Bachelor Computer Science & Data Science and Artificial Intelligence

Optimizing Algorithm Selection with Deep Learning-Based Meta-Feature Extraction

Khayri Hamza

First supervisor and (if already known) second supervisor:
Dr. J.N. van Rijn

RESEARCH PROPOSAL BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

05/03/2025

1 Background and theory

Meta-Learning, or 'learning to learn' is the study of principled methods that exploit metaknowledge to obtain efficient models and solutions by adapting machine learning processes [BGCVS22, Chapter 1]. Instead of training each model from scratch and choosing the best one, a Meta-Learning system leverages previous experiences to adjust the learning strategy, leading to more efficient and accurate implementation.

Meta-learning has a wide variety of applications, such as algorithm selection, hyperparameter optimization, few-shot learning, and transfer learning. This bachelor thesis will focus on algorithm selection, where the Meta-Learning model can predict the most suitable machine learning model for a machine learning task based on patterns learned from previous tasks. 1 visualises the Meta-Learning pipeline.

The Meta-Learning pipeline for algorithm selection consists of the following steps:

1. **Train Base Learners** – Train different algorithms on a diverse set of datasets and record performance metrics. This will be used to train the Meta-Model.
2. **Extract Meta-Features** – Compute meta-features from the datasets as inputs for the Meta-Model.
3. **Train Meta-Model** – Use the meta-features as inputs and the base learners' performance as targets to train a meta-model that learns the relationship between the dataset characteristics and the model performance.
4. **Algorithm Prediction** – Given a new dataset and its meta-features, the Meta-Model predicts which algorithm would be best suited for the task.

Meta-features are dataset characteristics that capture the properties of datasets. They are used to predict the performance of machine learning algorithms by learning from insights of past experiences. They are categorized into simple statistical measures (number of instances, number of attributes etc.), information-theoretic features (feature entropy, class entropy etc.), model-based characteristics (decision tree attributses like depth and number of nodes etc), and landmarkers (Decision stump accuracy, Naïve Bayes accuracy etc.) [BGCVS22, Chapter 4].

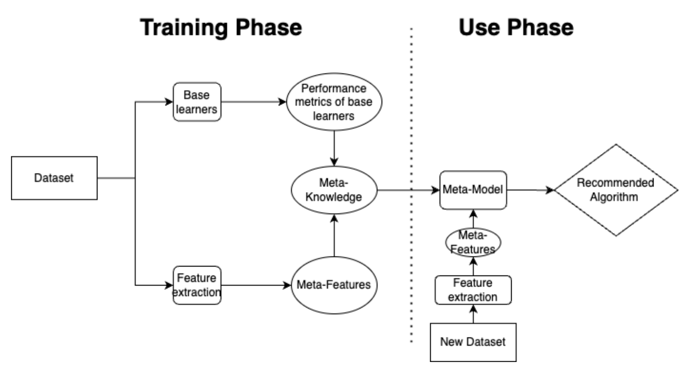

Figure 1: Meta-Learning Pipeline illustrating the process of training base learners, extracting meta-features, and predicting the most suitable algorithm using a meta-model.

Deep learning-based meta-feature extraction improves on traditional Meta-Feature extraction methods by learning more abstract and domain-agnostic dataset representations. It utilises deep neural networks to capture hierarchical representations of data and allow the models to recognize more complex relationships that traditional feature extraction may not [DT18]. However, traditional Meta-Feature extraction is constrained by the uneven distribution of Meta-Features across datasets [CGL23]. By leveraging deep-learning-based feature extraction, we can enhance generalization across unseen datasets and improve overall performance [DWL+22].

2 Research question

How can deep learning based feature extraction improve the effectiveness of Meta-Learning for algorithm selection compared to traditional feature extraction methods?

Sub-Research question

- Which deep learning models work best for feature extraction?
- How do deep learning Meta-Features affect Meta-Model performance and selection?
- How does deep learning-based feature extraction compare to traditional methods in terms of computational cost?
- How well do deep learning Meta-Features generalize across domains?

3 Study Design and Methodology

This study will evaluates deep learning-based Meta-Feature extraction against traditional Meta-Feature extraction for Meta-Learning in algorithm selection. This consists of extracting Meta-Features, training base learners, Meta-Learning system construction and evaluation.

Diverse datasets will be sourced from OpenML, UCI, and other repositories to ensure generalization across different tasks. These datasets will provide a broad benchmark for comparing meta-feature extraction methods.

A variety of machine learning algorithms, including Decision Trees, Random Forests, SVMs, and Neural Networks, will be trained on each dataset. Their performance metrics will be recorded to serve as training data for the meta-learning model, and ranked to establish hierarchy of algorithm effectiveness [PRAF23] that will be input as the target for the Meta-Model.

Three Meta-Learning pipelines will be developed: one using traditional feature extraction, another using deep learning-based extraction [DT18], and a hybrid approach [PdSdLFdC21] that combines both. The inclusion of these three approaches allows for a comprehensive comparison of trade-offs between interpretability, computational efficiency, and predictive performance in meta-learning for algorithm selection.

The evaluation will assess both performance and efficiency. Performance will be measured using accuracy and ranking correlation , while efficiency will be analyzed in terms of memory usage and training time.

Lastly, there will also be in-depth analysis to provide insights into the effectiveness of deep-learning based meta-feature extraction. This will consist of studying different deep learning architectures in relation to meta-feature extraction [EAAAM24], the impact of deep learning based meta-features on Meta-Model performance, and the generalizability of deep learning based Meta-Features to determine the robustness [LYSH17].

4 Global Planning

The total duration of the research is 20 weeks. Notable milestones include the proposal submissions (Week 3 and 5), midterm presentations (Week 12-16), and the complete thesis draft (Week 18). **??** visualises the thesis process.

The **research and planning** phase (Week 1 - 4) involves conducting a literature review to gather knowledge on the topic and refining the research approach through discussions with the supervisor and peers. The research proposal will be defined, and relevant datasets will be identified to ensure a strong foundation for experimentation.

In the **experimental preparation** phase (Week 4-5), selected datasets will be preprocessed and cleaned to ensure data quality. The appropriate target algorithms, deep learning models, and Meta-Models will be chosen. Feature extraction and feature selection techniques will be implemented to establish a structured experimental framework.

During **experimentation and training** (Week 6 - 14), base learners will be trained, and their performance metrics will be ranked. Meta-Models will then be trained and tested to assess their effectiveness in selecting optimal algorithms.

The **evaluation** phase (15-16) consists of documenting the experimental process and analyzing the results. Findings will be synthesized into discussion and evaluation sections to highlight key insights and performance comparisons.

Finally, in the **refinement** stage (Week 17-20), the conclusion will be written, the report will be refined based on feedback, and the final submission will be prepared.

Certain phases, such as literature review and supervisor / peer feedback, will persist throughout the process. This is because the research is an iterative process, and further revisions will be necessary throughout. However, they may be more concentrated at the start, to establish strong conceptual understanding, and the end, for clarity.

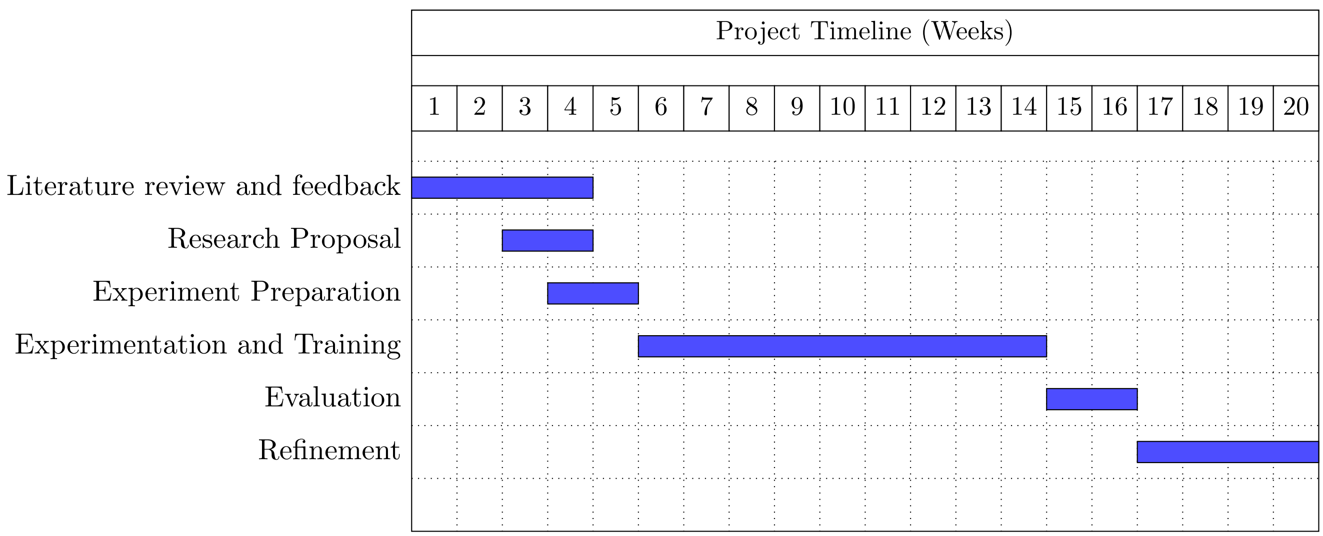

Figure 2: Gantt chart illustrating the research timeline, outlining key phases and milestones over the 20-week study period.

References

[BGCVS22] Pavel Brazdil, Christophe Giraud-Carrier, Ricardo Vilalta, and Carlos Soares. *Metalearning: Applications to Automated Machine Learning and Data Mining.* Lecture Notes in Artificial

Intelligence. Springer, 2nd edition, 2022. Open Access.

[CGL23] Nathan Carvalho, André Gonçalves, and Ana Lorena. Collecting meta-data from the openml public repository. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, pages 610–624, Porto Alegre, RS, Brasil, 2023. SBC.

[DT18] Suresh Dara and Priyanka Tumma. Feature extraction by using deep learning: A survey. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1795–1801, 2018.

[DWL+22] Yu Ding, Lei Wang, Bin Liang, Shuming Liang, Yang Wang, and Fang Chen. Domain generalization by learning and removing domain-specific features, 2022.

[EAAAM24] Omar Elharrouss, Younes Akbari, Noor Almadeed, and Somaya Al-Maadeed. Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. *Computer Science Review*, 53:100645, August 2024.

[LYSH17] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization, 2017.

[PdSdLFdC21] Geand Trindade Pereira, Moises Rocha dos Santos, and Andre Carlos Ponce de Leon Ferreira de Carvalho. Evaluating meta-feature selection for the algorithm recommendation problem, 2021.

[PRAF23] Namyong Park, Ryan Rossi, Nesreen Ahmed, and Christos Faloutsos. Metagl: Evaluation-free selection of graph learning models via meta-learning, 2023.