Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240487

Comparing nnU-Net and deepflash2 for Histopathological Tumor Segmentation

Daniel HIEBER [a,b,c,1], Nico HAISCH [d], Gregor GRAMBOW [d], Felix HOLL [a], Friederike LIESCHE-STARNECKER [b], Rüdiger PRYSS [c], Jürgen SCHLEGEL [b,e], and Johannes SCHOBEL [a]

[a] *DigiHealth Institute, Neu-Ulm University of Applied Sciences*
[b] *Pathology, Medical Faculty, University of Augsburg*
[c] *Institute of Medical Data Science, University Hospital Würzburg*
[d] *Dept. of Computer Science, Aalen University of Applied Sciences*
[e] *Technical University of Munich*

ORCiD ID: Daniel Hieber https://orcid.org/0000-0002-6278-8759, Felix Holl https://orcid.org/0000-0002-4020-9509, Friederike Liesche-Starnecker https://orcid.org/0000-0003-1948-1580, Rüdiger Pryss https://orcid.org/0000-0003-1522-785X, Johannes Schobel https://orcid.org/0000-0002-6874-9478

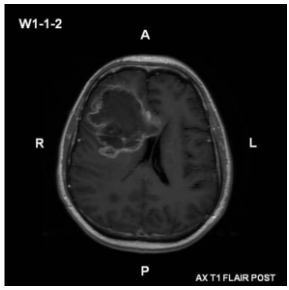
Abstract. Machine Learning (ML) has evolved beyond being a specialized technique exclusively used by computer scientists. Besides the general ease of use, automated pipelines allow for training sophisticated ML models with minimal knowledge of computer science. In recent years, Automated ML (AutoML) frameworks have become serious competitors for specialized ML models and have even been able to outperform the latter for specific tasks. Moreover, this success is not limited to simple tasks but also complex ones, like tumor segmentation in histopathological tissue, a very time-consuming task requiring years of expertise by medical professionals. Regarding medical image segmentation, the leading AutoML frameworks are *nnU-Net* and *deepflash2*. In this work, we begin to compare those two frameworks in the area of histopathological image segmentation. This use case proves especially challenging, as tumor and healthy tissue are often not clearly distinguishable by hard borders but rather through heterogeneous transitions. A dataset of 103 whole-slide images from 56 glioblastoma patients was used for the evaluation. Training and evaluation were run on a notebook with consumer hardware, determining the suitability of the frameworks for their application in clinical scenarios rather than high-performance scenarios in research labs.

Keywords. Machine Learning, AutoML, Segmentation, nnU-Net, deepflash2, Histopathology

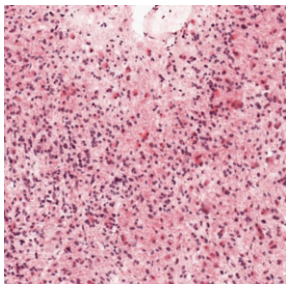
1. Introduction

Over the last few years, automated Machine Learning (AutoML) techniques have become popular alternatives to individually engineered Machine Learning (ML) models.

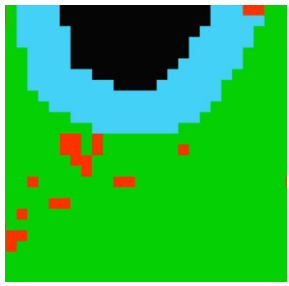
[1] Corresponding Author: Daniel Hieber, e-mail: daniel.hieber@hnu.de.



(a) MRI Image of a Glioblastoma



(b) Section of an HE Image of a Glioblastoma



(c) Annotations for the HE Image

Figure 1. Comparison of Radiological and Histopathological Images (Images from Ivy Glioblastoma Atlas [13], Slide W1-1-2-D.2)

Instead of handling the training and evaluation process of ML models manually, with AutoML, only data has to be provided, and the best resulting model must be chosen [1]. Depending on the framework, even data processing and model selection (based on predefined criteria) can be automated, drastically reducing the need for human intervention. Large companies such as Microsoft, Google, and Amazon provide AutoML frameworks both open-source [2,3] and as part of their cloud platforms [4,5,6].

Besides generic frameworks, AutoML frameworks have also been created for specialized areas, including healthcare [7,8]. For example, *AutoPrognosis2.0* allows for the automatic analysis of tabular data for classification and regression tasks or survival prediction [9] or the application of AutoML for the diagnosis of diabetes [10].

In this work, we conduct a first comparison between the leading AutoML frameworks for medical image segmentation, *nnU-Net* (v2) [11] and *deepflash2* [12] for histopathological image segmentation. An internal dataset of 103 Whole-Slide Images (WSIs) of Glioblastoma (GBM) patients is used as the use case. GBM is an especially hard-to-segment malignant brain tumor due to its high heterogeneity. This use case is selected for its difficult generalizability, as histopathological images often lack a clear structure (e.g., distinguishable organs) compared to radiology images. GBMs further have especially hard-to-determine borders. Figure 1 illustrates this problem by showing an MRI image of a brain with a GBM (Figure 1a upper left), a small section of a Hematoxylin and Eosin (HE) stained WSI (Figure 1b) and its annotations (Figure 1c).

The evaluation focuses on the general performance, usability for clinicians, and clinical relevance of *nnU-Net* and *deepflash2*. By running the training and inference on consumer-grade hardware, this research also investigates the accessibility and practicality of using such frameworks in a clinical setting, where resources may be limited compared to highly specialized research laboratories. However, the lab/clinic should still own the hardware, not individuals, to comply with privacy and data protection regulations.

2. Methods

Figure 2 shows the approach for evaluating both AutoML frameworks. The initial preparation of the dataset and the final evaluation are conducted manually. The frameworks

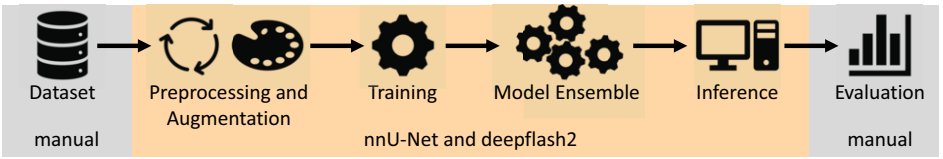


Figure 2. Workflow for Model Training and Evaluation

themselves handle the rest, including preprocessing and training with hyperparameter tuning. Inference can be run directly within the frameworks, requiring no additional code.

Both *nnU-Net* and *deepflash2* are available as open-source projects. The frameworks provide the same capabilities, as they only require input data (with proper labels) and can handle the complete learning process by themselves. A more precise and robust prediction is possible by generating a model ensemble instead of a single model.

While similar, the two frameworks use different approaches during training. The *nnU-Net* models are trained from scratch without pre-trained weights. On the contrary, *deepflash2* uses transfer learning with pre-trained weights.

Both frameworks require a specific format for the input data. For *deepflash2*, all images and masks are split into two folders. The masks must have the same name as the image and contain a suffix (e.g., `_mask`). For *nnU-Net*, three top-level dataset folders are required are needed. One is for the raw data (prior to preprocessing), one is for the processed data, and finally, one is for the results after training. The folder paths must be further exported as environment variables. In the `raw` folder, the data is provided similarly to *deepflash2*, with one folder for labels and another for the masks. The other folders are used during preprocessing and training. The *nnU-Net* setup further requires a `JSON` file containing metadata for the dataset. The background color in masks for both models must be set to a pixel value of 0.

The data for this evaluation consists of a GBM dataset with 103 WSIs from 56 patients. Each image is split into smaller tiles (1.024 x 1.024 pixels), which are resized to 256 x 256 pixels. For the training, 3.700 tiles were selected, each expressing tumor, healthy tissue, and/or background. Both frameworks use cross-validation during the training phase, automatically handling training and validation dataset assignments.

To evaluate the segmentation accuracy of both frameworks after training, the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), and recall were chosen as metrics. For these metrics, the mean value was calculated for all images in the test dataset (n=2023, with empty tiles and tiles containing only tumor). The required training time was calculated using a simple script logging the start and end timestamps. No cache clearing/sandboxing was used. Times are therefore provided as an approximated value. Further, the ease of use for non-IT-experts (i.e., clinicians) was ranked subjectively.

Both frameworks were evaluated using the same lab-owned Dell XPS 15 notebook, running Windows 10 with 64GB RAM, an RTX 4070M, and an Intel Core i9-13900H.

3. Results

Figure 3 shows the major results of the evaluation on the test dataset for both model ensembles. *deepflash2* was able to outperform *nnU-Net* in all metrics. The former achieved

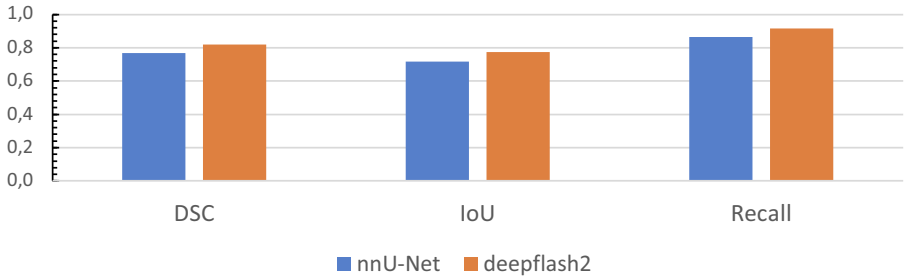


Figure 3. Results of *nnU-Net* and *deepflash2* Evaluation on Test Dataset. Higher Values are better.

a DSC of 81.9% (vs. 76.8%), an IoU of 77.3% (vs. 71.8%), and a Recall of 91.6% (vs. 86.5%). A model specifically trained for this task achieved an IoU of 84.5% [14].

Regarding the training time on consumer hardware, *deepflash2* was trained significantly faster due to its transfer learning approach (5 hours vs. 240 hours for *nnU-Net*).

Finally, regarding usability, *nnU-Net* can be controlled via command line prompts without coding. However, the dataset structure is more complex than the simple structure from *deepflash2*. Further, an extra `JSON` file must be created that describes the dataset. Also, *deepflash2* provides a GUI to interact with the framework. However, this GUI requires the setup of a Jupyter Notebook to work.

4. Discussion

While both frameworks worked very well for the presented task without any modifications or fine-tuning to the dataset, *deepflash2* produced overall better results.

Both frameworks require a low to moderate understanding of coding, as no installable application with a GUI is provided. The frameworks must be installed using Python's package manager. *nnU-Net* can be run via the command line, *deepflash2* provides an optional GUI. However, a Jupyter Notebooks setup is required for this. The simple dataset structure of *deepflash2* also contributes to its ease of use.

The limiting factor regarding *deepflash2* is its current development. While *nnU-Net* is actively developed by the Division of Medical Image Computing of the German Cancer Research Center (DKFZ), *deepflash2* is a small project with three main contributors. Its last commit to GitHub at the date of analysis was in June 2023 (almost 10 months ago). This makes *nnU-Net* a better foundation for projects that aim to be used in production.

As the scope of this work was to compare their applicability to medical use cases by non-IT experts (i.e., clinicians), the frameworks' parameters were not optimized. Such parameter tuning requires a deeper knowledge of the frameworks and ML in particular and would render the evaluation unrepresentable for clinicians. However, both frameworks most likely are able to achieve better results with parameter tuning.

5. Conclusion

This work provides a first comparison of two major AutoML frameworks, *nnU-Net* and *deepflash2*, for the segmentation of GBMs in histopathological images.

While both frameworks handled the task comparable to a model trained explicitly for this use case, *deepflash2* achieved higher scores in all evaluated metrics. The latter also required significantly less training time (approx. 5 hours vs approx. 240 hours) and was easier to use. However, *nnU-Net* is under active development by a larger community and backed by the DKFZ, making it a more reliable and safe foundation for new projects.

From the results of this work, one could argue that clinicians looking for a simple way to get good results should use *deepflash2*. While this tool is still not easy to use, the required technical setup is significantly easier than the nnU-Net setup. Researchers or companies with the necessary expertise and computing resources to create new models that can be used beyond a single study should rely on *nnU-Net*.

As future work, a large, comprehensive evaluation based on the complete dataset (20k images) is planned. To determine the generalizability and robustness of the models, this includes evaluating the inference results on the Ivy Glioblastoma Atlas, the largest public repository for annotated GBM WSIs worldwide [13].

References

[1] He X, Zhao K, Chu X. AutoML: A survey of the state-of-the-art. Knowledge-Based Systems. 2021 Jan;212. Available from: https://www.sciencedirect.com/science/article/pii/S0950705120307516.

[2] Erickson N, Mueller J, Shirkov A, Zhang H, Larroy P, Li M, et al.. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv; 2020. ArXiv:2003.06505 [cs, stat]. Available from: http://arxiv.org/abs/2003.06505.

[3] manashgoswami. What is automated ML? AutoML - Azure Machine Learning; 2023. Last accessed: 2024-03-06. Available from: https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml?view=azureml-api-2.

[4] AutoML tools and solutions from AWS;. Last accessed: 2024-03-06. Available from: https://aws.amazon.com/machine-learning/automl/.

[5] AutoML Solutions - Train models without ML expertise;. Last accessed: 2024-03-06. Available from: https://cloud.google.com/automl.

[6] Azure Automated Machine Learning - AutoML | Microsoft Azure;. Last accessed: 2024-03-06. Available from: https://azure.microsoft.com/en-gb/products/machine-learning/automatedml/.

[7] Waring J, Lindvall C, Umeton R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine. 2020 Apr;104:101822. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0933365719310437.

[8] Mustafa A, Rahimi Azghadi M. Automated machine learning for healthcare and clinical notes analysis. Computers. 2021;10(2):24.

[9] Imrie F, Cebere B, McKinney EF, van der Schaar M. AutoPrognosis 2.0: Democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. PLOS Digital Health. 2023;2(6):e0000276.

[10] Zhuhadar LP, Lytras MD. The application of autoML techniques in diabetes diagnosis: Current approaches, performance, and future directions. Sustainability. 2023;15(18):13484.

[11] Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203-11.

[12] Griebel M, Segebarth D, Stein N, Schukraft N, Tovote P, Blum R, et al. Deep learning-enabled segmentation of ambiguous bioimages with deepflash2. Nature Communications. 2023;14(1):1679.

[13] Puchalski RB, Shah N, Miller J, Dalley R, Nomura SR, Yoon JG, et al. An anatomic transcriptional atlas of human glioblastoma. Science. 2018;360(6389):660-3.

[14] Hieber D, Prokop G, Karthan M, Märkl B, Schobel J, Liesche-Starnecker F. Neural Network Assisted Pathology for Labeling Tumors in Whole-Slide-Images of Glioblastoma. In: (Abstractband der) 106. Jahrestagung der Deutschen Gesellschaft für Pathologie "Pathology - more than meets the eye"; 2023. p. 218f. Abstract P.13.13.