Deep learning models play a crucial role in document analysis by enabling computers to understand, classify, and extract structured data from documents such as PDFs, scanned images, and complex layouts. Here are some key deep learning models and techniques used for document analysis:

Key Deep Learning Models for Document Analysis
Convolutional Neural Networks (CNNs)

Used primarily for image-based document processing like OCR (optical character recognition) and layout analysis.

Models such as U-Net and Residual U-Net variants excel at detecting text regions, segmenting paragraphs, and recognizing handwriting with high precision.

Transformer Models

Transformer-based architectures like LayoutLM, Donut (Document Understanding Transformer), and BERT variants are state-of-the-art for understanding document content, structure, and semantic relationships.

These models combine textual and layout information to classify documents, extract key fields, and answer document-related questions.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

Used to capture sequential dependencies in text, useful in text line segmentation and recognition in documents with complex layouts.

Graph Neural Networks (GNNs)

Applied to model the relationships between different elements of a document (e.g., tables, figures, text blocks), improving the extraction of structured information.

Multi-modal Deep Learning

Integrates image features and textual content simultaneously, enabling models to interpret the visual and textual layout of documents cohesively.

Popular Libraries and Frameworks
DeepDoctection: A Python library built on PyTorch for document layout analysis, OCR, and document understanding using transformers like LayoutLM and Detectron2 for layout detection. It supports training, evaluation, and inference of document AI models.

Donut (Document Understanding Transformer): Enables end-to-end document understanding without explicit OCR, combining visual and textual features.

Microsoft Azure Document Intelligence: Offers pretrained models for OCR, layout analysis, and domain-specific extraction like tax forms or insurance cards, leveraging transformer architectures.

Nougat (Meta): Advanced layout analysis model designed for understanding document structure and text segments.

Applications
Document classification and routing

Form data extraction and automation

Invoice and receipt processing

Historical and handwritten document transcription

Contract analysis and compliance checks

These deep learning models have set new benchmarks in accuracy, flexibility, and understanding complex document structures and semantics, supporting more intelligent and automated document workflows.

If you want, I can share more details on specific models or toolkits for your use case.

Here are some of the popular and widely used datasets for document analysis, suitable for training and benchmarking deep learning models:

Popular Datasets for Document Analysis
DocLayNet

Contains 80,863 manually annotated document pages from diverse sources.

Provides detailed labeled bounding boxes for 11 layout classes such as text, title, table, figure, list, etc.

Covers varied and complex layouts improving model generalization beyond scientific papers.

PubLayNet

Large-scale dataset with 350,000+ training images and 11,000 validation images.

Layout annotations include 5 categories: text, title, list, table, figure.

Derived from scientific articles, widely used for layout segmentation tasks.

DocBank

Around 500,000 document pages with 12 semantic unit annotations (e.g., section, paragraph, figure, table, caption).

Combines textual and layout information for detailed token-level layout analysis.

Includes fine-grained annotations for NLP and computer vision usage.

TableBank

Focused on table detection and recognition in documents.

Contains nearly 270,000 annotated table images extracted from Word and LaTeX documents.

Useful for training table structure recognition and extraction models.

RVL-CDIP

Large dataset for document classification containing over 400,000 grayscale images in 16 categories like letter, form, memo, scientific report.

Commonly used for document type classification research.

ICDAR Datasets

Released as part of the ICDAR challenges, these datasets focus on tasks such as layout analysis, table recognition, and handwritten text recognition.

Where to Find & Explore
GitHub repositories for DocBank, PubLayNet, DocLayNet.

Kaggle (for document image classification datasets).

Academic datasets shared through research papers and communities.

These datasets are critical resources for developing state-of-the-art document layout analysis, classification, segmentation, and information extraction models at scale. If you want, I can help you find links or details for accessing any of these datasets.

Here is a comparison of notable datasets for table extraction, highlighting their features, scope, and unique strengths:

Dataset	Size & Scope	Key Features	Use Case Focus	Unique Points
PubTables-1M	Nearly 1 million tables from scientific articles	Table detection, structure recognition, functional analysis; detailed annotations for projected rows, headers, cells, including blank ones	Training and evaluating table extraction models on complex scientific documents	Largest dataset with a new evaluation metric (Grid Table Similarity — GriTS) addressing annotation consistency issues
PubLayNet	358K PDF pages, multiple layout regions annotated	Annotates regions like title, text, list, image, table	Document layout and table region detection	Integrates with PubTables-1M for multi-task learning, rich semantic layout diversity
DocBank	500K documents with token-level layout annotations	Semantic units like sections, paragraphs, figures, tables	Fine-grained token-level layout and table extraction	Extends TableBank, integrates textual and visual layout info for unified analysis
TableBank	417K document images from scientific/business docs	Table detection and structure recognition using weak supervision	Large scale table detection across multiple domains	Uses source documents' colored borders to generate bounding boxes; lacks textual ground truth in images
SynFinTabs	100K synthetic financial tables	Table detection, information extraction from financial domain tables	Financial document table extraction	Synthetic data generation enabling modeling of financial tables with realistic layouts and contents
CTE (Contextualized Table Extraction)	Merges PubLayNet and PubTables-1M	Multi-task: detection, structure, functional—plus contextual info like captions, authors	Scientific paper table extraction with contextual understanding	Encourages use of graph neural networks (GNNs) and multi-task learning for comprehensive extraction
TabRecSet	Large-scale dataset from “in-the-wild” real-world tables	Captures diverse table forms in natural settings	Real-world table recognition across scenarios	Provides high variability and real-life complexity for robust model training

Summary

PubTables-1M leads for scale and detailed annotations suitable for advanced table recognition and functional analysis.

PubLayNet and DocBank complement by providing comprehensive document layout context.

TableBank focuses on weakly supervised large-scale table detection, primarily for images.

SynFinTabs uniquely addresses financial domain challenges with synthetic realistic tables.

CTE dataset promotes contextual and multi-task extraction bridging multiple datasets.

TabRecSet targets real-world robustness by including tables from diverse natural environments.

This diversity in datasets allows researchers to choose according to their domain needs, annotation granularity, and scale of data for training and benchmarking table extraction models.

Dataset diversity is a critical factor in table recognition research as it helps develop robust models that can generalize well across various real-world images and document types. Here are some insights into dataset diversity and its importance in this research area:

Importance of Dataset Diversity in Table Recognition
Variety of Table Forms: Datasets with diverse table structures—including bordered, borderless, rotated, distorted, regular, irregular, and multi-language tables—prepare models to handle real-world complexities.

Multiple Document Types: Inclusion of tables from a range of document types such as scanned documents, digital-born PDFs, invoices, forms, financial reports, test papers, and camera-captured images increases model applicability.

Language Diversity: Datasets containing tables in multiple languages (e.g., English, Chinese) support multilingual table recognition models.

Annotation Completeness: Providing detailed spatial annotations using polygons, bounding boxes, or token-level labeling improves precision in detecting irregular tables and complex layouts.

Contextual Variety: Tables appearing in different contexts and with varied surrounding content (e.g., captions, paragraphs, images) enhance models' ability to distinguish table boundaries and contents accurately.

Examples of Diverse Table Recognition Datasets
TabRecSet:
Largest and first bilingual dataset for end-to-end table recognition, covering 38.1K tables with polygon-based annotations supporting irregular tables, sourced from scanned, camera photos, financial invoices, Excel sheets, and educational materials.

Diverse Table Detection Dataset (from arXiv paper):
Over 7,000 samples featuring a wide variety of table structures collected from diverse sources, addressing previous datasets’ limitations like lack of quality and structural complexity.

WikiDT:
Features 70,919 images and 159,905 tables addressing multi-level reasoning and question-answering tasks, emphasizing real-world diversity and hierarchical labels for detailed model diagnosis.

PubTabNet:

Large set of scientifically typeset tables with diverse styles from over 6,000 journals, combined with cell categorization into headers/body to enhance extraction accuracy.

Summary
Dataset diversity enables training of table recognition models that are:

Robust to real-world document conditions, noise, and layout variability.

Capable of recognizing complex and irregular tables.

Applicable across different languages and document types.

Effective in end-to-end tasks including detection, structure recognition, content extraction, and question answering.

Fostering dataset diversity continues to be a key research priority to improve reliability and usability of table recognition technologies across industries.