Deep Learning Benchmark Datasets

Each dataset includes:

- Non-deep learning baseline performance for comparison
- Current state-of-the-art (SOTA) deep learning results
- Recommended architectures students can implement
- Links to official leaderboards and datasets

Note on Tabular Data: Tabular datasets (e.g., Adult Census, California Housing) are **excluded** because tree-based methods (XGBoost, LightGBM) typically match or exceed deep learning performance on medium-sized tabular data. See Grinsztajn et al., NeurIPS 2022: "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?"

Please let me know if a link is broken.

1 Computer Vision Datasets

Table 1: Image Classification and Object Detection Benchmarks

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
CIFAR-10 Classification	60K images 32×32	SVM+HOG: 59% 1-NN: 35%	ViT-H/14: **99.5%** ResNet: 96%	MLP → CNN (LeNet, VGG) → ResNet → ViT fine-tuning	Papers With Code
CIFAR-100 Classification	60K images 100 classes	SVM+HOG: 34% RF: 38%	CoCa: **96.1%** ResNet-110: 74%	CNN with augmentation, transfer learning from ImageNet	Papers With Code
Fashion-MNIST Classification	70K images 28×28	SVM+RBF: 89% k-NN: 84%	WRN: **96.9%** CNN: 93%	MLP baseline → CNN progression, demonstrate inductive bias	Papers With Code

Continued on next page

Continued from previous page

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
ImageNet-1K Classification	1.28M images 1000 classes	SIFT+SVM: 26% (pre-AlexNet)	CoCa: **91.0%** ViT-G: 90.4%	Fine-tune pretrained ResNet/ViT, transfer to downstream tasks	ImageNet Challenge
MS COCO Detection	330K images 80 classes	DPM v5: 33% mAP@.5 (HOG+SVM)	RF-DETR: **60+ mAP** Co-DETR: 66 mAP	Faster R-CNN → YOLO family → DETR-based	COCO Leaderboard
Pascal VOC Detection	11K images 20 classes	DPM: 34% mAP	YOLOv8: **89+ mAP** Faster R-CNN: 73%	Classic detection pipeline, good intro before COCO	VOC Challenge
ADE20K Segmentation	25K images 150 classes	N/A (DL-native task)	Mask2Former: **57.7 mIoU**	UNet → DeepLabv3 → Segformer	Papers With Code
CelebA Face Attr.	202K images 40 attributes	SIFT+SVM: ∼80%	ResNet-50: **91%+**	Multi-task CNN, face-specific pretraining	MMLAB

2 Natural Language Processing Datasets

Table 2: NLP Benchmarks Where Transformers Dominate

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
GLUE Multi-task NLU	9 tasks 1.5M examples	BoW+SVM: 63.7 ELMo: 66.5	DeBERTa: **91.3** Human: 87.1	BERT fine-tuning, multi-task learning, adapter methods	GLUE Leaderboard
SuperGLUE Adv. NLU	8 tasks harder	BERT: 69.0 CBOW: 42.0	T5-11B: **90.4** Human: 89.8	Advanced fine-tuning, prompt engineering, few-shot	SuperGLUE

Continued on next page

Continued from previous page

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
SST-2 Sentiment	67K sentences	Naive Bayes: 83% SVM: 86%	RoBERTa: **97.5%** BERT: 93%	BoW → LSTM → BERT fine-tune progression	Part of GLUE
IMDB Sentiment	50K reviews	TF-IDF+SVM: 88% BoW+NB: 83%	XLNet: **97.2%** BERT: 95%	Word embeddings → LSTM/GRU → BERT	Papers With Code
SQuAD 2.0 Reading Comp.	150K QA pairs	BiDAF: 78 F1 (attention RNN)	Human: 89.5 F1 ALBERT: **92.2 F1**	Attention mechanisms, BERT-based QA, span extraction	SQuAD Explorer
CNN/DailyMail Summarization	312K articles	Lead-3: 40 R-1 TextRank: 36 R-1	BRIO: **47.8 R-1** BART: 44.2 R-1	Pointer-Gen → BART/T5 fine-tuning	Papers With Code
CoNLL-2003 NER	23K sentences 4 entity types	CRF: 88.3 F1 HMM: 74 F1	ACE+DeBERTa: **94.6 F1** BERT: 92.8 F1	BiLSTM-CRF → BERT+CRF, token classification	Papers With Code
WMT14 En-De Translation	4.5M pairs	Phrase-based SMT: 20.7 BLEU	mBART: **35+ BLEU** Transformer: 28.4	Seq2Seq+Attention → Transformer from scratch	Papers With Code
WikiText-103 Lang. Model	103M tokens	KN 5-gram: 145.5 PPL	Transformer-XL: **18.3 PPL**	LSTM LM → Transformer LM, analyze perplexity	Papers With Code

3 Audio and Speech Datasets

Table 3: Speech Recognition and Audio Classification

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
LibriSpeech ASR	960 hrs 1000 speakers	GMM-HMM: 8% WER (test-clean)	Whisper-v3: **1.8% WER** Wav2Vec2: 2.0%	CTC models → Wav2Vec2 fine-tuning → Whisper	Papers With Code

Speech Commands Keyword Spot.	105K clips 35 words	MFCC+GMM: 85% MFCC+SVM: 89%	EfficientNet: **98.7%** CNN: 96%	Spectrogram+CNN, 1D conv on raw audio	Papers With Code
UrbanSound8K Sound Events	8.7K clips 10 classes	MFCC+SVM: 68% MFCC+RF: 70%	VGGish+Aug: **85%+** CNN: 79%	Mel-spectrogram → 2D CNN, audio augmentation	Papers With Code
VoxCeleb1/2 Speaker ID	1M+ utterances 6K+ speakers	i-vector+PLDA: 5.0% EER	ECAPA-TDNN: **0.87% EER**	x-vector → ECAPA-TDNN, contrastive learning	Papers With Code

4 Video Understanding Datasets

Table 4: Video Classification and Temporal Reasoning

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
UCF101 Action Recog.	13K videos 101 classes	IDT+Fisher: 85.9% (hand-crafted)	ViViT: **98.8%** I3D: 95.1%	2D CNN+LSTM → 3D CNN (C3D, I3D) → ViViT	Papers With Code
Kinetics-400 Action Recog.	306K videos 400 classes	N/A (DL-scale dataset)	ViViT: **84.9%** SlowFast: 79.8%	Fine-tune pretrained I3D/SlowFast, video augmentation	Papers With Code
SSv2 Temporal	220K videos 174 classes	N/A (requires temporal reasoning)	VideoMAE: **77.4%** TimeSformer: 62.5%	Focus on temporal modeling; 3D CNNs struggle here	Papers With Code

5 Time Series Datasets

Important Note: Deep learning advantages in time series are **task-dependent**. For forecasting, simple linear models often match Transformers. For classification, deep learning shows clear benefits.

Table 5: Time Series Classification and Forecasting

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
UCR Archive Classification	128 datasets univariate	1-NN DTW: 72% avg HIVE-COTE: 85%	InceptionTime: **85.4% avg** FCN: 83%	MLP → 1D CNN (FCN) → InceptionTime	UCR Archive
UEA Archive Multivar. Class.	30 datasets 2-963 channels	1-NN DTW: 68% avg ROCKET: 83%	ROCKET: **85%** ResNet: 80%	Channel-independent CNN, multivariate attention	TSC Website
ETT (h1/h2/m1/m2) Forecasting	17K-70K rows 7 features	ARIMA: varies **DLinear often wins**	PatchTST: **∼0.37 MSE** (competitive only)	Compare DLinear vs LSTM vs Transformer; learn when DL helps	Papers With Code
M4 Competition Forecasting	100K series various freq.	ETS: 11.7 sMAPE Statistical: strong	ES-RNN: **9.4 sMAPE** (hybrid wins)	Hybrid statistical+neural approaches	M4 GitHub

6 Multimodal Datasets

Table 6: Vision-Language and Cross-Modal Benchmarks

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
VQA v2.0 Visual QA	1.1M questions 265K images	N/A (DL-native)	PaLI-X: **86.0%** BLIP-2: 82.2%	CNN+LSTM → ViLT → BLIP-2 fine-tuning	EvalAI
COCO Captions Image Caption	164K images 5 caps each	N/A (DL-native)	CoCa: **145.3 CIDEr**	Show-Attend-Tell → Transformer captioning	Papers With Code
DocVQA Document QA	50K QA pairs 12K documents	OCR+Rule: ∼40% ANLS	Qwen-VL: **93.1%** LayoutLM: 83%	OCR integration, layout-aware transformers	RRC Portal

Flickr30k Retrieval	32K images 5 caps each	BoW+SVM: ∼30% R@1	CLIP: **88% R@1** VSE++: 52%	Image-text matching, CLIP zero-shot vs fine-tuning	Papers With Code

7 Medical Imaging Datasets

Note: Many require data use agreements. IRB/ethics considerations apply.

Table 7: Medical Imaging Benchmarks

Dataset	**Size**	**Non-DL Baseline**	**DL SOTA**	**Methods (6 months)**	**Leaderboard**
ChestX-ray14 Multi-label	112K images 14 findings	SIFT+SVM: 0.65 AUC	DenseNet-121: **0.84 AUC**	Transfer learning from ImageNet, multi-label BCE loss	NIH Box
CheXpert Chest X-ray	224K images uncertainty labels	N/A	CheXzero: **0.89 AUC**	Handle uncertain labels, comparison with radiologists	Stanford
ISIC 2019 Skin Lesion	25K images 8 diagnoses	Hand-craft: 0.75 AUC	EfficientNet: **0.91 AUC**	Data augmentation critical, class imbalance handling	ISIC Archive

8 Quick Reference: Architecture Progression by Domain

Domain	**Week 1-5: Fundamentals**	**Week 6-12: Advanced**	**Week 13-18: SOTA**
Image	MLP on Fashion-MNIST LeNet on CIFAR-10	ResNet, VGG Transfer learning	Vision Transformers Fine-tuning CLIP
Text	BoW + MLP Word2Vec + LSTM	Attention mechanisms Transformers	BERT fine-tuning Prompt engineering
Audio	MFCC + MLP Spectrogram + CNN	CTC for ASR Wav2Vec2	Whisper fine-tuning Speaker verification
Video	Frame-by-frame CNN CNN + LSTM	3D CNNs (I3D) SlowFast	Video Transformers ViViT
Time Series	1D CNN FCN baseline	InceptionTime LSTM/GRU	Compare with DLinear Foundation models
Multimodal	Separate encoders Late fusion	Cross-attention Early fusion	CLIP, BLIP-2 LLaVA fine-tuning

9 Key Resources

9.1 Leaderboard Aggregators

- **Papers With Code**: `https://paperswithcode.com` – Comprehensive SOTA tracking
- **Hugging Face**: `https://huggingface.co/spaces` – Model demos and leaderboards
- **CodeSOTA**: `https://www.codesota.com` – Emerging alternative to Papers With Code

9.2 Dataset Repositories

- **Hugging Face Datasets**: `pip install datasets` – One-line data loading
- **TorchVision**: Built-in CIFAR, ImageNet, COCO loaders
- **UCR/UEA Archives**: `https://www.timeseriesclassification.com`
- **OpenML**: `https://www.openml.org` – Tabular benchmarks (where DL often loses)

9.3 Pretrained Model Hubs

- **Hugging Face Hub**: BERT, GPT, CLIP, Whisper, and more
- **timm**: `pip install timm` – PyTorch image models
- **TensorFlow Hub**: `https://tfhub.dev`

9.4 Key Papers on "When DL Fails"

- Grinsztajn et al. (2022). "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" *NeurIPS 2022*.

- Zeng et al. (2023). "Are Transformers Effective for Time Series Forecasting?" *AAAI 2023*. (DLinear often wins)
- Gorishniy et al. (2021). "Revisiting Deep Learning Models for Tabular Data." *NeurIPS 2021*.