MetaBench-7: A Multi-Modal Benchmark for Model and Hyperparameter Selection

Anonymous ACL submission

Abstract

Hyperparameter selection in machine learning remains a critical challenge often involving tedious trial-and-error or costly optimization processes. This paper presents “MetaBench-7”,a comprehensive metadata dataset that includes seven modalities (Image, Text, Tabular, Graph, Time-Series, Audio, and Video) with their optimal hyperparameter configurations.The collection includes 573 distinct models, each with standardized metadata like dataset size and class count,enabling detailed quantitative investigation of design trends Exploratory Data Analysis shows unique patterns for each modality.Text datasets often use Transformer-based models with large batch sizes and relatively few epochs,while Graph and Tabular datasets use larger batch sizes and more epochs as dataset size increases. The model-modality specialization statistics indicate that certain architectures, like ResNet50 and XGBoost, are specialized for specific modalities, whereas Transformer variants can work across multiple modalities The dataset provides “safe default” hyperparameter configurations tailored to each modality, offering reliable baselines for new datasets.The dataset serves as a valuable, reusable resource for meta-learning, AutoML and research into the hyperparameter performance dynamics.

1 Introduction

The process of selecting hyperparameters is one of the least automated but most important section of the modern machine learning workflow.Experts still invest considerable time tuning hyperparameters and training epochs through trial and error or expensive search procedures, Even for well-established tasks. This problem becomes more pronounced in multi modal environments,where the "ideal” configuration depends not just on dataset size and task type, but also on the modality (Liang et al., 2021) as images, text, graphs, and time series rarely behave the same way under the same hyperparameters.

In principle, data-driven systems could be designed to learn these patterns.If there were enough examples of real datasets together with the models and hyperparameters that performed well on them,it would be possible train meta models or LLM based advisors (Takayanagi et al., 2025) that recommend default configurations for new tasks. For this,a collection of past examples is needed, including dataset characteristics and the models and configurations that performed well. MetaBench 7 takes a step towards filling this gap in infrastructure. It arranges metadata for 1,200 datasets across seven modalities into a single framework. This includes dataset level statistics along with best reported models and their hyperparameters. Through this analysis, it’s possible to measure how training configurations change across modalities, task type, and dataset scale, and identify which model families behave as reliable specialists or versatile generalists.

”MetaBench 7” can also serve as training data for learned advisors,such as large language model based systems that take a new task description and a few dataset properties as input and recommend possible models and hyperparameter.Thus,the dataset is not just a valuable resource but also a foundation for future progress in data-driven hyperparameter recommendations, AutoML pipelines (Trirat et al., 2025), and LLM-powered configuration assistants (Neszlényi et al., 2024).

2 Literature Review

This section reviews recent advances in automated machine learning, focusing on LLM-driven frameworks and multimodal benchmarking systems. The following subsections summarize key methodologies, contributions, and limitations of these studies.

2.1 Previous Research Findings

Recent advances in large language models have transformed AutoML into flexible, conversational systems. UniAutoML (Guo et al., 2024) pioneered this approach, achieving 8.5% higher usability and 10–100× faster convergence by unifying discriminative and generative tasks through conversational interfaces. However, it lacks systematic guidance for hyperparameter transfer across modalities. AutoFlow (Li et al., 2024) introduced workflow generation through natural language, enabling LLMs to refine execution plans iteratively. While effective for workflow design, it does not tackle the hyperparameter selection challenge. AutoML-Agent (Trirat et al., 2025) proposed a multi-agent framework reaching 68–75% end-to-end automation across 14 datasets using retrieval-augmented planning.

Standardized benchmarks are essential for reproducible AutoML research, yet existing collections lack cohesion. MultiBench (Liang et al., 2021) unified evaluation across 15 datasets and 10 modalities, demonstrating that cross-domain techniques improved performance on 9/15 datasets. However, it focuses only on model architectures and omits optimal hyperparameters needed for AutoML initialization. HPOBench (Eggensperger et al., 2022) provides frameworks for comparing optimization algorithms but does not capture cross-modal hyperparameter patterns. Recent surveys (He et al., 2021) highlight benchmark fragmentation: vision datasets like CIFAR-10 and ImageNet, NLP benchmarks like GLUE, and domain-specific repositories like UCI and UCR operate independently with inconsistent metadata.

2.2 Gaps in Literature

Despite progress in LLM-based AutoML and multimodal benchmarking, critical gaps remain. First, existing AutoML systems rely on expensive task-specific hyperparameter searches without exploiting cross-modal patterns. There is no systematic study of whether optimal learning rates for image classification apply to audio or text domains. Second, current benchmarks evaluate models independently, missing crucial interactions between data modality, task type, and dataset scale. We lack understanding of whether small image datasets and large text datasets require different hyperparameter strategies. Third, while recent work shows LLMs (Fan et al., 2024) can plan workflows, no work leverages LLMs to directly predict optimal hyperparameters from task characteristics, missing an opportunity to learn generalizable optimization patterns from benchmark data. Fourth, evaluation metrics vary across domains accuracy for vision, F1-score for NLP, MAE for time series preventing systematic cross-modal analysis and transfer of hyperparameter insights.

3 Dataset Collection and Preprocessing

3.1 Collection

This work is based on a carefully selected dataset of hyperparameter information gathered from various public and academic sources, instead of raw instance-level data. Each entry in the dataset represents a unique "configuration on a dataset," which combines the details of the training setup (like batch size, learning rate, epochs, and model) with high-level descriptions of the dataset and the task being tackled. The metadata was collected from three key sources: **(i)** academic papers presented at top machine learning conferences that share detailed experimental setups, **(ii)** public platforms like Kaggle (Kaggle, 2021), Papers with Code, GitHub, and others focused on specific domains, **(iii)** benchmark and meta-learning studies that assess multiple models on the same datasets and publish their hyperparameter settings.

3.2 Attribute

The initial dataset was a mix of free-text fields, which needed significant cleaning and normalization to create a structured, analysis-friendly schema.

- **Data Modality**: Initially, there were many variants for the same type of data (e.g., "Image", "Images", "Image (Computer Vision)", "Image (Remote Sensing)"), which caused inconsistency. These variants were later standardized and grouped into a smaller set of categories. The same approach was applied to other modalities like Audio (Kaggle, 2021) , Graph (Hu et al., 2021), Image, Tabular, Text, Time-Series, (Trirat et al., 2025) and Video (Soomro et al., 2012), ensuring consistency and clarity across the dataset.

- **Downstream Task**: encodes the learning objective (classification, regression, detection, etc.), enabling comparisons of hyperparameter behavior across tasks and preventing inappro-

priate aggregation of fundamentally different objectives.

- **Features** column describes how the input data looks like for each dataset in a simple way. For images, it gives the image size and number (Chiu et al., 2020) of color channels (for example, 224×224 with 3 channels). For tabular data, it records how many input columns or attributes are used as features, and for other modalities it provides a short description of the main input shape when that information is available.

- **Train / Valid / Test** and the derived Total Size describe the sample size and split strategy, which strongly influence reasonable choices for batch size, number of epochs, and learning rate schedules.

- **Batch Size**, **Epoch**, **Learning Rate**, and **Best Models** form the core of the hyperparameter and model configuration space. This allows for an analysis of which settings perform well under specific data conditions and across different model families.

- **Evaluation Metric** (Dmitriev and Wu, 2016) column with the corresponding performance scores, serves as a common benchmark for meta-learning. It clearly defines how "best" performance is measured, making it easier to compare results across a variety of different benchmarks.

  Table 1 illustrates a sample from our dataset, showing the columns and the corresponding data for each entry.

3.3 Cleaning and Normalization

The raw spreadsheet was manually created,so many columns were messy and inconsistent. Data Modality had many different strings for the same type of data, which were merged into a small set of standard categories like Image, Text, Tabular, Graph, Time Series, Audio, and Video. Downstream Task contained over a hundred variants with different capitalization and spacing, so these were mapped to a clean task list such as classification, regression, detection, segmentation, and forecasting. The Features column mixed shapes and text (for example "224×224×3" or longer descriptions), so for images it was turned into separate fields for width, height, and channels, while for tabular data it kept

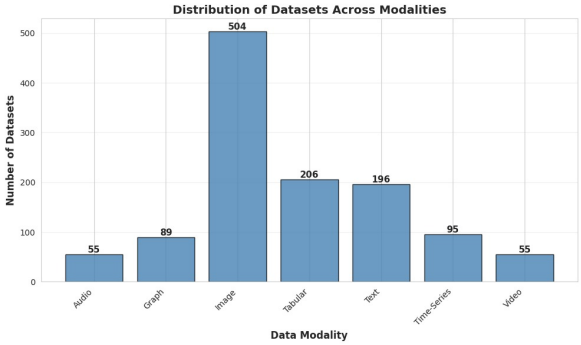

Figure 1: MetaBench-7 Dataset

a single numeric feature count. Train, Valid, and Test sizes were stored as strings with commas and were converted into plain integers, with non numeric values treated as missing. Batch Size, Epoch, and Learning Rate were often written as ranges or with comments, so numeric values were parsed and stored in extra numeric columns to make analysis easier. Model names in Best Models, such as "ResNet50", "ResNet-50", and "ResNet 50",(Liu et al., 2015) were normalized to a single canonical form, and Classes, License, and Evaluation Metric were cleaned into numeric counts or controlled vocabularies wherever possible.

3.4 Final Dataset Statistics

After removing duplicates and empty columns, the final per-modality dataset sizes are: Audio: 55 rows, 16 columns; Graph: 89 rows, 16 columns; Image: 504 rows, 16 columns; Tabular: 206 rows, 16 columns; Text: 196 rows (reduced to 16 columns after dropping 4 empty columns); Time-Series: 95 rows, 16 columns; Video: 55 rows, 16 columns.

Figure 1 illustrates the distribution of datasets across the seven supported modalities in our collection.

4 Methodology

The methodology consists of three stages.The first stage is to collect the configuration of hyperparameters (Becker et al., 2024) from a variety of real, world sources which include research papers, open databases, and benchmark studies that report models, datasets, training settings, and performance.The next step is to clean and prepare the raw data for the analysis. Aligning modality and task names, converting features and splits into numbers, and creating clean columns for hyperparameters are some of the steps involved in this process. Unifying model names, licenses, and met-

Table 1: MetaBench-7: Sample Datasets Collection with Hyperparameters

Data Modality	Downstream Task	Dataset Name	Features	Train	Valid	Test	Batch Size	Epoch	Best Models	Learning Rate	Classes	Source	Evaluation Metric
Image (CV)	Classification	Butterfly Image	224×224×3	4549	1299	651	32	50	InceptionV3	0.001	75	Kaggle	Accuracy
Text (NLP)	Classification	Ecommerce Text	N/A	35296	10084	5044	32	10	BERT	0.00001	4	Kaggle	Accuracy, F1
Tabular (Classic ML)	Clustering	Higher Education	31	101	29	15	16	100	Ensemble	0.05	8	UCI ML	RI
Time Series	Forecasting	Weather	21	36887	10539	5270	32	150	PatchTST	0.0001	N/A	TSLib	RMSLE
Graph	Classification	Cora	1433	140	500	1000	32	1000	GCNII	0.001	7	Planetoid	Accuracy
Video	Recognition	UCF101	112×112×3	9537	3783	3783	8	150	I3D	0.01	101	UCF	Accuracy
Audio Spectrogram	Classification	Birdsong	Spectrogram2D	15000	2000	2000	32	50	EfficientNetB0	0.001	397	Kaggle	Accuracy

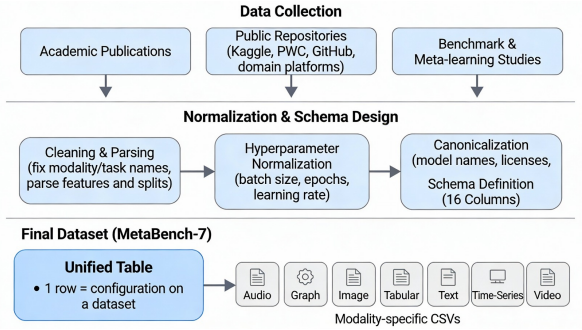

Figure 2: Methodology

rics into a 16, column format are also a few of the tasks performed in this stage.After this, all standardized configurations are joined into one table. This table is then divided into seven different CSV files corresponding to seven different modalities (Audio, Graph, Image, Tabular, Text, Time, Series, and Video). Each row in the dataset stands for a configuration of a dataset, thus, it is appropriate for the next analysis and meta, learning.

5 Conclusion and Future Work

We introduce the first comprehensive,cross-modal hyperparameter metadata repository—a resource compiled from 1,200 datasets spanning seven distinct modalities and 573 unique models.Our analysis systematically documents how dataset properties,model architectures, and training configurations interact across different data modalities. Rather than relying on guesswork, researchers can now input their dataset characteristics and receive evidence-backed recommendations for model selection, training hyperparameters, and appropriate evaluation metrics. Furthur, we will build an intelligent recommendation system that automates these suggestions in real time. We plan to expand our repository to handle multi-modal datasets, profile computational requirements (memory and training time), and add explanations via specialized language models. By maintaining a living, version-controlled knowledge base that adapts as new architectures emerge, we aim to democratize access to principled machine learning practices—leveling the playing field for researchers everywhere.

6 Limitations

The boundary of this dataset primarily stem from the methods used for collecting, cleaning, and maintaining the meta data information.Firstly,it only provides configurations that have been published,which means that popular datasets and models are overrepresented,where niche domains and flawed experiments are mostly missing.This can result in the patterns and recommendations being biased towards mainstream settings.Secondly,a lot of fine-grained labels(such as,different types of image data such as generic computer vision vs. remote sensing)are combined into broad modalities to keep the structure simple.This means,subdomain detail, which could enable more precise,domainspecific insights and recommendations,is sacrificed.Some hyperparameters and practical details are also missing or inconsistently reported,for this reason,the analysis is limited to a core set of fields that are reliably available.Lastly,the current version is essentially a static snapshot of around 1, 200 configurations.Therefore, without regular updates, it will slowly become outdated with respect to new architectures and training settingns. As a result, its guidance should be considered as trend-based rather than something that applies forever.

References

Marc Becker, Lennart Schneider, and Sebastian Fischer. 2024. Hyperparameter optimization. In Bernd Bischl, Raphael Sonabend, and 1 others, editors, *Applied Machine Learning Using mlr3 in R*, pages 85–115. Taylor & Francis, Boca Raton, FL.

Mang Tik Chiu, Xingqian Xu, Yunchao Wei, Zilong Huang, Alexander Schwing, Robert Brunner, Hrant Khachatrian, Hovnatan Karapetyan, Ivan Dozier,

Greg Rose, David Wilson, Adrian Tudor, Naira Hovakimyan, Thomas S. Huang, and Honghui Shi. 2020. Agriculture-vision: A large aerial image database for agricultural pattern analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pavel Dmitriev and Xian Wu. 2016. Measuring metrics. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 429–437, Indianapolis, IN, USA. ACM.

Katharina Eggensperger, Philipp Müller, Neeratyoy Mallik, Matthias Feurer, René Sass, Aaron Klein, Noor Awad, Marius Lindauer, and Frank Hutter. 2022. Hpobench: A collection of reproducible multi-fidelity benchmark problems for hpo.

Shengda Fan, Xin Cong, Yuepeng Fu, Zhong Zhang, Shuyan Zhang, Yuanwei Liu, Yesai Wu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Workflowllm: Enhancing workflow orchestration capability of large language models.

Jiayi Guo, Zan Chen, Yingrui Ji, Liyun Zhang, Daqin Luo, Zhigang Li, and Yiqin Shen. 2024. Uniautoml: A human-centered framework for unified discriminative and generative automl with large language models.

Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2021. Open graph benchmark: Datasets for machine learning on graphs.

Kaggle. 2021. Birdsong Recognition Dataset. Audio classification competition for identifying bird species from audio recordings.

Zelong Li, Shuyuan Xu, Kai Mei, Wenyue Hua, Balaji Rama, Om Raheja, Hao Wang, He Zhu, and Yongfeng Zhang. 2024. Autoflow: Automated workflow generation for large language model agents.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. *Preprint*, arXiv:1411.7766.

Kálmán Balázs Neszlényi, Alex Milos, and Attila Kis. 2024. Assistantgpt: Enhancing user interaction with llm integration.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. (CRCV-TR-12-01).

Takehiro Takayanagi, Kiyoshi Izumi, Javier Sanz-Cruzado, Richard McCreadie, and Iadh Ounis. 2025. Are generative ai agents effective personalized financial advisors? In *Proceedings of the ACM Conference*, pages 286–295.

Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. 2025. Automl-agent: A multi-agent llm framework for full-pipeline automl.