tables, multi-panel figures, and heterogeneous visual structures that are difficult to detect and reconstruct reliably. In OCR-based systems, recognition errors can propagate through downstream stages and reduce extraction accuracy.

Even when numerical values are recovered correctly, contextual information such as units, captions, variable relationships, and provenance links is frequently lost. As a result, extracted outputs may become difficult to interpret, validate, or reuse in downstream evidence synthesis workflows. In addition, differences in output representations across tools complicate integration and interoperability between extraction systems.

3 Taxonomy of Approaches

To better understand the field, approaches are grouped according to their role within the numerical extraction pipeline rather than treated as isolated tools. The taxonomy highlights differences in representational assumptions, extraction capabilities, context preservation, and integration behavior

across workflows. Figure 1 summarises the relationship between these categories.

3.1 Table Extraction Approaches

Table extraction methods can be broadly divided into three categories based on the level of structural reconstruction they provide.

Detection-based approaches treat tables primarily as visual regions within a document and focus on identifying their boundaries using object detection or segmentation techniques (Schreiber et al., 2017). While effective for locating tables, these methods usually do not recover internal structure or cell-level relationships.

Structure recognition approaches extend this by reconstructing rows, columns, headers, and cell boundaries, often using segmentation, graph representations, or transformer-based architectures. Models such as TableNet and related systems have shown strong performance for structural recovery (Paliwal et al., 2019). However, most approaches at this stage still produce intermediate representations that require additional processing before numerical evidence can be interpreted reliably.

End-to-end extraction systems move closer to usable outputs by directly generating structured representations such as HTML, JSON, or relational tables from document inputs (Nassar et al., 2022). These approaches improve downstream usability, but remain sensitive to layout complexity, particularly in tables containing merged cells, hierarchical headers, or irregular formatting. Across all categories, preserving semantic links between values, headers, units, and surrounding context remains a persistent limitation.

3.2 Chart and Figure Extraction

Compared to table extraction, chart and figure extraction remains more fragmented due to the diversity of visual encodings used in scientific figures. Existing approaches can be grouped into three main categories.

Rule-based and geometric methods rely on explicit assumptions about chart structure. They detect graphical elements such as bars, lines, or pie segments and use geometric relationships, often combined with OCR, to infer numeric values (Al-Zaidy and Giles, 2015). While effective for simple visualisations, these methods are highly sensitive to layout variation, image quality, and chart design.

Deep learning-based approaches model charts as visual understanding tasks. Systems such as ChartOCR detect components including axes, ticks, legends, and data points before reconstructing the underlying numeric representation (Luo et al., 2021). These approaches are more flexible than rule-based systems, but performance remains strongly dependent on chart complexity and clarity.

Vision-language models treat charts as multimodal inputs and translate them into textual or tabular representations. Models such as DePlot enable downstream reasoning by converting plots into structured intermediate forms (Liu et al., 2023). However, these systems may partially infer values rather than strictly recover them from the visual representation. Across all categories, maintaining reliable links between values, axes, legends, captions, and surrounding document context remains difficult, particularly for complex scientific figures.

3.3 Document-Level Processing Systems

In addition to specialised extraction methods, document-level systems aim to convert complete PDFs into structured machine-readable representations suitable for downstream processing. These pipelines typically combine layout analysis, OCR, and document parsing to identify text blocks, tables, figures, captions, and reading order within a unified document representation. Systems such as Nougat focus on end-to-end document conversion into structured textual formats (Blecher et al., 2023), while OCR-oriented frameworks including PaddleOCR and olmOCR2 emphasise text recognition and layout recovery for scanned or image-based documents (Cui et al., 2025; Poznanski et al., 2025). Their primary strength lies in preserving document structure and contextual metadata rather than directly extracting numerical evidence. As a result, document-level systems are best understood as upstream structuring components within larger extraction pipelines. Although they provide important contextual information such as figure-caption relationships and reading order, they typically require specialised table or chart extraction modules for precise numerical recovery.

3.4 LLM-Based Approaches

LLMs introduce a different capability within the extraction pipeline by focusing on semantic interpretation and context reconstruction rather than direct layout analysis or visual parsing. In practice, they are most effective as post-processing components operating on outputs generated by specialised extraction systems. LLMs can support linking values

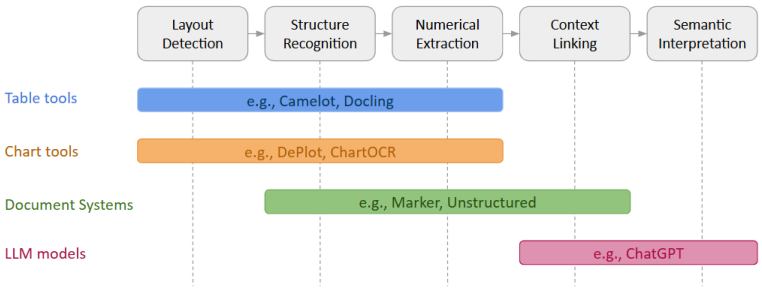

Figure 1: Capability coverage across numerical extraction approaches. Table and chart tools primarily support local extraction, document-level systems preserve layout and structural context, and LLM-based systems support interpretation. The gaps between extraction, context linking, and semantic interpretation motivate hybrid workflows for traceable numerical evidence extraction.

to variables, recovering missing context, summarising extracted evidence, and integrating information across tables, figures, and surrounding text. This makes them useful for higher-level reasoning tasks within evidence synthesis workflows. However, their generative nature also introduces important risks. Instead of strictly extracting grounded values, LLMs may infer, approximate, or hallucinate information, leading to incorrect numeric outputs or unsupported contextual associations (Ji et al., 2023). For this reason, LLMs are most reliable when used as interpretation and validation layers within hybrid workflows rather than as standalone numerical extraction systems.

4 Systematic Review Methodology

4.1 Research Questions

This systematic review was guided by the following research questions **RQ1**. What approaches and tools exist for extracting numerical values from tables, figures, and charts in scientific PDFs, and what types of structures and output formats do they support? **RQ2**. How do these approaches differ in their technical characteristics, including reliance on OCR, evaluation settings, and common failure modes? **RQ3**. How do representative tools perform in practice when applied to real-world scientific documents, particularly in terms of numerical extraction accuracy? **RQ4**. To what extent do existing approaches preserve contextual information required for downstream evidence synthesis, and how suitable are they for integration into practical workflows such as CIA?

4.2 Data Sources and Search Strategy

Relevant studies were identified through structured searches in Scopus and Web of Science, which provide broad coverage of research in document analysis, computer vision, and information extraction. The search targeted studies on table extraction, chart and figure extraction, and related document understanding methods for scientific PDFs. Searches were limited to English-language publications from 2010 onwards. Full search strings are provided in Appendix S1. The review process followed established systematic review guidance, including the PRISMA framework (Page et al., 2021) and recommendations by Kitchenham and Charters(Kitchenham and Charters, 2007).

4.3 Eligibility Criteria

Studies were included if they described implemented tools or systems for extracting tables, figures, or numerical values from scientific PDFs or document images. Peer-reviewed papers, preprints, and official technical reports were considered. To remain relevant to the study objective, included systems had to provide numerical recovery either directly or through structured table or chart reconstruction, and they had to report some form of evaluation evidence. Detailed inclusion and exclusion criteria are listed in Appendix S1.

4.4 Screening and Study Selection

Screening was conducted in two stages. First, titles and abstracts were reviewed to remove clearly irrelevant records. Second, full texts were assessed against the predefined eligibility criteria. Duplicate records were removed before screening. To improve coverage, reference lists of selected papers were also checked manually. In addition, openly available implementations and actively maintained public tools were reviewed, even when they were not introduced through a standalone academic pa-

per. This made it possible to include both research prototypes and practically relevant systems.

5 Comparative Analysis

5.1 Overview

The reviewed systems differ substantially in scope, extraction strategy, and output representation. Some tools focus on specialised tasks such as table extraction or chart digitisation, while others support broader document-level processing. To compare these approaches systematically, we analyse them across four dimensions: extraction capabilities, OCR dependence, context preservation, and practical deployment.

5.2 Comparison Dimensions

The comparison focuses on dimensions that directly affect practical evidence extraction workflows: what each system extracts, how much structure and context it preserves, whether it depends on OCR, and how easily it can be integrated into reproducible pipelines.

Extraction Capabilities. The reviewed systems differ substantially in the type and granularity of information they recover. Table-oriented approaches primarily focus on structure reconstruction and cell-level extraction, while chart extraction systems attempt to recover values from graphical encodings such as axes, legends, and visual marks. Compared to tables, chart extraction remains less robust due to visual variability and the absence of explicit structural organisation.

Document-level systems provide broader parsing capabilities by identifying text blocks, tables, figures, captions, and reading order within unified structured representations. Their primary strength lies in preserving layout and contextual metadata rather than directly extracting precise numerical evidence. As a result, they are most effective as upstream components within larger extraction pipelines.

OCR Dependence. OCR dependence represents another major distinction between systems. PDF-based approaches operating on embedded document text generally perform well on clean born-digital PDFs but fail on scanned or image-based inputs. In contrast, many chart and document-processing systems rely on OCR to recover textual elements such as axis labels, legends, and captions. In these workflows, OCR errors propagate into downstream extraction and semantic interpretation. OCR-oriented frameworks therefore function mainly as supporting infrastructure rather than complete numerical extraction solutions.

Context Preservation. Output representations vary considerably across systems. Many table extraction tools export flat formats such as CSV or Excel, which are easy to process but frequently lose document-level context. More advanced systems generate structured formats such as HTML, JSON, or Markdown that preserve relationships between tables, figures, captions, and surrounding text.

Despite these improvements, context preservation remains limited across most approaches. Numerical values are often detached from units, captions, variable descriptions, and source locations. While some systems attempt semantic linking between extracted content and document structure, reliable preservation of provenance and contextual meaning remains unresolved across the field.

Practical Deployment. Practical usability varies considerably across the reviewed systems. Some tools provide open-source implementations, structured outputs, and local deployment support, making them easier to integrate into reproducible workflows. In contrast, many research-oriented systems remain difficult to deploy or are not publicly maintained despite strong reported benchmark performance. For applications such as CIA, maintainability, interoperability, and reproducibility are often as important as extraction accuracy itself.

6 Empirical Evaluation

6.1 Experimental Setup

The evaluation dataset was manually constructed from scientific publications and includes three input categories: (i) standalone tables, (ii) standalone charts and figures, and (iii) full scientific PDF documents. [1] Table extraction was evaluated on 10 tables covering both simple and complex layouts, including merged cells, grouped rows, and multi-level headers. Chart extraction was tested on 6 figures representing different visualisation types, including bar charts, line plots, pie charts, and histograms. Full-document processing was assessed using five born-digital marine science PDFs relevant to evidence synthesis workflows. One document was used for quantitative comparison, while

[1] The benchmark dataset, ground-truth annotations, and evaluation resources are publicly available at https://github.com/idkcheckgooglebro/Numeric_Evidence_Extraction.

the remaining documents supported qualitative validation and error analysis. A representative subset of tools was selected to cover different stages of the extraction pipeline: table extraction (Camelot, Docling), chart extraction (DePlot), document-level processing (PaddleOCR, Unstructured.io, Marker, olmOCR2), and LLM-based interpretation (ChatGPT). Tool selection was based on public availability, practical usability, and relevance to scientific PDF processing.

All outputs were normalised into a unified JSON representation and compared against manually prepared ground truth. Extraction accuracy was defined as the proportion of correctly recovered numerical values. For table and chart extraction systems, results are reported as mean accuracy across multiple samples. For document-level systems, evaluation focused on a representative reference PDF combined with qualitative inspection across additional documents.

In addition to numeric accuracy, the evaluation examined *context preservation*, defined as the ability of a system to maintain links between extracted values and associated contextual information. Manual inspection assessed whether outputs preserved (i) structural relationships, (ii) captions and legends, (iii) units and labels, and (iv) provenance links to source locations within the document. Because the evaluation is intentionally small and practice-oriented, the results should be interpreted as comparative evidence of tool behaviour rather than as a large-scale benchmark.

Table Extraction Results. Table extraction systems showed the most stable performance among the evaluated categories. Docling achieved the highest average accuracy (80%), performing well on structured tables with clear boundaries and regular layouts. However, performance declined for complex structures where units and labels became partially misaligned. Camelot achieved lower average accuracy (56%) and was reliable mainly for simple born-digital tables. Performance degraded substantially for irregular layouts, merged cells, and hierarchical headers.

Chart Extraction Results. Chart extraction showed substantially higher variability. DePlot achieved moderate average accuracy (65%), with performance strongly dependent on chart type and visual complexity. Simple visualisations such as pie charts and basic line plots were handled reasonably well, whereas performance decreased for stacked bar charts, multi-series plots, and visually dense scientific figures.

Full-Document Processing Results. Full-document systems exhibited a different performance profile. Rather than optimising purely for numeric extraction, these tools prioritised document structuring and context preservation. Marker (72%) and Unstructured.io (70%) produced the most balanced outputs, preserving layout organisation, reading order, and relationships between document elements. olmOCR2 (68%) showed strong OCR performance but limited figure-level extraction capability. PaddleOCR (60%) performed reliably for text recovery but required additional downstream processing to generate structured outputs. Across additional validation documents, the overall behaviour remained

6.2 Error Analysis

Several recurring error patterns were observed across the evaluated systems.

Structural alignment errors were common in table extraction systems such as Docling and Camelot. Complex layouts containing grouped rows or hierarchical headers frequently caused incorrect alignment between numeric values and their associated variables, even when values themselves were extracted correctly.

Semantic context loss affected both table and chart extraction workflows. Units, labels, captions, and variable associations were often partially lost or incorrectly linked during extraction. As a result, otherwise correct numeric values became difficult to interpret reliably in downstream evidence synthesis tasks. Figure 2 in illustrates an example where Docling partially reconstructs table structure but misaligns headers and contextual information.

Approximation and reconstruction errors were particularly visible in chart extraction. DePlot occasionally generated approximate rather than exact numeric values, especially in visually complex figures where relationships between chart elements and numeric scales were ambiguous.

OCR propagation errors were also observed in OCR-dependent pipelines. Misrecognised axis labels, captions, or table text propagated into downstream stages, affecting both structural reconstruction and semantic interpretation.

6.3 Context Preservation Evaluation

Manual inspection showed that document-level systems preserved contextual structure more consistently than specialised extractors. Marker, Unstruc-

Taxon	Average abundance (# alga^{-1})			Contrib. to differences (%)	
	2012	2013	2014	2012–2014	2013–2014
T. fluviatilis	22.5	7.2	0.4	47.7	21.7
Hydrobia spp.	7.9	0.3	0.0	12.5	0.9
Amphipoda	6.9	12.9	1.4	13.7	37.6
Idotea spp.	4.4	0.9	0.04	10.4	3.0
M. trossulus	4.1	4.8	4.1	10.0	17.0
Cerastoderma/Parvicardium	1.6	3.0	1.0	4.3	10.1
Jaera spp.	0.04	1.8	0.0	0.1	5.5

```
| Taxon                    |   Average abundance (# alga-1) |   Average abundance (# alga-1) |   Average abundance (# alga-1) | Contrib. to differences (%)
| Contrib. to differences (%)   |
|--------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------
---|-----------------------------|
|                          |                           2012   |                           2013   |                           2014   | 2012-2014
| 2013-2014
| T. fuviatilis            |                           22.5   |                            7.2   |                            0.4   | 47.7
| 21.7
| Hydrobia spp.            |                            7.9   |                            0.3   |                              0   | 12.5
| 0.9
| Amphipoda                |                            6.9   |                           12.9   |                            1.4   | 13.7
| 37.6
| Idotea spp.              |                            4.4   |                            0.9   |                           0.04   | 10.4
| 3.0
| M.trossulus              |                            4.1   |                            4.8   |                            4.1   | 10.0
| 17.0
| Cerastoderma/Parvicardium |                           1.6   |                              3   |                              1   | 4.3
| 10.1
| Jaera spp.               |                           0.04   |                            1.8   |                              0   | 0.1
| 5.5
```

Figure 2: Example of structural reconstruction errors in Docling table extraction. While the system recovers most numerical values correctly, complex headers and grouped rows become partially misaligned, illustrating the difficulty of preserving semantic relationships between values, labels, and table structure.

tured.io, and Docling generally retained document organisation and relationships between content elements, whereas Camelot and DePlot provided weaker links between extracted values and captions, units, legends, or surrounding text. olmOCR2 primarily recovered textual content without deeper semantic relationships.

7 Discussion

Maturity Differences Across Tasks. Table extraction is the most developed area, with multiple tools achieving reliable performance on structured, born-digital inputs. In contrast, chart and figure extraction remains less mature and more fragmented, with performance highly dependent on chart type and visual complexity. This difference reflects the nature of the tasks. Tables provide explicit row–column structure, while charts require reconstruction of values from visual encodings such as axes, scales, and legends. As a result, chart extraction remains more sensitive to layout variation and harder to generalise.

Limits of Current Approaches. Many current approaches remain sensitive to document structure. Performance typically drops for irregular tables, multi-level headers, and complex figure layouts, where alignment between values and labels becomes unclear (Paliwal et al., 2019; Prasad et al., 2020; Nassar et al., 2022). In pipelines that rely on OCR, recognition errors further propagate into downstream extraction, affecting both structure reconstruction and numeric interpretation, especially in figures (Luo et al., 2018; Li et al., 2022c; Cui et al., 2025) . In addition, most tools focus on recovering structure rather than meaning, so extracted values are often not reliably linked to variables, units, or experimental conditions, which limits their usability for evidence synthesis (Zhong et al., 2020; Smock et al., 2022; Livathinos et al., 2025).

Role of LLMs. LLMs may support downstream interpretation and context reconstruction once numerical content has been extracted into a structured representation. However, LLM-based interpretation was not systematically benchmarked in this study, and their generative nature introduces risks of unsupported inference and numerical hallucination(Ji et al., 2023).We therefore consider LLMs as potential supporting components for validation and semantic interpretation rather than primary extraction systems.

Context as the Central Bottleneck. The main limitation across the study is not extraction accuracy, but context preservation. For applications such as CIA, a value is only useful if it remains linked to its meaning, including units, variables, study conditions, and source location. However,

most tools operate at the level of structure or local data recovery without consistently preserving these relationships. Full-document systems such as Docling, Unstructured, and Marker provide better structural context, but still do not fully capture semantic meaning. Table tools often lose links to captions and units, while chart tools rarely connect values to surrounding text. As a result, extracted outputs remain incomplete for decision-grade use. This suggests that numerical extraction is not only a perception problem, but also a representation problem requiring explicit schema alignment and provenance tracking (Clark and Divvala, 2016; Livathinos et al., 2025; Li et al., 2022a).

Evaluation Limitations. The empirical evaluation is intentionally small and practice-oriented, covering 10 tables, 6 figures, and five scientific PDFs selected to represent varied layouts and extraction challenges. It should therefore not be interpreted as a large-scale benchmark or as establishing statistically generalisable rankings between tools. Instead, the evaluation complements the systematic review by illustrating how representative systems behave on realistic scientific material and by identifying recurring failure modes relevant to evidence synthesis. Future evaluation should extend this analysis to larger domain-diverse datasets and assess numerical accuracy, context preservation, and provenance jointly.

Practical Considerations. Tool selection for evidence synthesis depends not only on extraction accuracy but also on availability, maintainability, structured output support, and local deployment. Open and actively maintained systems with interoperable outputs are easier to incorporate into reproducible workflows than research prototypes without accessible implementations. These considerations are particularly important for long-term evidence synthesis applications such as CIA.

8 Toward Hybrid Extraction Pipelines

The complementary capabilities identified in the review and evaluation motivate a modular pipeline for traceable numerical evidence extraction. It support a layered pipeline architecture in which different components handle specialised tasks, and effective workflows combine complementary tools rather than relying on a single system. PDFs are first converted into structured representations using layout-aware tools, which detect text blocks, tables, figures, captions, and their positions while preserving document structure and reading order. Specialised tools are then applied to extract numeric values from tables and figures, with methods selected based on content type. The extracted outputs are normalised into a common structured format (e.g., JSON) to ensure consistency and support comparison and downstream use. Throughout the process, it is important to preserve provenance by maintaining links to source elements such as captions, page numbers, and document locations to ensure traceability. Finally, LLMs are applied after extraction to interpret results, reconstruct context, and support validation. In this setting, they function primarily as a reasoning layer rather than as direct extraction systems. A central requirement for such pipelines is that each extracted value remains linked to its source location, structural context, and semantic interpretation, enabling validation and later auditing.

9 Conclusion

This study combined a systematic review with empirical evaluation to assess approaches for extracting numerical data from tables and figures in scientific PDFs. The results show that table extraction is relatively mature for structured inputs, while chart extraction remains less reliable and sensitive to visual complexity. Full-document systems preserve layout and contextual structure effectively, but do not provide reliable numerical extraction on their own. A central finding is that numerical evidence extraction is not only an extraction problem, but also a context-preservation challenge. Many systems recover values without reliably maintaining links to units, variables, captions, and provenance information. As a result, no single approach currently produces fully traceable, decision-ready outputs. The main practical implication is that hybrid pipelines combining document parsing, specialised extraction, structured normalisation, and LLM-supported validation provide the most reliable strategy for evidence extraction workflows. Future work should focus on schema-aware extraction, end-to-end evaluation benchmarks, and improved methods for provenance tracking and semantic grounding. The findings further suggest that future progress will depend less on isolated extraction accuracy and more on integrated approaches.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Moitinho de Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, and Valerie Balcom. 2023. GPT-4 technical report.

Madhav Agarwal, Ajoy Mondal, and CV Jawahar. 2021. CDEC-Net: composite deformable cascade network for table detection in document images. In *2020 25th international conference on pattern recognition (ICPR)*, pages 9491–9498. IEEE.

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Rabah A Al-Zaidy, Sagnik Ray Choudhury, and C Lee Giles. 2016. Automatic summary generation for scientific data charts. In *AAAI Workshop: Scholarly Big Data*.

Rabah A Al-Zaidy and C Lee Giles. 2015. Automatic extraction of data from bar charts. In *Proceedings of the 8th international conference on knowledge capture*, pages 1–4.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Mathias H Andersson and Marcus C Öhman. 2010. Fish and sessile assemblages associated with wind-turbine constructions in the baltic sea. *Marine and Freshwater Research*, 61(6):642–650.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Rabia Azzi, Sylvie Despres, and Gayo Diallo. 2020. KEFT: knowledge extraction and graph building from statistical data tables. In *International Conference on Computational Collective Intelligence*, pages 701–713. Springer.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Zewen Chi, Heyan Huang, Heng-Da Xu, Houjin Yu, Wanxuan Yin, and Xian-Ling Mao. 2019. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*.

Christopher Clark and Santosh Divvala. 2016. PDFfigures 2.0: mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Zhang Yue, Lv Wenyu, Huang Kui, Zhang Yichao, Zhang Jing, Zhang Jun, Liu Yi, Yu Dianhai, and Ma Yanjun. 2025. PaddleOCR 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Wenjing Dai, Meng Wang, Zhibin Niu, and Jiawan Zhang. 2018. Chart decoder: Generating textual and numeric information from chart images automatically. *Journal of Visual Languages & Computing*, 48:101–109.

Amanda Dash, Melissa Cote, and Alexandra Branzan Albu. 2023. Weathergov+ a table recognition and summarization dataset to bridge the gap between document image analysis and natural language generation. In *Proceedings of the ACM Symposium on Document Engineering 2023*, pages 1–10.

Jianxin Deng, Gang Liu, Rui Tang, Xiusong Wu, and Zheng Yin. 2025. An automatic selective pdf table-extraction method for collecting materials data from literature. *Advances in Engineering Software*, 204:103897.

Aarya Gawhane, Mitij Raul, Shrawani Terde, and Neha Surti. 2025. Automated sales data extraction and forecasting using deep learning. In *2025 International Conference on Electronics, AI and Computing (EAIC)*, pages 1–6. IEEE.

Mengyu Ge, Han Liu, Yilin Liu, and Changgang Zheng. 2019. Data mining techniques for the analysis and prediction in bioenergy yield. In *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pages 214–217. IEEE.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Muhammad Yusuf Hassan, Mayank Singh, and Shivasankaran. 2023. LineEX: data extraction from scientific line charts. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 6213–6221.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TAPas: weakly supervised table parsing via pre-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4320–4333.

Qiyu Hou and Jun Wang. 2025. TABLET: table structure recognition using encoder-only transformers. In *International Conference on Document Analysis and Recognition*, pages 253–278. Springer.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023.

Isaak Kavasidis, Carmelo Pino, Simone Palazzo, Francesco Rundo, Daniela Giordano, P Messina, and Concetto Spampinato. 2019. A saliency-based convolutional neural network for table and chart detection in digitized documents. In *International conference on image analysis and processing*, pages 292–302. Springer.

Pratik Kayal, Mrinal Anand, Harsh Desai, and Mayank Singh. 2023. Tables to latex: structure and content extraction from scientific tables. *International Journal on Document Analysis and Recognition (IJDAR)*, 26(2):121–130.

Sebastian Kempf, Markus Krug, and Frank Puppe. 2023. KIETA: key-insight extraction from scientific tables. *Applied Intelligence*, 53(8):9513–9530.

Iftakhar Ali Khandokar and Priya Deshpande. 2024. Computer vision-based framework for data extraction from heterogeneous financial tables: A comprehensive approach to unlocking financial insights. *IEEE Access*, 13:17706–17723.

Barbara Kitchenham and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering.

Chenxia Li, Ruoyu Guo, Jun Zhou, Mengtao An, Yuning Du, Lingfeng Zhu, Yi Liu, Xiaoguang Hu, and Dianhai Yu. 2022a. PP-structurev2: a stronger document analysis system. *arXiv preprint arXiv:2210.05391*.

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022b. DiT: self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3530–3539.

Shufan Li, Congxi Lu, Linkai Li, and Haoshuai Zhou. 2022c. Chart-RCNN: efficient line chart data extraction from camera images. *arXiv preprint arXiv:2211.14362*.

Yongzhi Li, Pengle Zhang, Meng Sun, Jin Huang, and Ruhan He. 2024. Pre-training transformer with dual-branch context content module for table detection in document images. *Virtual Reality & Intelligent Hardware*, 6(5):408–420.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. DePlot: one-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, Shubham Gupta, Rafael Teixeira de Lima, Valery Weber, Lucas Morin, Ingmar Meijer, Viktor Kuropiatnyk, and Peter Staar. 2025. Docling: An efficient open-source toolkit for ai-driven document conversion. *arXiv preprint arXiv:2501.17887*.

Daipeng Luo, Jing Peng, and Yuhua Fu. 2018. BioTable: a tool to extract semantic structure of table in biology literature. In *Proceedings of the 5th International Conference on Bioinformatics Research and Applications*, pages 29–33.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. ChartOCR: data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925.

Nam Tuan Ly and Atsuhiro Takasu. 2023. An end-to-end multi-task learning model for image-based table recognition. *arXiv preprint arXiv:2303.08648*.

Carin Magnhagen, Kajsa Johansson, and Peter Sigray. 2017. Effects of motorboat noise on foraging behaviour in eurasian perch and roach: a field experiment. *Marine Ecology Progress Series*, 564:115–125.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: a universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 14662–14684.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. PlotQA: reasoning over scientific plots. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, pages 1527–1536.

Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. 2022. TableFormer: table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623.

Danish Nazir, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. 2021. HybridTabNet: towards better table detection in scanned document images. *Applied Sciences*, 11(18):8396.

AB Nugumanova, KS Apayev, YM Baiburin, M Mansurova, and AG Ospan. 2022. QURMA: a

table extraction pipeline for knowledge base population. *Journal of Mathematics, Mechanics and Computer Science*, 114(2).

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, and Elie A Akl. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. 2019. TableNet: deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 128–133. IEEE.

Martha O Perez-Arriaga, Trilce Estrada, and Soraya Abad-Mota. 2016. TAO: system for table detection and extraction from pdf documents. In *FLAIRS*, pages 591–596.

Jorge Poco, Angela Mayhua, and Jeffrey Heer. 2017. Extracting and retargeting color mappings from bitmap images of visualizations. *IEEE transactions on visualization and computer graphics*, 24(1):637–646.

Jake Poznanski, Luca Soldaini, and Kyle Lo. 2025. olmOCR 2: Unit test rewards for document ocr. *arXiv preprint arXiv:2510.19817*.

Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. 2020. CascadeTabNet: an approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 572–573.

Liang Qiao, Zaisheng Li, Zhanzhan Cheng, Peng Zhang, Shiliang Pu, Yi Niu, Wenqi Ren, Wenming Tan, and Fei Wu. 2021. LGPMA:complicated table structure recognition with local and global pyramid mask alignment. In *International conference on document analysis and recognition*, pages 99–114. Springer.

Chunxia Qin, Zhenrong Zhang, Pengfei Hu, Chenyu Liu, Jiefeng Ma, and Jun Du. 2024. SemV3: a fast and robust approach to table separation line detection. *arXiv preprint arXiv:2405.11862*.

Nikitha Rachapudi, Lakshmipathy Ganesh, A Sekar, KS Anand, and R Sakthibalan. 2019. Discovery of structured data using unsupervised spatial clustering and human supervision. *International Journal of Machine Learning and Computing*, 9(5):586–591.

Chinmayee Rane, Seshasayee Mahadevan Subramanya, Devi Sandeep Endluri, Jian Wu, and C Lee Giles. 2021. Chartreader: Automatic parsing of bar-plots. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 318–325. IEEE.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Roya Rastan, Hye-Young Paik, John Shepherd, Seung Hwan Ryu, and Amin Beheshti. 2018. TEXUS: table extraction system for pdf documents. In *Australasian Database Conference*, pages 345–349. Springer.

Pau Riba, Anjan Dutta, Lutz Goldmann, Alicia Fornés, Oriol Ramos, and Josep Lladós. 2019. Table detection in invoice documents by graph neural networks. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 122–127. IEEE.

Pau Riba, Lutz Goldmann, Oriol Ramos Terrades, Diede Rusticus, Alicia Fornés, and Josep Lladós. 2022. Table detection in business document images by message passing networks. *Pattern Recognition*, 127:108641.

Gustav Rosén. 2019. Analysis of tabula: A pdf-table extraction tool.

Rohit Sahoo, Chinmay Kathale, Milind Kubal, and Shaveta Malik. 2020. Auto-table-extract: A system to identify and extract tables from pdf to excel. *Int. J. Sci. Technol. Res*, 9:217.

Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. 2017. DeepDeSRT: deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 1162–1167. IEEE.

Mauricio Vargas Sepúlveda, Thomas J Leeper, Tom Paskhalis, Manuel Aristarán, Jeremy B Merrill, and Mike Tigas. 2024. tabulaPDF: an r package to extract tables from pdf documents. *arXiv preprint arXiv:2409.14524*.

Lei Sheng and Shuai-Shuai Xu. 2024. PDFtable: a unified toolkit for deep learning-based table extraction. *arXiv preprint arXiv:2409.05125*.

Alexey Shigarov, Andrey Altaev, Andrey Mikhailov, Viacheslav Paramonov, and Evgeniy Cherkashin. 2018. TabbyPDF: web-based system for pdf table extraction. In *International Conference on Information and Software Technologies*, pages 257–269. Springer.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Chris Tensmeyer, Vlad I Morariu, Brian Price, Scott Cohen, and Tony Martinez. 2019. Deep splitting and merging for table structure decomposition. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 114–121. IEEE.

Fanghao Tian, Diego Bernal Cobaleda, and Wilmar Martinez. 2022. Automatic data extraction based on semiconductor datasheet for design automation of power converters. In *2022 International Power Electronics Conference (IPEC-Himeji 2022-ECCE Asia)*, pages 922–927. IEEE.

Mengshi Zhang, Daniel Perelman, Vu Le, and Sumit Gulwani. 2021. An integrated approach of deep learning and symbolic analysis for digital pdf table extraction. In *2020 25th international conference on pattern recognition (ICPR)*, pages 4062–4069. IEEE.

Zhenrong Zhang, Jianshu Zhang, Jun Du, and Fengren Wang. 2022. Split, embed and merge: An accurate table structure recognizer. *Pattern Recognition*, 126:108565.

Xinyi Zheng, Douglas Burdick, Lucian Popa, Xu Zhong, and Nancy Xin Ru Wang. 2021. Global table extractor (GTE): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 697–706.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer.

Mengxi Zhou and Rajiv Ramnath. 2022. A structure-focused deep learning approach for table recognition from document images. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 593–601. IEEE.

Arthur Zucker, Younes Belkada, Hanh Vu, and Van Nam Nguyen. 2021. ClusTi: clustering method for table structure recognition in scanned images. *Mobile Networks and Applications*, 26(4):1765–1776.

Table 1: Comparative summary of the reviewed tools across extraction and deployment dimensions.

Group	Tool	Table	Chart	Full-doc	OCR	CSV	JSON/HTML	Context	Deployment
Table	Camelot	✓			No	✓		Weak	✓
	TabtoLaTeX	∼			Req		∼	Weak	✓
	Tabula	✓			No	✓		Weak	✓
	tabulapdf	✓			No	✓		Weak	✓
	TableFormer	∼			No		✓	partial	✓
	TEXUS	✓			No	✓	✓	Weak	?
	TAO	✓			No		✓	Weak	?
	KIETA	∼			No		✓	Strong	✓
	GTE	∼			No		✓	Weak	✓
	TAPAS	∼			No			Weak	✓
	TableNet	∼			Yes			Weak	?
	CascadeTabNet	∼			No			Weak	✓
	MTL-TabNet	✓			No			Partial	✓
	SEM	∼			No			Weak	?
	SPLERGE	∼			No			Weak	?
	TABLET	∼			Req.			Partial	?
	LGPMA	∼			Opt.			Partial	✓
	PdfTable	✓			Cond.	✓	✓	Partial	✓
	PubTables1M	∼			No			Partial	✓
	GraphTSR	∼			No			Weak	✓
	DeepDeSRT	∼			No			Weak	✓
	SEMv3	∼			No			Weak	✓
	TabbyPDF	✓			No	✓	✓	Partial	✓
	TD-DCCAM	∼			No			Weak	✓
	TableExtractor	∼			No			Weak	?
	SAL-CRF-DocDet	∼	∼		No			Weak	✓
	BIOTABLE	✓			No			Partial	✓
	HybridTabNet	∼			No			Weak	?
	EDD	✓			No		✓	Partial	?
	Deng	✓			No	✓	✓	Partial	✓
	CluSTi	∼			Partial		✓	Weak	?
	KEFT	✓			No		✓	Strong	?
	Spatial clustering	✓			No		✓	Strong	?
	GNNF	∼			Req			Weak	?
	Bioenergy-Miner	∼			No	✓		Partial	✓
	CDeC-Net	∼			No			Weak	✓
	Structure CNN	∼			No			Weak	?
	QURMA	✓			No	✓	✓	Weak	?
	MPNN GNN	∼			Req.	✓	✓	Weak	?
	WEATHERGOV+	∼			Req.		✓	Partial	✓
	TableNetOCR	✓			Req.	✓		Partial	✓
	AutoTableExtract	✓			Cond	✓		Weak	✓
	DiT	∼			No	✓		Weak	✓
	CV FinTable	✓			Req.		✓	Partial	?
Chart	ChartOCR		✓		Req.	✓		Weak	∼
	ChartReader		✓		Req.		∼	Partial	?
	UniChart		∼		impl		∼	Partial	✓
	Chart-RCNN		✓		Req.			Weak	✓
	LINEEX		✓		Req.			Partial	✓
	DePlot		✓		No		∼	Partial	✓
	REV		∼		Req.			Weak	✓
	Chart-to-Text		∼		Req.		∼	Weak	✓
	PlotQA		∼		No			Weak	
	SemiDatasheet		✓		Req.			Weak	?
	BarChartSummary		✓		Req.			Partial	?
	Chart Decoder		✓		Req.			Partial	?
	BarChartExtraction		✓		Req.	✓		Weak	?
Full-doc	Docling	∼		✓	Cond.	✓	✓	Strong	✓
	PP-StructureV2	✓		✓	Req.	✓		Partial	✓
	PDFFigures 2.0			✓	No			Strong	✓
	Nougat	∼		✓	No		∼	Weak	✓
	Marker	✓	∼	✓	Req.		✓	Strong	✓
	Unstructured.io	∼	?	✓	Req.	∼	✓	Strong	✓
	PaddleOCR	∼	∼	✓	Req.		∼	Partial	✓
LLM/OCR	olmOCR2	∼		✓	Req.			Weak	✓
	ChatGPT	∼	∼		No			Partial	

✓= supported | ∼= partial/indirect support | blank = not supported | ? = unclear | Req. = required | Cond. = conditional

Table 2: Table 2: References and Resources for the Reviewed Numerical Evidence Extraction Tools

Group	Tool	Tool URL	Reference
Table	Camelot	https://github.com/camelot-dev/camelot	-
	TabtoLaTeX	https://tinyurl.com/ydbw9asp	Kayal, P., Anand, M., Desai, H., Singh, M. (2022). Tables to LaTeX: Structure and Content Extraction from Scientific Tables. arXiv:2210.17246.
	Tabula	https://github.com/tabulapdf/tabula	Rosén, G. (2019). Analysis of Tabula – a PDF-Table extraction tool. Uppsala University
	tabulapdf	-	Vargas Sepúlveda, M., Leeper, T. J., Paskhalis, T., Aristarán, M., Merrill, J. B., Tigas, M. (2024). tabulapdf: An R Package to Extract Tables from PDF Documents. arXiv:2409.14524v1.
	TableFormer	https://github.com/IBM/SynthTabNet	Nassar et al. (2022). TableFormer: Table Structure Understanding with Transformers. arXiv:2203.01017.
	TEXUS	-	Rastan, R., Paik, H.-Y., Shepherd, J., Ryu, S.H., Beheshti, A. (2018). TEXUS: Table Extraction System for PDF Documents. In: ADC 2018, LNCS 10837, pp. 345–349.
	TAO	-	Perez-Arriaga, M. O., Estrada, T., Abad-Mota, S. TAO: System for Table Detection and Extraction from PDF Documents. Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference (FLAIRS), 2016.
	KIETA	https://gitlab2.informatik.uni-wuerzburg.de/kieta/kieta	Kempf, S., Krug, M., Puppe, F. "KIETA: Key-insight extraction from scientific tables." Applied Intelligence (published online 9 Aug 2022; journal issue 2023).
	GTE	-	Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N. X. R. (2021). Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context. WACV 2021.
	TAPAS	https://github.com/google-research/tapas	Herzig, J. et al. TAPAS: Weakly Supervised Table Parsing via Pre-training. arXiv:2004.02349v2, 2020.
	TableNet	-	Paliwal et al., "TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images", ICDAR 2019 (conference).
	CascadeTabNet	https://github.com/DevashishPrasad/CascadeTabNet	Prasad et al., CascadeTabNet: An approach for end-to-end table detection and structure recognition from image-based documents, ICDAR 2020
	MTL-TabNet	https://github.com/namtuanly/MTL-TabNet	Ly, N.T. Takasu, A. "An End-to-End Multi-Task Learning Model for Image-based Table Recognition.
	SEM	-	Zhang et al., "Split, Embed and Merge: An accurate table structure recognizer", Pattern Recognition, 2022
	SPLERGE	-	Tensmeyer et al., "Deep Splitting and Merging for Table Structure Decomposition", ICDAR 2019
	TABLET	-	Paliwal et al., "TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images", ICDAR 2019 (conference).
	LGPMA	https://github.com/hikopensource/DAVAR-Lab-OCR/tree/main/demo/table_recognition/lgpma	Qiao et al., "LGPMA: Complicated Table Structure Recognition with Local and Global Pyramid Mask Alignment", arXiv:2105.06224v3, 2022.
	PdfTable	https://github.com/CycloneBoy/pdf_table	Sheng, L., Xu, S.-S. "PdfTable: A Unified Toolkit for Deep Learning-Based Table Extraction." arXiv:2409.05125v1, 2024.
	PubTables1M	https://github.com/microsoft/table-transformer	Smock, Pesala, Abraham. "PubTables-1M: Towards comprehensive table extraction from unstructured documents." arXiv:2110.00061v3, 2021
	GraphTSR	https://github.com/Academic-Hammer/SciTSR	Chi et al., "Complicated Table Structure Recognition", arXiv:1908.04729v2, 2019

Continued on next page

Table 2 continued from previous page

Group	Tool	Tool URL	Reference
	DeepDeSRT	-	Schreiber et al., "DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images," ICDAR 2017.
	SEMv3	https://github.com/Chunchunwumu/SEMv3	Qin et al., "SEMv3: A Fast and Robust Approach to Table Separation Line Detection," arXiv:2405.11862v1, 2024.
	TabbyPDF	https://github.com/cellsrg/tabbypdf	Shigarov et al., "TabbyPDF: Web-Based System for PDF Table Extraction."
	TD-DCCAM	https://github.com/YongZ-Lee/TD-DCCAM	Li et al., "Pre-training transformer with dual-branch context content module for table detection in document images," Virtual Reality Intelligent Hardware, 2024, 6(5): 408–420
	TableExtractor	–	Zhang et al., "An Integrated Approach of Deep Learning and Symbolic Analysis for Digital PDF Table Extraction," ICPR (Milan), DOI: 10.1109/ICPR48806.2021.9413069.
	SAL-CRF-DocDet	–	Kavasidis, I., Palazzo, S., Spampinato, C., et al. A Saliency-based Convolutional Neural Network for Table and Chart Detection in Digitized Documents. (Submitted to IEEE Trans. on Multimedia / arXiv 2018).
	BIOTABLE	–	Luo, D., Peng, J., Fu, Y. Biotable: A Tool to Extract Semantic Structure of Table in Biology Literature. ACM ICBRA 2018.
	HybridTabNet	–	Nazir, D.; Hashmi, K.A.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M.Z. HybridTabNet: Towards Better Table Detection in Scanned Document Images. Applied Sciences 2021, 11, 8396.
	EDD	–	Zhong, X., ShafieiBavani, E., Jimeno Yepes, A. (2020). Image-Based Table Recognition: Data, Model, and Evaluation. ECCV 2020 (LNCS 12366), pp. 564–580
	Deng	https://github.com/implyDavv/PDF-document-extraction/tree/master/code	Deng, J. et al. (2025). An automatic selective PDF table-extraction method for collecting materials data from literature. Advances in Engineering Software, 204, 103897.
	CluSTi	–	Zucker, A., Belkada, Y., Vu, H., Nguyen, V. N. (2021). ClusTi: Clustering Method for Table Structure Recognition in Scanned Images. Mobile Networks and Applications, 26:1765–1776
	KEFT	–	Azzi, R., Despres, S., Diallo, G. (2020). KEFT: Knowledge Extraction and Graph Building from Statistical Data Tables. ICCCI 2020, CCIS 1287, pp. 701–713
	Spatial clustering	–	Rachapudi, N., Ganesh, L., Sekar, A., Anand, K.S., Sakthibalan, R. Discovery of Structured Data Using Unsupervised Spatial Clustering and Human Supervision. International Journal of Machine Learning and Computing, Vol. 9, No. 5, Oct. 2019. doi: 10.18178/ijmlc.2019.9.5.844
	GNNF	–	Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., Lladós, J. Table Detection in Invoice Documents by Graph Neural Networks. 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019.
	Bioenergy-Miner	https://github.com/text-mining-project/Table-information-extraction	Ge, M., Liu, H., Liu, Y., Zheng, C. Data Mining Techniques for the Analysis and Prediction in Bioenergy Yield. 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), IEEE, 2019
	CDeC-Net	https://github.com/madhav1ag/CDeCNet	Agarwal, M., Mondal, A., Jawahar, C.V. CDeC-Net: Composite Deformable Cascade Network for Table Detection in Document Images. (ICPR 2021; DOI shown in PDF).

Continued on next page

Table 2 continued from previous page

Group	Tool	Tool URL	Reference
	Structure CNN	–	Zhou, M., Ramnath, R. A Structure-Focused Deep Learning Approach for Table Recognition from Document Images. 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC).
	QURMA	–	Nugumanova, A.B., Apayev, K.S., Baiburin, Y.M., Mansurova, M., Ospan, A. QURMA: A Table Extraction Pipeline for Knowledge Base Population. JMMCS №2(114), 2022. DOI: 10.26577/JMMCS.2022.v114.i2.08.
	MPNN GNN	–	Pau Riba, Lutz Goldmann, Oriol Ramos Terrades, Diede Rusticus, Alicia Fornés, Josep Lladós. Table detection in business document images by message passing networks. Pattern Recognition 127 (2022) 108641
	WEATHERGOV+	https://github.com/dash-uvic/WeatherGovPlus	Dash, A., Cote, M., Branzan Albu, A. (2023). WEATHERGOV+: A Table Recognition and Summarization Dataset to Bridge the Gap Between Document Image Analysis and Natural Language Generation. DocEng '23.
	TableNetOCR	–	Paliwal et al., "TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images", ICDAR 2019 (conference).
	AutoTableExtract	–	Sahoo, R., Kathale, C., Kubal, M., Malik, S. (2020). Auto-Table-Extract: A System To Identify And Extract Tables From PDF To Excel. IJSTR, 9(05)
	DiT	https://aka.ms/msdit	Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., Wei, F. (2022). DiT: Self-supervised Pre-training for Document Image Transformer. ACM MM 2022.
	CV FinTable	–	Iftakhar Ali Khandokar; Priya Deshpande. "Computer Vision-Based Framework for Data Extraction From Heterogeneous Financial Tables: A Comprehensive Approach to Unlocking Financial Insights.
Chart	ChartOCR	https://github.com/soap117/DeepRule/tree/master	Luo, J., Li, Z., Wang, J., Lin, C.-Y. "ChartOCR: Data Extraction from Charts Images via a Deep Hybrid Framework." WACV 2021.
	ChartReader	https://github.com/Cvrane/ChartReader	Rane, C., Subramanya, S. M., Endluri, D. S., Wu, J., Giles, C. L. (2021). ChartReader: Automatic Parsing of Bar-Plots. IEEE IRI 2021.
	UniChart	https://github.com/vis-nlp/UniChart	Liu et al. (2023). UniChart: A Universal Vision-Language Pretrained Model for Charts.
	Chart-RCNN	–	Li, S., Lu, C., Li, L., Zhou, H. (2022). Chart-RCNN: Efficient Line Chart Data Extraction from Camera Images. arXiv:2211.14362.
	LINEEX	https://github.com/Shiva-sankaran/LineEX	Shivasankaran V P; Muhammad Yusuf Hassan; Mayank Singh. LINEEX: Data Extraction from Scientific Line Charts. WACV 2023.
	DePlot	https://github.com/google-research/google-research/tree/master/deplot	Liu et al. (2023). DEPLOT: One-shot visual language reasoning by plot-to-table translation. arXiv:2212.10505.
	REV	–	Poco, J., Mayhua, A., Heer, J. "Extracting and Retargeting Color Mappings from Bitmap Images of Visualizations." IEEE Transactions on Visualization and Computer Graphics, 24(1), 2018. DOI: 10.1109/TVCG.2017.2744320.
	Chart-to-Text	https://github.com/vis-nlp/Chart-to-text	Kantharaj, Leong, Lin, Masry, Thakkar, Hoque, Joty. "Chart-to-Text: A Large-Scale Benchmark for Chart Summarization." arXiv:2203.06486v3, 2022.
	PlotQA	–	Methani, N. et al. PlotQA: Reasoning over Scientific Plots. ICCV 2019.
	SemiDatasheet	-	Tian, F., Bernal Cobaleda, D., Martinez, W. (2022). Automatic Data Extraction based on Semiconductor Datasheet for Design Automation of Power Converters. The 2022 International Power Electronics Conference (IPEC-Himeji 2022 -ECCE Asia-).

Continued on next page

Table 2 continued from previous page

Group	Tool	Tool URL	Reference
	BarChartSummary	-	Deng, J., Liu, G., Tang, R., Wu, X., Yin, Z. An automatic selective PDF table-extraction method for collecting materials data from literature. Advances in Engineering Software 204 (2025) 103897.
	Chart Decoder	-	Dai, W., Wang, M., Niu, Z., Zhang, J. Chart decoder: Generating textual and numeric information from chart images automatically. Journal of Visual Languages and Computing, 48 (2018) 101–109.
	BarChartExtraction	-	Al-Zaidy, R. A., Giles, C. L. (2015). Automatic Extraction of Data from Bar Charts. K-CAP 2015.
Full-doc	Docling	GitHub-docling-project/docling: GetyourdocumentsreadyforgenAI	Livathinos, N. et al. (2025). Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion. arXiv:2501.17887v1.
	PP-StructureV2	https://github.com/PaddlePaddle/PaddleOCR/tree/release/2.6/ppstructure	Li, C., Guo, R., Zhou, J., An, M., Du, Y., Zhu, L., Liu, Y., Hu, X., Yu, D. (2022). PP-StructureV2: A Stronger Document Analysis System. arXiv:2210.05391v2
	PDFFigures 2.0	https://github.com/allenai/pdffigures2	Clark, C. Divvala, S. (2016). PDFFigures 2.0: Mining Figures from Research Papers. JCDL '16.
	Nougat	https://github.com/facebookresearch/nougat	Blecher, L., Cucurull, G., Scialom, T., Stojnic, R. "Nougat: Neural Optical Understanding for Academic Documents." arXiv:2308.13418v1, 25 Aug 2023.
	Marker	https://github.com/datalab-to/marker	-
	Unstructured.io	https://github.com/Unstructured-IO/unstructured	-
	PaddleOCR	https://github.com/PaddlePaddle/PaddleOCR?utm_source=chatgpt.com	Cui, C., Sun, T., Lin, M., et al. (2025). PaddleOCR 3.0 Technical Report. arXiv preprint arXiv:2507.05595.
LLM/OCR	olmOCR2	https://github.com/allenai/olmocr	Poznanski, J., Soldaini, L., Lo, K. (2025). olmOCR 2: Unit Test Rewards for Document OCR. arXiv:2510.19817v1.
	ChatGPT	https://chatgpt.com/	-

Appendix S1. Database search strings used for the systematic review

The literature search was conducted primarily in Web of Science and Scopus to identify studies describing tools, systems, and methods for extracting numerical data from tables and figures in scientific PDF documents. The search strategy was designed to capture work related to table extraction, chart and figure understanding, document image analysis, and numerical or structured data recovery. Searches were limited to English-language publications published from 2010 onwards. The search strings were adapted to the syntax requirements of each database.

S1.1 Web of Science search string

The following search string was used in Web of Science:

```
TS = (
    (
        PDF OR "scientific figure" OR "
    scientific chart"
        OR "document image" OR "research article
    "
    )
    AND
    (
        "table extraction" OR "table recognition
    " OR "table detection"
        OR "table structure" OR "table
    segmentation" OR "table mining"
        OR "chart extraction" OR "chart
    understanding"
        OR "bar chart extraction" OR "figure
    extraction"
        OR "figure understanding" OR "diagram
    extraction"
        OR "plot digitization" OR "graph
    digitization"
        OR "scientific chart analysis" OR "data
    reconstruction"
    )
    AND
    (
        numeric OR numerical OR "data values" OR
     "quantitative data"
        OR "structured data" OR "information
    extraction"
    )
    AND
    (
        tool OR algorithm* OR framework* OR
    software OR pipeline*
        OR "open-source" OR "deep learning" OR "
    machine learning"
        OR "neural network*" OR CNN OR
    transformer OR GNN
    )
)
```

S1.2 Scopus search string

The following search string was used in Scopus:

```
TITLE-ABS-KEY(
    (
        PDF OR "portable document format" OR "
    scientific paper*"
        OR "research article*" OR "scientific
    publication*"
        OR "document image" OR "scientific
    figure" OR "scientific chart"
    )
    AND
    (
        "table extraction" OR "table recognition
    " OR "table detection"
        OR "table structure" OR "table
    reconstruction" OR "table segmentation"
        OR "table understanding" OR "table
    mining"
        OR "tabular data extraction" OR "tabular
     structure"
        OR "chart extraction" OR "chart
    recognition" OR "chart understanding"
        OR "chart data extraction" OR "bar chart
     extraction"
        OR "figure extraction" OR "figure
    understanding"
        OR "diagram extraction" OR "diagram
    understanding"
        OR "plot digitization" OR "graph
    digitization"
        OR "scientific chart" OR "scientific
    figure processing"
        OR ("data extraction" W/3 (table* OR
    chart* OR figure* OR graph* OR plot*))
    )
    AND
    (
        numeric OR numerical OR quantitative
        OR "data values" OR "data recovery" OR "
    data reconstruction"
        OR "structured data" OR "information
    extraction"
    )
    AND
    (
        software OR tool OR algorithm* OR
    framework* OR library OR pipeline*
        OR "open source" OR "open-source" OR
    code OR implementation
        OR "deep learning" OR "machine learning"
     OR "neural network*"
        OR CNN OR transformer OR "graph neural
    network"
        OR "object detection" OR "image
    segmentation"
    )
)
AND PUBYEAR > 2010
AND (LIMIT-TO(LANGUAGE, "English"))
```

S1.3 Study selection criteria

To ensure consistency and reproducibility of the review process, predefined inclusion and exclusion criteria were applied during the screening of retrieved records. These criteria were used during both title or abstract screening and full-text assessment.

Inclusion criteria

- Studies describing implemented tools or systems for extracting tables and figures from scientific PDFs or document images
- Systems capable of recovering numerical values, either directly or through structured table or chart reconstruction
- Peer-reviewed articles, preprints, or official technical reports
- Studies providing evaluation evidence, such as benchmark results, dataset descriptions, or validation procedures

- Publications written in English and published from 2010 onwards

Exclusion criteria

- Studies focusing only on general PDF text extraction without addressing tables or figures
- Purely theoretical works without an implemented system or method
- Tools designed exclusively for non-scientific domains (e.g., invoices or business documents), unless demonstrated on scientific-style layouts
- Studies lacking evaluation or validation evidence
- Duplicate records identified across databases

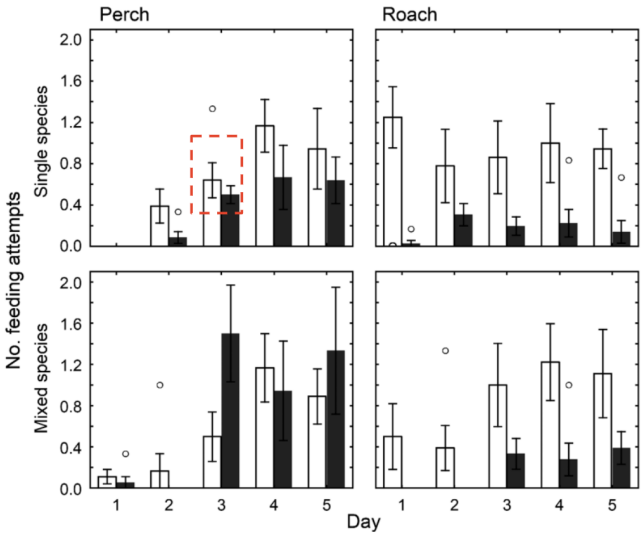

Fig. 3. Mean **(±SE) number of feeding attempts** of perch **Perca fluviatilis** and roach Rutilus rutilus in single- and mixed-species enclosures when subjected to either silent (control) conditions (open bars) or noise exposure (filled bars). Scores show the average **number of feeding attempts** for 6 individuals in the single-species enclosure and 3 individuals in the mixed-species enclosure.

Parameter	Value	Source of Information
Ecoregions Spalding	Baltic Sea	TEXT
Scale	local	TEXT
Country	Sweden	TEXT
Study type	Field Experiment	TEXT
Species / Community	**Perca fluviatilis**	TEXT/FIGURE
Habitat Directive	HD Sandbanks	
Response Variable	**number of feeding attempts**	CAPTION/FIGURE
Response Variable Unit	Count/Number	CAPTION
Extracted Source Detailed	Fig. 3.	CAPTION
Type of Data	A (Arithmetic Means)	FIGURE
Study ID	1	AUTOMATIC
Impact Value	0.502	AUTOMATIC
Control Value	0.638	AUTOMATIC
Error Impact Value	0.091	AUTOMATIC
Error Control Value	0.166	AUTOMATIC
Type of Error	±SE	FIGURE
N Impact	6	CAPTION/FIGURE
N Control	6	CAPTION/FIGURE

Figure 3: Example of context-aware numerical evidence extraction from a scientific figure and caption. The left side shows the original figure from Magnhagen et al. (2017), while the right side illustrates structured extraction outputs linking impact values to contextual metadata, including species, response variables, units, sample sizes, and provenance information. The example demonstrates that usable evidence extraction requires not only recovering numeric values, but also preserving semantic context and traceability.