Technische Universität München
TUM School of Computation, Information and
Technology

Towards Reproducible, Stable, and Robust Machine Learning Research in Clinical Environments

Sarthak Pati

Vollständiger Abdruck der von der TUM School of Computation, Information and Technology der Technischen Universität München zur Erlangung eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigten Dissertation.

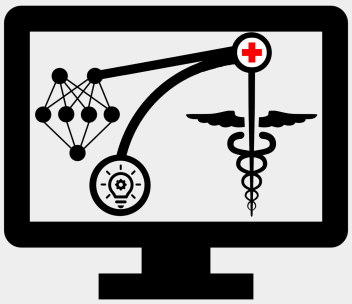
Vorsitz: Prof. Dr. Klaus Diepold

Prüfende der Dissertation:

1. Prof. Dr. Björn Menze
2. Prof. Dr. Dimitrios Makris
3. Prof. Dr. Shadi Albarqouni

Die Dissertation wurde am 12.02.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Computation, Information and Technology am 25.01.2025 angenommen.

Image-Based Biomedical Modeling
Prof. Dr. Bjoern Menze



Dissertation

Towards Reproducible, Stable, and Robust Machine Learning Research in Clinical Environments

Sarthak Pati

Fakultät für Informatik
Technische Universität München

Acknowledgments

I would like to express my deepest gratitude to all the people who have helped me in completing this thesis. First and foremost, I would like to thank my thesis mentor, ***Spyros*** who believed in me from the beginning, pushed me to do better, and gave me his unyielding support throughout this journey. He has been a constant source of inspiration, guidance, and encouragement for me. I am truly fortunate to be able to learn from him.

I would also like to thank my thesis supervisor, ***Bjoern***, who provided keen insights during the process of my exploring various topics. He has been very generous with his time and expertise, and always offered constructive feedback and suggestions. His knowledge and experience have been invaluable for me.

I am grateful to ***Mocha***, for his unending patience while his walk was delayed because of my work. He has been a source of joy and comfort for me. He always greeted me with a wagging tail regardless of how late I wrapped up work or how stressed I was.

I would like to thank my ***in-laws***, who have supported me throughout this journey.

I would like to express my gratitude to my son, ***Samriddh***, who was instrumental in redefining my life's priorities. He has been the motivation behind everything I do, and the reason why I strive to be a better person. He has filled my life with love, happiness, and wonder. I am immensely proud of him, and I hope he will be proud of me too.

I owe a debt of gratitude to my ***parents*** and my ***brother***, whose unwavering support has allowed me to achieve everything I have in life. They have always been there for me, through thick and thin, and have given me the best education, upbringing, and opportunities possible. They have taught me the values of hard work, honesty, and perseverance. They have sacrificed a lot for me, and I can never thank them enough.

Finally, I would like to express my heartfelt appreciation to my wife, ***Shweta***, for her endless patience throughout this entire process, and without whom this endeavor would have neither materialized nor concluded. She has been my pillar of strength, my best friend, and my soul-mate. She has supported me in every way possible, emotionally, financially, and intellectually. She has endured my long hours of work, my frequent travels, and my (more than) occasional frustrations. She has celebrated my successes, consoled me in my failures, and cheered me on in my challenges. She has made countless sacrifices for me and our family, and I am eternally grateful to her.

I dedicate this thesis to my family...

You have been my source of inspiration, motivation, and support throughout my academic journey. You have always encouraged me to pursue my dreams and aspirations, and you have never doubted my abilities or potential. You have sacrificed so much for me, and I am forever grateful for your unconditional love and care. You have taught me the values of hard work, perseverance, and resilience, and you have shown me the meaning of happiness, joy, and gratitude. You are the best gift I have ever received.

Table of Contents

Abstract

Towards Reproducible, Stable, and Robust Machine Learning Research in Clinical Environments is a dissertation that guides researchers to analyze their work through the lens of clinical deployment. It highlights the challenges a research project faces when considering clinical translation and the considerations needed to overcome them. It emphasizes the importance of healthcare data in medicine and showcases academic contributions and relevant publications. Finally, it summarizes the results of the work and provides future directions, thus providing a broad outlook.

Zusammenfassung

Towards Reproducible, Stable, and Robust Machine Learning Research in Clinical Environments ist eine Dissertation begleitet Forschende dabei, ihre Arbeit am Maßstab der klinischen Anwendung zu prüfen. Sie beleuchtet die Herausforderungen der klinischen Translation und zeigt Strategien zu deren Bewältigung. Kernpunkte sind die Bedeutung klinischer Daten in der Medizin sowie akademische Beiträge und relevante Publikationen. Abschließend fasst die Arbeit die Ergebnisse zusammen und weist Zukunftsperspektiven auf, wodurch ein umfassender Ausblick entsteht.

Acronyms

Abbreviation	Elaboration
2D	Two Dimensional
3D	Three Dimensional
AI	Artificial Intelligence
COFE	Comprehensive Open Federated Ecosystem
COLLAGE	Co-occurrence of Local Anisotropic Gradient Orientations
CT	Computed Tomography
DL	Deep Learning
DSC	Dice Similarity Coefficient
EHR	Electronic Health Records
FL	Federated Learning
GaNDLF	Generally Nuanced Deep Learning Framework
GLCM	Gray-Level Co-occurrence Matrix
GLRLM	Gray-Level Run-Length Matrix
GLSZM	Gray-Level Size Zone Matrix
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NGTDM	Neighborhood Gray Tone Difference Matrix

Part I

Introduction and Background

1 Motivation

1.1 Role of Imaging in Modern Healthcare

Imaging has a crucial role in modern healthcare as it provides physicians with the ability to observe the internal structural, pathological, and physiological properties of the body and enables diagnostic, prognostic, therapeutic, as well as interventional assessments [122]. Some of the common medical imaging modalities are radiology (such as ultrasonography, x-rays, mammography, computed tomography (*CT*), magnetic resonance imaging (*MRI*), and nuclear medicine) and microscopy (histopathology). Each imaging modality specializes in a specific type of structure or function of the body. For example, x-rays are useful for imaging calcified structures such as bones, magnetic resonance images are good for soft tissues, and microscopy images are good for understanding the morphological characteristics of diseases. However, together, these technologies have ensured that they are a vital tool for improving the overall quality and safety of healthcare.

In a drive to personalize medicine, and with technical milestones in healthcare leapfrogging over each other, healthcare data has become increasingly complicated and feature-rich (for example, generation of high-density gene maps, and usage of dictation tools to generate medical reports) [49]. With every passing year, the radiology and microscopy scanning resolution is getting better, thus producing higher-resolution images with more information [31]. Individual samples from each source of healthcare information can be thought of as individual **data points**. Due to its sensitive nature and potential impact, each **data point** requires careful evaluation, and analyzing the increasing breadth and depth of the generated data can be exhausting to clinical experts. To address this issue, many initiatives (both in terms of academic research community and commercial products) have been undertaken to develop, test, and ultimately apply computational methods to discover meaningful connections among these **data points**, thus potentially easing the workload of clinical experts.

1.2 Overview of Healthcare Data Analysis

Statistical modeling, machine learning, and deep learning are three interconnected fields that play a crucial role in data analysis and prediction (see Figure 1.2 for an illustration). **Statistical modeling** (which is termed as **artificial intelligence** (*AI*) in recent years [**russell2010artificial**]) is a fundamental approach that involves the use of rule-based systems to perform predictive analysis. These systems rely on predefined rules and mathematical equations to analyze data and make predictions. On the other hand, **machine learning** (*ML*) takes a more dynamic approach. It involves training statistical models that learn from a dataset, which can be either labeled or unlabeled, to make predictions or decisions on previously

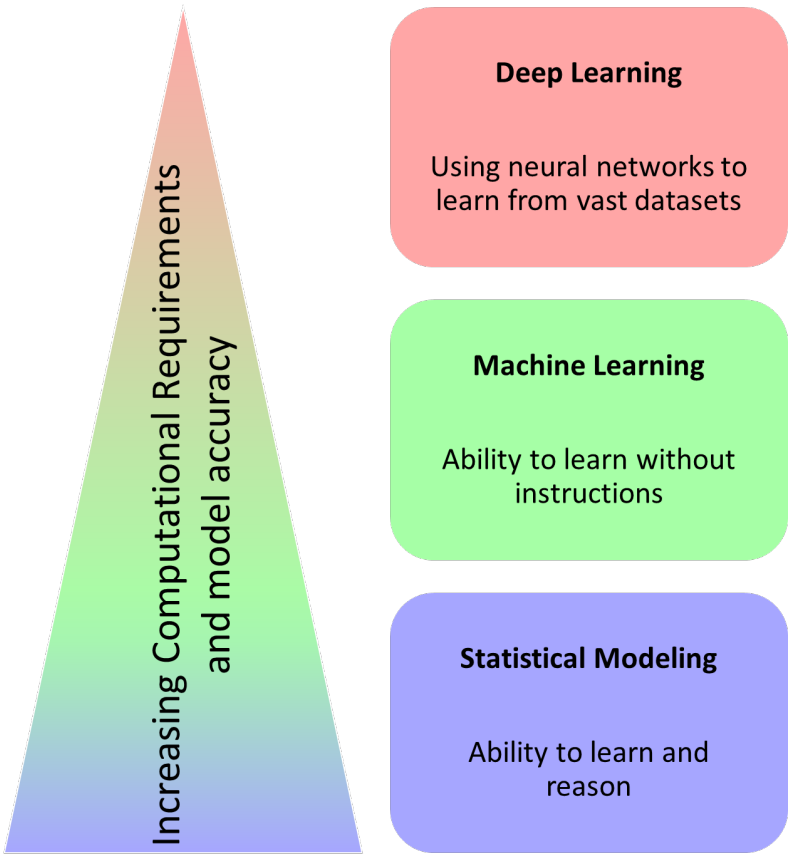


Fig. 1.1. An illustration of the various computational techniques used for data analysis, in association to their increasing computational requirements and model utility. While **statistical modeling** is one of the first class of methodologies to have been developed that were given the moniker of *artificial intelligence*, strides in this research area has given rise to a multitude of techniques which can be broadly classified as **machine learning**. **Deep learning** is a unique subset of machine learning techniques because of the way it learns from training data.

unseen data samples based on engineered or hand-crafted "features" in the dataset. This learning process involves the use of algorithms that can learn and improve from experience. Lastly, **deep learning** (*DL*) is a special category within *ML*. It utilizes artificial neural networks to automatically tune the features used for training a model. This automatic feature tuning is what sets *DL* apart, allowing it to deliver superior performance in a myriad of tasks such as image and speech recognition. Together, these three fields form the backbone of data analysis and predictive modeling tasks in computational healthcare.

One of the earliest mechanisms that enabled computational analysis of healthcare imaging data was defining characteristic "features" in images via the use of **radiomics** [**haralick1973textural**]. Radiomic features describe the extraction of quantitative feature attributes that capture textural and morphological characteristics within and outside specific regions of interest, such as healthy white matter or necrotic tumor tissue in *MRI* [47, 116]. They can quantify shape, size, texture, and intensity of tissue and lesions in an image, and can reveal information that cannot be clearly discerned by the naked eye [**traverso2018repeatability**]. They can help doctors garner increased insight from images, and thus help diagnose diseases, predict patient out-

comes, and plan treatments based on the quantitative characteristics of each patient's imaging characteristics. By combining these features with other types of health data, such as genetics, histology, or clinical records, physicians can obtain a more comprehensive understanding of the pathology and the patient [61, 62, 87].

Radiomic features have allowed algorithmic analysis of a large number of features from multiple images at the same time, thus enabling population or cohort-based analysis [91]. This, in turn, has facilitated the ability to generate *ML* models that correlate with clinical outcomes, such as diagnosis, prognosis, or treatment response [86], thereby furthering the goals of precision medicine. These models can be based on "more traditional" (*ML* algorithms [**khan2020review**, **kansara2020comparison**, **chowdhury2022prediction**] (such as support vector machines [**vapnik2006estimation**], random forests [**breiman2017classification**], so on) to perform regression or classification tasks, or clustering approaches that learn from the radiomic features and other relevant clinical data (such as clinical or genomic information) [24]. Another paradigm to process healthcare data, and especially imaging data, at scale, is via *DL* models, which have been applied in a variety of domains and demonstrated great potential. Some of them are quantum physics [17], image registration [21, 57], predictive modeling [5], semantic segmentation [36, 51, 55, 102], segmentation of regions of interest (such as tumors) in medical images [12, 13, 14, 60, 65, 96, 103], medical landmark detection [37, 121], among many others [83, 98, 99].

One of the key differences between *DL* and traditional *ML* lies in their approach to how they use relevant features for the problem at hand. *DL* utilizes multiple filtering layers, with each layer learning distinct attributes from the data. These layers progressively extract higher-level features, allowing for increased specialization as we move deeper into the network [56]. The filters within these layers adjust themselves during the entire training process, fine-tuning their weights to optimize performance, eventually resulting in a cascade of increasingly specialized feature representations. In contrast, traditional (*ML*) approaches rely on either hand-crafted radiomic features [125] or features extracted from pre-trained *DL* models or a combination of both to adapt a predefined algorithms to the problem or task at hand. Statistical modeling can be used to perform feature selection to select the most appropriate features for the specific task [**sun2019comparison**].

Both *ML* and *DL* have their unique merits and drawbacks. *DL* is computationally expensive but excels in tasks such as image recognition, natural language processing, and complex pattern recognition. On the other hand, traditional (*ML*) and statistical modeling techniques can be tailored to focus on specific clinically relevant features, making them interpretable and easier to understand. An illustration of the relationship of these methods can be seen in Figure 1.1 and Figure 1.2.

1.3 Criteria for Meaningful Computational Analysis

Reproducibility and stability in computational healthcare research have been a long-standing concern. This is especially true for *DL*, where the training process tends to be highly stochastic in nature [34], and the training mechanism for a model is usually not very intuitive and explainable [41]. This is especially egregious in the healthcare domain, where the **robustness,**

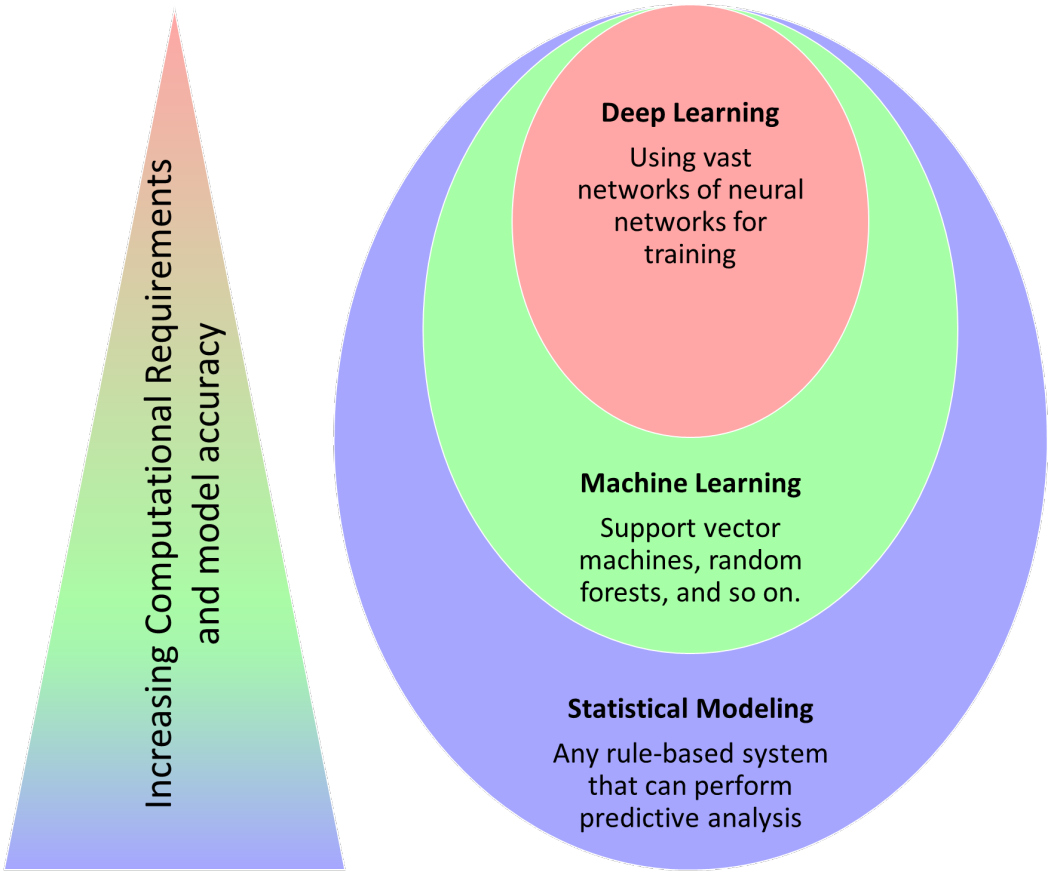


Fig. 1.2. An illustration of the inter-relationship between the various computational techniques used for data analysis. **Statistical modeling** can be interpreted as the foundational technique, which describes any method that can perform rule-based predictions. **Machine learning** techniques are more specialized techniques that require prior information from the data to "learn". **Deep learning** is a unique subset of machine learning techniques because it uses vast networks of neural networks to automatically tune the features of the data it is provided with.

stability, and reproducibility of the model is of prime importance with regards to regulatory bodies [8, 113], and is necessary to improve standard of care [18, 63]. As illustrated in Figure 1.3, the **ideal** algorithm should balance all these fundamental principles:

1. **Reproducibility** is the ability of an *AI* model to generate the same results when it is run multiple times with the same data and parameters. For example, a reproducible (*ML*) model should produce the same accuracy and predictions when it is trained and tested on the same dataset with the same hyperparameters and random seeds. Reproducibility can be ensured by following best practices in software development, such as documentation, version control, unit testing, and code review [18, 63].

2. **Stability** or **Generalizability** is the property of an *AI* model to produce consistent and reliable predictions while preserving the "model utility" (defined as the performance of the model by taking the specific task into account [67]) when the input data is slightly perturbed or disturbed. For example, a stable image classification model should not

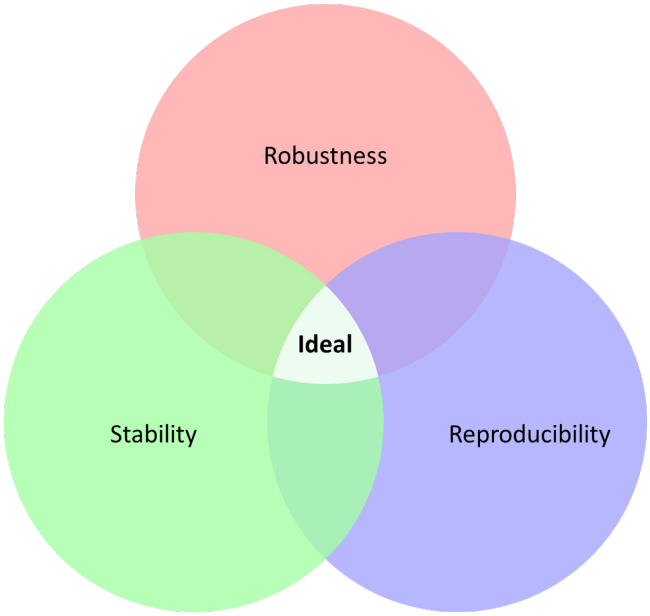


Fig. 1.3. The **ideal** (*ML*) model is one that is able to perform with an acceptable level of model utility (i.e., *stable*) regardless of any perturbations or issues in the data (i.e., *robust*), and giving the same outputs every single time (i.e., *reproducible*).

change its prediction drastically when a pixel is changed or added. A stable regression model should not produce very different outputs when the input values are rounded or truncated. Stability can be measured by using various metrics and techniques, such as sensitivity analysis, Lipschitz continuity, and smoothness [107].

3. **Robustness** is the ability of an *AI* model to maintain its performance and accuracy when faced with changes in the data, such as noise, outliers, or adversarial attacks. For example, a robust face recognition model should be able to recognize faces regardless of lighting, angle, or occlusion. For example, a robust natural language processing model should be able to understand sentences regardless of spelling, grammar, or slang. Robustness can be improved by using various methods and approaches in different phases of the (*ML*) pipeline, such as data preprocessing, augmentation, regularization, and adversarial training [81, 106].

In addition to a model being **reproducible** and **robust**, it needs to ensure **stability** by being generalizable to new data to ensure that the promise of precision medicine gets fulfilled. This, in turn, is inextricably linked to data diversity, both in terms of the imaging protocols used and the demographics of the patient populations included in the training dataset. It has been shown that models trained on diverse data are more robust and stable [72], leading to predictions that are consistent across different scanners, protocols, modalities, and patient demographic profiles. This characteristic is crucial for clinical translation of the model, as it allows for the development of tools that can be confidently applied in real-world settings and have a notable impact on improving the standard of care.

Achieving diversity in the training data is the first step towards developing a reproducible, robust, and stable model. For precision medicine to truly flourish, we need further avenues to

obtain open and accessible data. **Open data** refers to data that is freely available for anyone to use, analyze, and build upon, with as few restrictions as possible [109]. This promotes collaboration and innovation, allowing researchers from around the world to work together to develop new and improved diagnostic and therapeutic approaches. **Accessible data**, on the other hand, refers to data that is readily available to researchers, regardless of their technical expertise or access to resources. This can involve providing data in user-friendly formats, accompanied by clear documentation and tutorials. Accessible data empowers a wider range of researchers to contribute to the field, leading to a more diverse and inclusive research landscape [26]. This is essential for several reasons, such as:

- Open data ensures that researchers can avoid repeating the same experiments and focus on building upon previous work, which leads to reduction of redundancy and acceleration of research.

- Open data allows independent verification and validation of research findings, which builds trust in the scientific process and encourages wider acceptance of precision medicine approaches.

- Open data enables computational researchers to develop and test new algorithms and tools on a larger scale, leading to more robust and generalizable solutions.

- By making data accessible, it can be ensured that a much wider array of researchers have the opportunity to benefit from the advancements in precision medicine, regardless of their location or resources.

However, data accessibility is often hampered by restrictive licensing agreements. Data licensing can be complex and time-consuming, and it can impose unnecessary barriers to research. To truly realize the potential of precision medicine, we need to move towards more open and flexible data licensing models that encourage collaboration and knowledge sharing. In all these endeavors, we should keep the privacy of patients in mind [8, 113].

Technical Background

2

2.1 Software Nomenclature and Taxonomy

Within the landscape of computational research and software development, there are a lot of different approaches to designing and writing software tools, and each has a unique and crucial role in advancing open science within the community. At the foundational layer lie *libraries*, which provide the low-level building blocks and interact directly with the hardware systems. They can be interpreted as the "nuts and bolts" of computational research, and enable basic functionalities needed for further research, such as reading files or performing calculations. Although libraries offer access to raw power, their true scientific potential is unlocked through *toolkits*. These are higher-level abstractions built upon libraries, offering pre-packaged functionalities for common tasks while making easy functional or modular interfaces for developers. They can be interpreted as specialized instruments that save researchers' time and effort by providing pre-built solutions for various tasks such as data analysis, visualization, or matrix manipulations. Researchers can customize the appropriate toolkits for their own needs. However, when it comes to deployment (regardless of whether they are in a clinical setting or not), specific easy-to-use user interfaces (either graphical or command-line) are required. When toolkits are given this capability, they can be termed *applications*. They can be interpreted as the microscopes or spectrometers of the digital world, allowing researchers to directly interact with their data and perform complex analyses. Finally, there is a class of hybrid software known as *frameworks*, which straddle the line between toolkits and applications, offering both pre-built components and the flexibility to customize them. Frameworks provide a structured environment for developers to build custom applications, often focusing on specific research domains or methodologies. It is important to note that these are just general categories, and there can be significant overlap between them. Some libraries might be quite complex, while some applications might be highly customizable. An illustration of these concepts can be seen in Figure 2.1.

2.2 Medical Image Analysis Based on Radiomics

The digitization of healthcare imaging has opened up the ability of large-scale computational analyses techniques to be applied to these datasets, thus unlocking previously unimaginable avenues of research. By leveraging *radiomic analysis* and the support of dedicated software applications developed alongside advancements in computational imaging, the analysis of medical images has seen a great deal of research activity. Early on, platforms like the Medical Imaging Interaction Toolkit (*MITK*) [115] and 3D Slicer [52] paved the way, offering user-friendly interfaces and robust radiomic pipelines to democratize the technology for non-computational researchers. These were followed by tools like the Cancer Imaging Phenomics

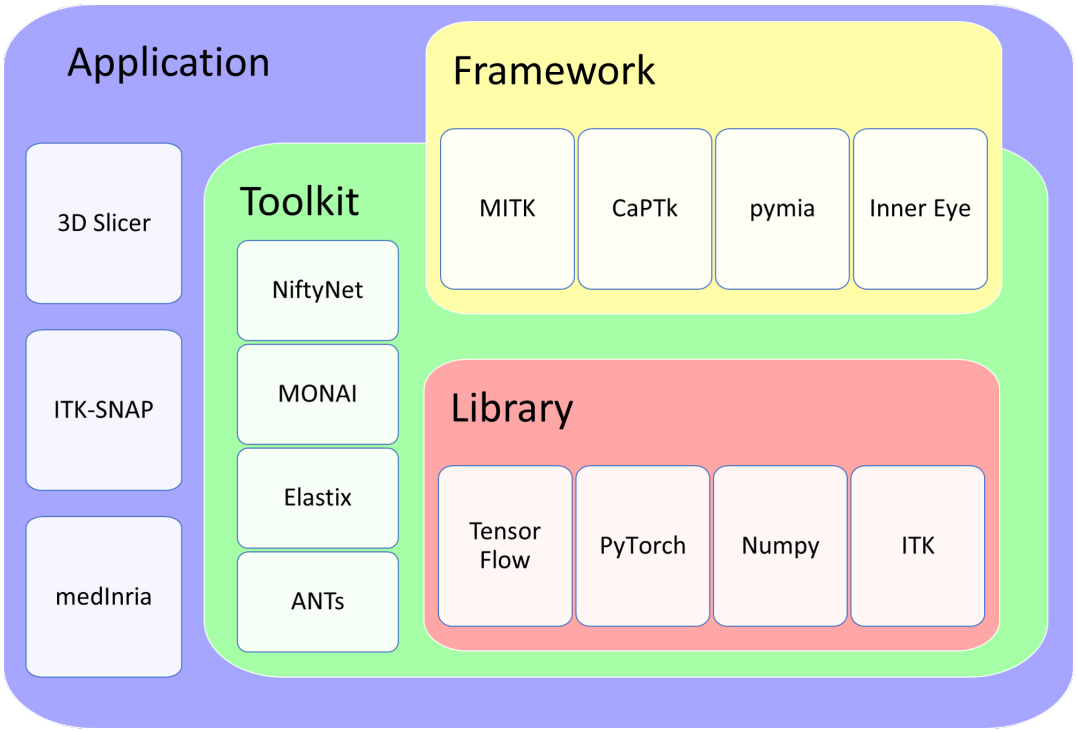


Fig. 2.1. Illustration of the various software terminologies used in open science and their inter-relationships. *Libraries* provide access to low-level machine functionality. *Toolkits* provide abstraction to libraries and general-purpose functionalities to improve the developer experience. *Applications* focus on the end-user, with powerful user interfaces which can be either command line or graphical. *Frameworks* straddle the line between toolkits and application.

Toolkit (*CaPTk*) [27, 76], which further streamlined the process with customizable parameter configurations for radiomic feature extraction which could be applied in a standardized manner across entire patient cohorts while respecting intrinsic imaging characteristics (such as voxel resolution). While these tools leveraged easy-to-use graphical interfaces to expand the usage of radiomics to non-computational experts, the researchers with a strong computational background favored tools such as PyRadiomics [110], Medical Image Radiomics Processor [111], and MATLAB-based radiomics [124] that are purely based on the command line, thus making their application across population cohorts easier. The ability of the same radiomic feature parameter configuration to run across entire cohorts of subjects paved the way for rigorous quantitative analysis of various types of variability analyses [78], advancing the field's reliability. Meanwhile, initiatives such as the Image Biomarker Standardization Initiative (IBSI) [125] emerged, addressing the crucial issue of standardization. IBSI employed both generic "synthetic" images and real medical data to establish standardized radiomic parameters and computational configurations. Additional dedicated efforts focused specifically on x-ray computed tomography (*CT*) data [64], further solidifying standardization within the field. With these studies, a standardized nomenclature for radiomic features was established, which could then be used across the community to give precise definitions and descriptions of their radiomic extraction pipelines, thus showcasing their applicability for potential in clinical applications.

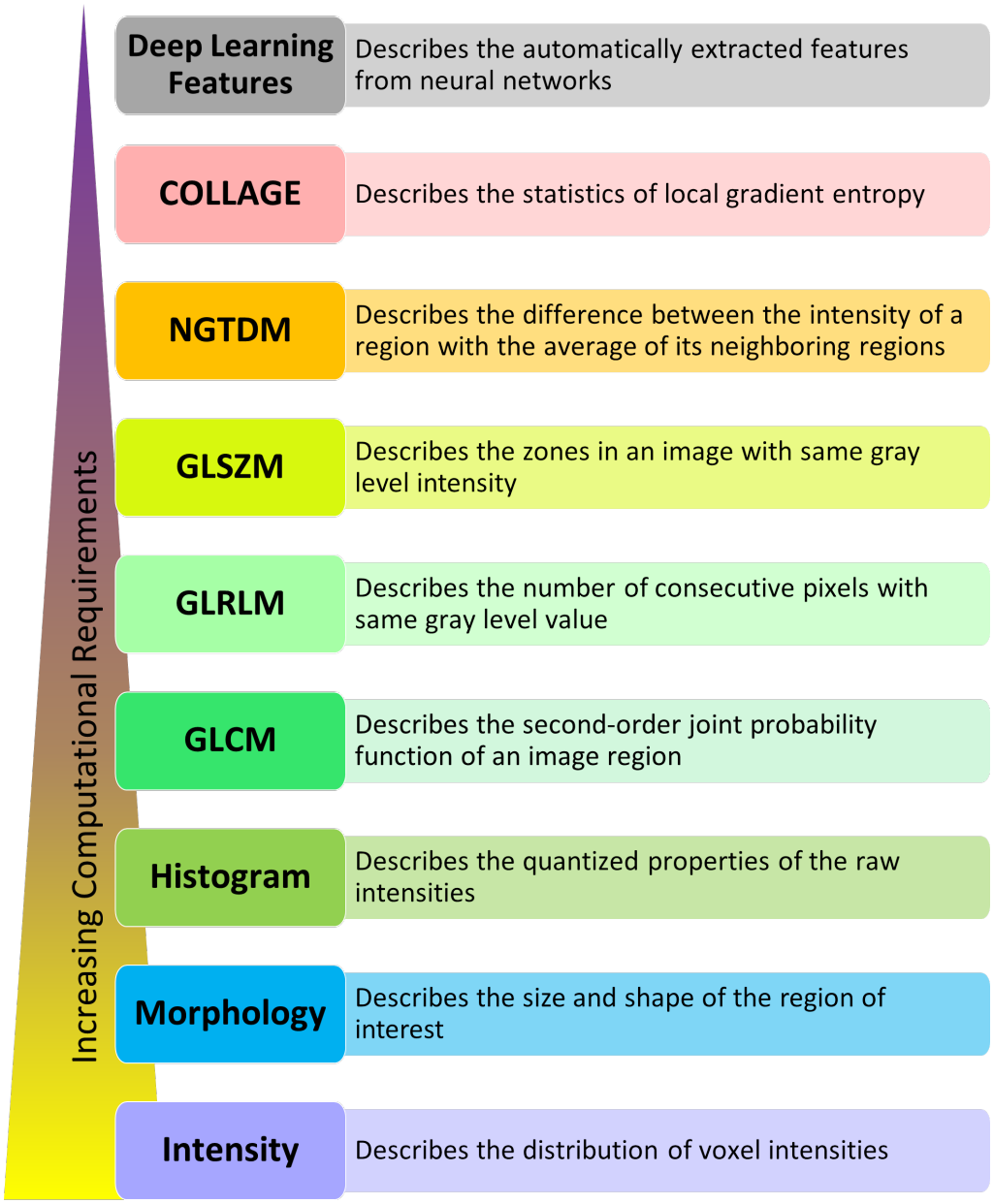


Fig. 2.2. An illustration of the 8 different radiomic feature families along with deep learning based features that are considered in this thesis, in association to their increasing computational requirements [84, 125]: intensity-based statistical features (20 descriptors), morphological features (19 descriptors), histogram features (135 descriptors), Gray-level co-occurrence matrix (*GLCM*) (6 descriptors), Gray-level run-length matrix (*GLRLM*) (16 descriptors), Gray-level size zone matrix (*GLSZM*) (16 descriptors), Neighborhood gray tone difference matrix (*NGTDM*) (5 descriptors), and Co-occurrence of Local Anisotropic Gradient Orientations (*COLLAGE*) (52 descriptors). The number of deep learning based features vary with the type of network and its output channels.

2.3 Medical Image Analysis Based on Deep Learning

As open-source *DL* libraries such as TensorFlow (2016) [1] and PyTorch (2019) [71] started entering the sphere of computational imaging, the performance and versatility of algorithms started increasing dramatically, which was seen in virtually every community-driven computational challenge [14, 65, 74]. However, all *DL* algorithms were built as "one-off" solutions that targeted specific problems, and there was still no way to standardize their highly complex workflows, which is a critical first step towards ensuring *reproducibility* in algorithms. Not only are there significantly more parameters to be considered in comparison to traditional *ML* and statistical modeling, but the training and inference pipelines themselves tend to be more complicated since processing usually happens on image patches rather than full images because of resource limitations [38]. An additional parameter that increases the complexity is the fact that *DL* tended to be highly stochastic in general.

2.4 Considerations for Robust, Stable, and Reproducible Computational Analysis

There are various considerations that need to be taken into account during a computational study. Specifically for ones involving imaging, the effect of resolution needs to be understood and accounted for in a meaningful way. This is a necessity when conducting research that involves data from multiple scanner protocols [27, 76] to ensure that the extracted features are properly normalized. This effect has been illustrated in Figure 2.3. There are a few different mechanisms to take care of this, and each has its corresponding pros and cons, and this has been showcased in Table 2.1.

For studies leveraging *DL*, the considerations are a bit different. *DL* is a highly computationally expensive method to analyze data, and it requires large quantities of (ideally well-annotated) data for model training. This limits their applicability in studies which do not have access to significant quantities of well-curated annotated datasets [112]. This issue can be overcome by using weights from models trained on public datasets like ImageNet [28], which is known as **transfer learning**. This approach potentially reduces convergence time and concludes the training at a superior state, while utilizing otherwise insufficient data [39, 119]. While this technique had been used to process healthcare images in 2D (usually by processing individual slices), a novel method of performing convolutions, called **axial-coronal-sagittal convolutions** (*ACSConv*) [117], allowed transfer learning of model weights trained in 2D to be used for 3D tasks. This has been illustrated in Figure 2.4.

Tab. 2.1. Various methods of mitigating the effects of resolution in a cohort.

Solution	Pros	Cons
Resample all images in a cohort to the "highest" resolution.	All images will be defined in the same physical space (i.e., images from Protocol 1 & 2 will both have 10 pixels). Extracted features will have consistency (bin count = 10). Post-resample quality control of the data will be possible.	Extrapolation would result in introduction of "new" data. This can be potentially mitigated by using linear or b-spline interpolation during resampling
Resample all images in a cohort to the "lowest" resolution.	All images will be defined in the same physical space (i.e., images from Protocol 1 & 2 will both have 1 pixels). Extracted features will have consistency (bin count = 1). Post-resample quality control of the data will be possible.	Loss of data fidelity.
Rescale features and process based on "real" resolution	Extracted features will have consistency (bin count will be normalized by resolution). Maintain data fidelity and no interpolation.	Difficult to implement. Not defined in any current radiomic standard.

2.5 Deployment of Computational Healthcare Software Tools

In the last couple of decades, there have been multiple efforts by the open scientific community towards packaging and distribution of applications that can handle end-to-end processing

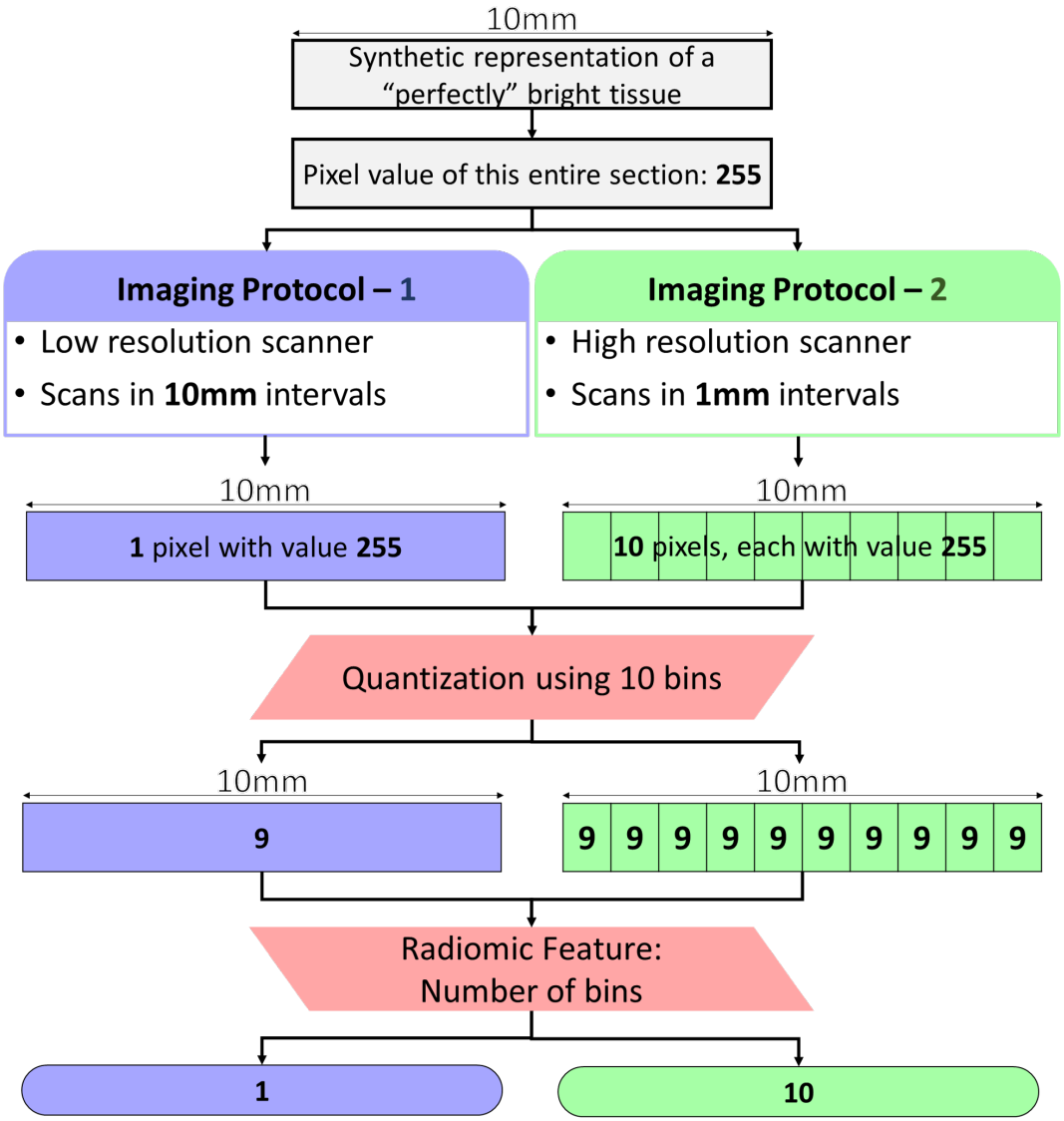


Fig. 2.3. An illustration of the effect of scanner resolutions on extracted features using a synthetic example and a simple imaging feature, i.e., the *histogram count*.

of healthcare data, as well as deployment of trained *ML* models. The applications that have specifically focused on non-*DL* based deployments have seen tremendous success in bringing the tremendous gains in the *ML* community to the clinical research field, such as the *MITK* [115], 3D Slicer [52], ITK-SNAP [120], and *CaPTk* [27, 76]. The *ML* models deployed with these applications have been lauded for their generalizability, but unfortunately fall short when it comes to competitive performance for specific tasks, where *DL* excels. However, *DL* models are notoriously difficult to deploy, since they usually require specialized hardware that can perform *DL* acceleration in order to make the computation more efficient. Numerous efforts have been put forth in the community to design *DL* toolkits using TensorFlow [1], such as NiftyNet [38], DeepNeuro [19], ANTsPyNet [108], and *DL*TK [80], as well as those written in PyTorch [71] pymia [48], InnerEye [70], and MONAI [23]. In addition to these, there are specialized toolkits and libraries that cater to specific workloads, such as segmentation [46, 73, 103], registration [35], or specialized imaging domains, such as PathML [94], and

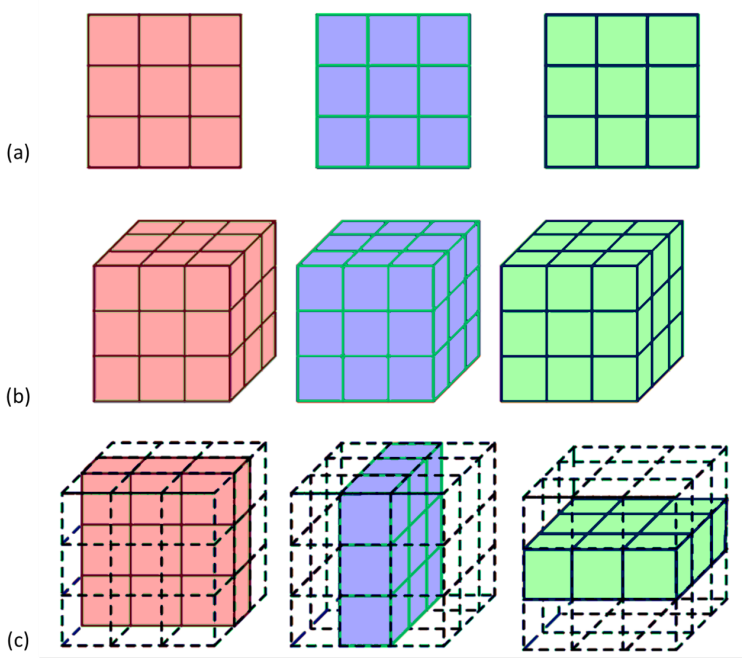


Fig. 2.4. Comparison of various types of convolution for 3D medical data: *(a)* represents native 2D convolutions, *(b)* represents native 3D convolutions, and *(c)* represents axial-sagittal-coronal convolutions [117].

Tissue Image Analytics Toolbox [82], that focus on data engineering and enabling *ML* in computational pathology. All these applications and toolkits have some specific drawbacks, such as they either *i)* showcase developer-focused tools targeting experienced members of the computational healthcare research community; *ii)* are difficult to understand and conceptualize by researchers without sufficient experience in the computational aspect of *DL*; *iii)* not provide enough simplistic application programming interfaces to make it easy for computational scientists to write their methods in a mechanism that allows them to be used on problems spanning across domains; *iv)* make it difficult to design training pipelines that are reproducible while being able to work across various problem domains; *v)* put the responsibility of training robust models on the user's knowledge and experience of dealing with the training mechanism and the dataset in question; *vi)* lack an easy to leverage application programming interface for both the training and inference portions of the pipeline that can work across various problem domains; or *vii)* do not provide acceptable level of explainability or interpretability for researchers to garner meaningful clinical insights into the training or inference process.

In a typical workflow of a healthcare research study, the structured steps below need to be considered (see Figure 2.6 for an illustration):

1. The research process begins with the *conceptualization and design*. This is a critical stage where the researcher identifies the specific use case and problem that the study aims to address. The researcher must have a clear understanding of the issue at hand and the context in which it exists. This understanding is crucial in shaping the direction of the research. In addition to identifying the problem, the researcher also needs to determine the data that will be used in the study. This involves a careful evaluation of the data already available and whether it is sufficient and relevant for the study. If the existing

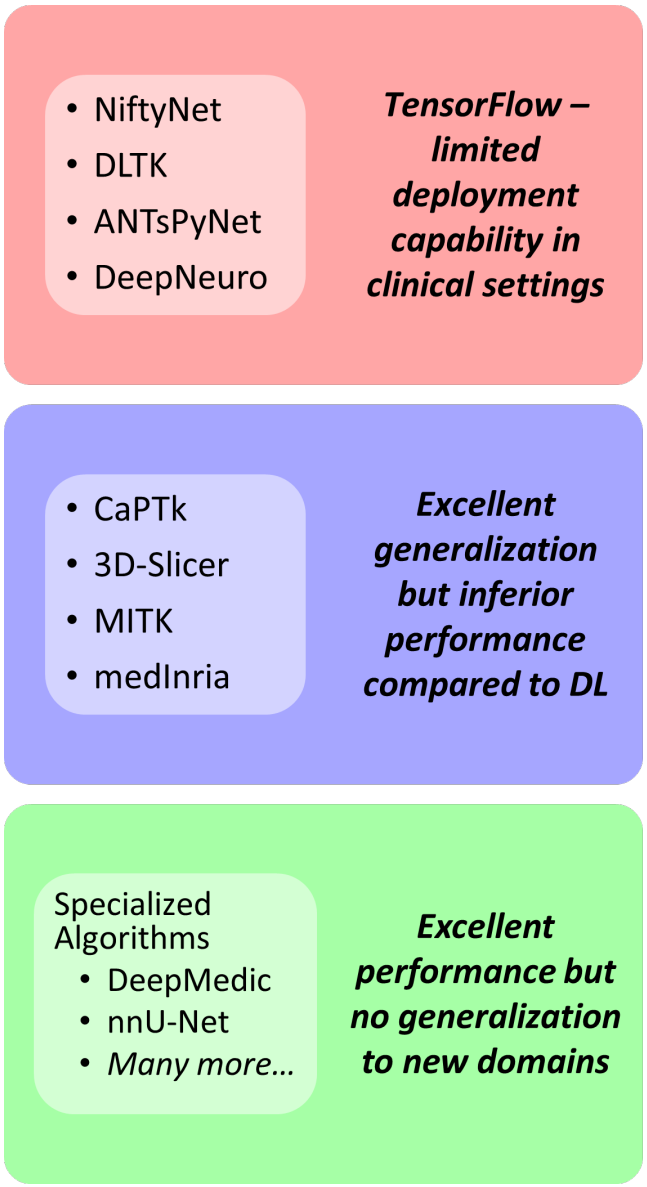


Fig. 2.5. Illustration of some of the previous applications, toolkits, and frameworks that paved the way for the development of the Generally Nuanced Deep Learning Framework (*GaNDLF*).

data is inadequate or irrelevant, the researcher may need to acquire new data. This could involve designing surveys or experiments, conducting interviews, or using other methods of data collection. Once the problem has been identified and the data requirements have been determined, the next step is to evaluate the idea for its potential impact. This involves assessing the potential significance of the research and its implications. The researcher needs to consider how the findings of the study could contribute to the existing body of knowledge, and how they could be applied in practical contexts. This evaluation helps to ensure that the research will be valuable and worthwhile. The final step in this phase is designing the entire experimental protocol. This involves planning the methods and procedures that will be used in the research. The researcher needs to decide on the research design, the sample size, the data collection methods, and the data analysis techniques. This plan serves as a roadmap for the research, guiding the researcher through the subsequent stages of the study. They provide a structured approach to research, helping to ensure that the study is well-planned, rigorous, and meaningful. This thought process is not a one-time activity but a continuous one, with the researcher constantly reflecting on and refining the research as it progresses. This iterative process helps to ensure the quality and integrity of the research, ultimately leading to more reliable and valid results.

2. The next phase in the research process is the *technical development* phase. This phase involves a series of complex and critical tasks that are essential for the successful execution of the research study. The first task in this phase is to design, formulate, and code the various data input/output (*I/O*) protocols. These protocols define how data will be read in into and written out from the computational system. The researcher needs to carefully design these protocols to ensure that they are efficient, reliable, and secure. This might involve choosing appropriate data formats, designing user interfaces for data input, and implementing error checking procedures to ensure data integrity. Next, the researcher needs to determine how the data curation should occur for the study. Data curation involves organizing, integrating, and maintaining the data to ensure its quality and usability. This might involve tasks such as data cleaning, data integration, and data annotation. The researcher needs to plan these tasks carefully to ensure that the curated data is accurate, consistent, and suitable for the study. Following this, the researcher needs to establish the preprocessing and harmonization protocols to be used. Preprocessing involves preparing the data for analysis, which might include tasks such as data cleaning, data transformation, and data normalization. Harmonization involves ensuring that the data is consistent and comparable across different sources or datasets. The researcher needs to carefully plan these protocols to ensure that the preprocessed and harmonized data is suitable for the subsequent analysis. Finally, the researcher needs to design and implement the actual *ML* algorithm. This involves choosing an appropriate *ML* model, training the model with the curated and preprocessed data, and testing the model to evaluate its performance. The researcher needs to carefully consider factors such as the complexity of the model, the computational resources required, and the interpretability of the model's outputs. All these tasks together form the technical development phase of the research process. They require a high level of technical expertise and careful planning to ensure that the research study is conducted efficiently and effectively. This phase is crucial for the success of the research study, as it directly impacts the quality of the research findings and the validity of the research conclusions.

3. The research process then proceeds to the *algorithmic evaluation phase*. This phase is crucial as it involves the assessment of the ML model's utility and the validation of the results obtained. One of the key tasks in this phase is ensuring correct cross-validation in the dataset [7]. Cross-validation is a statistical method used to estimate the skill of ML models. It involves partitioning the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. Correct cross-validation is essential to prevent data leakage, a common problem in ML where information from outside the training dataset is used to create the model. This can lead to overly optimistic performance estimates. Another important task in this phase is the choice of appropriate evaluation metrics for the task at hand [90]. The choice of metric depends on the specific objectives of the study. For example, if the task is a classification problem, metrics such as accuracy, precision, recall, F1 score, or area under the ROC curve might be appropriate. If the task is a regression problem, metrics such as mean squared error, root mean squared error, mean absolute error, or R squared might be used. The chosen metric should accurately reflect the goals of the study and should be robust to the specific characteristics of the data, such as class imbalance or outliers.

The conceptualization and design form the "thought process" for a research study, and the technical development and algorithmic evaluation can be thought of as the basis for determining "reproducibility" and "potential clinical translation" of the research study.

2.6 Performance Evaluation

Regardless of the complexity or ease (or lack) of interpretability of a *ML* algorithm, choosing appropriate metrics for performance evaluation is a critical final step [90]. There are well-accepted metrics in literature, and it is imperative for a study to be able to quantify the performance of all models both during and after training, and mechanisms to incorporate new validated recommendations [89] as needed. Specifically, for segmentation workloads, the **Dice Similarity Coefficient** (*DSC*) [123], and is mathematically represented in the Equation 2.1. *DSC* is a common metric used to evaluate the performance of segmentation workloads. It measures the extent of spatial overlap, while taking into account the intersection between the predicted label (PL) and the provided ground truth (GT), hence handles over- and under-segmentation.

$$DSC = \frac{2|GT \cap PL|}{|GT| + |PL|} \tag{2.1}$$

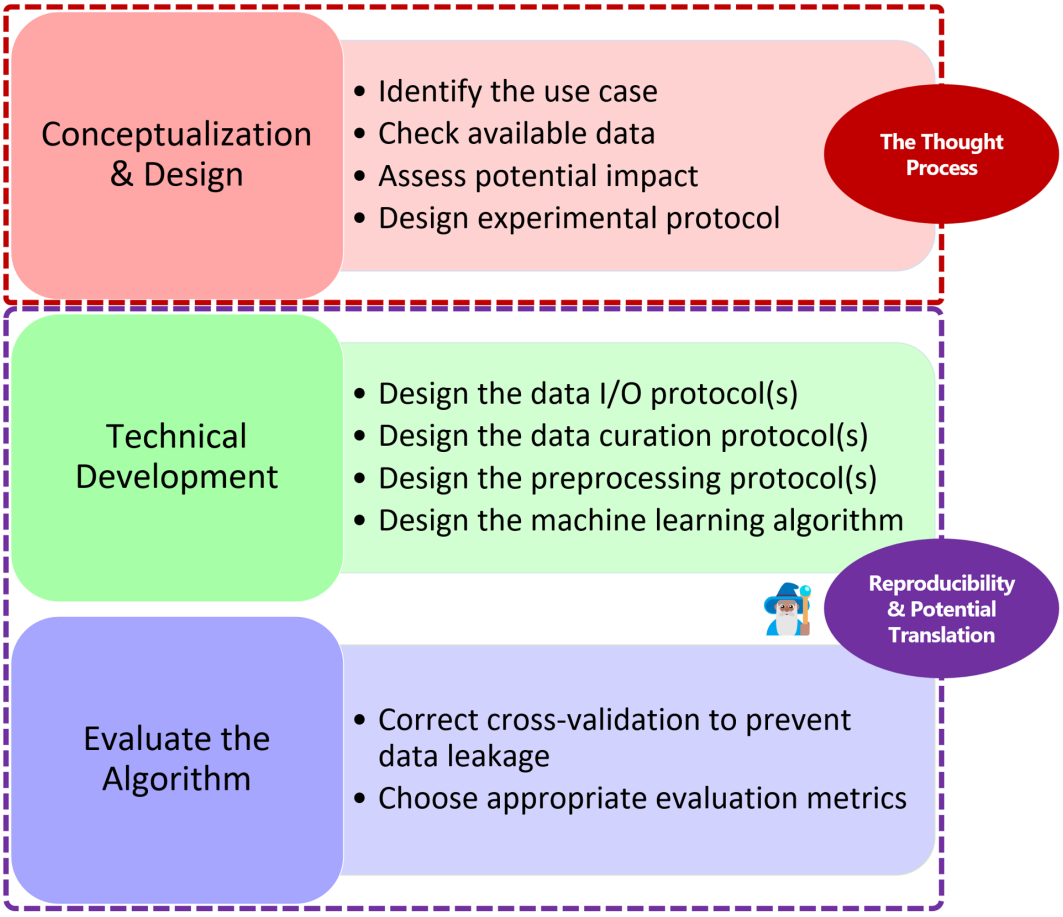


Fig. 2.6. Illustration of the multiple steps in a research project life-cycle [77]. Starting with the *conceptualization and design* of the study, researchers need to think about the *technical components and development*, followed by appropriate *evaluation of the algorithm*. The Generally Nuanced Deep Learning Framework (*GaNDLF*) was designed to help the technical development and algorithmic evaluation, thereby enabling *reproducibility* and potential *clinical translation*.

Additionally, the **Hausdorff Distance** (*Hausdorff*) [92] is a metric for segmentation workloads, and is mathematically represented in the Equation 2.2. This metric quantifies the distance between the boundaries of the ground truth labels against the predicted label. It is sensitive to local differences, as opposed to *DSC*, which represents a global measure of overlap. Specifically, this thesis uses the 95$^{\text{th}}$ percentile of the distance measure when referring to *Hausdorff*. *DSC* is useful when the size and shape of the object are important, whereas *Hausdorff* becomes critical when the accuracy of boundary localization is critical.

$$H_{95}(PL, GT) = \max \left\{ \underset{p \in PL}{P_{95\%}}\, d(p, GT), \underset{g \in GT}{P_{95\%}}\, d(g, PL) \right\} \tag{2.2}$$

where $d(x, Y) = \min_{y \in Y} ||x - y||$ is the distance of x to set Y.

Sensitivity (also known as the **true positive rate**) is the probability of a positive test result, given that the individual truly has the condition. It measures how well a test can identify true positives. It's mathematically represented by Equation 2.3.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.3}$$

where TP is the number of true positives, and FN is the number of false negatives when comparing each pixel of the PL with GT.

Specificity (also known as the **true negative rate**) is the probability of a negative test result, given that the individual truly does not have the condition. It measures how well a test can identify true negatives. It's mathematically represented by Equation 2.4.

$$Specificity = \frac{TN}{TN + FP} \tag{2.4}$$

Either $Sensitivity$ and $Specificity$ alone cannot adequately describe the picture during a *ML* task (specifically, for classification workloads) [90]. Thus, an added metric called **balanced accuracy** (*Acc*) [22] is considered, which can be useful for both binary and multi-class classification, and is defined the arithmetic mean of sensitivity and specificity (see Equation 2.5). This metric is especially useful when dealing with imbalanced data, i.e., when one of the target classes appears a lot more than the other [22].

$$Acc = \frac{Sensitivity + Specificity}{2} \tag{2.5}$$

3 Aims & Objectives

3.1 Main Aim

The primary objective of this thesis is to provide precise and easy-to-use computational healthcare software solutions that can effectively address the pressing demand for enhanced reproducibility, stability, and robustness in various clinical and computational research projects. These software solutions are designed to facilitate the analysis, interpretation, and visualization of complex biomedical data, as well as to support the development and validation of novel computational models and methods for healthcare applications.

3.2 Objectives

The objectives of this work are the following:

1. Implementation, systematic evaluation, and identification of computational imaging features, towards robust and reproducible research.

2. Development of a stable codebase that can support long-term monitoring of the validity of computational model utility.

3. Generation of baseline results using the provided data in a quick and reliable manner.

4. Reduce the amount of required data required to train robust models.

3.3 Contributions

The interdisciplinary nature of this study facilitates a dual perspective assessment of its academic contributions, primarily through its technical novelty involving the development and implementation of computational solutions, and secondarily through its potential clinical relevance. The technical novelty introduced in this study directly aligns with the predefined project objectives, enhancing the fulfillment of key research goals. These novel computational solutions, paramount to achieving the study's objectives, concurrently serve as foundational elements supporting the potential clinical value of its research outcomes. The relationship between the thesis' objectives, the technical novelty, and clinical relevance is visually illustrated in Figure 3.1, providing a comprehensive overview of the interplay among these critical components.

The key technical novelties of this thesis are the following:

1. Systematic evaluation of radiomics robust to segmentation variability across readers/sites.

2. Develop a "zero-/low-code" framework for computational healthcare model development.

3. Develop a framework incorporating "good" *ML* practices.

4. Develop a software tool that leverages the unprecedented size of computer vision 2D data and facilitates native 3D medical image analysis.

The potential clinical value of this thesis comprises the following:

1. Provide reproducible imaging features that can capture the relevant information from medical images and facilitate the analysis and interpretation of complex data.

2. Produce computational models that can generalize well to unseen data and maintain their utility and robustness across different settings.

3. Expedite the model development process by using efficient and scalable algorithms that can handle large and complex datasets, while being able to optimize the model parameters in a fast and reliable manner.

4. Lower the resource requirements for training (by leveraging transfer learning in a scalable manner), and inference (by optimizing the models so that they can run without any need for specialized hardware).

5. Enable researchers to develop *DL* models without coding by providing user-friendly and intuitive software tools that can automate the various steps in the model creation, evaluation, and deployment process.

3.4 Thesis Structure

Part I gives an introduction to the complete thesis, including both the motivation (Chapter 1) and the necessary technical background (Chapter 2). Chapter 1 contains information essential to introduce the reader to *i)* the role that imaging holds in modern healthcare (Section 1.1), *ii)* a historic overview of healthcare data analysis (Section 1.2), and *iii)* major criteria considered for clinically-impactful and clinically-relevant computational analysis (Section 1.3). Chapter 2 is then diving into the technical background necessary to follow the academic contributions of this thesis. The technical background specifically touches upon nomenclature and a taxonomy of computational healthcare software (Section 2.1) followed by the two major paradigms for medical image analysis (Sections 2.2-2.3). Considerations for robust, stable, and reproducible computational analysis are then described (Section 2.4) to allow for smoother deployment of

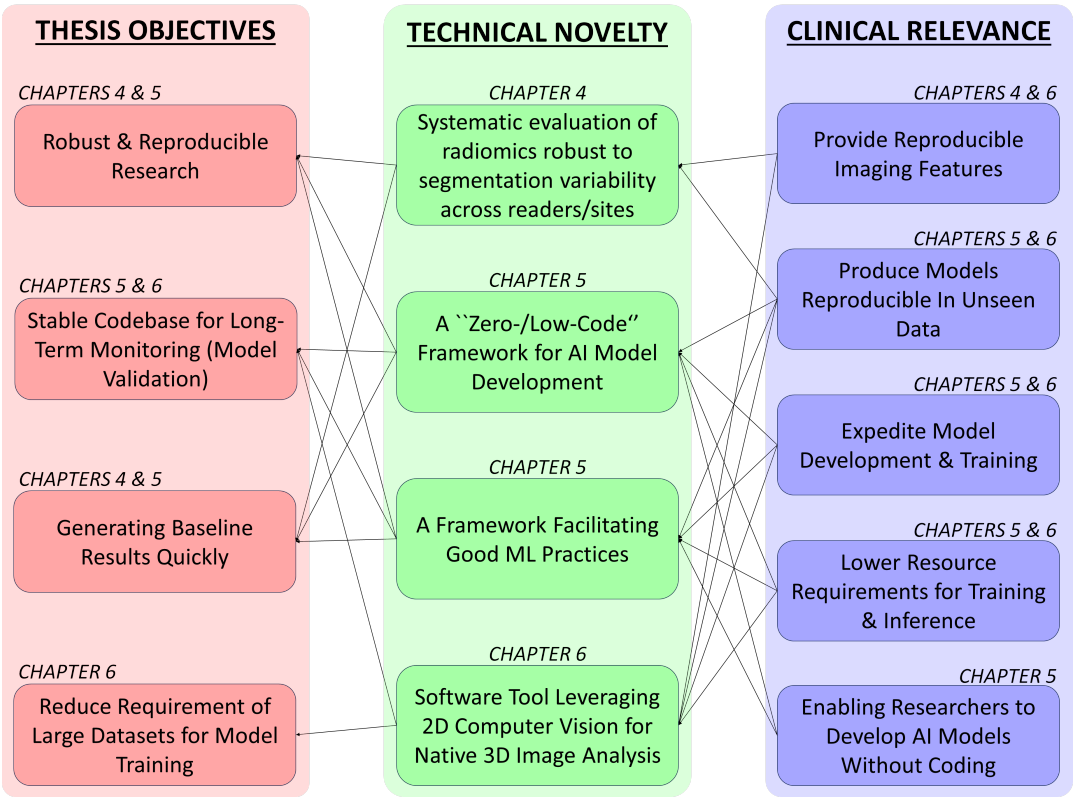


Fig. 3.1. Illustration of the interplay among thesis objectives, technical novelty, and clinical relevance. Arrows linking the initial two columns signify the objectives to which each novelty contributes. Likewise, arrows connecting the second and third columns indicate the technical innovations that underpin each clinical value.

these tools (Section 2.5), as well as appropriate metrics for their quantitative performance evaluation (Section 2.6).

Part II consists of three peer-reviewed publications [9, 77, 78], which make up the main academic contributions of this thesis. Each of these journal and conference papers is presented in a self-contained section, starting with a summary of the publication.

Part III then provides a summary of the thesis and discussion in Chapter 7 and outlook in Chapter 8.

Finally, Part IV is the Appendix, and provides the abstracts from publications [25, 27, 42, 50, 75, 76, 104] that are not directly relevant to the evaluation of this thesis, but act as a complement to the main publications by presenting initial findings and projects related to main themes of the thesis in Chapter A, and the complete bibliography in Chapter B.

Part II

Academic Contributions

4

Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset

Authors

Sarthak Pati*, Ruchika Verma*, Hamed Akbari, Michel Bilello, Virginia B Hill, Chiharu Sako, Ramon Correa, Niha Beig, Ludovic Venet, Siddhesh Thakur, Prashant Serai, Sung Min Ha, Geri D Blake, Russell Taki Shinohara, Pallavi Tiwari, Spyridon Bakas

* Joint first authors

Publication Information

Medical physics, volume: 47, issue: 12, pages: 6039-6052, year: 2020. DOI: $10.1002/mp.14556$.

Abstract

Purpose: The availability of radiographic magnetic resonance imaging (MRI) scans for the Ivy Glioblastoma Atlas Project (Ivy GAP) has opened up opportunities for development of radiomic markers for prognostic/predictive applications in glioblastoma (GBM). In this work, we address two critical challenges with regard to developing robust radiomic approaches: (a) the lack of availability of reliable segmentation labels for glioblastoma tumor sub-compartments (i.e., enhancing tumor, non-enhancing tumor core, peritumoral edematous/infiltrated tissue) and (b) identifying "reproducible" radiomic features that are robust to segmentation variability across readers/sites. **Acquisition and validation methods:** From TCIA's Ivy GAP cohort, we obtained a paired set (n = 31) of expert annotations approved by two board-certified neuroradiologists at the Hospital of the University of Pennsylvania (UPenn) and at Case Western Reserve University (CWRU). For these studies, we performed a reproducibility study that assessed the variability in (a) segmentation labels and (b) radiomic features, between these paired annotations. The radiomic variability was assessed on a comprehensive panel of 11 700 radiomic features including intensity, volumetric, morphologic, histogram-based, and textural parameters, extracted for each of the paired sets of annotations. Our results

demonstrated (a) a high level of inter-rater agreement (median value of DICE ≥ 0.8 for all sub-compartments), and (b) $\approx 24\%$ of the extracted radiomic features being highly correlated (based on Spearman's rank correlation coefficient) to annotation variations. These robust features largely belonged to morphology (describing shape characteristics), intensity (capturing intensity profile statistics), and COLLAGE (capturing heterogeneity in gradient orientations) feature families. **Data format and usage notes:** We make publicly available on TCIA's Analysis Results Directory (`https://doi.org/10.7937/9j41-7d44`), the complete set of (a) multi-institutional expert annotations for the tumor sub-compartments, (b) 11 700 radiomic features, and (c) the associated reproducibility meta-analysis. **Potential applications:** The annotations and the associated meta-data for Ivy GAP are released with the purpose of enabling researchers toward developing image-based biomarkers for prognostic/predictive applications in GBM.

Contributions of S.P.

Study conceptualization, algorithm development and implementation, data acquisition & organization, interpretation of results, and writing & editing of manuscript.

Copyright

Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset

Sarthak Pati*
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA
Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Ruchika Verma*
Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106, USA

Hamed Akbari
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA
Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Michel Bilello
Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Virginia B. Hill
Department of Radiology, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA

Chiharu Sako
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA
Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Ramon Correa and Niha Beig
Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106, USA

Ludovic Venet and Siddhesh Thakur
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA

Prashant Serai
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA
Department of Computer Science and Engineering, The Ohio State University, OH 43210, USA

Sung Min Ha
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA

Geri D. Blake
University of Arkansas for Medical Sciences, Little Rock, AR, USA

Russell Taki Shinohara
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA
Penn Statistical Imaging and Visualization Endeavor (PennSIVE), University of Pennsylvania, Philadelphia, PA 19104, USA

Pallavi Tiwari[a)†]
Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH 44106, USA

Spyridon Bakas[‡]
Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA 19104, USA
Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

(Received 14 April 2020; revised 26 July 2020; accepted for publication 26 August 2020; published 4 December 2020)

Purpose: The availability of radiographic magnetic resonance imaging (MRI) scans for the Ivy Glioblastoma Atlas Project (Ivy GAP) has opened up opportunities for development of radiomic markers for prognostic/predictive applications in glioblastoma (GBM). In this work, we address two critical challenges with regard to developing robust radiomic approaches: (a) the lack of availability of reliable segmentation labels for glioblastoma tumor sub-compartments (i.e., enhancing tumor, non-enhancing tumor core, peritumoral edematous/infiltrated tissue) and (b) identifying "reproducible" radiomic features that are robust to segmentation variability across readers/sites.
Acquisition and validation methods: From TCIA's Ivy GAP cohort, we obtained a paired set (n = 31) of expert annotations approved by two board-certified neuroradiologists at the Hospital of the University of Pennsylvania (UPenn) and at Case Western Reserve University (CWRU). For these studies, we performed a reproducibility study that assessed the variability in (a) segmentation labels

and (b) radiomic features, between these paired annotations. The radiomic variability was assessed on a comprehensive panel of 11 700 radiomic features including intensity, volumetric, morphologic, histogram-based, and textural parameters, extracted for each of the paired sets of annotations. Our results demonstrated (a) a high level of inter-rater agreement (median value of DICE ≥0.8 for all sub-compartments), and (b) ≈24% of the extracted radiomic features being highly correlated (based on Spearman's rank correlation coefficient) to annotation variations. These robust features largely belonged to morphology (describing shape characteristics), intensity (capturing intensity profile statistics), and COLLAGE (capturing heterogeneity in gradient orientations) feature families.

Data format and usage notes: We make publicly available on TCIA's Analysis Results Directory (https://doi.org/10.7937/9j41-7d44), the complete set of (a) multi-institutional expert annotations for the tumor sub-compartments, (b) 11 700 radiomic features, and (c) the associated reproducibility meta-analysis.

Potential applications: The annotations and the associated meta-data for Ivy GAP are released with the purpose of enabling researchers toward developing image-based biomarkers for prognostic/predictive applications in GBM. © *2020 American Association of Physicists in Medicine* [https://doi.org/10.1002/mp.14556]

Key words: glioblastoma, IvyGAP, MRI, radiomics, reproducibility, segmentation

1. INTRODUCTION

Glioblastoma (GBM) is the most aggressive and heterogeneous brain tumor. Despite multimodal treatment consisting of maximal safe surgical resection, radiation, and chemotherapy, median survival has only slightly improved to approximately 15 months, with less than 10% of patients surviving for over 5 yr.[1] This poor prognosis is largely on account of the underlying disease heterogeneity inherent in GBM tumors, which ultimately leads to treatment resistance, and thus dismal patient outcomes.

Radiographic imaging (i.e., magnetic resonance imaging (MRI)) is the modality of choice for routine clinical diagnosis and response assessment in GBM. Recently, computational analysis of these routine MRI scans, also known as *radiomics*, has enabled the extraction of quantitative feature attributes that capture textural and morphologic diversity,[2-4] within and outside the enhancing GBM tumor. These radiomic attributes describe subvisual cues reflecting the underlying biological processes of the tumor and its microenvironment, which otherwise are not visually discernible. Radiomic analysis in GBM has also greatly benefited from the availability of large multi-institutional publicly available data repositories, such as The Cancer Imaging Archive (TCIA)[5] with its Ivy Glioblastoma Atlas Project (Ivy GAP) collection.[6,7] These rich anonymized data repositories have enabled research groups to develop imaging phenotypes toward identifying tumor molecular characteristics,[8-11] predicting overall survival[12-15], and progression-free survival,[16] as well as the location of recurrence[17] and response to chemotherapy.[18] These radiomic approaches have involved capturing radiomic attributes from different tumor sub-compartments including non-enhancing tumor core (NET), enhancing tumor (ET), and peritumoral edematous/infiltrated (ED) regions obtained from multiparametric MRI (mpMRI) scans including native (T1) and gadolinium-enhanced T1-weighted (T1Gd), T2-weighted (T2), and T2-weighted-Fluid-Attenuated Inversion Recovery (FLAIR) scans, for tumor characterization.

However, in order to leverage multi-institutional repositories, such as TCIA's Ivy GAP, for development of robust radiomic approaches, two key challenges need to be carefully accounted for. First is the lack of availability of reliable segmentation labels of the different GBM tumor sub-compartments (NET, ET, ED).[19,20] Lesion segmentation is a foremost step for downstream radiomic analysis. However, obtaining reliable annotations is a manual, tedious, and time-consuming process. While efforts to make expert-annotated segmentation labels publicly available for other TCIA collections have previously been undertaken by our group,[21] such labels and the associated metadata are currently missing for the Ivy GAP collection. The second challenge is to account for the variability in radiomic features across segmentation labels obtained from different experts/institutions. Along with radiomic variability with respect to image acquisition protocols, and reconstruction kernels, the variability in radiomic features with respect to segmentation is well recognized in the field.[22] While a few studies have recently explored the issue of segmentation variability in radiomic analysis for the TCGA-GBM cohort,[23-25] to our knowledge, none of these studies have comprehensively explored the reproducibility of radiomic features in the context of multi-institutional paired expert annotations.

Toward addressing these challenges, in this work, we have three objectives. First, we investigate the variability in segmentation labels signed off by two experienced board-certified neuroradiologists (M.B. and V.B.H.) performed at two different institutions [University of Pennsylvania (UPenn), and Case Western Reserve University (CWRU), respectively] for the publicly available Ivy GAP collection. Second, we seek to investigate the reproducibility of radiomic features across the set of segmentation labels obtained from the two institutions (CWRU, UPenn). Lastly, the segmentation labels for the three tumor sub-compartments (NET, ET, ED), the

corresponding subcompartment-specific radiomic features (including intensity, volumetric, morphologic, histogram-based, textural, and COLLAGE), as well as the associated metadata collected as a part of this study are made publicly available through the TCIA Analysis Results portal (https://doi.org/10.7937/9j41-7d44).[26] Our overarching purpose is to (a) provide an online resource of multi-institutional paired segmentation labels for evaluation of segmentation-variability for the publicly available Ivy GAP cohort as well as (b) enable imaging and non-imaging researchers to be able to leverage the Ivy GAP cohort for development of robust and reproducible computational approaches for GBM characterization.

2. ACQUISITION AND VALIDATION METHODS

2.A. Data description

The Ivy Glioblastoma Atlas Project (Ivy GAP)[6,7] is a freely accessible online data resource, comprising a comprehensive cohort of radiological scans (i.e., MRI, CT), digitized tissue pathology slides, and corresponding transcriptomic data of 41 GBM patients.[6] This data collection is a collaborative effort between the *Ben and Catherine Ivy Foundation*, the *Allen Institute for Brain Science*, and the *Ben and Catherine Ivy Center for Advanced Brain Tumor Treatment*. The radiographic scans for Ivy GAP are available through TCIA (wiki.cancerimagingarchive.net/display/Public/Ivy+GAP), the RNA sequencing data, *in situ* hybridization, and digitized histology slides, along with corresponding anatomic annotations are available through the Allen Institute (glioblastoma.alleninstitute.org), while the genomic and clinical data are available through the Swedish Institute (ivygap.org).

The retrospectively collected 41 subjects as a part of the Ivy GAP collection were triaged in our work to a total of $n = 31$ subjects following the inclusion criteria that comprised the availability of (a) the four structural mpMRI scans, that is, T1, T1Gd, T2, and FLAIR, and (b) baseline preoperative timepoint of acquisition (i.e., prior to any instrumentation). We further excluded two subjects (i.e., W32 and W42) on account of obvious registration failures, as illustrated in Fig. 10(b). Finally, one subject (i.e., W50) was also excluded from the analysis of radiomic feature robustness due to an observed disagreement across the two expert readers (M.B and V.B.H) for the annotations corresponding to tumor core.

2.B. Preprocessing

The four structural baseline pre-operative mpMRI protocols, that is, T1Gd, T1, T2, and T2-FLAIR (FLAIR) were downloaded from TCIA in DICOM format and converted to the NIfTI format. Different preprocessing pipelines were followed at each institution (UPenn, CWRU), as shown in (Fig.1) described below.

Preprocessing at UPenn. All four modalities were first placed in a common orientation (the chosen orientation is "LPS" in the radiological convention, which is the same as "RAI" in the neurological convention). Then, to ensure cross-subject consistency, the T1Gd scan for every subject was registered to the SRI24 anatomical atlas space (www.nitrc.org/projects/sri24).[27] To facilitate registration, the T1Gd scan was first bias-corrected using the N4 Bias correction method[28] from ITK, using the Cancer Imaging Phenomics Toolkit (CaPTk),[29] and then registered to the SRI24 space using the Greedy registration framework[30] (https://github.com/pyushkevich/greedy, available in CaPTk and ITK-SNAP[31]). The generated transformation was then applied to the original T1Gd scan (i.e., prior to bias correction) to ensure minimal loss of signal. Bias correction was not included in the pre-processing pipeline at UPenn since the group previously reported that this process obliterates the MRI signal, particularly that of the FLAIR modality, and may have a negative impact on the downstream segmentation.[21] Subsequently, the remaining scans of each subject were registered to the transformed T1Gd scan resulting in co-registered MRI volumes of $1\,mm^3$ isotropic resolution in the SRI space. The brain was then extracted from all co-registered scans using a pretrained DeepMedic model, available through CaPTk,[32] and the resulting brainmask was manual revised when needed ensuring that the complete abnormal hyper-intense FLAIR signal was always included within the brainmask.

Pre-processing at CWRU. Registration of the T1Gd MRI scan of each Ivy GAP subject to the Montreal Neurological Institute (MNI - http://brainmap.org/training/BrettTransform.html) $1\,mm^3$ isotropic brain atlas[33] was performed using three-dimensional (3D) rigid and affine transformation via 3D Slicer 4.8.[34] Furthermore, to account for the resolution variability across studies from across protocols, the T1, T2, and FLAIR MRI scans were co-registered with the registered T1Gd sequence to ensure all MRI sequences are isotropic with $1\,mm^3$ dimensions. Following registration, the Swiss skull stripper[35] module of 3D slicer was used to strip the skull across the three MRI protocols (T1Gd, T2, and FLAIR). Every skull stripped MRI scan was corrected for bias field inhomogeneity using N4 bias-correction module available in 3D slicer.[28]

2.C. Segmentation of tumor sub-compartments

All the tumors included in the Ivy GAP data were segmented at UPenn (M.B) and CWRU (V.B.H) following a consistent annotation protocol as defined by the International Brain Tumor Segmentation (BraTS) Challenge.[19-21] The segmentation labels were performed/approved by two expert board-certified neuroradiologists with over 10 yr of experience. The tumor subcompartment labels comprised the ET, NET, and ED. *ET* is radiographically defined by the hyperintense signal in T1Gd scans not only when compared to T1, but also when compared to "healthy" white matter in T1Gd. *NET* is typically defined radiographically by hypointense scans in T1-Gd scans when compared to their corresponding areas in the T1 scan. The combination of *ET* and *NET* describes the bulk of the tumor, which is what is typically

resected, and here onwards defined as the tumor core (*TC*). Beyond the boundaries of *TC*, the complete extent of the disease is typically depicted radiographically as the area enclosed by the abnormal/hyperintense signal in the T2-FLAIR scan. This area, defined as the "whole tumor" (*WT*), entails the *TC* and the *ED*.

UPenn segmentations. The expert tumor annotations from UPenn, for the three tumor sub-compartments (i.e., *ET*, *NET*, and *ED*), were a product of a computer-aided segmentation using an in-house software[36] followed by manual revisions including corrections for (a) obvious under- or over-segmented sub-compartments, (b) voxels classified as *ED* within the *TC*, (c) unclassified voxels within the *TC*, (d) voxels classified as NET outside the *TC*, and (e) voxels corresponding to vessels within the *ED* that were either classified as *ED* or *ET*. Finally, contralateral, periventricular, and noncontiguous *WT* areas with hyperintense signal in the FLAIR scans were considered to represent chronic microvascular changes, or age-associated demyelination, rather than tumor infiltration,[37] and hence were excluded from the *WT*.

CWRU segmentations. Expert tumor annotations from CWRU for the three tumor sub-compartments (i.e., *ET*, *NET*, and *ED*), were performed manually by a collaborating neuroradiologist (V.B.H.) with over 10 years of experience in neuroradiology, after carefully considering three structural MRI scans, that is, T1Gd, T2, FLAIR.

2.D. Transforming annotations to a common atlas space

Since the annotations were performed in two different atlas spaces at each institution (i.e., SRI for UPenn, and MNI for CWRU), they needed to be brought to a common atlas space. To ensure consistency with the BraTS,[19-21] datasets, the SRI space was chosen as the reference atlas onto which the MNI labels were transformed. Four different registration solutions were explored for this transformation using Greedy with Normalized Mutual Information called from CaPTk:

1. MNI atlas to the SRI atlas and transformations applied to the CWRU segmentation labels.
2. Each MNI-registered T1Gd scans to the SRI atlas and apply corresponding transformations to the CWRU segmentation labels.
3. One MNI-registered T1Gd scan to the SRI atlas and apply the transformation to all CWRU segmentation labels.
4. Each MNI-registered T1Gd scan to the corresponding T1Gd SRI-registered volume and apply the corresponding transformation to CWRU segmentation labels.

After generating transformed labels following these four approaches (each using both skull-stripped and non-skull-

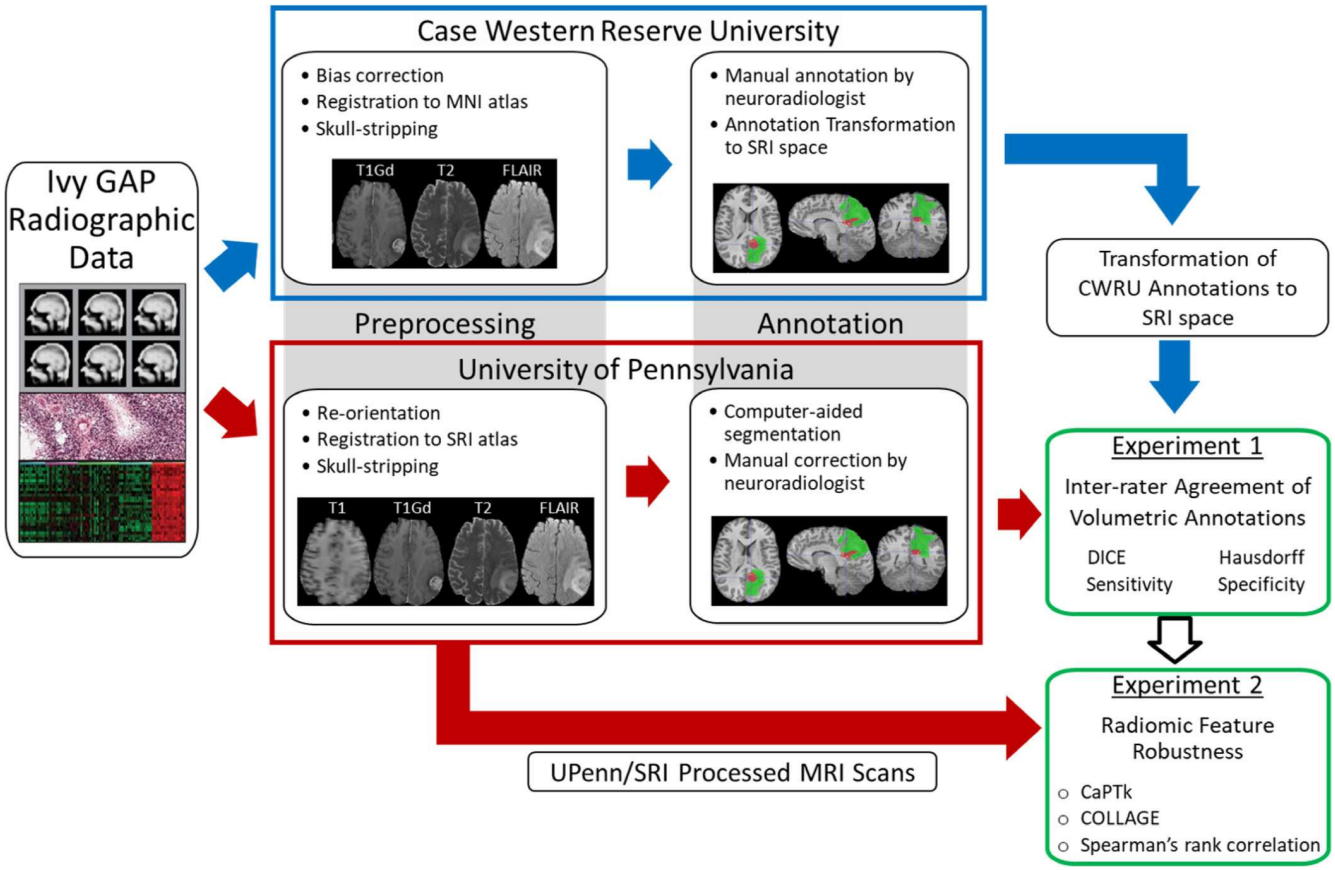


FIG. 1. Overall workflow of the present work. [Color figure can be viewed at wileyonlinelibrary.com]

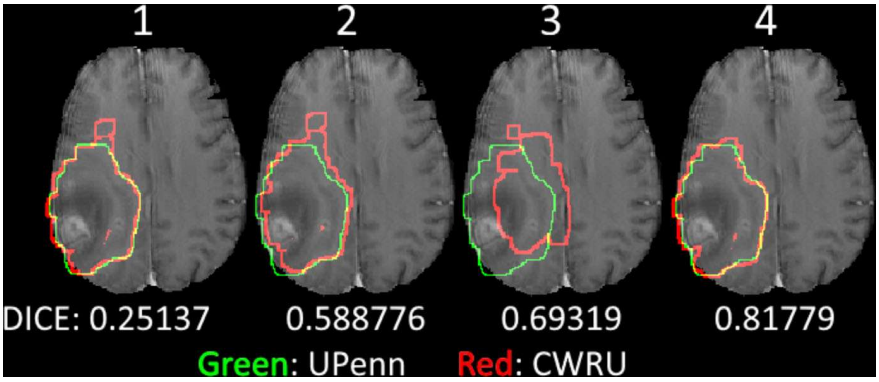


FIG. 2. Screenshots of Subject W8, showcasing the various registration transformations between CWRU and UPenn annotation we have used in Section 2.D, with the corresponding overall DICE scores. Green represents the UPenn tumor annotations and red represents the CWRU transformed annotations. [Color figure can be viewed at wileyonlinelibrary.com]

stripped images and different registration kernels), the most optimal alignment, based on the qualitative assessment of the end results (where we observed notable differences as shown in Fig. 2), was obtained using the last approach on skull-stripped images using Greedy[6] with normalized mutual information called from CaPTk.[29,38]

2.E. Radiomic analysis

Following standardization of the segmentation labels across the two institutions, a comprehensive array of 975 unique radiomic features (Table I) were obtained from eight different feature families including intensity-based statistical features (20 descriptors), morphological features (19 descriptors),[39] histogram features (503 descriptors), Gray-level co-occurrence matrix (GLCM)[40] (72 descriptors), Gray-level run-length matrix (GLRLM)[41-44] (90 descriptors), Gray-level size zone matrix (GLSZM) (162 descriptors), Neighborhood gray tone difference matrix (NGTDM) (5 descriptors), and co-occurrence of Local Anisotropic Gradient Orientations (COLLAGE) (104 descriptors). These feature sets were extracted per tumor subcompartment (*NET*, *ET*, *ED*) for every MRI scan (i.e., T1, T1Gd, T2, and FLAIR), for every subject using the same set of input images obtained from UPenn, that is, 11 700 features per patient. Since the preprocessing steps were different across the two institutions, for radiomic comparison we chose a single set of input images, processed using the UPenn pipeline which was consistent with the popular BraTS pipeline,[19-21] to ensure that the feature differences are on account of segmentation variability and not due to the varying pre-processing steps across institutions. Our radiomic feature set was extracted using open source tools comprising the Cancer imaging Phenomics Toolkit (CaPTk, www.cbica.upenn.edu/captk)[29] and a 3D Slicer extension for the COLLAGE feature (https://github.com/ccipd/CoLlAGeSlicerExtension).[45] CaPTk is an open-source software toolkit, which offers functionalities to extract a wide array of radiomic features compliant with the image biomarker standardization initiative (IBSI),[22] and has been extensively used in radiomic analysis studies.[12,14,16,21,44] Similarly, COLLAGE is a new open-source radiomic feature set, which has shown promise in disease prognosis and prediction for different solid tumors including brain, breast, lung, and prostate cancer.[13,18,46,47] Both CaPTk and COLLAGE were configured with a varying set of input parameters during feature extraction, including varying binning values ($\mathscr{B} \in \{16, 32, 64\}$) for quantization, radii ($\mathscr{R} \in \{1, 2, 3\}$) around the center voxel under consideration, and the window sizes (w) of 3 and 5 for computation of COLLAGE features. A complete set of extracted features can be found in the data repository available through TCIA,[26] as well as in supplementary documentation.

2.F. Experimental design

We quantitatively evaluated reproducibility for the Ivy GAP cohort with regard to two distinct endpoints: (a) the inter-reader agreement of the volumetric annotations across the three tumor sub-compartments (*NET*, *ET*, *ED*), and (b) the reproducibility of the extracted radiomic features across the three sub-compartments as well as across four MRI protocols (i.e. T1, T1Gd, T2, and FLAIR), as described below.

Inter-rater Agreement of Volumetric Annotations. We used the four most-commonly used metrics for semantic segmentation, including Dice Similarity Coefficient (DICE), Hausdorff distance, sensitivity, and specificity, to quantitatively compare the segmentation labels obtained from the two experts (M.B, V.B.H). For completeness, we have performed the analysis by first considering the CWRU rater as ground truth and comparing UPenn rater and then considering the UPenn rater as the ground truth and comparing the CWRU rater; both done on a per-voxel manner. Specifically, DICE was used to evaluate the extent of spatial overlap between the two sets of annotations and sensitivity and specificity are used to assess the overall agreement of the raters between all the sub-compartments. Furthermore, the 95th percentile of the Hausdorff distance was used to measure the maximum distance of the point set of one annotation boundary to the

TABLE 1. Summary of the radiomic features extracted in this study and the associated input parameters.

Feature family	Total features	Description	Parameters
Morphology	19	Geometric properties of the ROI	–
Intensity	20	Intensity distribution within the ROI	–
Histogram	503	Intensity distribution within the ROI after bin quantization	$\mathfrak{B} \in \{16,32,64,128\}$
COLLAGE	104	Quantifies heterogeneity of local gradient orientations within w	$w \in \{3,5\}$
GLCM	72	Distribution of discretized intensities of neighboring voxels along all directions within the ROI	$\mathfrak{B} \in \{16,32,64,128\}$ $\mathfrak{R} \in \{1,2,3\}$
GLRLM	90	Distribution of discretized intensities in all directions across run lengths within the ROI	$\mathfrak{B} \in \{16,32,64,128\}$ $\mathfrak{R} \in \{1,2,3\}$
GLSZM	162	Number of groups (or zones) of neighboring discretized voxels within the ROI	$\mathfrak{B} \in \{16,32,64,128\}$ $\mathfrak{R} \in \{1,2,3\}$
NGTDM	5	Number of groups of neighboring discretized voxels within the ROI, within a Chebyshev distance	$\mathfrak{B} \in \{16,32,64,128\}$ $\mathfrak{R} \in \{1,2,3\}$

nearest point in the other. In addition, the sensitivity and specificity metrics that describe the *true positive rate* and the *true negative rate* across the pair of segmentations were evaluated. Notably, these metrics were estimated for every tumor region, that is, *ET*, *NET*, *ED*, *TC*, and *WT*.

Radiomic Feature Robustness. To assess the robustness of the extracted radiomic features across the two sets of expert annotations, different correlation metrics were considered including the intraclass correlation coefficient (ICC),[48] which has been extensively used in the literature for assessing segmentation variability[24,49,50] as well as Spearman rank correlation[51]. Spearman's rank correlation coefficient (r_s) allows for sensitivity to nonlinear relationships in assessing the statistical dependence between the rankings of each feature across the two experts, and hence was used as the method of choice for our analysis. Additionally, along with Spearman correlation coefficient, we found intraclass correlation coefficient ($ICC(3,1)$) in Ref. [48] to be applicable in the case of our study[52] and thus calculated $ICC(3,1)$ measure for our analysis.

3. RESULTS

Inter-rater Agreement of Volumetric Annotations. The inter-reader agreement across different tumor sub-compartments was obtained using 3D volumetric analysis, as illustrated in Fig. 3 and Fig. 10(a). The high overall values of DICE, sensitivity, and specificity, combined with the low 95th percentile Hausdorff distances demonstrate the high rate of agreement between the UPenn and CWRU raters across various labels for the included Ivy GAP subjects. Specifically, the composite tumor regions of TC and WT consistently demonstrated the best inter-rater agreement in terms of their spatial overlap, when compared with the individual tumor sub-compartments of ET, NET, and ED. Particularly the agreement for the TC area, which represents the bulk of the mass under consideration for resection, obtained a median DICE > 0.85, followed by WT with a median DICE slightly above 0.7. When observed in tandem with the DICE score of all tumor sub-compartments, the lower agreement of WT appeared to be driven by the tumor region of ED that had the lowest DICE scores.

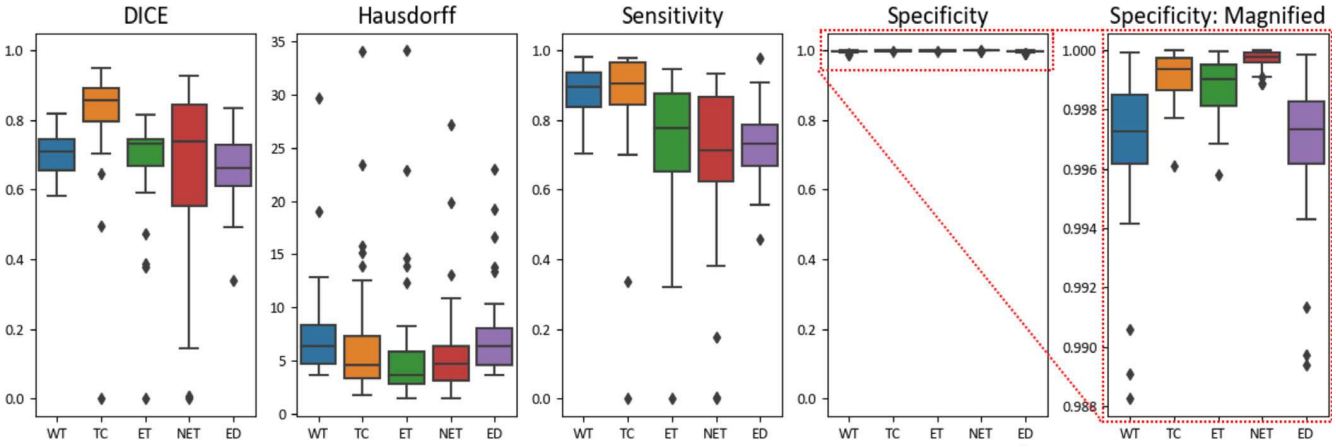


FIG. 3. Inter-rater agreement using three-dimensional volumetric analysis comparing CWRU rater with UPenn rater using different metrics (DICE, Sensitivity, Specificity, Hausdorff) across labels. Note that Specificity has also been plotted on a magnified scale to better highlight differences between the various sub-compartments. [Color figure can be viewed at wileyonlinelibrary.com]

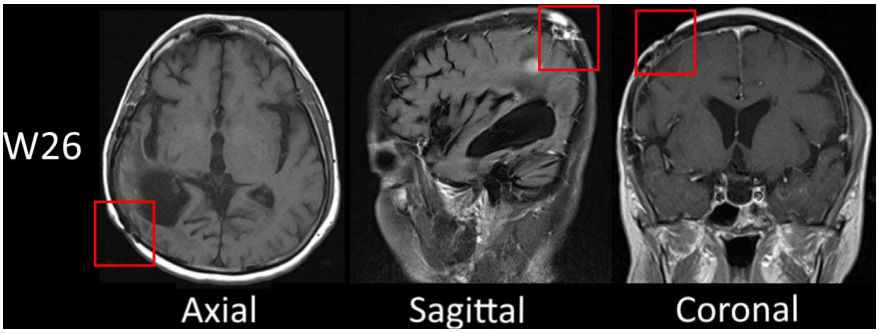


FIG. 4. Subject W26, showing screenshots of the axial, sagittal and coronal views, where the points of entry for prior instrumentation are visible and highlighted with a red square in the image. [Color figure can be viewed at wileyonlinelibrary.com]

The extreme outliers for the NET region (Fig. 3), belonged to cases W26 and W50. W26 shows an apparent previous instrumentation (Fig. 4). W50 is another exceptional case, where the annotations of the two expert raters were in disagreement, especially with respect to TC, which was identified in completely different locations (Fig. 5).

Radiomic Feature Robustness. Fig. 6 and Fig. 11 shows the Spearman's rank correlation coefficients and intraclass correlation (specifically, ICC(3,1)) obtained for different feature families, across different tumor sub-compartments (i.e., ET, NET, and ED), as well as across the four MRI protocols (T1, T1Gd, T2, and FLAIR). Interestingly, for ET, and NET sub-compartments, we observed consistent patterns across different radiomic feature families, with high correlation values observed for morphology (also reported lowest variance), intensity, and COLLAGE features across the four MRI protocols and highly variable correlation values for Histogram, GLRLM and GLSZM feature families (Fig. 6). For the ET region, while intensity tended to have high correlation values, lowest variance was observed in NGTDM features, across all four MRI protocols.

In order to identify the most correlated features, we used a threshold of $\geq$0.8 for the correlation coefficient measure across the segmentation set, obtained for every feature. After imposing the threshold, a small percentage (24.3%) of the overall feature set was identified as "reproducible" across the paired segmentation sets, as elucidated in Fig. 7; Figs. 9 and 12. The largest number of robust features was obtained for the morphology feature family across NET and ET sub-compartments, across T1, T1Gd, and T2 MRI protocols. For ET subcompartment, the COLLAGE features were found to have the largest number of robust features for T1, and FLAIR MRI protocols, while morphology feature family had slightly higher percentage of features being picked up for T1Gd and T2 protocols.

Overall, the highest correlations were consistently observed for intensity-based, and COLLAGE features, aside from the morphology feature family. Interestingly, the COLLAGE entropy, sum variance, and energy features were found to be most stable ($r_s \geq 0.8$) across all MRI protocols and tumor sub-compartments. In contrast, low correlations were observed for most of the other texture features obtained from GLCM, GLRLM, GLSZM, and NGTDM feature families, across all sub-compartments, as well as feature families.

4. DATA FORMAT AND USAGE NOTES

In accordance with the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR principles),[53] all the data and the associated meta-data generated as a part of this study is made publicly available through the TCIA's Analysis Results Directory (https://doi.org/10.7937/9j41-7d44).[26] Specifically, the released data comprises of (a) the available expert segmentation labels of the various tumor sub-compartments performed at each institution (i.e. 34 subjects segmented at UPenn, 34 subjects segmented at CWRU, with a total of 37 subjects (including 31 paired segmentations performed at both UPenn and CWRU), in the original space they were created (i.e., SRI for UPenn and MNI for CWRU), with (b) their corresponding co-registered and skull-stripped structural mpMRI scans (i.e., in SRI for UPenn and in MNI for CWRU), (c) the paired expert segmentation labels that were available for the 31 subjects, all being co-registered in the SRI atlas, (d) the corresponding SRI and MNI anatomical atlas files that we employed, (e) the complete set of 11 700 extracted radiomic features per subject, for each of the 31 included subjects, (f) the metadata relating to the metrics we utilized for the evaluation of the inter-rater agreement, as well as (g) the parameters used for the radiomic feature extraction and the correlation analysis results for identifying robust radiomic features, for the 28 subjects, and finally, (h) the specific identified robust/reproducible radiomic features. All image related files are provided in NIfTI format, while the metadata files are provided in tabular formats (.xlsx and .csv).

5. DISCUSSION

The availability of large data repositories such as TCIA's Ivy GAP cohort has opened up tremendous possibilities with the use of radiomics (i.e., quantitative feature analysis) for applications in prognosis and prediction in GBM tumors. However, in order to develop robust noninvasive image-based markers using the TCIA's Ivy GAP, there are two significant challenges that need to be accounted for: (a) the lack of

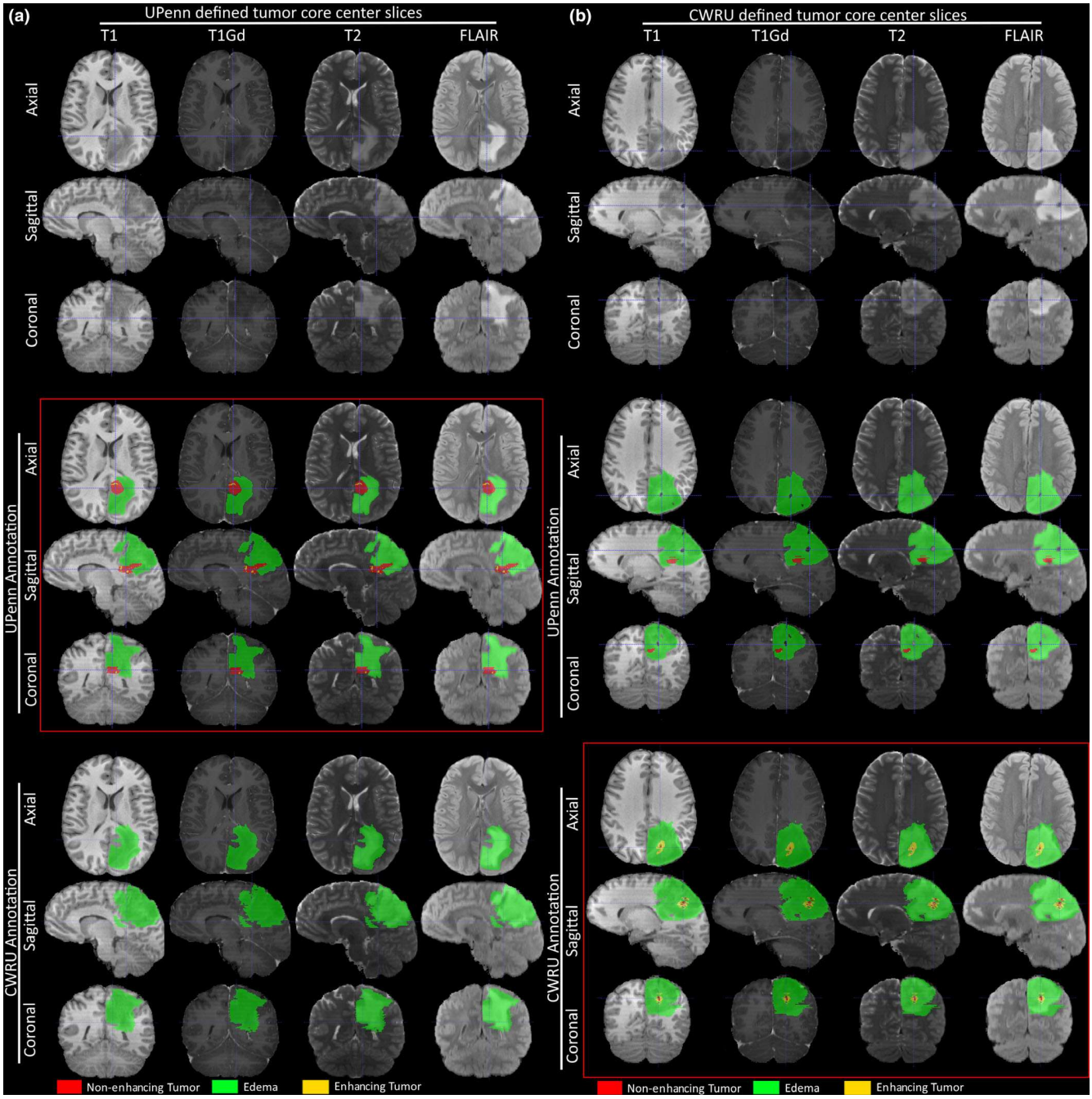


FIG. 5. Screenshots of Subject W50, where the raters' agreement regarding the site of *NET* and *ET* was different (locations with largest diameter of Non-enhancing part of tumor highlighted for each annotation). [Color figure can be viewed at wileyonlinelibrary.com]

availability of reliable segmentation labels for different tumor sub-compartments (NET, ET, and ED) and (b) identification of "reproducible" radiomic features that are robust to variability in segmentation labels obtained from different institutions. In this study, we sought to address these challenges via, (a) evaluating inter-rater agreement in volumetric annotations of tumor sub-compartments obtained from two institutions (UPenn and CWRU), (b) identifying robust/stable radiomic features across the two sets of segmentations obtained from UPenn and CWRU, and (c) the public release of the multi-institutional paired expert segmentation labels, the identified robust radiomic features, as well as the associated analysis,[26] through TCIA.

Most notable among previous related works, the work of Tixier et al.[24] has explored the robustness of radiomic features extracted from the TCGA-GBM dataset. However, there are four key differences between the two studies, particularly in terms of the comparative analysis. First and foremost, Tixier et al. compared the radiomic features extracted from a single tumor region, by considering non-enhancing tumor, enhancing tumor, and peritumoral edematous/invaded tissue as a single lesion habitat. Our work, in contrast, provides a

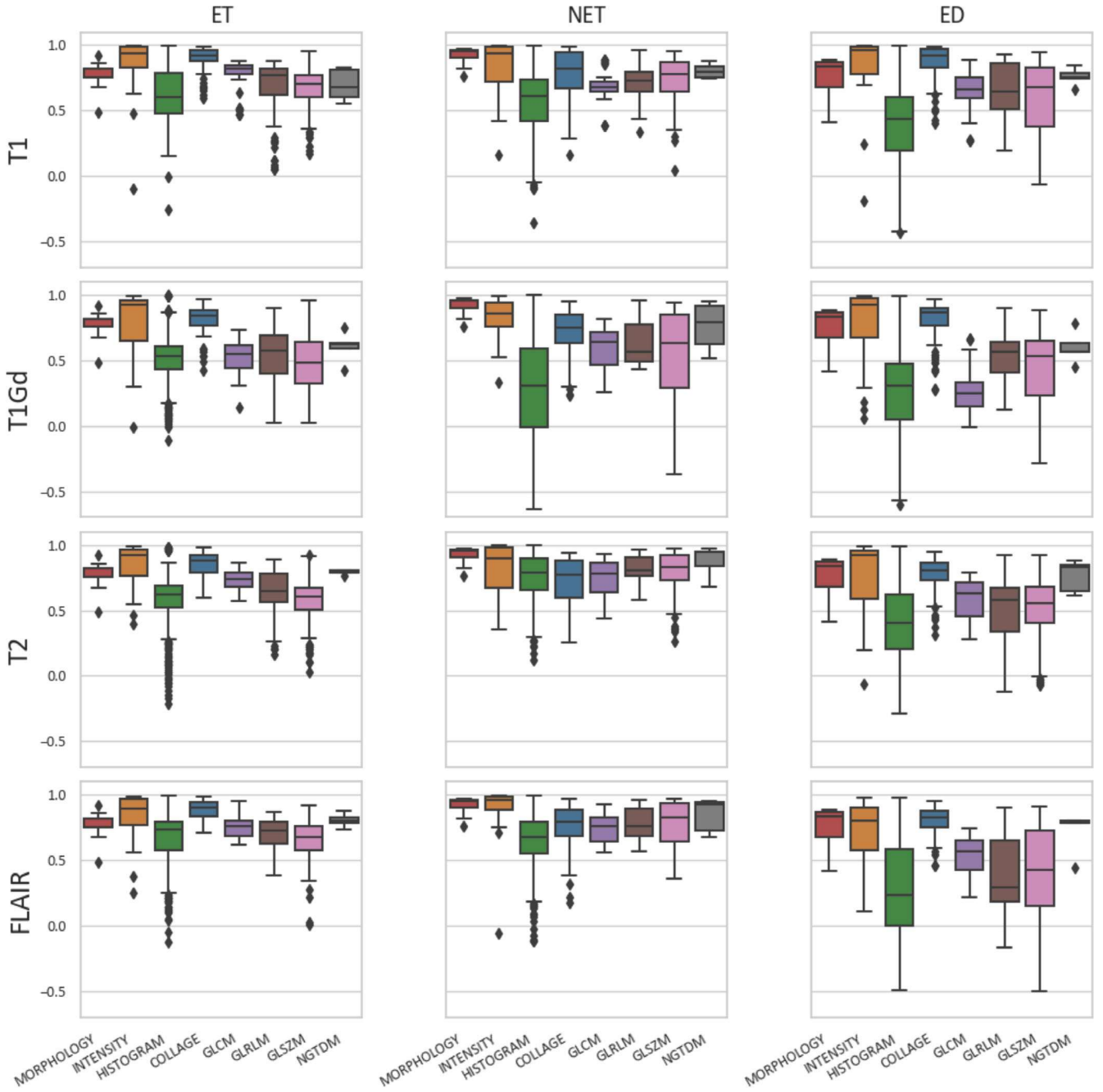


FIG. 6. Inter-rater agreement analysis using Spearman's rank correlation coefficient for the UPenn and CWRU raters across the 8 feature families, as well as across T1, T1Gd, T2, and FLAIR protocols. [Color figure can be viewed at wileyonlinelibrary.com]

more comprehensive comparative analysis following the most widely accepted convention (used by the International BraTS challenge[19-21]) wherein we consider (a) each tumor subcompartment (non-enhancing tumor, enhancing tumor, peritumoral edematous/invaded tissue) separately, (b) the enhancing and non-enhancing tumor as a single "tumor core" region (i.e., the potentially resectable tumor), as well as (c) the union of all the three tumor sub-compartments as a single habitat ("whole tumor"). Second, another notable difference between the two studies include consideration of only FLAIR and T1Gd scans by Tixier et al., in contrast to the present study that considers all four structural MRI modalities, that is, T1, T1Gd, T2, and FLAIR. Third, a major difference was in terms of the radiomic features considered across the two analyses, where Tixier et al. evaluated a total of 108 features (extracted using the open-source CERR package[54]), whereas we extracted a total of 11 700 radiomic descriptors from various different feature families (Table I) (extracted using open-source packages, COLLAGE[15] and CaPTk[29,38]). Finally, we performed our statistical analysis based on Spearman's correlation coefficient. Spearman's correlation coefficient is a nonparametric measure of the degree of association between

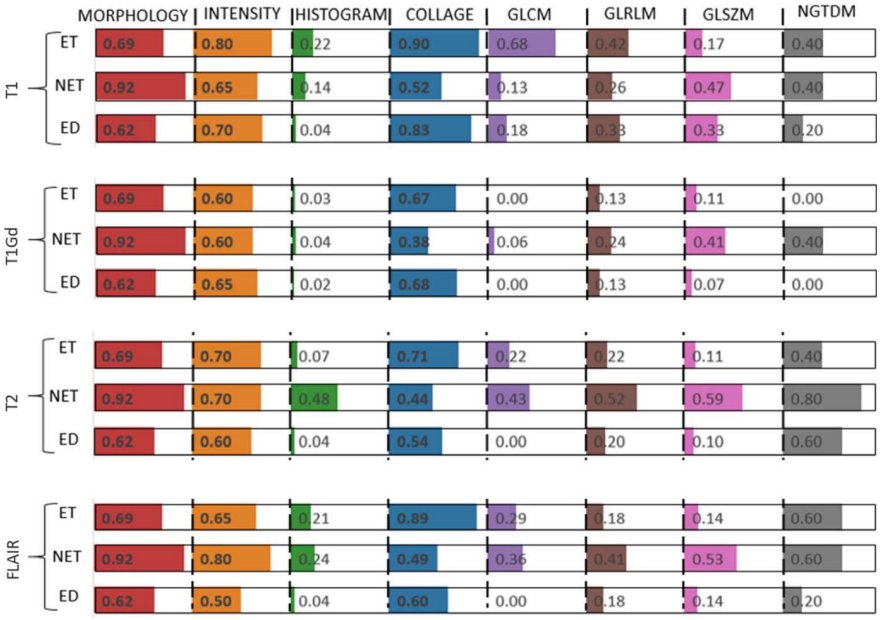


FIG. 7. Thermometer plot highlighting the percentage of robust features across UPenn and CWRU segmentations, (with Spearman's correlation coefficient of ≥0.8) for the 8 feature families across T1, T1Gd, T2, FLAIR protocols. [Color figure can be viewed at wileyonlinelibrary.com]

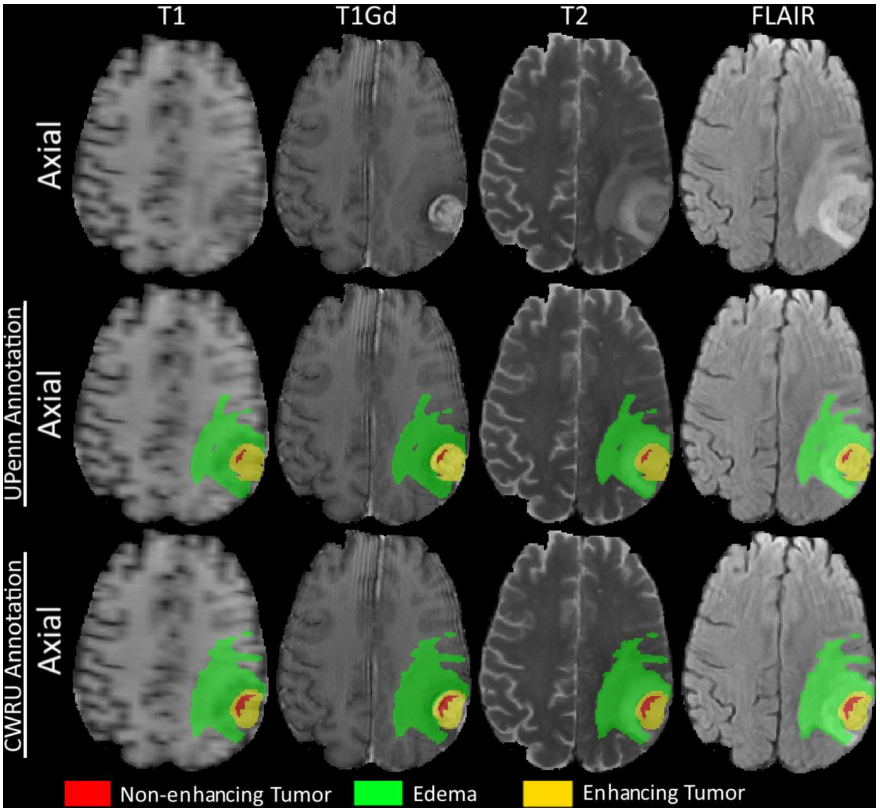


FIG. 8. Screenshots of Subject W8, showcasing maximal agreement between UPenn and CWRU raters (with regard to the whole tumor). Each image shows the axial slice from all 4 structural modalities in the top row with the annotations of UPenn and CWRU raters in the bottom 2 rows. [Color figure can be viewed at wileyonlinelibrary.com]

two variables, and unlike ICC[48] (that was used by[24]), it does not require the assumption that the relationship between the variables is linear.[49] For completeness, we also assessed ICC (3,1) (Fig. 11) metric[48] and found the results to be comparable to using Spearman's coefficient (Fig. 6), except for the NGTDM feature family, where more number of features were

identified as stable using the ICC measure for T2 and FLAIR as compared to using the Spearman measure.

Our volumetric analysis across the segmentation labels obtained from the two institutions indicated a high level of agreement between the two raters, especially for TC region as evidenced by the relatively high values of sensitivity (median value ≥ 0.85) and specificity (median value ≥ 0.95), which is of vital clinical importance as it defines the region that is considered for surgical resection. Similar levels of agreement can be seen for the WT (median sensitivity ≥ 0.85), ET (median sensitivity ≥ 0.8), NET (median sensitivity ≥ 0.7), and ED (median sensitivity ≥ 0.7) with median specificity ≥ 0.9 for all, highlighting the correlation between the two raters. The standard deviation and median values of the evaluation metrics for the inter-rater agreement across the GBM subcompartments in our work were found to be consistent with previously reported results on other similar TCIA and BRATS studies.[19-21]

Our results for radiomic feature reproducibility across the pair of segmentation labels identified 24.3% of 11 700 extracted radiomic features to be robust to segmentation changes across the two sites. A substantial proportion of these selected features belonged to morphology (describing shape characteristics), intensity (capturing statistics across intensity. profiles), and COLLAGE feature (capturing heterogeneity in local gradient orientations) families (Fig. 7 and Fig. 12). The high correlations obtained for the morphology and intensity feature families were likely on account of the high inter-reader agreement observed across the tumor regions, especially across NET and ET. Similarly, high correlations obtained for the COLLAGE feature family could be attributed to the fact that COLLAGE features are not directly computed on the intensity measurements but are rather derived from the gradient orientations within a local neighborhood window. The gradient orientations seem to be less impacted by the variability in segmentation labels across sites. Further, it was observed that the maximum number of total stable features from these three feature families ($r_s \geq 0.8$) belonged to the T1, protocol followed by T2, FLAIR, and T1Gd respectively.

Based on our feasibility study, most of the Morphological features were not found to be dependent on the differences in segmentations themselves, rather on segmentation characteristics (such as elongation, sphericity, eccentricity, and flatness), which were found to be fairly similar across the two raters and thus robust to per-pixel segmentation variations. Intensity statistics features capture the aggregated measures (i.e., mean, median) of the intensity profile of the modalities in the specified tumor compartment and hence were not found to be dependent on local differences in intensities across the two segmentations. Most of the intensity statistics features demonstrated a high degree of correlation between the two raters. Strikingly, the histogram feature family, and by extension, GLCM, GLRLM, GLSZM, and NGTDM feature families (which are known to capture local image heterogeneity) demonstrated low correlation values across segmentations for the majority of their features. This may be since these features are computed across multiple binning values (16, 32, 64, and 128), thereby making the feature set highly dependent on intensity changes, which may be reflected in lower correlation values across the patients. Additionally, these features include contrast, coarseness, homogeneity, and busyness, which have been previously been indicated to present large variations in their correlation values, therefore may need to be carefully investigated for robustness across segmentations before being employed in radiomic analysis for GBM tumors. Interestingly, while Haralick texture measurements across GLCM, GLSZM, and NGTDM feature families were sensitive to segmentation variability, COLLAGE texture features, which are also considered measures of local image heterogeneity, demonstrated high correlations measures, across all three sub-compartments and MRI protocols. This may be on account of the fact that COLLAGE computes measurements such as energy/entropy from the local intensity gradients rather than local intensity differences, and hence rendered more resilient to local differences in image intensities across segmentations. Previous studies[55] have similarly demonstrated that the features which are driven by entropy and energy exhibit lesser variations due to variability in acquisition variations and reconstruction parameters.

It was noted that the brain extraction (also known as skull-stripping) approaches employed across the two sites, may have caused issues in the transformation of the respective annotations due to parts of the head (e.g., eyeballs) that were not removed during skull stripping. Examples of this issue can be found in the uploaded data for subjects W32 and W42 in the MNI created annotations by CWRU. However, even in the cases where registration did not fail, we observed that the tumor segmentation can be affected when part of a tumor or peritumoral area adjacent to the skull is removed during the brain extraction process, (Fig. 8). This highlights the need for a robust brain extraction method optimized for pathological brains that could work consistently across modalities and clinical sites.[32]

Interestingly, during our segmentation analysis, we observed an exceptional case (subject W50), for which the TC was annotated in two completely different locations by the two expert raters, as shown in detail in Fig. 5. It was noted that the CWRU rater had demarcated the center of the ED region (within the superior parietal lobule) as the TC, whereas the UPenn rater had highlighted the edge of the ED, closer to the ventricles (within the more inferior parasagittal precuneus) as the TC. One possible reason for this might be the fact that there are minimally enhancing foci in both these locations in the T1Gd scan, without a distinct central TC. There is also infiltrative non-enhancing or poorly enhancing tumor throughout the abnormal FLAIR hyper-intense signal (in a gliomatosis cerebri pattern), which is seen best on T2 through a slightly less hyper intense envelope, than the rest of the FLAIR hyper intense signal, reflecting highly cellular tumor compatible with the pathologically proven GBM. This case points to the difficulty and variability involved in the task of tumor region delineation, even by experienced

clinicians. Another subject of particular interest was W26, where radiologic assessment indicated that it was a non-baseline scan (the points of entry for a resection are visible, Fig. 4). We still included it in our analysis for segmentation agreement as well as radiomic feature analysis because the tumor that was being assessed did not seem to be affected by previous instrumentation.

Our work did have limitations. Our study was limited to investigating inter-reader agreement and did not consider intra-reader variability across segmentation labels. Further, segmentations were obtained from a single reader per institution. Allocating more than 2 raters would have allowed for a consensus analysis. While comprehensive (with over 11 000 radiomic features analyzed), the radiomic variability analysis was limited to 8 feature families. Future work will include interrogating intra-, as well as multiple-inter-reader segmentation variability, as well as including additional feature families (i.e., Laws, local binary patterns) for radiomic feature variability. We will also consider interrogating reproducibility of radiomic features across variations in slice thickness, image reconstruction methods, magnetic field strengths, echo times, and repetition times.

6. CONCLUSIONS

Radiomics has recently provided a surrogate mechanism for capturing GBM tumor heterogeneity using routine non-invasive MRI scans.[56] However, radiomic features are known to be susceptible to variations in annotation protocols across sites. In this work, we presented a feasibility study to (a) evaluate inter-reader agreement obtained for tumor segmentation labels, and (b) identify reproducible radiomic features across variations in tumor segmentations, in a multi-institutional setting, for the TCIA's[5] Ivy GAP dataset.[6] First, we quantified the inter-reader agreement using the most-commonly used metrics (DICE, Sensitivity, Specificity, and Hausdorff). Higher value of the DICE, Sensitivity and Specificity while, lower value of Hausdorff indicates better inter-reader agreement, between the two segmented regions. Our results demonstrated that there was a high amount of overall correlation between the two raters for all sub-compartments. Second, our radiomic variability analysis experiment suggested that (a) certain features and feature families such as intensity statistics (mean, median, standard deviation, and kurtosis), morphologic (flatness, elongation, and sphericity), and COLLAGE (statistics of local gradient entropy) may be more robust to variability in segmentation labels obtained from different readers, and (b) GLCM and GLRLM feature families, which are dependent on local intensity differences, showed lower correlation across features extracted from the segmented tumor regions demarcated by two different raters. While GLCM and GLRLM features have previously shown to be prognostic of GBM,[8,12,45] our results indicated that most of these features represented large variations across the two segmentations (Fig. 6 and Fig. 11), and may need to be carefully investigated for robustness across segmentations for prognostic modeling in GBM tumors. However, in contrast, majority of morphology and intensity statistics-based features seemed to be resilient to segmentation differences across the two readers. We further made the multi-institutional segmentations as well as associated meta-data collected as a part of this analysis available on the TCIA web-portal as a community resource,[26] with the purpose of enabling imaging and non-imaging researchers to leverage the Ivy GAP cohort for developing image-based biomarkers for prognosis and prediction of GBM tumors.

ACKNOWLEDGMENTS

Research reported in this publication was partly supported by the National Institutes of Health (NIH) under award number NCI/ITCR:U01CA242871, as well as by the Department of Defense (DoD) Peer Reviewed Cancer Research Program (W81XWH-18-1-0404), Dana Foundation David Mahoney Neuroimaging Grant, the CCCC Brain Tumor Pilot Award, the CWRU Technology Validation Start-Up Fund (CTP), and The V Foundation Translational Research Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, U.S. Department of Veterans Affairs, the DoD, or the United States Government. Niha Beig is an employee of Tempus Labs, Inc.

*Equally contributing first author.
†Equal senior corresponding author: pxt130@case.edu.
‡Equal senior corresponding author: sbakas@upenn.edu.
a)Author to whom correspondence should be addressed. Electronic mail: pxt130@case.edu.

REFERENCES

1. Ostrom QT, Gittleman H, Fulop J, et al. CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2008–2012. *Neuro-oncology*. 2015;17:iv1–iv62.
2. Amadasun M, King R. Textural features corresponding to textural properties. *IEEE Trans Syst*. 1989;19:1264–1274.
3. Jaffe CC. Imaging and genomics: Is there a synergy? *Radiology*. 2012;264:329–331. PMID: 22821693.
4. Xiao T, Hua W, Li C, Wang S. Glioma grading prediction by exploring radiomics and deep learning features. In *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*. 2019:208–213.
5. Clark K, Vendt B, Smith K, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Dig Imaging*. 2013;26:1045–1057.
6. Puchalski RB, Shah N, Miller J, et al. An anatomic transcriptional atlas of human glioblastoma. *Science*. 2018;360:660–663.
7. Shah N, Feng X, Lankerovich M, Puchalski RB, Keogh B. Data from Ivy GAP. *The Cancer Imaging Archive*. 2016. https://doi.org/10.7937/K9/TCIA.2016.XLwaN6nL
8. Akbari H, Bakas S, Pisapia JM, et al. In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. *Neuro-oncology*. 2018;20:1068.
9. Bakas S, Akbari H, Pisapia J, et al. In vivo detection of EGFRvIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral infiltration: the φ-index. *Clinical Cancer Res*. 2017;23:4724–4734.
10. Ellingson BM, Lai A, Harris RJ, et al. Probabilistic radiographic atlas of glioblastoma phenotypes. *Am J Neuroradiol*. 2012;34:533.

11. Zinn PO, Majadan B, Sathyan P, et al. Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme. *PloS One*. 2011;6:e25451.
12. Bakas S, Shukla G, Akbari H, et al. Overall survival prediction in glioblastoma patients using structural magnetic resonance imaging (MRI): advanced radiomic features may compensate for lack of advanced MRI modalities. *Journal of Medical Imaging*. 2020;7(3):1.
13. Beig N, Patel J, Prasanna P, et al. Radiogenomic analysis of hypoxia pathway is predictive of overall survival in glioblastoma. *Sci Rep*. 2018;8:1–11.
14. Macyszyn L, Akbari H, Pisapia JM, et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology*. 2015;18:417–425.
15. Prasanna P, Patel J, Partovi S, Madabhushi A, Tiwari P. Radiomic features from the peritumoral brain parenchyma on treatment-naive multiparametric MR imaging predict long versus short-term survival in glioblastoma multiforme: preliminary findings. *Eur Radiol*. 2017;27:4188–4197.
16. Fathi Kazerooni A, Akbari H, Shukla G, et al. Cancer imaging phenomics via CAPTK: multi-institutional prediction of progression-free survival and pattern of recurrence in glioblastoma. *JCO Clin Cancer Inform*. 2020;4:234–244.
17. Akbari H, Macyszyn L, Da X, et al. Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. *Neurosurgery*. 2016;78:572–580.
18. Verma R, Correa R, Hill V, et al. Radiomics of the lesion habitat on pretreatment MRI predicts response to chemo-radiation therapy in glioblastoma. In: *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. International Society for Optics and Photonics; 2019:109500B.
19. Bakas S, Reyes M, Jakab A, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*; 2018.
20. Menze BH, Jakab A, Bauer S, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2014;34:1993–2024.
21. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117.
22. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high throughput image-based phenotyping. *Radiology*. 2020;295:191145.
23. Shiri I, Hajianfar G, Sohrabi A, et al. Repeatability of radiomic features in magnetic resonance imaging of glioblastoma: test-retest and image registration analyses. *Med Phys*. 2020;47:4265.
24. Tixier F, Um H, Young RJ, Veeraraghavan H. Reliability of tumor segmentation in glioblastoma: impact on the robustness of MRI-radiomic features. *Med Phys*. 2019;46:3582–3591.
25. Um H, Tixier F, Bermudez D, Deasy JO, Young RJ, Veeraraghavan H. Impact of image preprocessing on the scanner dependence of multi-parametric MRI radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Phys Med Biol*. 2019;64:165011.
26. Pati S, Verma R, Akbari H, et al. Multi-institutional paired expert segmentations and radiomic features of the Ivy GAP dataset. *The Cancer Imaging Archive*. 2020. https://doi.org/10.7937/9j41-7d44
27. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Human Brain Mapping*. 2010;31:798–819.
28. Tustison NJ, Avants BB, Cook PA, et al. N4itk: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29:1310–1320.
29. Davatzikos C, Rathore S, Bakas S, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J Med Imaging*. 2018;5:011018.
30. Yushkevich PA, Pluta J, Wang H, Wisse LE, Das S, Wolk D. Fast automatic segmentation of hippocampal subfields and medial temporal lobe subregions in 3 tesla and 7 tesla T2-weighted MRI. *Alzheimer's & Dementia*. 2016;12:P126–P127.
31. Yushkevich PA, Piven J, Cody Hazlett H, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage*. 2006;31:1116–1128.
32. Thakur SP, Doshi J, Pati S, et al. Skull-stripping of glioblastoma MRI scans using 3D deep learning. In *International MICCAI Brainlesion Workshop*. Springer; 2019:57–68.
33. Talairach J. Co-planar stereotaxic atlas of the human brain-3-dimensional proportional system. *An Approach to Cerebral Imaging*; 1988.
34. Kikinis R, Pieper SD, Vosburgh KG. 3D slicer: a platform for subject-specific image analysis, visualization, and clinical support. In:*Intraoperative Imaging and Image-Guided Therapy*. Berlin: Springer; 2014:277–289.
35. Bauer S, Fejes T, Reyes M. A Skull-Stripping Filter for ITK. *The Insight Journal*. 2012. http://doi.org/10.5281/zenodo.811812
36. Bakas S, Zeng K, Sotiras A, et al. Glistrboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: *BrainLes 2015*. Berlin: Springer; 2015:144–155.
37. Haller S, Kövari, E, Herrmann FR, et al. Do brain T2/flair white matter hyperintensities correspond to myelin loss in normal aging? A radiologic-neuropathologic correlation study. *Acta Neuropathol Commun*.2013;1:14.
38. Rathore S, Bakas S, Pati S, et al. Brain cancer imaging phenomics toolkit (brain-CaPTk): an interactive platform for quantitative analysis of glioblastoma. In *International MICCAI Brainlesion Workshop*. Springer; 2017:133–145.
39. Max J. Quantizing for minimum distortion. *IRE Trans Inform Theory*. 1960;6:7–12.
40. Haralick RM, Shanmugam K, and Dinstein IH. Textural features for image classification. *IEEE Trans Syst*.1973;SMC-3:610–621.
41. Chu A, Sehgal CM, Greenleaf JF. Use of gray value distribution of run lengths for texture analysis. *Pattern Recogn Lett*. 1990;11:415–419.
42. Dasarathy BV, Holder EB. Image characterizations based on joint gray level—run length distributions. *Pattern Recogn Lett*. 1991;12:497–502.
43. Galloway M. Texture analysis using gray level run lengths. *Comput Graphics Image Process*. 1975;4:172–179.
44. Tang X. Texture information in run-length matrices. *IEEE Transactions on Image Processing*. 1998;7(11):1602–1609.
45. Prasanna P, Tiwari P, Madabhushi A. Co-occurrence of local anisotropic gradient orientations (collage): distinguishing tumor confounders and molecular subtypes on MRI. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2014:73–80.
46. Braman NM, Etesami M, Prasanna P. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res*. 2017;19:57.
47. Shiradkar R, Ghose S, Jambor I, et al. Radiomic features from pretreatment biparametric MRI predict prostate cancer biochemical recurrence: preliminary findings. *J Magn Reson Imaging*. 2018;48:1626–1636.
48. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bullet*. 1979;86:420.
49. Liu R, Elhalawani H, Radwan Mohamed AS. Stability analysis of CT radiomic features with respect to segmentation variation in oropharyngeal cancer. *Clin Translat Radiat Oncol*. 2020;21:11–18.
50. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J Appl Clin Med Phys*. 2020;21:179–190.
51. Rockafellar RT, Wets RJ-B. *Variational Analysis*. Vol. 317. Berlin: Springer Science & Business Media; 2005.
52. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
53. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The fair guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:160018.
54. Apte AP, Iyer A, Crispin-Ortuzar M, et al. Extension of CERR for computational radiomics: a comprehensive matlab platform for reproducible radiomics research. *Med Phys*. 2018;45:3713–3720.
55. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol*. 2010;49:1012–1016.

56. Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumor imaging: A systematic review. *PloS One*. 2014;9: e110300.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Data S1. Multi-institutional paired expert segmentations and radiomic features of the Ivy GAP dataset.

Reprint Permissions

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Jan 03, 2024

This Agreement between Sarthak Pati ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5701381183422
License date	Jan 03, 2024
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Medical Physics
Licensed Content Title	Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset
Licensed Content Author	Spyridon Bakas, Pallavi Tiwari, Russell Taki Shinohara, et al
Licensed Content Date	Dec 4, 2020
Licensed Content Volume	47
Licensed Content Issue	12
Licensed Content Pages	14

Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	Yes, including English rights
Number of languages	1
Title of new work	Reproducibility of Machine Learning Research in Clinical Environments
Institution name	Technical University of Munich
Expected presentation date	Mar 2024
Specific Languages	English
Requestor Location	Sarthak Pati 1051 Harriman Ct WEST CHESTER, PA 19380 United States Attn: Sarthak Pati
Publisher Tax ID	EU826007151
Total	0.00 USD

Terms and Conditions

TERMS AND CONDITIONS

no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a

waiver of or consent to any other or subsequent breach by such other party.

- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.
- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.
- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.
- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.
- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

5

GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows

Authors

Sarthak Pati*, Siddhesh P Thakur, İbrahim Ethem Hamamcı, Ujjwal Baid, Bhakti Baheti, Megh Bhalerao, Orhun Güley, Sofia Mouchtaris, David Lang, Spyridon Thermos, Karol Gotkowski, Camila González, Caleb Grenko, Alexander Getka, Brandon Edwards, Micah Sheller, Junwen Wu, Deepthi Karkada, Ravi Panchumarthy, Vinayak Ahluwalia, Chunrui Zou, Vishnu Bashyam, Yuemeng Li, Babak Haghighi, Rhea Chitalia, Shahira Abousamra, Tahsin M Kurc, Aimilia Gastounioti, Sezgin Er, Mark Bergman, Joel H Saltz, Yong Fan, Prashant Shah, Anirban Mukhopadhyay, Sotirios A Tsaftaris, Bjoern Menze, Christos Davatzikos, Despina Kontos, Alexandros Karargyris, Renato Umeton, Peter Mattson, Spyridon Bakas

Publication Information

Communications Engineering 2, no. 1 (2023): 23. DOI: $10.1038/s44172 - 023 - 00066 - 3$.

Abstract

Deep Learning (DL) has the potential to optimize machine learning in both the scientific and clinical communities. However, greater expertise is required to develop DL algorithms, and the variability of implementations hinders their reproducibility, translation, and deployment. Here we present the community-driven Generally Nuanced Deep Learning Framework (GaNDLF), with the goal of lowering these barriers. GaNDLF makes the mechanism of DL development, training, and inference more stable, reproducible, interpretable, and scalable, without requiring an extensive technical background. GaNDLF aims to provide an end-to-end solution for all DL-related tasks in computational precision medicine. We demonstrate the ability of GaNDLF to analyze both radiology and histology images, with built-in support for k-fold cross-validation, data augmentation, multiple modalities and output classes. Our quantitative performance evaluation on numerous use cases, anatomies, and computational tasks supports GaNDLF as a robust application framework for deployment in clinical workflows.

Contributions of S.P.

Study conceptualization, algorithm development and implementation, interpretation of results, and writing & editing of manuscript.

Copyright

communications engineering

ARTICLE

https://doi.org/10.1038/s44172-023-00066-3 OPEN

GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows

Sarthak Pati[1,2,3,4,5], Siddhesh P. Thakur[2,3,4], İbrahim Ethem Hamamcı[2,6], Ujjwal Baid[2,3,4], Bhakti Baheti[2,3,4], Megh Bhalerao[2,4], Orhun Güley[5], Sofia Mouchtaris[2,7], David Lang[2,7,8], Spyridon Thermos[9], Karol Gotkowski[10], Camila González[10], Caleb Grenko[2,3,4], Alexander Getka[2,4], Brandon Edwards[11], Micah Sheller[1,11], Junwen Wu[11], Deepthi Karkada[11], Ravi Panchumarthy[11], Vinayak Ahluwalia[2,4], Chunrui Zou[2,4], Vishnu Bashyam[2,4], Yuemeng Li[2,4], Babak Haghighi[2,4], Rhea Chitalia[2,4], Shahira Abousamra[12], Tahsin M. Kurc[13], Aimilia Gastounioti[2,4,14], Sezgin Er[6], Mark Bergman[2,4], Joel H. Saltz[13], Yong Fan[2,4], Prashant Shah[11], Anirban Mukhopadhyay[10], Sotirios A. Tsaftaris[9], Bjoern Menze[5,15], Christos Davatzikos[2,4], Despina Kontos[2,4], Alexandros Karargyris[1,16], Renato Umeton[1,17,18,19,20], Peter Mattson[1,21] & Spyridon Bakas[1,2,3,4]✉

Deep Learning (DL) has the potential to optimize machine learning in both the scientific and clinical communities. However, greater expertise is required to develop DL algorithms, and the variability of implementations hinders their reproducibility, translation, and deployment. Here we present the community-driven Generally Nuanced Deep Learning Framework (GaNDLF), with the goal of lowering these barriers. GaNDLF makes the mechanism of DL development, training, and inference more stable, reproducible, interpretable, and scalable, without requiring an extensive technical background. GaNDLF aims to provide an end-to-end solution for all DL-related tasks in computational precision medicine. We demonstrate the ability of GaNDLF to analyze both radiology and histology images, with built-in support for *k*-fold cross-validation, data augmentation, multiple modalities and output classes. Our quantitative performance evaluation on numerous use cases, anatomies, and computational tasks supports GaNDLF as a robust application framework for deployment in clinical workflows.

[1] MLCommons, Medical Working Group, San Francisco, CA, USA. [2] Center For Artificial Intelligence And Data Science For Integrated Diagnostics (AI2D) and Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA. [3] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [5] Department of Informatics, Technical University of Munich, Munich, Bavaria, Germany. [6] International School of Medicine, Istanbul Medipol University, Istanbul, Marmara, Turkey. [7] Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA. [8] Department of Mathematics, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA. [9] Institute for Digital Communications, School of Engineering, The University of Edinburgh, Scotland, UK. [10] Department of Computer Science, Technical University of Darmstadt, Darmstadt, Hesse, Germany. [11] Intel Corporation, Santa Clara, CA, USA. [12] Department of Computer Science, Stony Brook University, Stony Brook, New York, NY, USA. [13] Department of Biomedical Informatics, Stony Brook University, Stony Brook, New York, NY, USA. [14] Mallinckrodt Institute of Radiology, Washington University School of Medicine, St. Louis, MO, USA. [15] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland. [16] Institute of Image-Guided Surgery of Strasbourg, Strasbourg, France. [17] Department of Informatics & Analytics, Dana-Farber Cancer Institute, Boston, MA, USA. [18] Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA. [19] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [20] Department of Biological Engineering, Department of Mechanical Engineering, Massachusetts Institute of Technology, Boston, MA, USA. [21] Google, Menlo Park, CA, USA. ✉email: sbakas@upenn.edu

Deep Learning (DL) describes a subset of Machine Learning (ML) algorithms built upon the concepts of neural networks[1]. Over the last decade, DL has shown great promise in various problem domains such as semantic segmentation[2–5], quantum physics[6], segmentation of regions of interest (such as tumors) in medical images[7–13], medical landmark detection[14,15], image registration[16,17], predictive modelling[18], among many others[19–21]. The majority of this vast research was enabled by the abundance of DL libraries made open-source and publicly available, with some of the major ones being TensorFlow (developed by Google) and PyTorch (by Facebook - originally developed as Caffe by the University of California at Berkeley), which represent the most widely used libraries facilitating DL research. Among the currently available libraries, PyTorch has demonstrated itself to be one of the most customizable and easily deployable through its robust and efficient C++ backend.

There have been various efforts by the medical imaging community towards addressing the clinical end-points of academic research, and packaging pre-coded/pre-trained models for data scientists to leverage and address clinical requirements (Fig. 1). However, all these efforts, resulting in numerous software packages, can confuse the less experienced user and result in endless hours of searching for the appropriate tool to use. To alleviate this situation, we hereby stratify these efforts into a set of well-defined categories to deepen the community's understanding (Fig. 2). Some of these efforts lie on one side of the spectrum and can be classified as "applications", since they focus on the end-user, with powerful user interfaces (either graphical, or otherwise). Software packages on the other end of the spectrum can be stratified as "libraries", since they are built as a mechanism to access low-level machine functionality, while "toolkits" fall in between these two ends, and provide a layer of abstraction to enable research. Finally, "frameworks" fulfil various roles and attempt to provide a multitude of functions targeting both developers and end-users. Examples of such packages are the Medical Imaging Interaction Toolkit (MITK)[22] and the Cancer Imaging Phenomics Toolkit (CaPTk)[23]. GaNDLF is also a framework with a notably unique emphasis to DL. Figure 2 illustrates this stratification, while also providing some pertinent examples.

Some of these prior efforts are non-DL based, such as MITK[22], 3D Slicer[24], ITK-SNAP[25], and CaPTk[23]. While they have been lauded for their generalizability, they may fall short when it comes to competitive performance for specific tasks. Towards obtaining superior performance, various efforts concentrating on DL have been devised recently by the community, such as NiftyNet[26], DeepNeuro[27], ANTsPyNet[28], and DLTK[29], that are implemented in TensorFlow, as well as pymia[30], InnerEye[31], and MONAI[32], that are implemented in PyTorch. Additionally, there are specialized DL-based tools that cater to specific problems, such as segmentation[11,33–35], registration[36], or specific imaging domains, like PathML[37], TIAToolbox[38], HistomicsML[39], that focus on data engineering and enabling ML in computational pathology. However, all these applications and toolkits either (i) describe developer-focused tools targeting members of the

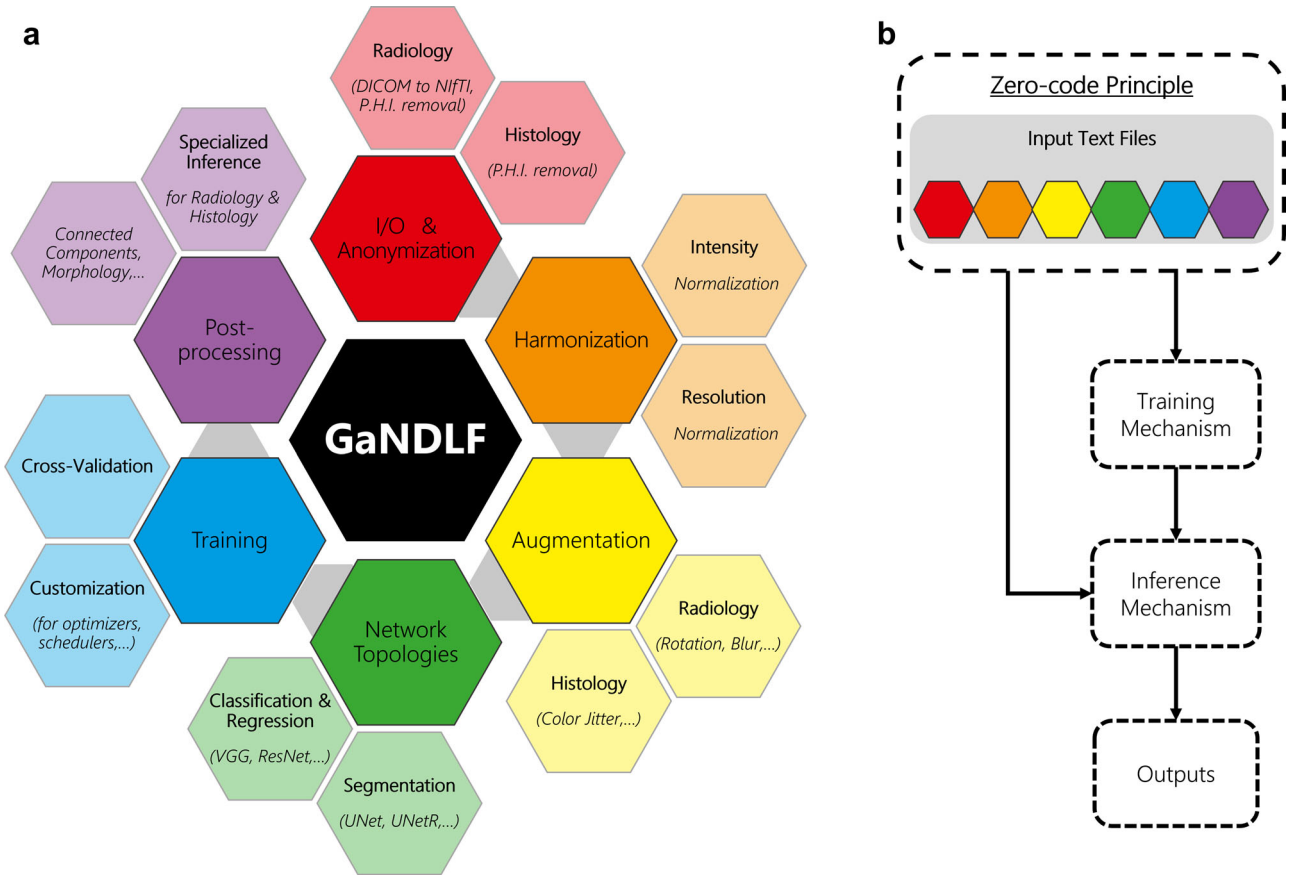


Fig. 1 Current amalgamation of the functionality of GaNDLF. a The entire functionality palette is focused to promote "zero/low-code" principles, and at the same time, each component in the major color groups (i.e., anonymization, harmonization, augmentation, network topologies, training, and post-processing) can be used independently to create customized solutions. The grey arrows represent the flow of operations for a user towards a "zero/low-code" principle for an entire computational training pipeline, starting with data I/O and ending with post-processing. **b** A high-level flowchart highlighting the "zero-code principle" entry point for the entire functionality palette of GaNDLF and their interactions throughout an AI clinical workflow, using the "zero/low-code" principle. A more comprehensive flowchart version is given in Fig. 4.

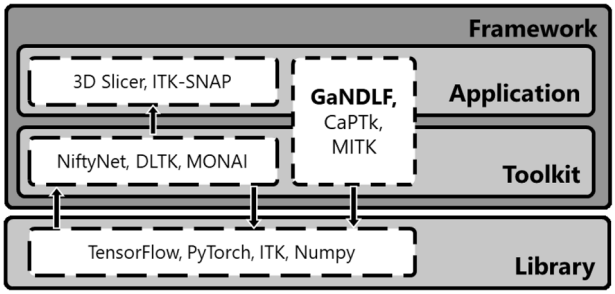


Fig. 2 Schematic categorization of open-source software, with representative examples. 'Libraries' focus on software developers offering access to low-level machine functionality. 'Toolkits' target computational experts and provide a layer of abstraction to enable research, by requiring users to write code to enable their functionality. 'Applications' focus on the non-computational end-user offering their functionality via user interfaces. 'Frameworks' fulfil both roles of 'Applications' and 'Toolkit', and provide a multitude of functions targeting both computational and non-computational end-users. Light gray represents software that a user interacts with on a lower level, and dark gray represents interaction using a command line or graphical interface.

advanced computational research community; (ii) can be difficult to grasp by researchers without sufficient experience in DL; (iii) do not make it easy for DL scientific developers to write their architectures in a generalizable way, allowing their application on problems spanning across domains; (iv) make it difficult to write reproducible training pipelines for different problem domains; (v) put the onus of training robust and generalizable models to the user's knowledge of the training mechanism and the dataset in question; (vi) lack a single end-to-end application programming interface (API) for training and inference that can span across various problem domains; or (vii) do not have appropriate level of interpretability or explainability functionality for researchers to garner meaningful insights into the training.

Here, we introduce the Generally Nuanced Deep Learning Framework (GaNDLF) as a community-driven open-source framework by MLCommons, which is an industry-academic partnership aiming to accelerate the adoption of machine learning innovation to benefit the larger community, to enable both clinical and computational researchers address various AI workloads (such as segmentation, regression, and classification), while producing robust AI models without requiring extensive computational experience. This is done by focusing on ensuring that AI algorithms and pipelines follow paradigms adhering to best practices established by the greater ML community, and leveraging existing collaborative efforts in the space (such as the MLCommons' MedPerf[40]). Such practices include: (i) nested cross-validation[41]; (ii) handling class imbalance[42]; and (iii) artificial augmentation of training data. Additionally, GaNDLF incorporates capabilities to handle end-to-end processing (i.e., pre- and post-processing steps) in a cohesive and reproducible manner to contribute towards democratizing AI in healthcare, while these best ML practices are at the forefront during training and inference. GaNDLF has been developed in PyTorch/Python as an abstraction layer that incorporates widely used open-source libraries and toolkits (such as MONAI[32]) that can help researchers generate robust AI models quickly and reliably, facilitating reproducibility and being consistent with the criteria of findability, accessibility, interoperability, and reusability (FAIR). Furthermore, the flexibility of its codebase permits GaNDLF to be used across modalities (e.g., 2D/3D radiology scans, and 2D multi-level histology whole slide images (WSI)), and has scope and functionality for integrating other clinical data (such as genomics and electronic health records) in the future, thus taking current clinical diagnostics to the next frontier of quantitative integration.

Results

To highlight the generalizability of the framework, GaNDLF was applied on both radiology and histology data for a variety of DL workloads/tasks (i.e., segmentation, regression, and classification) on multiple organ systems, imaging modalities, and various applications using numerous DL architectures. For each workload, we performed extensive performance evaluation using dedicated testing (or holdout[43]) datasets by averaging each model's training run in a cross-validated schema, ensuring stable model performance reporting without overfitting to a specific data split. Details regarding the experimental design of each application are shown in the Methods' Experimental Design section. The reported results for all the performed experiments are on the unseen testing (or holdout[43]) cohorts for each application, and collectively shown in Table 1.

Segmentation workloads. We applied GaNDLF to solve various segmentation problems on imaging acquired during standard clinical practice for multiple anatomical sites, comprising of brain, eyes, breast, lung, maxillofacial region, and colon. Numerous DL architectures, designed for segmentation workloads, were evaluated for multiple applications. These architectures include UNet, UNet with residual connections (ResUNet), Fully Convolutional network (FCN), and UNet with inception modules (see Methods section for details and Supplementary Figs. 1–10 for illustrations). Respective results are reported after quantitative performance evaluation based on Dice Similarity Coefficient ("Dice"). Note that GaNDLF offers the ability to generate other segmentation-specific metrics, such as the Hausdorff distance.

Several applications used brain Magnetic Resonance Imaging (MRI) scans, focusing on brain extraction (also known as skull-stripping)[11,44], boundary detection of histologically distinct brain tumor sub-regions[7–10], as well as comprehensive brain parcellation[45]. For brain extraction, we used each structural MRI volume as a separate independent input, with the goal of training a computational model that can segment the brain tissue region regardless of the input modality, and remove all non-brain tissue (e.g., neck, fat, eyeballs, and skull). In our analysis, we observed that the ResUNet architecture gave the best results, with average "Dice" of 0.98 ± 0.01. For brain tumor sub-regions, we considered the areas of necrosis, enhancing tumor, and peritumoral edematous/infiltrated tissue, following the convention of the International Brain Tumor Segmentation (BraTS) challenge[7–10]. To train these models, we used all four structural MRI volumes in tandem as input. For this application, the ResUNet architecture was again observed to give the best results with an average "Dice", across all the 3 sub-regions, of 0.71 ± 0.05. For brain parcellation, we segmented 133 fine-grained brain regions from the whole brain MRI scans[45]. In our analysis, we observed that ResUNet gave the most satisfactory results for the problem, with average "Dice" of 0.68 ± 0.15.

For the anatomical site of breast, we had two distinct applications. Firstly, we segmented the background, fatty breast tissue, and dense breast tissue from digital breast tomosynthesis scans[46]. Our experimentation resulted in the most optimal "Dice" scores using ResUNet, with an average of 0.94, 0.89, and 0.49, for each of the aforementioned regions, respectively, with an overall performance of 0.78 ± 0.09. Secondly, we segmented the structural tumor volume region from T1-weighted pre-contrast, peak-contrast and post-contrast injection scans using the ISPY-1 cohort[47]. We observed the best performance using ResUNet with an average "Dice" of 0.74 ± 0.01.

Table 1 Results of various DL workloads using GaNDLF for multiple anatomies.

Task	Organ	Application	Dims	Input modalities (number): type	Output classes	Architecture	Metric	
							Type	Average value
Segmentation	Brain	Brain extraction	3	(1): T1, T1Gd, T2, T2-FLAIR as individual inputs	1	UNet	Dice	0.97 ± 0.01
						ResUNet	Dice	0.98 ± 0.01
						FCN	Dice	0.97 ± 0.01
		Tumor sub-region segmentation		(4): T1, T1Gd, T2, T2-FLAIR	3	UNet	Dice	0.65 ± 0.05
						ResUNet	Dice	0.71 ± 0.05
						FCN	Dice	0.62 ± 0.05
						UInc	Dice	0.64 ± 0.05
		Brain parcellation		(1): T1	133	ResUNet	Dice	0.68 ± 0.15
						UNet	Dice	0.57 ± 0.26
	Breast	Breast segmentation	3	(1): Digital breast tomosynthesis	3	UNet	Dice	0.78 ± 0.09
		Tumor segmentation		(3): T1 pre, peak, and post-contrast injection	1	ResUNet	Dice	0.74 ± 0.01
	Lung	Lung field segmentation	3	(1): CT [Lung Cancer Screening]	1	ResUNet	Dice	0.95 ± 0.02
				(1): CT [COVID-19]		ResUNet	Dice	0.97 ± 0.01
	Eye	Fundus segmentation	2	(1): RGB Fundus Images	1	UNet	Dice	0.85 ± 0.04
						ResUNet	Dice	0.90 ± 0.05
						FCN	Dice	0.81 ± 0.04
						UInc	Dice	0.83 ± 0.03
	Dental	Quadrant segmentation	2	(1): X-Ray	4	UNet	Dice	0.91 ± 0.01
						ResUNet	Dice	0.88 ± 0.01
						FCN	Dice	0.85 ± 0.02
	Colon	Colorectal cancer segmentation	2	(1): Histology H&E	1	ResUNet	Dice	0.78 ± 0.03
Regression	Brain	Age prediction	2	(1): T1 slices	1	Specialized VGG	MSE	0.0141 ± 0.01
Classification	Brain	EGFRvIII status prediction	3	(4): T1, T1Gd, T2, T2-FLAIR	2	VGG11	Acc	0.74 ± 0.08
	Foot	Diabetic foot ulceration	2	(1): RGB Foot Images	4	VGG11	Acc	0.92 ± 0.01
						VGG16	Acc	0.90 ± 0.01
						VGG19	Acc	0.89 ± 0.01
						DenseNet121	Acc	0.87 ± 0.01
	Pan-Cancer	TIL Prediction	2	(1): Histology H&E	2	ImageNet_ VGG16	Acc	0.89 ± 0.01

The "Task" showcases the workload type, "Organ" describes the organ system of the data, "Application" describes the use case for the trained model(s), "Dims" describe the dimensionality for each input modality, "Input Modalities" describes the total number of input modalities for the model to train on, "Output Classes" shows the number of classes the model should be predicting, "Architecture" describes the network topology, and "Metric" describes the type and average value of the selected metric on the testing/holdout dataset, and is "Dice" for segmentation tasks, Mean squared error or "MSE" for regression, and Balanced accuracy or "Acc" for classification.

For lung, we used low-dose Computed Tomography (CT) scans acquired for both lung cancer screening and COVID-19 assessment, with the intention to segment the lung field incorporating apparent healthy and abnormal tissue. Application of GaNDLF's ResUNet architecture on scans for both these applications (i.e., cancer screening and COVID-19 assessment), we observed "Dice" scores of 0.95 and 0.97, respectively.

For the anatomical site of the eyes, we segmented the fundus region in Red-Green-Blue (RGB) retinal scans[48], and observed that the ResUNet architecture gave the best results, with the average "Dice" coming to 0.71 ± 0.05.

For the anatomical region of maxillofacial, we have used panoramic dental x-ray images, with the goal of distinguishing and accurately segmenting the four quadrants considered in dental practice[49]. After training various DL architectures, we observed that UNet yielded the best results, with the average Dice coming to 0.91 ± 0.01.

Last but not least, and to evaluate GaNDLF's performance beyond radiology scans, we utilized histology digitized tissue sections (e.g., whole slide images (WSI)), stained for Hematoxylin and Eosin (H&E), of colorectal cancer by leveraging the publicly available dataset of the DigestPath challenge[50], with the intention of delineating the cancerous regions. Our results yield and average "Dice" of 0.78 ± 0.03 using ResUNet for a pre-defined testing data split.

Regression. For the DL workload of regression, we have used GaNDLF to solve a specifically targeted regression problem in brain MRI scans, focusing on predicting a surrogate index for brain age[51]. By virtue of the inherent flexibility in GaNDLF's design, we modified the VGG16 architecture to predict the age of a brain from a single MRI slice, and replicated previously reported results[51]. The input for this use case was based on 2D MRI slices of T1-weighted scans, and the output was the brain age. With an average mean squared error ("MSE") of 0.0141, the prediction quality of the models trained by GaNDLF was in line with the original publication[51], showcasing the flexibility of GaNDLF to successfully adapt to various problem domains.

Classification. We have further used GaNDLF to solve multiple classification problems, spanning different domains (e.g., radiology and histology), as well as various organ systems, including feet, brains, and pan-cancer histology images.

Specifically, for the anatomical location of brain, we have applied GaNDLF on 3D MRIs of patients diagnosed with *de novo* glioblastoma, to predict the EGFRvIII mutation status. The inputs for model training were structural MRI scans in tandem (passing all the scans together at once) to a VGG11 customized to perform computations in 3D directly, resulting in a best accuracy of 0.74 ± 0.08.

Furthering the application towards 2D RGB data, we have predicted different ulceration status of diabetic foot images from the DFU challenge[52] by passing each image as an input along with its ground truth label. We observed the best performance on VGG11[53] (that was randomly initialized instead of being pretrained on ImageNet), with a macro-F1 score of 0.561. Notably, the defined approach[53] was among the top-performing ones (ranked 5th) in the International in the DFU Challenge 2021 leaderboard (dfu-2021.grand-challenge.org/evaluation/challenge/leaderboard).

Finally, we used a dataset of histology digitized tissue sections stained for H&E, spanning across 12 anatomical sites. The problem at hand was to predict patches containing tumor-infiltrating lymphocytes (TIL)[54]. We observed the best-balanced classification accuracy of 0.89 using a VGG16 that was pretrained on ImageNet[55] and customized for the specific problem.

Discussion

We have introduced the Generally Nuanced Deep Learning Framework (GaNDLF), as an end-to-end solution for scalable clinical workflows, currently focused on (bio)medical imaging. GaNDLF provides a "zero/low-code" solution enabling both computational and non-computational experts to train robust DL models to tackle a variety of workloads/tasks in both 2D and 3D radiology and histology data, without worrying about details such as appropriate data splitting for training, validation, and testing, tackling class imbalances, and implementing various training strategies (e.g., loss functions, optimizers). Specifically, GaNDLF's contribution spans across its ability to: (i) process images of various domains, including both radiology scans and digitized histology WSIs; (ii) enable work on various workloads (i.e., segmentation, regression, and classification); (iii) offer built-in general-purpose functionality for augmentations and cross-validation; (iv) be evaluated on a multitude of applications; (v) enable parallel training by using generic high-performance computing protocols; (vi) integrate tools to promote the interpretability and explainability of DL networks, via M3D-CAM[56].

Our overarching goal is to enable clinical translation and applicability of AI, since specialized hardware (e.g., DL accelerator cards) is usually not considered for purchase by clinical entities in higher income countries, and altogether out of reach for clinics in lower-income countries. Towards this end, we have developed built-in model optimization support in GaNDLF to automatically generate optimized models after the training process is complete, allowing inference of these models on machines without requiring any specialized hardware, or large amounts of memory. We further envision the "model library" in GaNDLF to potentially be a phenomenal resource for pre-trained models and corresponding configurations to replicate training parameters for the scientific community in general. By ensuring that the model library contains information beyond just the trained model weights, but also additional metadata, trained models through GaNDLF will remain reproducible through code changes. GaNDLF is a fully self-contained DL framework that has various abstraction layers to enable researchers to produce and contribute robust DL models with absolutely zero knowledge of DL or coding experience.

The concepts of "zero-" and "low-" code principles in software development have recently been introduced, targeting different user groups. In essence, the "zero-code" principle revolves around allowing users to build solutions without writing any code, whereas the "low-code" principle allows customization of the provided solution with minimal programming. GaNDLF follows these zero/low-code principles and enables targeting a dual audience type: (i) non-computational experts, by providing building blocks for conducting DL analyses by leveraging their domain expertise without the need for any programming skills; (ii) DL researchers, allowing for harmonized I/O (i.e., common data loaders enabling the main focus be kept on the algorithmic development), as well as leveraging or extending existing capabilities to create custom solutions. For a non-computational researcher, GaNDLF ensures the easy creation of robust models using various DL architectures, and built-in automatically triggered ML principles, that can be used for scientific research and method discovery, including the potential for aggregating results from various models, which has been shown to provide greater accuracy[7,10]. For DL researchers/developers, GaNDLF provides a mechanism for creating customized solutions, robust evaluation of their methods across a wide array of medical datasets that span across dimensions, channels/modalities, and prediction classes, as well as to conduct a comparative quantitative performance evaluation of their algorithm against well-established built-in network architectures, including, but not limited to, UNet[57], UNetR (UNet with transformer encoding)[58], ResNet[59], and EfficientNet[60]. Furthermore, GaNDLF provides the means to DL researchers/developers to distribute their methods in a reproducible way to the wider community, thereby expanding their application across various problem domains with relative ease, and providing re-usable components (Fig. 1) that can be combined to create customized solutions. Ideally, we anticipate the best results when both these groups of the scientific and clinical community bring their expertise together to further our understanding of healthcare. Towards this end, GaNDLF can provide a common frame of reference for both these user groups. By creating tools standardized within the same infrastructure (GaNDLF) for the entire community to leverage, we anticipate the cost and time of creating algorithms to be substantially reduced and hence put efforts in meaningfully translating methods into the clinical practice rather than trying to identify and/or make a tool to work.

The modularity of the software stack is highlighted by large-scale studies of specific focus on federated learning (FL) that GaNDLF has facilitated, beyond the results shown in this manuscript. The FL-specific functionality is provided by its integration to work in conjunction with the Open Federated Learning (OpenFL) library[61]. Further integration with other community-driven efforts, such as MedPerf[40] (medperf.org) of MLCommons (mlcommons.org), would increase the applicability of GaNDLF towards federated learning applications[35,62]. GaNDLF has notably been used to orchestrate the Federated Tumor Segmentation (FeTS) Challenge[62], which represents the first-ever computational challenge on FL, targeting (i) the development of novel aggregation methods for federated training, and (ii) the federated evaluation of algorithms "in-the-wild", to assess algorithmic robustness to distribution shifts between medical institutions. Moreover, GaNDLF's codebase has facilitated components of the largest to-date real-world FL study (i.e., the FeTS Initiative[35] - www.fets.ai), involving data from 71 geographically-distinct collaborating sites to develop a DL model to detect boundaries of intrinsic sub-regions for the rare disease of glioblastoma in mpMRI scans. Finally, indicating its joint

ability with OpenFL to address workloads in various domains, the GaNDLF-OpenFL integration has enabled an FL histology study on identifying TILs in WSIs from numerous anatomical sites[63].

One of GaNDLF's core tenets is to enable work across domains, currently spanning radiology (e.g., MRI, CT) and histology (e.g., H&E-stained slides), including specialized pre-processing functionalities for each. The notable difference between these images is the relatively small resolution and size of radiology scans (typically occupying a few megabytes of disk space), compared with the histology WSI that are described by relatively large resolution (150 K × 150 K pixels) and size, where a single WSI can occupy 40–50 gigabytes. GaNDLF enables researchers to use a single framework across virtually all medical imaging modalities without performing any additional coding, thereby enabling future studies that rely on integrative diagnostics. Owing to the flexibility of the data loading mechanism in GaNDLF, it could also be possible to integrate other data types (such as genomic or healthcare records) into a model towards further contributing in the field of personalized medicine.

Although GaNDLF has been evaluated across imaging modalities using single inputs (i.e., either a single radiology or histology image) or with multi-channel support (i.e., multiple MRI sequences considered in-tandem), so far, its application has been limited to workloads related to segmentation, regression, and classification, but not towards synthesis, semi/self-supervised training, or physics-informed modeling. Expanding the application areas would further bolster the applicability of the framework. Additionally, application to datasets representing analysis of 4D images (such as dynamic sequences or multi-spectral imaging) has not yet been evaluated. Also, a mechanism to enable aggregation of various models (i.e., train/infer models of different architectures concurrently) is not present, which have generally shown to produce better results[7–10,33]. Mechanisms that enable AutoML[64] and other network architecture search (NAS) techniques[65] are tremendously powerful tools that create robust models, but are currently not supported in GaNDLF. Finally, application of GaNDLF to other data types, such as genomics or electronic health records (EHR), which would allow GaNDLF to further inform and aid clinical decision-making by training multi-modal models, has not been fully explored yet but it is considered as current work in progress.

To facilitate clinical applicability, reproducibility, and translation, in the domain of healthcare AI, published research is essential to adhere to well-accepted reporting criteria. Some of these criteria are: i) CLAIM (Checklist for Artificial Intelligence in Medical Imaging)[66], which outlines the information that authors of medical-imaging AI articles should provide, ii) STARD-AI, which is the AI-specific version of the Standards for Reporting of Diagnostic Accuracy Study (STARD) checklist[67], and aims to address challenges related to the original STARD checklist related to the utilization of AI models, iii) TRIPOD-AI and PROBAST-AI, which are the AI versions of the TRIPOD (Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis) statement and the PROBAST (Prediction model Risk Of Bias ASsessment Tool)[68], and aim to provide standards both for reporting but also for Risk of Bias assessment, raising awareness of the importance in meta-analyses dealing with AI studies, iv) CONSORT-AI and SPIRIT-AI, which are the AI extensions of the CONSORT (Consolidated Standards of Reporting Trials) and SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials), providing guidance for reporting randomized clinical trials[69], v) MI-CLAIM (Minimum Information about Clinical Artificial Intelligence Modelling)[70], which focuses on the clinical impact and the technical reproducibility of clinically relevant AI studies, vi) MINIMAR (MINimum Information for Medical AI Reporting)[71], which sets the reporting standards for medical AI applications by specifying the minimum information that AI manuscripts should include, and vii) Radiomics Quality Score (RQS)[72], which outlines 16 criteria by which to judge the quality of a publication on radiomics[73].

In conclusion, this manuscript describes GenerAlly Nuanced Deep Learning Framework ("GaNDLF"), a stand-alone package that provides end-to-end functionality facilitating transparent, robust, reproducible, and deployable DL research. Due to its flexible software architecture, it is possible to either leverage certain parts of GaNDLF in other applications/toolkits, or leverage functions of other toolkits (e.g., MONAI) and libraries to incorporate them within the holistic functionality of GaNDLF. Furthermore, GaNDLF could partner with container-based platforms beyond MedPerf[40] (such as the BraTS algorithmic repository, or ModelHub.AI) towards a structured dissemination of DL models to the research community. As all development is open-sourced in github.com/mlcommons/GaNDLF, with robust continuous integration and code vulnerability testing through Dependabot, contributions from the community will ensure that this framework continues building ties to other packages quickly and reliably for end users. Finally, by creating tools standardized within the same infrastructure (GaNDLF) for the entire community to leverage, we anticipate the cost and time of creating algorithms to be substantially reduced and hence put efforts in meaningfully translating methods into the clinical practice rather than trying to make a tool to work.

Methods

Pre-processing. Providing robust pre-processing techniques that are widely applicable to (bio)medical data, is critical for such a general-purpose framework to succeed. GaNDLF offers most of the pre-processing techniques already reported in the literature, leveraging the capabilities of basic standardized pre-processing routines from ITK[74,75], and advanced pre-processing functionality from the CaPTk[23,76–79]. The main pre-processing steps for data curation (including harmonization and normalization) are described below.

1. Anonymization:
   - Radiology Images: Since the DICOM format[80] is the standard for radiology images, GaNDLF has provisions to remove all identifiable fields from the DICOM metadata, as well as a conversion to the Neuroimaging Informatics Technology Initiative (NIfTI) file format[81], which completely removes all extraneous metadata fields.
   - Histology Images: Most WSIs include metadata which could contain protected health information, and GaNDLF can remove such fields from the file header. This works for multiple formats defined by the Open Microscopy Environment standard[82], such as TIFF, SVS, and MRXS.
2. Data harmonization:
   - Voxel-resolution harmonization: To ensure that the physical definition of the input data is in a common space (for example, all images can have the voxel resolution of $I_{res} = [1.0, 1.0, 2.0]$).
   - Image-resolution harmonization: To ensure that the input data has the same image dimensions (for example, all images can be resampled to $I_{dim} = [240, 240, 155]$).
3. Intensity normalization:
   - Thresholding: To consider pixel/voxel values that belong to a specific intensity range and ignore values below/above this range, by making them equal to zero (Eq. (1)):

$$x_i = \begin{cases} 0 & x_i < threshold_{min} \\ 0 & x_i > threshold_{max} \\ x_i & otherwise \end{cases} \quad (1)$$

   - Clipping: To consider pixel/voxel values that belong to a specific intensity range and convert values below/above this range, by making them equal to the minimum/maximum threshold, respectively (Eq. (2)):

$$x_i = \begin{cases} threshold_{min} & x_i < threshold_{min} \\ threshold_{max} & x_i > threshold_{max} \\ x_i & otherwise \end{cases} \quad (2)$$

   - Rescaling: To consider all pixel/voxel values after converting them to a common profile (for example, all input images are rescaled to [0, 1]).

- Z-score normalization: A widely used technique for data normalization in medical imaging[83,84], that preserves the complete signal of the input image by subtracting the mean and then dividing by the standard deviation of the complete intensity range found in this image. Notably, the application of z-score normalization through GaNDLF can occur either on the full image or only within a masked region of interest, adding to the overall flexibility of this transform.
- Histogram Standardization[85] ensures harmonization of intensity profiles of input images based on a template (or reference) image. Different options are available to the user, such as histogram matching[85], adaptive histogram equalization[86] and global histogram equalization. Normalization methods specifically designed for WSIs that calculate stain vectors are also available, and these include methods from Vahadane[87], Ruifork[88], and Macenko[89].

Data augmentation. DL methods are well-known for being extremely data hungry[90,91] and in medical imaging, data is scarce because of various technical, privacy, cultural/ownership concerns, as well as data protection regulatory requirements, such as those set by the Health Insurance Portability and Accountability Act (HIPAA) of the United States[92] and the European General Data Protection Regulation (GDPR)[93]. This necessitates the addition of robust data augmentation techniques[94] into the training data, so that models can gain knowledge from larger datasets and hence be more generalizable to unseen data[95].

GaNDLF leverages existing robust data augmentation packages, such as TorchIO[96] and Albumentations[97]), to provide augmentation transformations in a PyTorch-based mechanism. GaNDLF also stores image metadata (such as affine transform, origin, resolution), which is critical for maintaining correct physical coordinate definition of radiology scans. More details on the available types of augmentations through GaNDLF are shown in the Supplementary Notes 1: Details of Data Augmentation (Supplementary Table 1), and examples of their effects are illustrated in Fig. 3, using a brain tumor T2-FLAIR MRI scan from the BraTS challenge dataset[7–10].

Training mechanism. The overall pipeline of the training procedure offered in GaNDLF is illustrated in Fig. 4a, and focuses on stability and robustness for the user to generate reproducible results, and clinically-deployable models. Figure 4b showcases the overall software stack. The data flow of GaNDLF leverages 2 main ideas that allow efficient processing of large datasets (such as histology images or large 3D volumes): (i) patch-based training and inference, which allows the model to operate on smaller "chunks" of the data at a single instance, and hence on the full gamut of images - the size and overlap of these chunks can be customized by the user, (ii) lazy loading of the datasets themselves, allowing GaNDLF to only read the datasets into the memory during computation, and immediately deallocate the memory once it is used.

Cross-validation. *k*-fold cross validation[98,99] is a useful technique in ML that ensures reporting unbiased quantitative performance evaluation estimates of algorithmic generalizability on new datasets, i.e., by evaluating results on new unseen data discretized from an entire given data cohort. GaNDLF offers a nested *k*-fold cross-validation schema[100], where initially, cases of the complete cohort are randomly divided into k non-overlapping, equally-sized subsets and during each fold, $k-1$ of these subsets are considered as the "retrospective"/"discovery" cohort and 1 as the "prospective"/"replication" cohort, which is unseen during training for this specific fold. Note that during each fold, the "prospective"/"replication" cohort is a different subset. This cross-validation scheme is analyzing the given data as if it had independent discovery and replication cohorts, but in a more statistically robust manner by randomly permuting across all given data. The number of folds for each level of split is specified in the configuration file, and the models for different folds can be trained in parallel (in accordance with the user's computation environment). GaNDLF also offers the option of specifying single fold training, if so desired.

Zero/low-code principle. The main entry point of GaNDLF's training mechanism follows a zero/low-code principle[101,102], where a dual file input is provided by the user, through the command line interface - a comma-separated-value (CSV) file and a text file (YAML) with intuitive indications of where to enter the training configuration parameters. The expected CSV file should comprise the subject identifiers along with the corresponding full paths of all required input images and masks (i.e., for segmentation workloads) and the values required for training and follow-up predictions (i.e., for regression and classification workloads). The subject identifiers are used to randomly split the entire dataset into training, validation, and testing subsets, using nested *k*-fold cross-validation[103]. The training can be configured to run on multiple DL accelerator cards, such as GPU or Gaudi. Furthermore, a YAML-based configuration file is used to control and parameterize all aspects of the training, such as the subject-based split of the cross-validation, data pre-processing, data augmentations (e.g., type, parameters, and probabilities), model parameters (e.g., architecture, list of classes, final convolution layer, optimizer type, loss function, number of epochs, scheduler, learning rate, batch size), along with the training queue parameters (i.e., samples to extract per volume, maximum queue length, and number of threads to use). The YAML-based configuration file requires an indication of the GaNDLF version used to create the trained model, and the actual trained model, with the intention of ensuring coherence between these two.

Monitoring & debugging. GaNDLF also supports mixed precision training[104] to save computational resources and reduce training time. A single epoch comprises training the model using the training portion of the data and backpropagation of the generated loss, followed by evaluating the model performance on the validation portion of the data. In addition to saving the model trained after every epoch, each model corresponding to the best global losses for the training, validation, and testing datasets is also saved. These saved models can be used for subsequent inference, either using a single independent model or in a aggregated fashion utilizing label fusion[53,105]. Training statistics (such as the "Dice" similarity coefficient and loss) are stored for each epoch, for the training, validation and the testing data in the form of a CSV file, with the intention of facilitating simplified results reporting and detailed debugging.

Handling class imbalance. Class imbalance, i.e., where the presence of one class is significantly different in proportion to another, is a common problem in healthcare informatics[106,107]. To address this issue, GaNDLF allows the user to set a penalty for the loss function[108], which is inversely proportional to the classes being trained on. The penalty weights for the loss function will be defined as:

$$p_c = 1 - \frac{n_c}{N} \tag{3}$$

where p_c is the penalty for class 'c', and n_c is the number of instances of the presence of class 'c' in the total number of samples N.

For example, for a classification workload using 100 cases, if there are 10 from class 0 and 90 from class 1, the weighted loss will get calculated to 0.9 for class 0 and 0.1 for class 1. This basically means that the misclassification penalty during loss back-propagation for class 0 (i.e., the "rarer" class) will be higher than that of class 1 (i.e., the more "common" class). The analogous process can be done for segmentation workloads as well. We recognize that this approach might not work for all problem types, and thus we have mechanisms for the user to specify a predetermined loss penalty for greater customization.

Inference mechanism. GaNDLF's inference mechanism follows the same "zero/low-code" principle as its training mechanism, where the user needs a CSV file comprising of the subjects' identifiers and the full paths of images, along with a YAML configuration file and the location of the trained models. For each trained model, the corresponding estimated output is stored and (depending on the user's parameterization) a final predicted output is generated by aggregating the outputs of the independent models. This aggregation happens through different approaches, subject to the prediction task, e.g., a label fusion approach may be used for segmentation workloads, averaging for regression workloads, and majority voting for classification. If the full paths of the ground truth labels are given in the input CSV, then the overall metrics (e.g., "Dice" and loss) of the model's performance are also calculated and stored.

For radiology scans. As soon as the data is read into memory, GaNDLF applies the pre-processing steps defined in the configuration file to each input dataset (see Section u for examples of these steps). Then TorchIO's[96] inference mechanism is used to enable patch-based inference for radiology images. This entails patch extraction, usually of the same size as the one that the corresponding model has been trained on, from the image(s) on which the model needs to infer. The forward pass of the model is then applied, and the result is stored in the corresponding location (Fig. 5a). This enables models to be trained and inferred on varied patch sizes based on the available hardware resources. Overlapping patches can be stitched by either cropping or taking an average of the predictions at the overlapping area, and the amount of overlap can be specified to ensure that dense inference can occur[96]. Although patch-based training and inference is being widely used, we note that various potential adverse effects of this process have been reported[109], requiring the operator's attention.

For histology WSIs. Histology WSIs need a different inference mechanism, than that for training, primarily due to their increased hardware requirements, i.e., WSIs can require more than 50GB when loaded completely on-memory. Fig. 5b illustrates this inference mechanism, which starts with the extraction of a WSI's imaging component at the maximum magnification/resolution (e.g., ×40) and its conversion to a TIFF with 9–10 layers of tiled images with different magnification levels (i.e., Fig. 5b(ii)-(iii) - "Data Fixing Pipeline"). The background area is then filtered out through the generation of a 'tissue mask' (Fig. 5b(iv)), using Red-Blue-Green (RGB) and Otsu-based thresholding[110,111], which is necessitated by the need to correctly tackle image reading issues occurring when trying to buffer any magnification level other than the lowest. This 'tissue mask' reduces the search space for downstream analyses, and hence reduces the overall computational footprint. This mask is further used to calculate foreground coordinates (Fig. 5b (v)), around which patches are extracted on-the-fly by leveraging TiffSlide's[112] dynamic read region property (Fig. 5b(vi)). This produces a

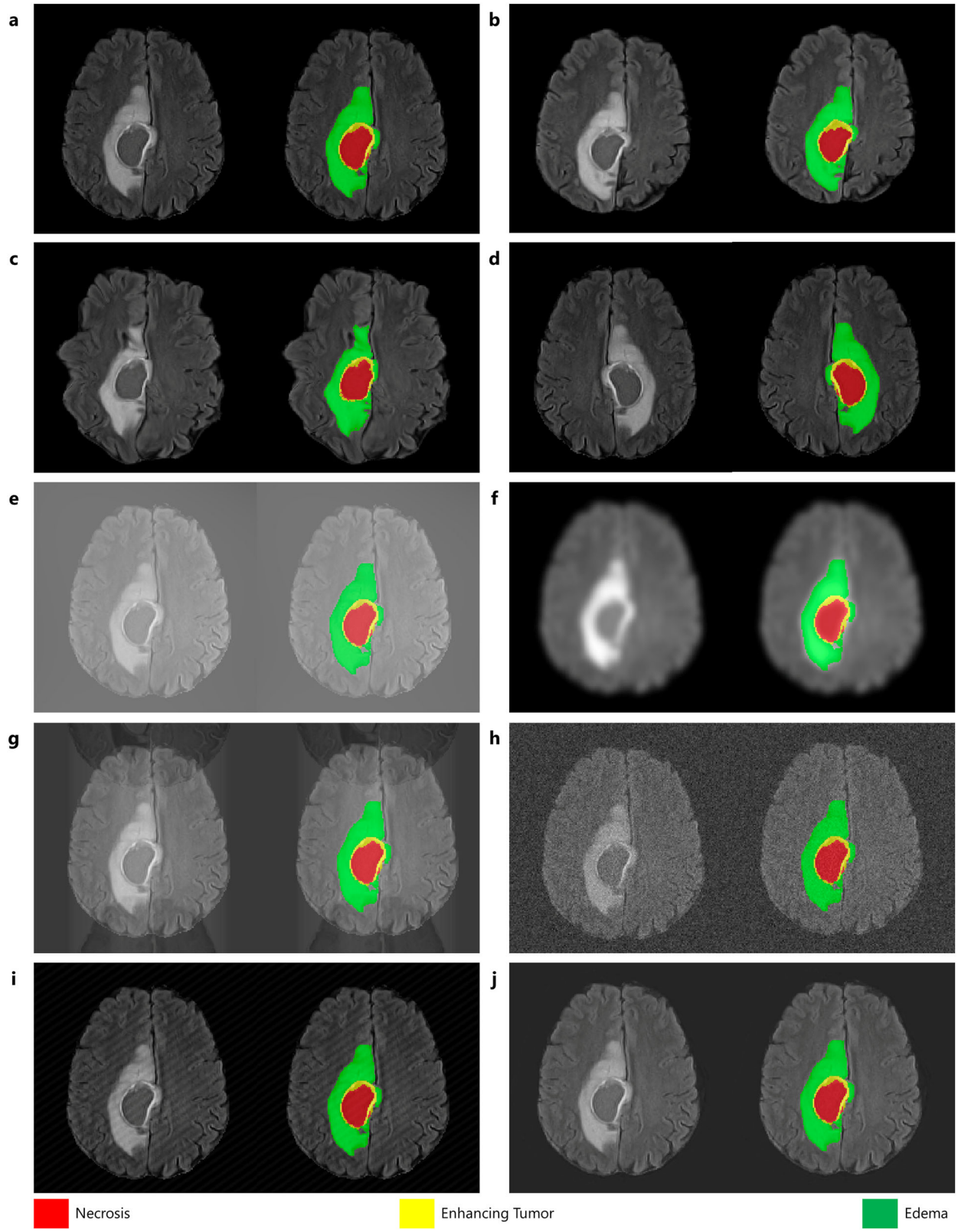


Fig. 3 Illustration of various data augmentation techniques available in GaNDLF showing the image and overlaid segmentation. a Original image, **b** affine augmentation, **c** elastic augmentation, **d** flip augmentation, **e** bias augmentation, **f** blur augmentation, **g** ghosting augmentation, **h** noise augmentation, **i** spike augmentation, and **j** motion augmentation.

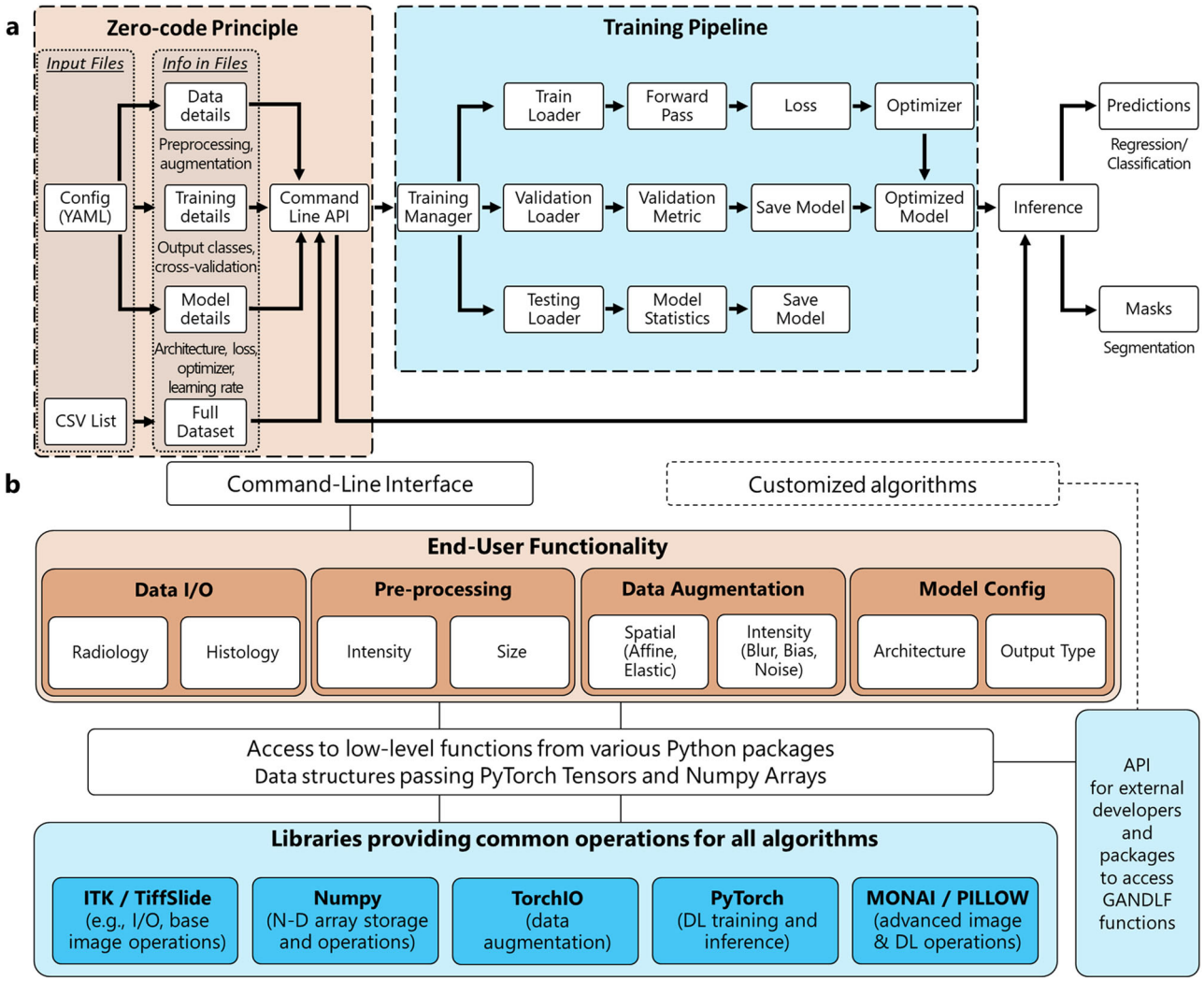


Fig. 4 Overall structure of GaNDLF. a Flowchart depicting the overall training procedure pipeline offered in GaNDLF. **b** GaNDLF 's software stack, highlighting the use of various low and high-level libraries to facilitate the creation of a flexible framework with an easy-to-use end-user interface. The orange pipeline represents the functionalities that can be changed using the "zero/low-code" principle, and blue pipeline represents the lower-level interactions of the codebase.

'count map' (Fig. 5b(viii)), which accounts for the contribution of overlapping patches for a tissue region ensuring probabilities are always between 0 and 1. The trained model is then used for a forward pass on each of these patches, producing an independent prediction for each. These predictions are then stitched together to form a 'segmentation probability map' (Figure 5b(ix)). The 'segmentation probability map' and the 'count map' are then multiplied to generate the 'final segmentation' output (Fig. 5b(x)).

Post-processing. It is conceivable that post-processing of a prediction would be required to get the most accurate result. GaNDLF provides a post-processing module that includes common image processing tasks, such as morphological operations (i.e., dilation, erosion, closing, opening), and the ability to map predicted labels from one value to another. The former is useful in cases where segmentation predictions are generated with holes and need to be closed, and the latter can be used to assign the desired final label values to a prediction.

Modularity and extendibility. A description of GaNDLF's software stack, modularity, and extendibility is hereby provided, as well as how the lower-level libraries are utilized to create an abstract user interface, which can be customized based on the application at hand. Following this, the flexibility of the framework from a technical point-of-view is chronicled, which illustrates the ease with which new functionality can be added. Further details on customizing the entire processing pipeline (including hyper-parameter tuning and optimization) can be found in the software documentation at: mlcommons.github.io/GaNDLF.

Software Stack. The software stack of GaNDLF, illustrated in Figure 4b, depicts the interconnections between the lower level libraries and more abstract functionalities exposed to the user via the command line interface. This ensures that a researcher can perform DL training and/or inference without having to write a single line of code. Furthermore, the flexibility of the stack is demonstrated by the ease with which a new component (e.g., a pre-processing step, or a new network architecture) can be incorporated into the framework, and subsequently applied to new types of data/applications with minimal effort. Specifically, the framework's flexibility affects components listed in the following subsections.

Dimensions. To ensure maximum flexibility and applicability across various types of data, GaNDLF supports both 3D and 2D datasets. Using the same codebase, GaNDLF has the ability to apply various architectures across diverse modalities such as MRI, CT, retinal, and digitized histology WSI, including immunohistochemical (IHC), In Situ Hybridization (ISH), and H&E stained tissue sections.

Input channels/modalities & output classes. GaNDLF supports multiple input channels/modalities/sequences and output classes, for either segmentation, classification, or regression, to ensure maximal applicability across various problem domains, whether it involves a binary task (e.g., brain extraction) or multi-class outputs (e.g., brain tumor sub-region segmentation).

- Radiology images require the ability to process both 2D and 3D data. Although imaging examples that GaNDLF has been applied and evaluated so far describe CT, MRI, and tomosynthesis scans, it offers support for almost every radiology image via ITK.

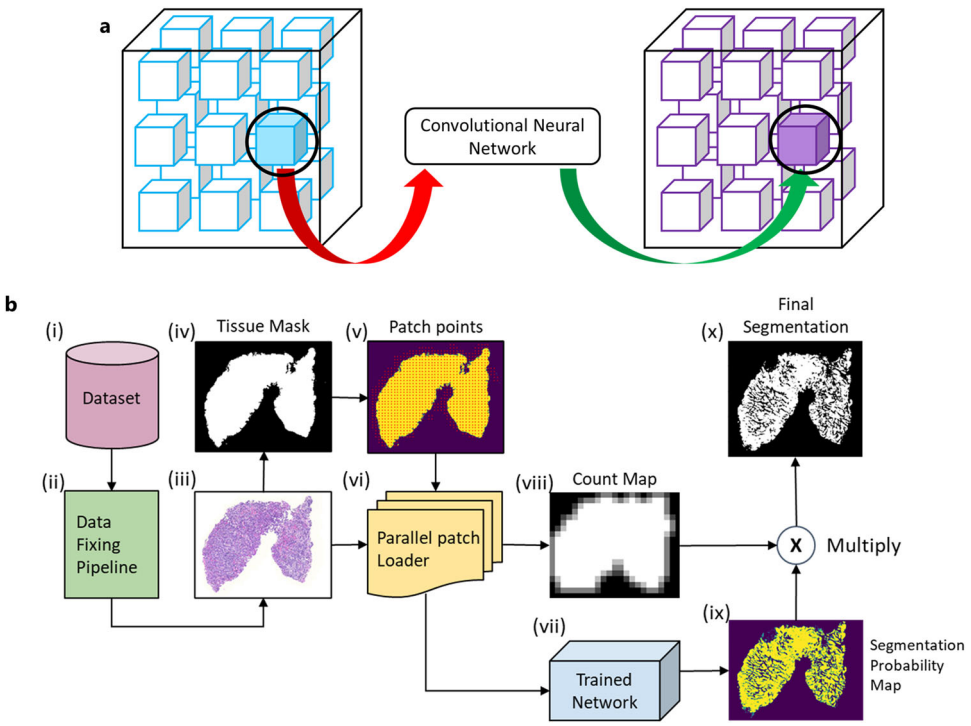


Fig. 5 GaNDLF's inference mechanism for medical images. a Illustration of the patch-based inference mechanism for radiology images. This process is repeated till the complete image gets processed. **b** Steps to perform specialized inference for histology images. Starting with the raw whole slide image (WSI), multiple specialized pre-processing steps (ii-vi) are performed before a patch can be given as input to a trained model. The coordinates of each patch need to be saved along with the overlap information in order to obtain the final result.

- Histology images, on the contrary, require specialized handling along the following criteria:
  - Input: The use of TiffSlide[112] allows GaNDLF to read a fraction of the entire WSI data at the resolution closest to the requested magnification level, thereby ensuring memory-efficiency.
  - Patch-extraction: Since a WSI cannot always be processed on its entirety due to hardware constraints, a patch-based mechanism considering multiple resolutions is essential. This mechanism is offered through our open-source Open Patch Miner (OPM, github.com/CBICA/OPM), which has been integrated within GaNDLF for simple and rapid batch-processing of patches. OPM can automatically mask tissue in a WSI and convert the highest available resolution to square patches, given a pre-defined overlap amount and patch dimensions. Specifically, it extracts patches with the pre-defined overlap using a pseudo-grid and parallel sampling adjustable for tissue inclusion, in proportion to different tissue classes (for classification workloads), and while omitting the background region.

Network architectures. GaNDLF seeks to provide both well-established and state-of-the-art network architectures showing promise in the field of healthcare. The currently available (and ever expanding) architectures offered through GaNDLF, and their detailed descriptions are provided in the Supplementary Methods: Network Architectures as well as their illustrations in Supplementary Figs. 1–10.

Applications. As previously stated, GaNDLF can train DL models to target various workloads, including segmentation, regression, and classification. Depending on available resources, most models can be extended for all these workloads (such as UNet), and there are workload-specific models, such as the brain age prediction model[51], which modifies a VGG-16 model pre-trained on ImageNet weights and is only defined for regression. The flexibility of GaNDLF's framework makes it possible for all these models to co-exist and to leverage the robustly designed data loading and augmentation mechanisms for future study extensions. Having a common API for all these workloads also makes it relatively easy for researchers to start applying well-defined network architectures towards various problems and datasets, thereby contributing in getting DL-based pipelines into clinical workflows.

Performance evaluation. We provide different options to evaluate the model performance during training, and mechanisms to incorporate new validated recommendations[113] as needed. Below definitions of the metrics used in the results section of this manuscript are provided. Specifically, for segmentation workloads, the "Dice Similarity Coefficient"[114] (Eq. (4)) is mostly used as the performance evaluation metric, and all related models were trained to maximize it. "Dice" is a common metric used to evaluate the performance of segmentation workloads. It measures the extent of spatial overlap, while taking into account the intersection between the predicted masks (PM) and the provided ground truth (GT), hence handles over- and under-segmentation.

$$Dice = \frac{2|GT \cap PM|}{|GT| + |PM|} \tag{4}$$

Additionally, the "Hausdorff Distance"[115] is a metric for segmentation workloads (Eq. (5)). This metric quantifies the distance between the boundaries of the ground truth labels against the predicted label. It is sensitive to local differences, as opposed to "Dice", which represents a global measure of overlap.

$$H_{95}(PM, GT) = \max\left\{ P_{\substack{95\% \\ p \in PM}} d(p, GT), P_{\substack{95\% \\ g \in GT}} d(g, PM) \right\} \tag{5}$$

where $d(x, Y) = \min_{y \in Y} ||x - y||$ is the distance of x to set Y.

For regression workloads, we used the Mean Squared Error ("MSE")[116] as our evaluation metric and all models were trained to minimize it. "MSE" measures the statistical difference between the target prediction T and the output of the model P for the entire sample size n (illustrated by Eq. (6)). The same mechanism has been used for accuracy, macro-averaged F1-score, and area-under-the-curve, among others by leveraging TorchMetrics[117].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(T_i - P_i)^2 \tag{6}$$

For classification workloads, we used the balanced accuracy ("Acc")[118] as an evaluation metric and trained models to minimize the cross entropy loss[119]. "Acc" can be used for both binary and multi-class classification, and is defined the arithmetic mean of sensitivity and specificity. This metric is especially useful when dealing with imbalanced data, i.e. when one of the target classes appears a lot more than the other[118].

$$Acc = \frac{\left(\frac{TP}{TP+FN}\right) + \left(\frac{TN}{TN+FP}\right)}{2} \tag{7}$$

where TP & TN are the number of true and false positives, and FP & FN are the number of false positives and negatives, respectively.

Interpretability tools. It is an ongoing problem that deep neural networks lack the interpretability or explainability necessary for medical practitioners to trust into the

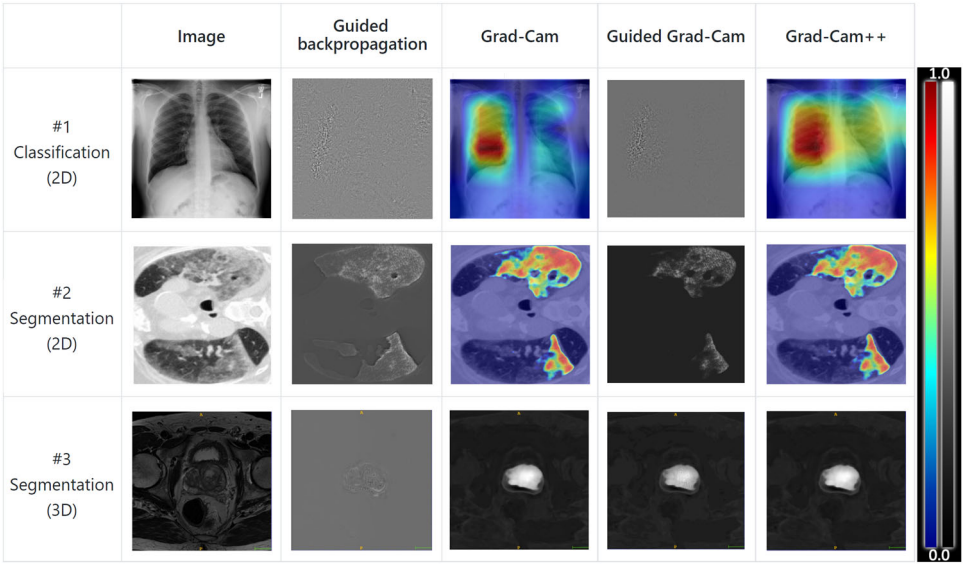


Fig. 6 Examples of generated M3D-CAM attention maps[56] with the Grad-CAM backend. The top row of images show the attention from a 2D classification network, the middle row from a 2D segmentation network and the bottom row from a 3D segmentation network.

networks decisions, hindering the practical application of such models in clinical practice[120,121]. To counter this, GaNDLF integrated the PyTorch library M3D-CAM[56], which enables the easy generation of attention maps of CNN-based models for both 2D and 3D data, and is applicable to both classification and segmentation models (examples illustrated in Fig. 6). The attention maps can be generated with multiple methods: Guided Back-propagation[122], Grad-CAM[123], Guided Grad-CAM[123] and Grad-CAM++[124]. The maps visualize the regions in the input data that most heavily influenced the model prediction at a certain layer.

Model optimization. Typical clinical environments do not have access to specialized hardware (such as DL acceleration cards) and increased memory, which are necessary for practical on-premise deployment of DL models. This situation is further exacerbated in developing regions, where clinical environments are even more limited in resources. The question of training, or even inference/execution, of DL models has not received sufficient attention in current literature, hindering clinical translation of state-of-the-art models. One of the reasons that this clinical applicability is typically not considered during the life-cycle of a research project is because of the added complexity. Thus, to further the use of models trained using GaNDLF in clinical/low-resource settings, GaNDLF incorporates post-training optimization of all models using OpenVINO[125] by default, and provides the optimized model as an additional output at the conclusion of any model training procedure. This enables inference of DL models to be deployed to low-resource machines[126], which magnifies their impact in precision medicine.

Software development practices. GaNDLF incorporates several well-known robust software development measures[127] to ensure ongoing software quality in the presence of community contributions. These measures include the following:

- "Unit testing" refers to tests of individual functional components of the software, to ensure that implementation changes do not break the usage contract established by that component. These units are the smallest relevant units of functionality, and testing these helps ensure that bugfixes, feature additions, and performance optimizations do not cause breaking changes to basic calculations made by the software, such as those that would impact model training. GaNDLF includes extensive unit tests for all custom functionality which is built atop other libraries.
- "System testing" refers to larger-scale tests of software functionality, to test the usage of the software in a broader way that more closely correlates to real usage. GaNDLF's test suite includes extensive system tests, including tests of each major usage mode (training, inference, data preparation, and so on), and tests for each model architecture across types of data (such as radiology and histology images) and types of workloads (such as classification, regression, and segmentation) as appropriate. GaNDLF's test suite requires all tests to pass before code can be committed to the repository, and changes cannot be committed to the code repository if any tests fail for any reason.
- Automated and publicly-declared vulnerability testing of code dependencies via Dependabot[128], which ensures that GaNDLF stays up-to-date on security patches.
- "Automated test coverage reports" are a metric collected during testing, reflecting how much of the codebase is traversed by tests. Higher code coverage indicates that more individual components, functions, and conditional branches of the software have been tested. GaNDLF automatically reports code coverage changes on any incoming contribution and flags changes that decrease code coverage for further review.
- "Continuous deployment" via containerization using the Docker, Singularity, and MLCube standards.

While the above tests cover code-level reliability, it is difficult to infer reliability regarding performance of the models produced by GaNDLF, in part due to stochasticity of the training process. We are actively working on additions to the automated test suite that would measure performance of each model on small sample datasets, and flag contributions that cause drops in performance for further review.

Experimental design. For each application, multiple models are trained in accordance with the cross-validation schema described in Methods Section. For performance evaluation, we use the model with the best validation score as defined in the application-specific evaluation criteria and apply this model to the test dataset for each fold, giving us the average performance of an architecture for the specific problem. To maintain reproducibility and prevent overfitting, we have trained each architecture with a 20/16/64 split, which results in the training of 25 models in total, for each architecture. Specifically, the 20/16/64 split comprises 5 non-overlapping splits (i.e., each containing 20%) of the complete dataset. Each one of these splits is set aside as the testing cohort for each fold. From the remaining 80% of the complete data during this fold, 5 further splits are done, each containing 16% of the full data, and used for validation. Finally, the remaining data for this fold, which represent 64% of the full cohort, are used for training.

Segmentation of brain in MRI. Brain extraction is an essential pre-processing step in the realm of neuroimaging, and has an immediate impact on the quality of all subsequent processing and analyses steps. We have used a multi-institutional dataset of 2520 MRI scans along with their corresponding manually annotated brain masks. We trained on 1320 scans in a modality-agnostic manner (i.e., each structural MRI scan was treated as a separate input) as described in ref. [11] and setting a internal validation set of 180 scans, with an independent testing cohort of 360 scans to ascertain the model performance. We trained by resampling the data from an isotropic resolution of 1 mm^3 with a shape of $240 \times 240 \times 160$ to a anisotropic resolution of $1.825 \times 1.825 \times 1.25$ mm^3 with a shape of $128 \times 128 \times 128$[11]. The reason for this resampling was GPU memory limitations, i.e., 11GB VRAM. We trained multiple architectures (UNet, ResUNet, FCN) with only z-score normalization by discarding the zero-voxels, with no augmentations enabled.

Segmentation of brain tumor sub-regions in MRI. Gliomas are among the most common and aggressive brain malignancies and accurate delineation of these regions can provide valuable clinical insights. We have used the publicly available MRI data from the International Brain Tumor Segmentation (BraTS) challenge of 2020[7–10,129,130] to train multiple models to segment the various brain tumor sub-regions. Specifically, we used the full cohort of 371 training subjects, which we iteratively split it into 74 testing, 60 validation, and 237 training subjects following the k-fold cross-validation schema mentioned in the Cross-Validation sub-section in Methods, with all the 4 structural MRI sequences making up a single input

data-point[11]. In total, 25 models are trained for each architecture (UNet, ResUNet, UInc, and FCN). For each model, we used a set of common hyperparameters that runs in a GPU with 11GB of memory, namely, patch size of $128 \times 128 \times 128$, 30 base filters, "Dice" loss, with stochastic gradient descent (SGD) as the optimizer. For pre-processing, we used z-score normalization by discarding the zero-voxels and cropping of the zero-planes. For data augmentation, we used noise, flipping, affine, rotation and blur, each with a probability of getting picked as 0.35. In each case, the model is trained to maximize the performance evaluation criteria, which is constructed by following the instructions in the BraTS challenge[7,10], i.e., averaging the "Dice" across the enhancing tumor, the tumor core (formed by combining necrosis and enhancing tumor), and the whole tumor (formed by combining the tumor core and the peritumoral edematous/infiltrated tissue).

Whole brain parcellation in MRI. Whole brain structural segmentation could provide richer neuroanatomy information in neuroimaging studies where those structures are relatively small and thus it becomes a more challenging task to accurately segment them in the similar image appearances[45]. We have used the publicly available MRI data from the Multi-Atlas Labelling Challenge (MALC) of 2012 to train multiple models to segment the whole brain into 133 fine-grained sub-regions[131] from T1 weighted scans. Specifically, this challenge dataset contains in total of 30 scans, where a training list of 15 scans and a testing list of 15 scans are provided from the challenge. We trained by resampling the data into an isotropic resolution of 1 mm^3 with a shape of $256 \times 256 \times 256$ as referred from prior work[45]. Particularly, a ResUNet model and an UNet model are implemented for training with a set of common hyperparameters that runs in a GPU with 24 GB of memory, 30 base filters, Dice loss, with SGD as the optimizer. Differently, ResUNet used a learning rate of 0.02 and a patch size of $64 \times 64 \times 64$, whereas UNet used a learning rate of 0.01 and a patch size of $96 \times 96 \times 96$. For pre-processing, the dataset are normalized into range of $[-1, 1]$[45] with no augmentations enabled.

Segmentation of fatty and dense breast tissue using DBT. Breast density has been widely demonstrated to be an independent risk factor for breast cancer[132–134]. Given the rise of digital breast tomosynthesis (DBT) in breast cancer screening compared to traditional 2D mammography[46], there is potential to estimate volumetric breast density (VBD) routinely using machine learning methods. We retrospectively analyzed 1080 negative DBT screening exams completed between 2011 and 2016 at the Hospital of the University of Pennsylvania that contained both 2D raw DBT and 3D reconstructed images. Using the available cranio-caudal and mediolateral-oblique views for each patient, a total of 7850 DBT views were available. We created a convolutional neural network that employed the U-Net architecture for a 3-label image segmentation problem (background, fatty breast tissue, dense breast tissue). Training, validation, and testing sets comprised 70%, 15%, and 15% of the original dataset, respectively. Corresponding ground truth segmentations were generated from a previously validated software that generated VBD metrics based on both 3D reconstructed slices and raw 2D DBT data. 24 models were trained, each using a unique combination of learning rates, batch sizes, patch sizes, and optimizers. Data augmentation during training included affine transformations, blur transformations, and noise transformations, with probabilities of 0.25, 0.5, and 0.5, respectively. The performance of each model was based on weighted and unweighted Dice scores and the final model was selected based on validation set performance.

Segmentation of structural tumor volume from breast MRI. The ACRIN 6657/I-SPY1 TRIAL[47,135] enrolled 237 women from May 2002 to March 2006. From these cases, after applying the inclusion/exclusion criteria, we were left with 163 subjects which contained the 3 time-points of interest with regards to contrast injection. These were pre-injection, and 2 post-injection scans. The first-post contrast image for each case was used by the radiologist to delineate the entire 3-D primary tumor segmentation for each patient, also known as the "structural tumor volume", since it contained peak excitation of the contrast agent[47,135]. We trained the ResUNet using all the 3 time-points using an initial and minimum learning rates of 0.01 and 10^{-4}, respectively, driven using the SGD optimizer. We observed an average "Dice" of 0.74 across 5 fold cross-validation[136].

Segmentation of lung field in CT. An accurate volumetric estimation of the lung field would be crucial towards furthering the clinical goals of tackling respiratory illnesses, such as influenza, pneumonia, and COVID-19 pathologies. However, manual segmentation of the lung field is time-intensive and subjective with low inter-individual reliability, especially for large-scale datasets. Automatic segmentation algorithms can substantially accelerate the analytical procedure. We trained 3D lung field segmentation models with two internal datasets from two independent cohorts based on the ResUNet structure. The first dataset was identified within the lung cancer screening cohort at the University of Pennsylvania Health System (UPHS), and consisted of 500 low-dose CT scans in which 25 were diagnosed with lung cancer. Their corresponding ground truth segmentations for the lung field were generated under a semi-automatic procedure leveraging 2-cluster k-means, followed by manual qualitative refinements. The second dataset contains 673 low-dose CT scans identified within COVID-19 patients admitted to UPHS. Because of the difficulties posed by pathological presentations of COVID-19 in scans, the ground truth was obtained by manually choosing scans with correct segmentations generated by the algorithm that worked on individual slices and accounted for the presence of severe pathologies[137]. We trained our models on the two datasets separately. We split both datasets into training, validation and test sets. For the first dataset, there are 254 scans in the training set, 64 scans in the validation set and 182 scans in the test set. For the second dataset, there are 360 scans in the training set, 98 scans in the validation set and 215 scans in the test set. We performed windowed pre-processing and clipped the intensities between $[-900,-300]$ Hounsfield Units (HU). We also resampled the data down to $[128 \times 128 \times 128]$ in order to consider the entire chest region and to ensure that the trained model remained agnostic to the original image resolution. We trained the ResUNet architecture with clipping and z-score normalization by discarding the zero-voxels with no augmentations enabled. The "Dice" score was employed as our evaluation metric and the model was trained to maximize the "Dice" score.

Segmentation of retinal fundus. We used the dataset from the PALM challenge[48], which consists of segmentation of lesions in retinal fundus images and replicated the results for a ResUNet architecture from[138]. Additionally, we trained on FCN, UNet, and UInc to show results from a diversified set of architectures from the same dataset. We used the full cohort of 400 training subjects, and iteratively split into 80 testing, 64 validation, and 256 training subjects following the k-fold cross-validation schema mentioned in the Cross-Validation sub-section in Methods. In total, 25 models are trained for each architecture (UNet, ResUNet, UInc, and FCN). For each model, we used a set of common hyperparameters options that runs in a GPU with 11GB of memory, namely, patch size of 2048×1024, 30 base filters, "Dice" loss, with SGD as the optimizer. For pre-processing, we used full-image normalization, and data augmentation was performed using flipping, rotation, noise and blur, each with a probability of 0.5. The performance is evaluated in comparison with the ground truth binary masks of the fundus in the testing set.

Segmentation of quadrants in panoramic dental X-ray images. Dental enumeration from panoramic dental X-Ray images has a crucial role in the identification of dental diseases. Performing that task with deep learning provides an extensive advantage for the clinician to number the dentition quickly and point out the teeth that need care more accurately. Quadrant segmentation from those panoramic images is the first and the most critical step of numbering the dentition accurately, and a previous study has used an UNet model to achieve that task[49]. Here, we replicated those results by training a segmentation model with GaNDLF that extracts quadrants from the dental X-ray images. To do that, we have used 900 dental X-ray images with their corresponding five classes (one for each quadrant plus the background). Class annotations have been generated by the experts and the images were resized down to 128×128 in order to consider the entire mouth region. We trained the UNet, the ResUNet, and the FCN architectures with 30 base filters with z-score normalization with no augmentations enabled. We used "Dice" as the evaluation metric and trained the model to maximize it.

Segmentation of colorectal cancer in WSI. Colonoscopy pathology examination can find cells of early-stage colon tumor from small tissue slices, and pathologists need to examine hundreds of tissue slices on a day-to-day basis, which is an extremely time consuming and tedious work. The DigestPath challenge[50] motivated participants to automate this process and thereby contribute to potentially improved diagnostics. The data provided in the DigestPath challenge includes slides containing colorectal cancer in JPEG format. The dimensions of the provided images range from 3000×3000 to 30000×30000. 180000 patches of the shape 512×512 at 10× resolution were extracted for training and 30000 for validation, with a set of 30 WSIs being kept separate as independent testing dataset. We trained the ResUNet architecture, and prior to training we normalized the training values to $[0-1]$ by dividing each pixel by the maximum possible intensity, i.e., 255. To account for model generalizability, we employed the flip, rotate, blur, noise, gamma, and brightness data augmentations. We used "Dice" as our evaluation metric and trained the model to maximize it. Inference was then done on the testing dataset and the output of the model was evaluated against the ground truth binary masks to calculate the "Dice" score.

Brain age prediction from MRI. The human brain ages differently because of various environment factors. Quantifying the difference between actual age and predicted age can provide a useful insight into the overlap of aging signatures with various neurodegenerative pathologies. A 2020 study[51] has used common 2D CNN architectures, borrowed from the computer vision community, to predict brain age from T1-weighted MRI scans across a wide age range. Methodologically, the original fully connected layers of the VGG-Net was replaced by a global average pooling, followed by a new fully connected layer of size 1024 with 80% dropout, and then a single output node with a linear activation was added. The network was then trained with the Adam optimizer[139], while using MSE. This study was evaluated on 10,000 diverse structural brain MRI scans, pooling data from various studies, including the UK Biobank[140] and a multisite schizophrenia consortium[141], thereby representing various subject populations and acquisition protocols. This inherent variability of the collective dataset allowed to successfully learn a regression model generalizable across sites. The study in question study[51] goes on to examine using the learned age

prediction weights as a starting point for transfer learning to other neuroimaging workflows. It is shown that the age prediction weights serve as a superior basis for transfer learning compared to ImageNet, particularly in neuroimaging problems, where the new training data is limited[51].

Leveraging the modular nature of GaNDLF, we replicated the age prediction results of that study[51] using the same model architecture, training schema, and dataset as in the original study, while following GaNDLF's procedures. Using the VGG-16 model architecture and GaNDLF's built-in cross-validation functionality, we trained regression models using the intermediate 80 axial slices of each subject, with input data being split on the subject level. The same network hyperparameters were used, as those specified in the original study[51].

Classification of diabetic foot ulcer images. Diabetic foot ulceration (DFU) is a serious complication of diabetes, which poses a major problem for health systems around the world. DFU can further lead to infection and ischemia, which can result in the amputation of limbs, with more severe cases being terminal illnesses. Diabetic Foot Ulcers Grand Challenge (DFUC) 2021[52] is conducted to help early detection of DFU, which can prevent turning into more serious cases, improve care and reduce the burden on healthcare systems.

DFUC 2021 required its participants to solve a multi-class classification problem through DFU images. The dataset contain DFU images of 4 different classes, labelled as 1) "infection", 2) "ischemia", 3) "both infection and ischemia", and 4) "controls" (i.e., neither infection, nor ischemia). The original resolution of the images are 224×224. The dataset consist of 15,863 images, partitioned into three distinct independent subsets. The training set includes 5955 images, where 2555 cases with only "infection", 227 cases with only "ischemia", 621 cases with "both infection and ischemia", and 2552 "control" cases.

Through the challenge, ease of conducting different experiments through GaNDLF assisted us to train well-known generalizable models. Three different versions of the VGG architecture[142,143] and one version of DenseNet architecture are experimented with GaNDLF, namely VGG11, VGG16, VGG19 and DenseNet121. We utilized *k*-fold cross validation functionality of GaNDLF to prevent overfitting. We applied patching with size of 128×128. We set the batch size as $b = 256$ for VGG11 and VGG16 architectures, $b = 128$ for VGG19, $b = 32$ for DenseNet121 (high GPU resources were not available for this experiment) to ensure maximal utilization of the available hardware resources. Adding bias, blur, noise, and swapping techniques are used as data augmentation with probability $p = 0.5$. Z-scoring normalization is used for data pre-processing. Cross-entropy loss is used as the loss function, which is shown work well for multi-class classification problems. We also experimented weighted cross entropy loss[144], which generally works better for imbalanced classes. The Adam optimizer[139] was used with an initial learning rate of $lr = 0.001$.

Classification of TILs using histology scans. Electronic capture (digitisation) and analyses of whole slide images (WSIs) of tissue specimens are becoming ubiquitous. Digital Pathology interpretation is becoming increasingly common, where many sites are actively scanning archived glass tissue slides with commercially available high-speed scanners to generate high-resolution gigapixel WSIs. Alongside these efforts, a great variety of AI algorithms have been developed to extract many salient tissue and tumor characteristics from WSIs. Examples include segmentation of tumor regions, histologic subtypes of tumors, microanatomic tissue compartments; detection and classification of immune cells to identify tumor-infiltrating lymphocytes (TILs); and the detection and classification of cells and nuclei. TILs are lymphoplasmacytic cells that are spatially located in tumor regions, where their role as an important biomarker for the prediction of clinical outcomes in cancer patients is becoming increasingly recognised[145–147]. Identification of the abundance and the patterns of spatial distribution of TILs in WSI represent a quantitative approach to characterizing important tumor immune interactions. We created a cohort of predefined training and validation cases consisting of patches extracted from WSIs of cancer from 12 anatomically distinct sites, comprising of breast, cervix, colon, lung, pancreas, prostate, rectum, skin, stomach, uterus, and uvea of the eye. All cases are publicly available in The Cancer Genome Atlas (TCGA)[148].

We have used a VGG-16 architecture[142,143] that has been pretrained using the ImageNet dataset[55]. We updated the architecture's first and final layers to be able to process input images of any size, and only output the 2 relevant classes for this problem, respectively[54]. We then proceed with training this architecture using different schedulers and optimizers along with varying learning rates to get an average performance of 0.89. The best results were seen with the step scheduler on Adam optimizer[139] using a learning rate of 0.001.

Prediction of EGFRvIII using structural MRI. Glioblastoma (GBM) is the most common and aggressive primary malignant adult brain tumor and epidermal growth factor receptor variant III (EGFRvIII) mutation has been considered a driver mutation and therapeutic target in GBM[149–151]. Usually, EGFRvIII presence is determined by analysis of surgically resected or biopsy-obtained tissue specimens. We are conducting experiments towards prediction of the EGFRvIII status non-invasively, by analyzing the preoperative and pre-processed structural multiparametric (mp)MRI sequences (T1, T2, T1-Gd and T2-Flair). We identified a cohort of 146 patients containing these four scans acquired at the Hospital of the University of Pennsylvania.

We trained the VGG11 classification architecture utilizing the k-fold cross validation functionality to classify the EGFRvIII status as positive or negative based on the four structural modalities as well as the segmentation map of tumor core. The patch size was set to $128 \times 150 \times 131$ the various experiments were carried out to find the optimal set of hyperparameters utilizing the various options available in GaNDLF. Baseline results were obtained without using any additional data augmentation techniques. Best performance was achieved with cross entropy loss function, SGD optimizer and step scheduler with learning rate of 0.1.

Reporting summary. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used for each of the experiments are available as follows:

Segmentation of Brain in MRI: The data used was a combination of a publicly available dataset[8,11], augmented with scans from private collections of multiple institutions, namely the University of Pennsylvania Health System (UPHS), Thomas Jefferson University, and MD Anderson Cancer Center. The data that support the findings of this study are available from the individual hospitals, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the aforementioned clinical sites.

Segmentation of Brain Tumor Sub-regions in MRI: The data used was from the Brain Tumor Segmentation (BraTS) challenge of 2020[7–10].

Whole Brain Parcellation in MRI: The data used was from the Multi-Atlas Labelling challenge (MALC) of 2012[131].

Segmentation of Breast Tissue using DBT: The data that support the findings of this study are available from the UPHS, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the University of Pennsylvania.

Segmentation of Structural Tumor Volume Breast MRI: The data used in this study was obtained from the ACRIN 6657/I-SPY1 TRIAL[47,135] and can be downloaded from https://wiki.cancerimagingarchive.net/display/Public/ISPY1.

Segmentation of Lung Field in CT: The data that support the findings of this study are available from the UPHS, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the University of Pennsylvania.

Segmentation of Retinal Fundus: The data used was from the PALM challenge[48].

Segmentation of Quadrants in Panoramic Dental X-Ray Images: The data that support the findings of this study are available from the Istanbul Medipol University, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Istanbul Medipol University.

Segmentation of Colorectal Cancer in WSI: The data used was from the DigestPath challenge[50].

Brain Age Prediction from MRI: The data used was from the UK Biobank[140] and a multisite schizophrenia consortium[141].

Prediction of the EGFRvIII mutation in brain tumors using structural mpMRI: The data that support the findings of this study are available from the UPHS, but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the University of Pennsylvania.

Classification of Diabetic Foot Ulcer Images: The data used was from the Diabetic Foot Ulcer Grand Challenge (DFUC) of 2021[52].

Classification of Tumor Infiltrating Lymphocytes: The data used is available in The Cancer Genome Atlas (TCGA)[148].

Code availability

To encourage reproducibility, all the code used for this work is open-sourced at github.com/mlcommons/GaNDLF, and it can be installed as detailed in mlcommons.github.io/GaNDLF/setup.

Received: 2 August 2022; Accepted: 27 March 2023;
Published online: 16 May 2023

References

1. Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Transactions Pattern Analysis Machine Intelligence*. **12**, 993–1001 (1990).

2. Szegedy, C. et al. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9 (2015) https://doi.org/10.1109/CVPR.2015.7298594.
3. Garcia-Garcia, A. et al. A survey on deep learning techniques for image and video semantic segmentation. *Appl. Soft Comput.* **70**, 41–65 (2018).
4. Lateef, F. & Ruichek, Y. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* **338**, 321–348 (2019).
5. Kemker, R., Salvaggio, C. & Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J Photogrammetry Remote Sens.* **145**, 60–77 (2018).
6. Baldi, P., Sadowski, P. & Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nat. Commun.* **5**, 1–9 (2014).
7. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions Medical Imaging.* **34**, 1993–2024 (2014).
8. Bakas, S. et al. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific Data.* **4**, 1–13 (2017).
9. Bakas, S. et al. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. The cancer imaging archive 286 (2017) https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q.
10. Bakas, S. et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018) https://doi.org/10.48550/arXiv.1811.02629.
11. Thakur, S. et al. Brain extraction on mri scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *NeuroImage* **220**, 117081 (2020).
12. Rudie, J. D. et al. Multi-disease segmentation of gliomas and white matter hyperintensities in the brats data using a 3d convolutional neural network. *Front. Comput. Neurosci.* **13**, 84 (2019).
13. Maghsoudi, O. H. et al. O-net: An overall convolutional network for segmentation tasks. In *International Workshop on Machine Learning in Medical Imaging*, 199-209 (Springer, 2020) https://doi.org/10.1007/978-3-030-59861-7_21.
14. Ghesu, F. C. et al. An artificial agent for anatomical landmark detection in medical images. In *International conference on medical image computing and computer-assisted intervention*, 229-237 (Springer, 2016) https://doi.org/10.1007/978-3-319-46726-9_27.
15. Zhang, J., Liu, M. & Shen, D. Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. *IEEE Transactions Image Process.* **26**, 4753–4764 (2017).
16. Borovec, J. et al. Anhir: automatic non-rigid histological image registration challenge. IEEE Transactions on Medical Imaging (2020) https://doi.org/10.1109/TMI.2020.2986331.
17. Li, H. & Fan, Y. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1075-1078 (IEEE, 2018) https://doi.org/10.1109/ISBI.2018.8363757.
18. Akbari, H. et al. Histopathology-validated machine learning radiographic biomarker for noninvasive discrimination between true progression and pseudo-progression in glioblastoma. *Cancer* **126**, 2625–2636 (2020).
19. Pouyanfar, S. et al. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surveys.* **51**, 1–36 (2018).
20. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J. & Bakas, S. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, 92-104 (Springer, 2018) https://doi.org/10.1007/978-3-030-11723-8_9.
21. Sheller, M. J. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports.* **10**, 1–12 (2020).
22. Wolf, I. et al. The medical imaging interaction toolkit (mitk): a toolkit facilitating the creation of interactive software by extending vtk and itk. In *Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display*, vol. 5367, 16-27 (International Society for Optics and Photonics, 2004) https://doi.org/10.1117/12.535112.
23. Davatzikos, C. et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. Med. Imaging.* **5**, 011018 (2018).
24. Kikinis, R., Pieper, S. D. & Vosburgh, K. G. 3d slicer: a platform for subject-specific image analysis, visualization, and clinical support. In *Intraoperative imaging and image-guided therapy*, 277-289 (Springer, 2014) https://doi.org/10.1007/978-1-4614-7657-3_19.
25. Yushkevich, P. A., Gao, Y. & Gerig, G. Itk-snap: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 3342-3345 (IEEE, 2016) https://doi.org/10.1109/EMBC.2016.7591443.
26. Gibson, E. et al. Niftynet: a deep-learning platform for medical imaging. *Computer Methods Programs Biomed.* **158**, 113–122 (2018).
27. Beers, A. et al. Deepneuro: an open-source deep learning toolbox for neuroimaging. Neuroinformatics 1–14 (2020) https://doi.org/10.1007/s12021-020-09477-5.
28. Tustison, N. J. et al. Antsx: A dynamic ecosystem for quantitative biological and medical imaging. *medRxiv* (2020) https://doi.org/10.1101/2020.10.19.20215392.
29. Pawlowski, N. et al. Dltk: State of the art reference implementations for deep learning on medical images. *arXiv preprint arXiv:1711.06853* (2017) https://doi.org/10.48550/arXiv.1711.06853.
30. Jungo, A., Scheidegger, O., Reyes, M. & Balsiger, F. pymia: A python package for data handling and evaluation in deep learning-based medical image analysis. *Computer Methods Programs Biomed.* **198**, 105796 (2021).
31. Oktay, O. et al. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. *JAMA Network Open* **3**, e2027426–e2027426 (2020).
32. Cardoso, M. J. et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701* (2022) https://doi.org/10.48550/arXiv.2211.02701.
33. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Meth.* 1–9 (2020) https://doi.org/10.1038/s41592-020-01008-z.
34. Iyer, A., Locastro, E., Apte, A., Veeraraghavan, H. & Deasy, J. O. Portable framework to deploy deep learning segmentation models for medical images. *bioRxiv* (2021) https://doi.org/10.1101/2021.03.17.435903.
35. Pati, S. et al. Federated learning enables big data for rare cancer boundary detection. *Nat. Commun.* 13 (2022) https://doi.org/10.1038/s41467-022-33407-5.
36. Fu, Y. et al. Deepreg: a deep learning toolkit for medical image registration. *arXiv preprint arXiv:2011.02580* (2020) https://doi.org/10.21105/joss.02705.
37. Rosenthal, J. et al. Building tools for machine learning and artificial intelligence in cancer research: best practices and a case study with the pathml toolkit for computational pathology. *Mol. Cancer Res.* **20**, 202–206 (2022).
38. Pocock, J. et al. Tiatoolbox: An end-to-end toolbox for advanced tissue image analytics. bioRxiv (2021) https://doi.org/10.1101/2021.12.23.474029.
39. Nalisnik, M. et al. Interactive phenotyping of large-scale histology imaging data with histomicsml. *Scientific Reports* **7**, 1–12 (2017).
40. Karargyris, A. et al. Medperf: Open benchmarking platform for medical artificial intelligence using federated evaluation. *arXiv preprint arXiv:2110.01406* (2021) https://doi.org/10.48550/arXiv.2110.01406.
41. Efron, B. & Tibshirani, R. Improvements on cross-validation: the 632+ bootstrap method. *J. American Statistical Association.* **92**, 548–560 (1997).
42. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks.* **106**, 249–259 (2018).
43. Mårtensson, G. et al. The reliability of a deep learning model in clinical out-of-distribution mri data: a multicohort study. *Med. Image Analysis* **66**, 101714 (2020).
44. Han, X. et al. Brain extraction from normal and pathological images: A joint pca/image-reconstruction approach. *NeuroImage* **176**, 431–445 (2018).
45. Li, Y., Li, H. & Fan, Y. Acenet: Anatomical context-encoding network for neuroanatomy segmentation. *Med. Image Analysis.* **70**, 101991 (2021).
46. Niklason, L. T. et al. Digital tomosynthesis in breast imaging. *Radiology* **205**, 399–406 (1997).
47. Newitt, D., Hylton, N. et al. Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials. *Cancer Imaging Arch.* (2016) https://doi.org/10.7937/K9/TCIA.2016.HdHpgJLK.
48. Fu, H. et al. Palm: Pathologic myopia challenge. In *Proc. IEEE Dataport*, 1 (2019) https://doi.org/10.21227/55pk-8z03.
49. Yüksel, A. E. et al. Dental enumeration and multiple treatment detection on panoramic x-rays using deep learning. *Scientific Reports.* **11**, 1–10 (2021).
50. Li, J. et al. Signet ring cell detection with a semi-supervised learning framework. In *International Conference on Information Processing in Medical Imaging*, 842–854 (Springer, 2019) https://doi.org/10.1007/978-3-030-20351-1_66.
51. Bashyam, V. M. et al. Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).
52. Yap, M. H. et al. Analysis towards classification of infection and ischaemia of diabetic foot ulcers. *arXiv preprint arXiv:2104.03068* (2021) https://doi.org/10.1109/BHI50953.2021.9508563.
53. Güley, O., Pati, S. & Bakas, S. Classification of infection and ischemia in diabetic foot ulcers using vgg architectures. In *Diabetic Foot Ulcers Grand Challenge*, 76-89 (Springer, 2021) https://doi.org/10.1007/978-3-030-94907-5_6.
54. Baid, U. et al. Federated learning for the classification of tumor infiltrating lymphocytes. *arXiv preprint arXiv:2203.16622* (2022) https://doi.org/10.48550/arXiv.2203.16622.

55. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248-255 (Ieee, 2009) https://doi.org/10.1109/CVPR.2009.5206848.
56. Gotkowski, K., Gonzalez, C., Bucher, A. & Mukhopadhyay, A. M3d-cam217-222 (2021) https://doi.org/10.1007/978-3-658-33198-6_52.
57. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234-241 (Springer, 2015) https://doi.org/10.1007/978-3-319-24574-4_28.
58. Hatamizadeh, A. et al. Unetr: Transformers for 3d medical image segmentation. In *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 574-584 (2022) https://doi.org/10.48550/arXiv.2103.10504.
59. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016) https://doi.org/10.1109/CVPR.2016.90.
60. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105-6114 (PMLR, 2019) https://doi.org/10.48550/arXiv.1905.11946.
61. Foley, P. et al. Openfl: the open federated learning library. *Phys. Med. Biol.* **67**, 214001 (2022).
62. Pati, S. et al. The federated tumor segmentation (fets) challenge. *arXiv preprint arXiv:2105.05874* (2021) https://doi.org/10.48550/arXiv.2105.05874.
63. Baid, U. et al. The federated tumor segmentation (fets) initiative: The first real-world large-scale data-private collaboration focusing on neuro-oncology. In *NEURO-ONCOLOGY*, vol. 23, 135-135 (OXFORD UNIV PRESS INC JOURNALS DEPT, 2001 EVANS RD, CARY, NC 27513 USA, 2021) https://doi.org/10.1093/neuonc/noab196.532.
64. Waring, J., Lindvall, C. & Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence Med.* **104**, 101822 (2020).
65. Elsken, T., Metzen, J. H. & Hutter, F. et al. Neural architecture search: A survey. *J. Mach. Learn. Res.* **20**, 1–21 (2019).
66. Mongan, J., Moy, L. & Kahn Jr, C. E. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. Radiology. Artificial Intelligence 2 (2020) https://doi.org/10.1148/ryai.2020200029.
67. Sounderajah, V. et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the stard-ai protocol. *BMJ Open* **11**, e047709 (2021).
68. Collins, G. S. et al. Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).
69. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *BMJ* 370 (2020) https://doi.org/10.1038/s41591-020-1034-x.
70. Norgeot, B. et al. Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. *Nat. Med.* **26**, 1320–1324 (2020).
71. Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P. & Shah, N. H. Minimar (minimum information for medical ai reporting): developing reporting standards for artificial intelligence in health care. *J. American Med. Informatics Association.* **27**, 2011–2015 (2020).
72. Lambin, P. et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Reviews Clinical Oncol.* **14**, 749–762 (2017).
73. Zwanenburg, A. et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology* **295**, 328–338 (2020).
74. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of simpleitk. *Front. Neuroinform.* **7**, 45 (2013).
75. McCormick, M. M., Liu, X., Ibanez, L., Jomier, J. & Marion, C. Itk: enabling reproducible research and open science. *Front. Neuroinform.* **8**, 13 (2014).
76. Pati, S. et al. The cancer imaging phenomics toolkit (captk): Technical overview. In *International MICCAI Brainlesion Workshop*, 380-394 (Springer, 2019) https://doi.org/10.1007/978-3-030-46643-5_38.
77. Rathore, S. et al. Brain cancer imaging phenomics toolkit (brain-captk): an interactive platform for quantitative analysis of glioblastoma. In *International MICCAI Brainlesion Workshop*, 133-145 (Springer, 2017) https://doi.org/10.1007/978-3-319-75238-9_12.
78. Rathore, S. et al. Multi-institutional noninvasive in vivo characterization of idh, 1p/19q, and egfrviii in glioma using neuro-cancer imaging phenomics toolkit (neuro-captk). *Neuro-oncology Adv.* **2**, iv22–iv34 (2020).
79. Fathi Kazerooni, A. et al. Cancer imaging phenomics via captk: multi-institutional prediction of progression-free survival and pattern of recurrence in glioblastoma. *JCO Clinical Cancer Inform.* **4**, 234–244 (2020).
80. Pianykh, O. S.Digital imaging and communications in medicine (DICOM): a practical introduction and survival guide (Springer, 2012) https://doi.org/10.2967/jnumed.109.064592.
81. Cox, R. et al. A (sort of) new image data format standard: Nifti-1. In: Proc. 10th Annual Meeting of the Organization for Human Brain Mapping (OHBM 2004), Vol. 25, Budapest, Hungary, June 13–17. Available at: http://nifti.nimh.nih.gov/nifti-1/documentation/hbm_nifti_2004.pdf.
82. Goldberg, I. Open microscopy environment. In *2005 IEEE Computational Systems Bioinformatics Conference Workshops and Poster Abstracts*, 380-380 (IEEE Computer Society, 2005) https://doi.org/10.1109/CSBW.2005.100.
83. Ellingson, B. M. et al. Comparison between intensity normalization techniques for dynamic susceptibility contrast (dsc)-mri estimates of cerebral blood volume (cbv) in human gliomas. *J. Magnetic Resonance Imaging.* **35**, 1472–1477 (2012).
84. Reinhold, J. C., Dewey, B. E., Carass, A. & Prince, J. L. Evaluating the impact of intensity normalization on mr image synthesis. In *Medical Imaging 2019: Image Processing*, vol. 10949, 109493H (International Society for Optics and Photonics, 2019) https://doi.org/10.1117/12.2513089.
85. Nyul, L., Udupa, J. & Zhang, X. New variants of a method of mri scale standardization. *IEEE Transactions Med. Imaging.* **19**, 143–150 (2000).
86. Stark, J. A. Adaptive image contrast enhancement using generalizations of histogram equalization. *IEEE Transactions Image Process.* **9**, 889–896 (2000).
87. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions Med. Imaging.* **35**, 1962–1971 (2016).
88. Ruifrok, A. C., Katz, R. L. & Johnston, D. A. Comparison of quantification of histochemical staining by hue-saturation-intensity (hsi) transformation and color-deconvolution. *Appl. Immunohistochem. Mol. Morphology.* **11**, 85–91 (2003).
89. Macenko, M. et al. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, 1107-1110 (IEEE, 2009) https://doi.org/10.1109/ISBI.2009.5193250.
90. Chartrand, G. et al. Deep learning: a primer for radiologists. *Radiographics* **37**, 2113–2131 (2017).
91. Marcus, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631* (2018) https://doi.org/10.48550/arXiv.1801.00631.
92. Annas, G. J. et al. Hipaa regulations-a new era of medical-record privacy? *New England J. Med.* **348**, 1486–1490 (2003).
93. Voigt, P. & Von dem Bussche, A. The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing (2017) https://doi.org/10.1007/978-3-319-57959-7.
94. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data.* **6**, 60 (2019).
95. Perez, L. & Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621* (2017) https://doi.org/10.48550/arXiv.1712.04621.
96. Pérez-García, F., Sparks, R. & Ourselin, S. Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Comp. Meth. Prog. Biomed.* **208**, 106236 (2021).
97. Buslaev, A. et al. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
98. Allen, D. M. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **16**, 125–127 (1974).
99. Molinaro, A. M., Simon, R. & Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307 (2005).
100. Cawley, G. C. & Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Machine Learning Res.* **11**, 2079–2107 (2010).
101. Di Sipio, C., Di Ruscio, D. & Nguyen, P. T. Democratizing the development of recommender systems by means of low-code platforms. In *Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, 1–9 (2020) https://doi.org/10.1145/3417990.3420202.
102. ElBatanony, A. & Succi, G. Towards the no-code era: a vision and plan for the future of software development. In *Proceedings of the 1st ACM SIGPLAN International Workshop on Beyond Code: No Code*, 29-35 (2021) https://doi.org/10.1145/3486949.3486965.
103. Hastie, T., Tibshirani, R., Friedman, J. H. & Friedman, J. H.The elements of statistical learning: data mining, inference, and prediction, vol. 2 (Springer, 2009) https://doi.org/10.1007/978-0-387-21606-5.
104. Micikevicius, P. et al. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017) https://doi.org/10.48550/arXiv.1710.03740.
105. Pati, S. & Bakas, S. LabelFusion: Medical Image label fusion of segmentations (2021). https://doi.org/10.5281/zenodo.4633206.
106. Rahman, M. M. & Davis, D. N. Addressing the class imbalance problem in medical datasets. *Int. J. Machine Learning Comput.* **3**, 224 (2013).
107. Chen, P.-H. C., Liu, Y. & Peng, L. How to develop machine learning models for healthcare. *Nat. Mater.* **18**, 410–414 (2019).
108. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Jorge Cardoso, M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and*

multimodal learning for clinical decision support, 240–248 (Springer, 2017) https://doi.org/10.1007/978-3-319-67558-9_28.
109. Reina, G. A., Panchumarthy, R., Thakur, S. P., Bastidas, A. & Bakas, S. Systematic evaluation of image tiling adverse effects on deep learning semantic segmentation. *Front. Neurosci.* **14**, 65 (2020).
110. Niethammer, M., Borland, D., Marron, J., Woosley, J. & Thomas, N. E. Appearance normalization of histology slides. In *International Workshop on Machine Learning in Medical Imaging*, 58–66 (Springer, 2010) https://doi.org/10.1007/978-3-642-15948-0_8.
111. Vahadane, A. & Sethi, A. Towards generalized nuclear segmentation in histological images. In *13th IEEE International Conference on BioInformatics and BioEngineering*, 1–4 (IEEE, 2013) https://doi.org/10.1109/BIBE.2013.6701556.
112. Poehlmann, A. & Villalba, S. TiffSlide - A drop-in replacement for openslide-python (2022). https://doi.org/10.5281/zenodo.6327079.
113. Reinke, A. et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642* (2021) https://doi.org/10.48550/arXiv.2104.05642.
114. Zijdenbos, A. P., Dawant, B. M., Margolin, R. A. & Palmer, A. C. Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE Transactions on Med. Imaging.* **13**, 716–724 (1994).
115. Rockafellar, R. T. & Wets, R. J.-B. Variational analysis, vol. 317 (Springer Science & Business Media, 2005) https://doi.org/10.1007/978-3-642-02431-3.
116. Berger, J. O. Statistical decision theory and Bayesian analysis (Springer Science & Business Media, 2013) https://doi.org/10.1007/978-1-4757-4286-2.
117. Detlefsen, N. S. et al. Torchmetrics-measuring reproducibility in pytorch. *J. Open Source Software.* **7**, 4101 (2022).
118. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, 3121–3124 (IEEE, 2010) https://doi.org/10.1109/ICPR.2010.764.
119. Cybenko, G., O'Leary, D. P. & Rissanen, J. The Mathematics of Information Coding, Extraction and Distribution, vol. 107 (Springer Science & Business Media, 1998) https://doi.org/10.1007/978-1-4612-1524-0.
120. Holzinger, A. From machine learning to explainable ai. In *2018 world symposium on digital intelligence for systems and machines (DISA)*, 55-66 (IEEE, 2018) https://doi.org/10.1109/DISA.2018.8490530.
121. Gastounioti, A. & Kontos, D. Is it time to get rid of black boxes and cultivate trust in ai? *Radiology: Artificial Intelligence.* **2**, e200088 (2020).
122. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014) https://doi.org/10.48550/arXiv.1412.6806.
123. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE international conference on computer vision*, 618–626 (2017) https://doi.org/10.48550/arXiv.1610.02391.
124. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839-847 (IEEE, 2018) https://doi.org/10.1109/WACV.2018.00097.
125. Gorbachev, Y. et al. Openvino deep learning workbench: Comprehensive analysis and tuning of neural networks inference (2019) https://doi.org/10.1109/ICCVW.2019.00104.
126. Thakur, S. P. et al. Optimization of deep learning based brain extraction in mri for low resource environments. In *International MICCAI Brainlesion Workshop*, 151-167 (Springer, 2022) https://doi.org/10.1007/978-3-031-08999-2_12.
127. Juristo, N., Moreno, A. M. & Strigel, W. Guest editors' introduction: Software testing practices in industry. *IEEE Software.* **23**, 19–21 (2006).
128. Alfadel, M., Costa, D. E., Shihab, E. & Mkhallalati, M. On the use of dependabot security pull requests. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*, 254–265 (IEEE, 2021) https://doi.org/10.1109/MSR52588.2021.00037.
129. Bakas, S. et al. Glistrboost: combining multimodal mri segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In *BrainLes 2015*, 144–155 (Springer, 2015) https://doi.org/10.1007/978-3-319-30858-6_13.
130. Zeng, K. et al. Segmentation of gliomas in pre-operative and post-operative multimodal magnetic resonance imaging volumes based on a hybrid generative-discriminative framework. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 184–194 (Springer, 2016) https://doi.org/10.1007/978-3-319-55524-9_18.
131. Landman, B. A. & Warfield, S. K. MICCAI 2012: Workshop on Multi-atlas Labeling (CreateSpace Independent Publishing Platform, 2012).
132. McCormack, V. A. & dos Santos Silva, I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiol. Prevention Biomarkers.* **15**, 1159–1169 (2006).
133. Boyd, N. F. et al. Breast tissue composition and susceptibility to breast cancer. *J. National Cancer Institute.* **102**, 1224–1237 (2010).
134. Brentnall, A. R., Cuzick, J., Buist, D. S. & Bowles, E. J. A. Long-term accuracy of breast cancer risk assessment combining classic risk factors and breast density. *JAMA Oncol.* **4**, e180174–e180174 (2018).
135. Hylton, N. M. et al. Neoadjuvant chemotherapy for breast cancer: functional tumor volume by mr imaging predicts recurrence-free survival-results from the acrin 6657/calgb 150007 i-spy 1 trial. *Radiology* **279**, 44–55 (2016).
136. Chitalia, R. et al. Expert tumor annotations and radiomics for locally advanced breast cancer in dce-mri for acrin 6657/i-spy1. *Scientific Data.* **9**, 440 (2022).
137. Hofmanninger, J. et al. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Exp.* **4**, 1–13 (2020).
138. Baid, U., Baheti, B., Dutande, P. & Talbar, S. Detection of pathological myopia and optic disc segmentation with deep convolutional neural networks. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, 1345-1350 (IEEE, 2019) https://doi.org/10.1109/TENCON.2019.8929252.
139. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014) https://doi.org/10.48550/arXiv.1412.6980.
140. Sudlow, C. et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos Med.* **12**, e1001779 (2015).
141. Rozycki, M. et al. Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals. *Schizophrenia Bulletin.* **44**, 1035–1044 (2018).
142. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014) https://doi.org/10.48550/arXiv.1409.1556.
143. Ben-Cohen, A., Diamant, I., Klang, E., Amitai, M. & Greenspan, H. Fully convolutional network for liver segmentation and lesions detection. In *Deep learning and data labeling for medical applications*, 77–85 (Springer, 2016) https://doi.org/10.1007/978-3-319-46976-8_9.
144. Fernando, K. R. M. & Tsokos, C. P. Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems (2021) https://doi.org/10.1109/TNNLS.2020.3047335.
145. Mlecnik, B. et al. Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction. *J. Clinical Oncol.* **29**, 610–618 (2011).
146. Badalamenti, G. et al. Role of tumor-infiltrating lymphocytes in patients with solid tumors: Can a drop dig a stone? *Cellular Immunol.* **343**, 103753 (2019).
147. Idos, G. E. et al. The prognostic implications of tumor infiltrating lymphocytes in colorectal cancer: a systematic review and meta-analysis. *Scientific Reports.* **10**, 1–14 (2020).
148. Abousamra, S. et al. Learning from thresholds: fully automated classification of tumor infiltrating lymphocytes for multiple cancer types. *arXiv preprint arXiv:1907.03960* (2019) https://doi.org/10.48550/arXiv.1907.03960.
149. Akbari, H. et al. In vivo evaluation of egfrviii mutation in primary glioblastoma patients via complex multiparametric mri signature. *Neuro-oncology* **20**, 1068–1079 (2018).
150. Bakas, S. et al. In vivo detection of egfrviii in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral infiltration: the φ-index. *Clinical Cancer Res.***23**, 4724–4734 (2017).
151. Binder, Z. A. et al. Epidermal growth factor receptor extracellular domain mutations in glioblastoma present opportunities for clinical imaging and therapeutic development. *Cancer Cell.* **34**, 163–177 (2018).

Acknowledgements

GaNDLF is now primarily maintained and supported by MLCommons (mlcommons.org). Research reported in this publication was partly supported by the National Cancer Institute (NCI), the National Institute of Neurological Disorders and Stroke (NINDS), the National Institute on Aging (NIA), the National Institute of Mental Health (NIMH), and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) of the National Institutes of Health (NIH), under award numbers NCI:U01CA242871, NCI:U24CA189523, NINDS:R01NS042645, NIA:RF1AG054409, NIA:U01AG068057, NIMH:R01MH112070, NCI:R01CA161749 and NIBIB:R01EB022573. S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819\8\25). B.M. acknowledges support by the Helmut-Horten-Foundation. A.K. receives funding from IHU Strasbourg under award number ANR-10-IAHU-02. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH or any other funding body.

Author contributions

Idea Conception: S.P., S.P.T., M.S., A.K., R.U., P.M., S.B. Development of software: S.P, S.P.T, I.E.H, U.B., B.B., Me.B., O.G., S.M., S.T., K.G., Cam.G., Cal.G., A.G., B.E., M.S., J.W., D.K., R.P. Data Acquisition, Processing, and Analysis: S.P, D.L., V.A., C.Z., V.B., Y.L., B.H., R.C., S.A., T.M.K., J.H.S., Y.F., A.G., S.E., S.B. Review and Edit of manuscript:

A.G., M.B., J.H.S., Y.F., P.S., A.M., S.A.T., B.M., C.D., D.K., A.K., R.U., P.M., S.B. Writing the Original Manuscript: S.P., S.P.T., U.B., B.B., S.B. Review, Edit, & Approval of the Final Manuscript: All authors.

Competing interests
The authors declare no Competing Interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-023-00066-3.

Correspondence and requests for materials should be addressed to Spyridon Bakas.

Peer review information *Communications Engineering* thanks Shuoong Wangand the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Mengying Su and Miranda Vinay. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reprint Permissions

GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows

Author: Sarthak Pati et al
Publication: Communications Engineering
Publisher: Springer Nature
Date: May 16, 2023

6

Leveraging 2D Deep Learning ImageNet-trained Models for Native 3D Medical Image Analysis

Authors

Bhakti Baheti*, **Sarthak Pati***, Bjoern Menze, Spyridon Bakas

* Joint first authors

Publication Information

International MICCAI Brainlesion Workshop, pp. 68-79. Cham: Springer Nature Switzerland, 2022. DOI: $10.1007/978-3-031-33842-7\_6$

Abstract

Convolutional neural networks (CNNs) have shown promising performance in various 2D computer vision tasks due to availability of large amounts of 2D training data. Contrarily, medical imaging deals with 3D data and usually lacks the equivalent extent and diversity of data, for developing AI models. Transfer learning provides the means to use models trained for one application as a starting point to another application. In this work, we leverage 2D pre-trained models as a starting point in 3D medical applications by exploring the concept of Axial-Coronal-Sagittal (ACS) convolutions. We have incorporated ACS as an alternative of native 3D convolutions in the Generally Nuanced Deep Learning Framework (GaNDLF), providing various well-established and state-of-the-art network architectures with the availability of pre-trained encoders from 2D data. Results of our experimental evaluation on 3D MRI data of brain tumor patients for i) tumor segmentation and ii) radiogenomic classification, show model size reduction by $\sim$22% and improvement in validation accuracy by $\sim$33%. Our findings support the advantage of ACS convolutions in pre-trained 2D CNNs over 3D CNN without pre-training, for 3D segmentation and classification tasks, democratizing existing models trained in datasets of unprecedented size and showing promise in the field of healthcare.

Contributions of S.P.

Study conceptualization, algorithm development and implementation, interpretation of results, and writing & editing of manuscript.

Copyright

Leveraging 2D Deep Learning ImageNet-trained Models for Native 3D Medical Image Analysis

Bhakti Baheti[1,2,3], Sarthak Pati[1,2,3,4], Bjoern Menze[4,5], and Spyridon Bakas[1,2,3(✉)]

[1] Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA
sbakas@upenn.edu
[2] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[3] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4] Department of Informatics, Technical University of Munich, Munich, Germany
[5] Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland

Abstract. Convolutional neural networks (CNNs) have shown promising performance in various 2D computer vision tasks due to availability of large amounts of 2D training data. Contrarily, medical imaging deals with 3D data and usually lacks the equivalent extent and diversity of data, for developing AI models. Transfer learning provides the means to use models trained for one application as a starting point to another application. In this work, we leverage 2D pre-trained models as a starting point in 3D medical applications by exploring the concept of Axial-Coronal-Sagittal (ACS) convolutions. We have incorporated ACS as an alternative of native 3D convolutions in the Generally Nuanced Deep Learning Framework (GaNDLF), providing various well-established and state-of-the-art network architectures with the availability of pre-trained encoders from 2D data. Results of our experimental evaluation on 3D MRI data of brain tumor patients for i) tumor segmentation and ii) radiogenomic classification, show model size reduction by ∼22% and improvement in validation accuracy by ∼33%. Our findings support the advantage of ACS convolutions in pre-trained 2D CNNs over 3D CNN without pre-training, for 3D segmentation and classification tasks, democratizing existing models trained in datasets of unprecedented size and showing promise in the field of healthcare.

Keywords: Deep learning · ImageNet · Transfer learning · MRI · segmentation · classification

B. Baheti and S. Pati contributed equally for this work.

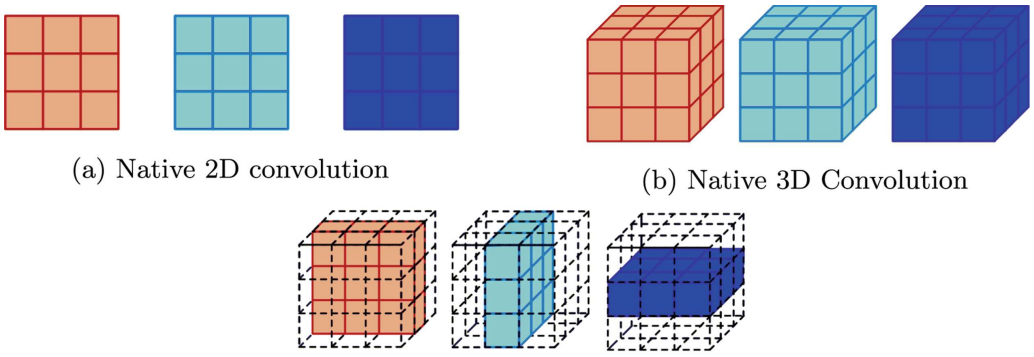
S. Bakas et al. (Eds.): BrainLes 2022, LNCS 13769, pp. 68–79, 2023.
https://doi.org/10.1007/978-3-031-33842-7_6

1 Introduction

Deep learning (DL) based approaches are continuously being developed for various medical imaging tasks, including segmentation, classification, and detection, for a wide range of modalities (i.e., MRI, CT, X-Ray), regularly outperforming earlier approaches [1,2]. However, DL is computationally expensive and requires large amounts of annotated data for model training limiting their applicability in problems where large amounts of annotated datasets are unavailable [3]. Transfer learning (TL) is a popular approach to overcome this issue by initializing a DL model with pre-trained weights, thereby reducing convergence time and concluding at a superior state, while utilizing otherwise insufficient data [4,5]. The basic idea of TL involves re-using model weights trained for a problem with a large available dataset as the initialization point for a completely different task. The foundation behind this idea is that convolutional layers extract general, lower-level features (such as edges, patterns, and gradients) that are applicable across a wide variety of images [6]. The latter layers of a convolutional neural network (CNN) learn features more specific to the image of the particular task by combining the previous lower-level features. Leveraging weights of trained models has proven to be a better initialisation point for DL model training, when compared to random initialization [4,7–10].

There are numerous pre-trained models available for applications on 2D imaging data, such as ImageNet [11], YOLO [12], and MS-COCO [13], however, universally applicable pre-trained models are not available for utilization on 3D data like medical images due to the lack of associated large and diverse data. Current application of pre-trained CNN for 3D medical image segmentation and classification can be divided in three categories depending on the dimensionality of the input data:

- **2D Approaches**
  Here, a 3D input volume is considered as a stack of 2D slices, and a multi-slice planar (2D) network is applied on each 2D slice independently [14,15]. Some earlier approaches considered 3D medical images as tri-planar representation where axial, coronal, and sagittal views are considered as 3 channels of the input data. But such 2D representation learning is fundamentally weak in capturing 3D contexts. Some DL based approaches for classification of brain cancer MRI images use representative 2D slices as the input data rather than utilizing full 3D volume [16,17].
- **3D Approaches**
  In this case, a 3D network is trained using native 3D convolution layers that are useful in capturing spatial correlations present along the 3^{rd} dimension, in order to capture 3D contextual information [18–21]. Significant improvement in classification accuracy was observed in [22] with the use of native 3D convolutions compared to 2D convolutions. Although data in adjacent slices, across each of the three axes, are correlated and can be potentially used to yield a better model, this suffers from two weaknesses: a) reduced model stability due to random weight initialization (since there are no available pre-trained models) and b) unnecessarily high memory consumption.



(a) Native 2D convolution (b) Native 3D Convolution

(c) ACS convolution

Fig. 1. Comparison of various types of convolution for 3D medical data (Figure adapted from [28]).

– **Hybrid Approaches**
There are few studies using a hybrid of the two aforementioned approaches, i.e. 2D & 3D. An ensemble-based learning framework built upon a group of 2D and 3D base learners was designed in [23]. Another strategy is to train multiple 2D networks on different viewpoints and then generate final segmentation results by 3D volumetric fusion net [24]. A similar approach was proposed in [25], which consists of a 2D DenseUNet for intra-slice feature extraction and its 3D counterpart for aggregating volumetric contexts. Finally, Ni *et al.* trained a 2D deep network for 3D medical image segmentation by introducing the concept of elastic boundary projection [26].

Current literature shows inadequate exploration on the application of 2D pre-trained models in native 3D applications. As medical datasets are limited when compared with those from the computer vision domain, TL of models trained in the latter can be beneficial in medical applications.

In this paper, we explore the concept of Axial-Coronal-Sagittal (ACS) convolution to utilize pre-trained weights of models trained on 2D datasets to perform natively 3D operations. This is achieved by splitting the 2D kernels into 3 parts by channels and convolving separately across Axial-Coronal-Sagittal views to enable development of native 3D CNNs for both classification and segmentation workloads. This way, we can take advantage of the 3D spatial context, as well as the available pre-trained 2D models to pave the way towards building better models for medical imaging applications. Multiple options of pre-trained models for use in 3D datasets have been made publicly available through the Generally Nuanced Deep Learning Framework (GaNDLF) [27][1].

2 Methods

In this work, we leverage the concept of Axial-Coronal-Sagittal (ACS) proposed in [28] and incorporate it into the Generally Nuanced Deep Learning Framework

[1] https://github.com/CBICA/GaNDLF.

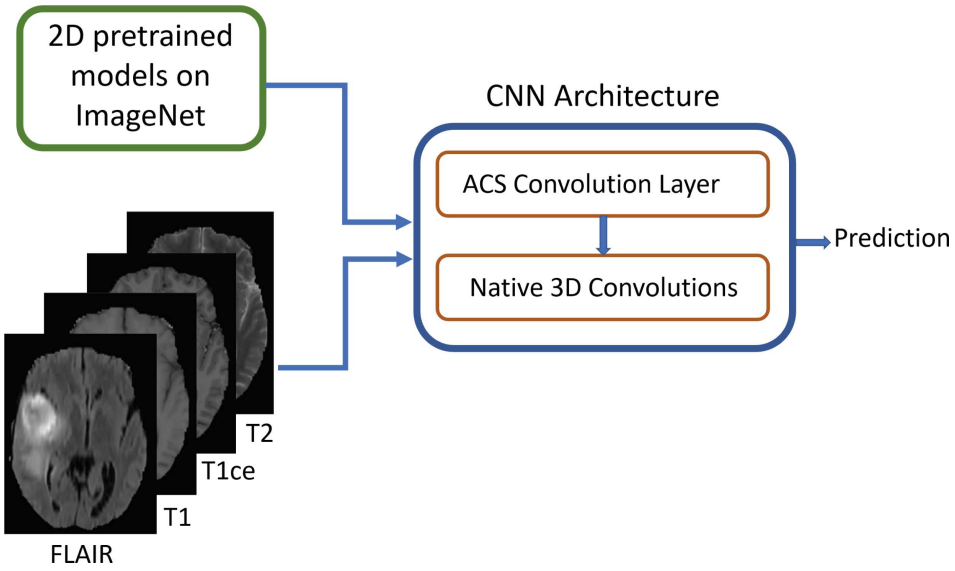


Fig. 2. Idea of integrating ACS convolutions with pre-trained 2D model weights to enable native 3D convolutions on 3D medical data (e.g., MRI)

(GaNDLF) [27][2] which supports a wide variety of model architectures, loss functions, pre-processing, and training strategies.

2.1 ACS Convolutions

Convolution operations in CNNs can be classified as either 2D or 3D. The 2D convolutional layers use 2D filter kernels ($K \times K$) and capture 2D spatial correlation, whereas 3D convolutional kernels ($K \times K \times K$) are used in native 3D convolutional layers capturing 3D context (Fig. 1). As mentioned earlier, each of these approaches have their own advantages and disadvantages.

Yang *et al.* [28] introduced the concept of Axial-Coronal-Sagittal (ACS) convolutions to learn the spatial representation of three dimensions from the combination of each of the three (A-C-S) views (Fig. 1(c)). The basic concept of the ACS convolutions is to split the kernel into three parts $(K \times K \times 1), (K \times 1 \times K)$ and $(1 \times K \times K)$ and run multiple 2D convolution filters across the three views (axial, coronal, and sagittal). For any convolution layer, let us consider the number of input channels as C_i and number of output channels as C_o. The number of output channels in ACS convolution are then set as:

$$C_o^{Axial} \approx C_o^{Coronal} \approx C_o^{Sagittal} \approx \lfloor \frac{C_o}{3} \rfloor \tag{1}$$

Thus 2D convolutions are transformed into 3 dimensions by simultaneously performing computations across axial, coronal, and sagittal axes. The final output

[2] https://github.com/CBICA/GaNDLF.

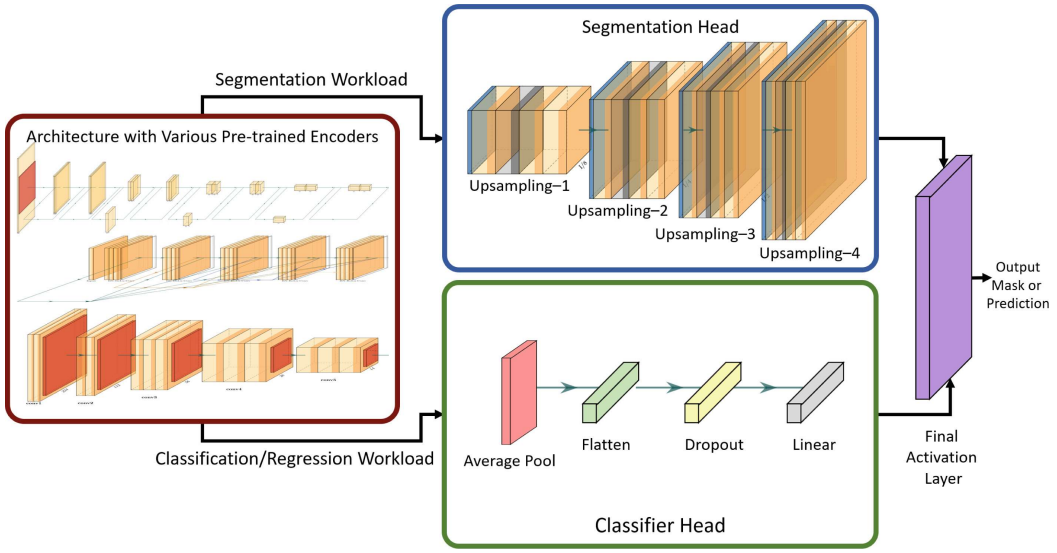


Fig. 3. The architecture that allows using different pre-trained encoders with either a segmentation or classifier head for specific workloads.

is then obtained by the concatenation of three convolved feature maps without any additional fusion layer.

The concept of ACS convolutions can be used as a generic plug-and-play replacement of 3D convolution enabling development of native 3D CNNs using 2D pre-trained weights as illustrated in Fig. 2.

2.2 Architecture Design

We have incorporated the concept of ACS convolutions in GaNDLF, which is a framework for training models for segmentation, classification, and regression in a reproducible and deployable manner [27]. GaNDLF has several architectures for segmentation and classification, as well as a wide range of data pre-processing and augmentation options along with the choice of several training hyper parameters and loss functions. We integrated several encoders from [29], pre-trained on ImageNet [11] into this framework, including variants of VGG [30], ResNet [31], DenseNet [32], and EfficientNet [33]. We have created a mechanism to combine the outputs of these encoders with either a segmentation or a classification head depending on the task, as shown in Fig. 3. The segmentation head consists of a set of upsampling layers similar to the decoder mechanism of the UNet network topology/architecture [34], where the user has the flexibility to choose the number of upsampling layers and the number of feature maps in each layer. The classification head consists of a average pooling layer applied on the top of feature maps obtained from the encoder. Dropout can be set between range 0 to 1 to reduce overfitting to the training data before the final classification layer.

While the 2D pre-trained weights could be directly loaded for applications on 2D data, we have replaced the usual convolution layer with an ACS convolution

layer in GaNDLF, enabling their use for training on 3D medical data in a native manner, regardless of the number of input modalities. In comparison with the 2D models, ACS convolution layers do not introduce any additional computation cost, memory footprint, or model size.

2.2.1 Design for Segmentation Gliomas are among the most common and aggressive brain tumors and accurate delineation of the tumor sub-regions is important in clinical diagnosis. We trained two different architectures for segmentation through GaNDLF. UNet [34] with residual connections (ResUNet) is one of the famous architectures for 2D and 3D medical segmentation. It consists of encoder and decoder modules and feature concatenation pathways. The encoder is a stack of convolutional and downsampling layers for feature extraction from the input images, and the decoder consists of a set of upsampling layers (applying transpose convolutions) to generate the fine-grained segmentation output.

We trained two different models using the publicly available multi-parametric magnetic resonance imaging (mpMRI) data of 369 cases from training set of the International Brain Tumor Segmentation [35–37] (BraTS2020) challenge. This dataset consists of four multi-parametric magnetic resonance imaging (mpMRI) scans per subject/case, with the exact modalities being: a) native (T1) and b) post-contrast T1-weighted (T1-Gd), c) T2-weighted (T2), and d) T2 fluid attenuated inversion recovery (T2-FLAIR). These models are evaluated on 125 unseen cases from the BraTS2020 validation dataset. We first trained a standard ResUNet architecture of depth = 4 and base filters = 32 such that weights of all the layers were randomly initialized. We then built another architecture by using pre-trained ResNet50 as an encoder with depth = 4 and the standard UNet decoder. For each of these experiments, 40 patches of $64 \times 64 \times 64$ were extracted from each subject. Various training parameters were also kept constant, like the choice of optimizer (we used SGD), scheduler (modified triangular) with learning rate of 0.001, and loss function based on the Dice similarity coefficient (DSC) [38]. Maximum number of epochs was set to 250 with patience of 30 for early stopping. The performance is evaluated on clinically-relevant tumor regions, i.e., whole tumor (considered for radiotherapy), tumor core (considered for surgical resection) as well as enhancing tumor.

2.2.2 Design for Classification Glioblastoma (GBM) is the most aggressive and common adult primary malignant brain tumor and epidermal growth factor receptor variant III (EGFRvIII) mutation is considered a driver mutation and therapeutic target in GBM [39–41]. Usually, the presence of EGFRvIII is determined by the analysis of actual tissue specimens and is stated as positive or negative. We focus on non-invasive prediction of EGFRvIII status by analysis of these pre-operative and pre-processed MRI data. Residual Networks (ResNets) [31] introduced the idea of skip connections which enabled design of much deeper CNNs. GaNDLF supports variants of ResNet, including ResNet18,

ResNet34, ResNet50, ResNet101, and ResNet152, each having different number of layers.

We use an internal private cohort of 146 patients containing four structural mpMRI modalities (T1, T2, T1-Gd and T2-FLAIR) such that the positive and negative classes were equally distributed. These 146 cases were distributed in Training (80%) and Validation (20%) sets for experimentation. We used cross entropy loss function, adam optimiser and cosine annealing scheduler with learning rate of 0.0001. As the dataset is smaller, we set the maximum number epochs to 100 and patience of 30 epochs for early stopping.

3 Results

In this section we present the quantitative results of the segmentation and classification workloads described above, to showcase the feasibility and performance of ACS convolutions on 3D medical imaging data. Specifically, we compare the 2D pre-training approach with the random initialization to evaluate the superiority of the ACS convolutions over usual 3D convolution operations.

3.1 Brain Tumor Segmentation Workload

The segmentation model is trained on the publicly available training data of BraTS2020 challenge. We then quantitatively evaluate the performance of the final models on the unseen BraTS2020 validation data by submitting results to the online evaluation platform (CBICA Image Processing Portal). Table 1 lists the number of parameters of each model, as well as the comparative performance, in terms of Dice Similarity Coefficient (DSC) and the 95^{th} percentile of the Hausdorff distance between the predicted ground truth labels.

3.2 Binary Classification of Brain Tumor Molecular Status

For the performance evaluation of the classification workload, we have used the structural mpMRI scans in-tandem as input (i.e., passing all the scans together at once as separate channels) similar to the segmentation workload. The classification model performance on training and validation data is summarized in Table 2, illustrating the effectiveness of pre-trained weights.

4 Discussion

In this work, we have assessed the functionality of transfer learning for 3D medical data based on the available 2D models pre-trained on ImageNet for segmentation and classification. The framework that this functionality is evaluated is designed such that deep learning network architecture's first and last layers are flexible to be able to process input images of any size with varying number of channels or modalities, and provide the final prediction based on the relevant

Table 1. Results on Brain Tumor Segmentation (BraTS2020) validation dataset

Metric	Region	Standard ResUNet	ResNet50+UNet (Random init.)	ResNet50+UNet (Pre-trained)
DSC	Whole Tumor	0.8771	0.8775	0.8736
	Tumor Core	0.7735	0.7458	0.7719
	Enhancing Tumor	0.7138	0.69508	0.7017
Hausdorff95	Whole Tumor	13.2425	7.6747	9.5384
	Tumor Core	14.7492	8.6579	15.4840
	Enhancing Tumor	34.8858	41.00332	40.2053
#Parameters	-	33.377 Million	25.821 Million	25.821 Million
#Epochs for convergence	-	104 epochs	250 epochs	95 epochs

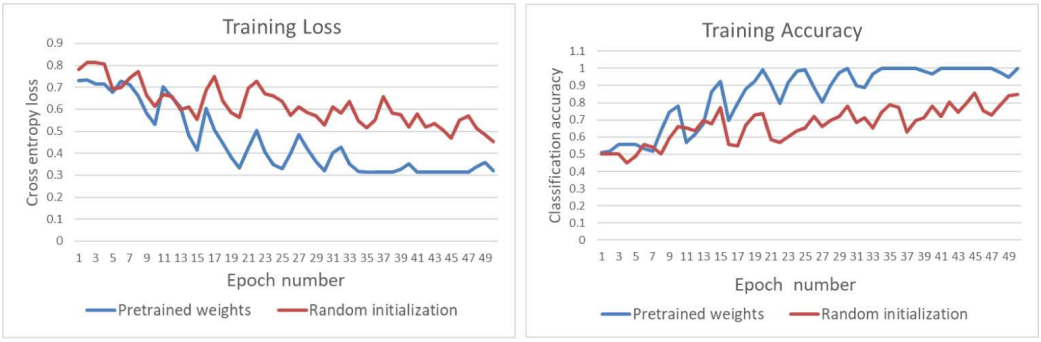
number of classes for the specified task. The rest of the layers are initialized with pre-trained weights from the ImageNet models and are further fine-tuned.

The results of brain tumor segmentation using i) 3D U-Net with residual connections, ii) randomly initialized ResNet50 encoder & UNet decoder, and iii) pre-trained ResNet50 encoder & UNet decoder are shown in Table 1. In these architectures, the obvious difference was in the encoders being randomly initialised in the former i) 3DUNet and ii) ResNet50 and pre-trained in the latter ResNet50-UNet (iii). As the pre-trained decoders are not available from ImageNet, the decoder was initialised with random weights in all the three architectures. We hypothesize that this might be the reason for comparable segmentation performance in terms of dice and hausdorff95 scores, while the difference in number of parameters is significant. It should be observed that the ResNet50-UNet (ii and iii) has only 25.821 Million parameters, which is around 22% less compared to 33.377 Million parameters of the standard ResUNet (i), with the same encoder-decoder depth. The randomly initialised ResNet50-UNet model oscillates around the same performance and did not converge in the specified maximum number of epochs (250). On the other hand, the same model initialised with pre-trained weights converged within 95 epochs. Thus models initialized with pre-trained weights have advantage of better convergence speed as well as smaller model size. Importantly, smaller models are more preferable in the clinical setting due to their higher feasibility for deployment in low-resource environments.

Baseline results of binary classification for the determination of the EGFRvIII mutational status are reported in Table 2, with ResNet50 architecture. We did not use any additional data augmentation techniques. As the data in this task were limited, the effect of pre-trained weights are clearly observed resulting in better accuracy. Figure 4 shows the plots of cross entropy loss and accuracy in training with respect to epochs. Similar performance is observed for validation set as well. The weights of the model with lowest validation loss are stored for reproducibility and the accuracy and loss values reported in Table 2 are for the saved model with lowest validation loss.

Table 2. Results on EGFR Classification

	ResNet50 (Random initialization)	ResNet50 (Pre-trained on ImageNet)
Training Accuracy	0.7203	0.9915
Training Loss	0.5736	0.3292
Val Accuracy	0.5357	0.7142
Val Loss	0.6912	0.5758



(a) (b)

Fig. 4. Comparison plots of training loss and accuracy for binary classification of EGFRvIII mutation status. These plots are for ResNet50 architecture with and without use of 2D pre-trained weights from ImageNet

Our findings support the incorporation of 2D pre-trained models towards improving the performance on 3D medical image segmentation and classification workloads, with demonstrably smaller model size (Table 1). Large increase in accuracy is specially observed in those applications where sufficient labelled data are not available. Incorporating this functionality in GaNDLF provides a readily available solution to researchers towards an end-to-end solution for several computational tasks, along with support for pre-trained encoders, making it a robust application framework for deployment and integration in clinical workflows. Future studies can explore this mechanism by applying it to compare randomly initialized and pre-trained models for convergence speed (in both centralized and federated learning settings [8,9,42,43]), performance gains in applications requiring 3D datasets, model optimization allowing deployment in low-resource environments, and privacy analysis.

Acknowledgments. Research reported in this publication was partly supported by the National Institutes of Health (NIH) under award numbers NIH/NCI:U01CA242871 and NIH/NINDS:R01NS042645. The content of this publication is solely the responsibility of the authors and does not represent the official views of the NIH.

References

1. Zhou, S.K., et al.: A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc. IEEE **109**(5), 820–838 (2021)
2. Chen, X., et al.: Recent advances and clinical applications of deep learning in medical image analysis. Med. Image Anal. 102444 (2022)
3. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit. Med. **5**(1), 1–8 (2022)
4. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems, vol. 27 (2014)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambbridge (2016)
6. Aloysius, N., Geetha, M.: A review on deep convolutional neural networks. In: 2017 International Conference on Communication and Signal Processing (ICCSP), pp. 0588–0592. IEEE (2017)
7. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
8. Pati, S., et al.: Federated learning enables big data for rare cancer boundary detection. arXiv preprint arXiv:2204.10836 (2022)
9. Sheller, M.J., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci. Rep. **10**(1), 1–12 (2020)
10. Baid, U., et al.: NIMG-32. The federated tumor segmentation (FETS) initiative: the first real-world large-scale data-private collaboration focusing on neuro-oncology. Neuro-Oncology. **23**, pp. vi135-vi136 (2021)
11. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
12. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
13. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
14. Chen, J., Yang, L., Zhang, Y., Alber, M., Chen, D.Z.: Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
15. Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L.: Recurrent saliency transformation network: incorporating multi-stage visual cues for small organ segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8280–8289 (2018)
16. Díaz-Pernas, F.J., Martínez-Zarzuela, M., Antón-Rodríguez, M., González-Ortega, D.: A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network. In: Healthcare, vol. 9, p. 153, MDPI (2021)
17. Ismael, S.A.A., Mohammed, A., Hefny, H.: An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artif. Intell. Med. **102**, 101779 (2020)

18. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
19. Zhao, W., et al.: 3d deep learning from CT scans predicts tumor invasiveness of subcentimeter pulmonary adenocarcinomas. Cancer Res. **78**(24), 6881–6889 (2018)
20. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
21. Baid, U., et al.: A novel approach for fully automatic intra-tumor segmentation with 3d u-net architecture for gliomas. Front. Comput. Neurosci. 10 (2020)
22. Trivizakis, E., et al.: Extending 2-d convolutional neural networks to 3-d for advancing deep learning cancer classification with application to mri liver tumor differentiation. IEEE J. Biomed. Health Inform. **23**(3), 923–930 (2019)
23. Zheng, H., et al.: A new ensemble learning framework for 3d biomedical image segmentation. Proc. AAAI Conf. Artif. Intell. **33**, 5909–5916 (2019)
24. Xia, Y., Xie, L., Liu, F., Zhu, Z., Fishman, E.K., Yuille, A.L.: Bridging the gap between 2D and 3D organ segmentation with volumetric fusion net. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11073, pp. 445–453. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00937-3_51
25. Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., Heng, P.-A.: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging **37**(12), 2663–2674 (2018)
26. Ni, T., Xie, L., Zheng, H., Fishman, E.K., Yuille, A.L.: Elastic boundary projection for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2109–2118 (2019)
27. Pati, S., et al.: GandLF: a generally nuanced deep learning framework for scalable end-to-end clinical workflows in medical imaging. arXiv preprint arXiv:2103.01006 (2021)
28. Yang, J., et al.: Reinventing 2d convolutions for 3d images. IEEE J. Biomed. Health Inform. **25**(8), 3009–3018 (2021)
29. Yakubovskiy, P.: Segmentation models Pytorch. https://github.com/qubvel/segmentation_models.pytorch (2020)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
32. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
33. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
34. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

35. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
36. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**(1), 1–13 (2017)
37. Bakas, S., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
38. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in MR images: method and validation. IEEE Trans. Med. Imaging **13**(4), 716–724 (1994)
39. Binder, Z.A., et al.: Epidermal growth factor receptor extracellular domain mutations in glioblastoma present opportunities for clinical imaging and therapeutic development. Cancer cell **34**(1), 163–177 (2018)
40. Bakas, S., et al.: In vivo detection of EGFRVIII in glioblastoma via perfusion magnetic resonance imaging signature consistent with deep peritumoral infiltration: The ϕ-indexin vivo EGFRVIII detection in glioblastoma via MRI signature. Clin. Cancer Res. **23**(16), 4724–4734 (2017)
41. Akbari, H., et al.: In vivo evaluation of EGFRVIII mutation in primary glioblastoma patients via complex multiparametric MRI signature. Neuro-oncology **20**(8), 1068–1079 (2018)
42. Rieke, N., et al.: The future of digital health with federated learning. NPJ Digit. Med. **3**(1), 1–7 (2020)
43. Baid, U., et al.: Federated learning for the classification of tumor infiltrating lymphocytes. arXiv preprint arXiv:2203.16622 (2022)

Reprint Permissions

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Nov 20, 2023

This Agreement between Sarthak Pati ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5673040832097
License date	Nov 20, 2023
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Springer eBook
Licensed Content Title	Leveraging 2D Deep Learning ImageNet-trained Models for Native 3D Medical Image Analysis
Licensed Content Author	Bhakti Baheti, Sarthak Pati, Bjoern Menze et al
Licensed Content Date	Jan 1, 2023
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	electronic
Portion	full article/chapter
Will you be translating?	no

Circulation/distribution	1 - 29
Author of this Springer Nature content	yes
Title of new work	Reproducibility of Machine Learning Research in Clinical Environments
Institution name	Technical University of Munich
Expected presentation date	Mar 2024
Requestor Location	Sarthak Pati 1051 Harriman Ct WEST CHESTER, PA 19380 United States Attn: Sarthak Pati
Billing Type	Invoice
Billing Address	Sarthak Pati 1051 Harriman Ct WEST CHESTER, PA 19380 United States Attn: Sarthak Pati
Total	0.00 USD

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning,

intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia

Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature*'.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE

CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Part III

Conclusions and Outlook

7 Discussion

7.1 Overview

This thesis is meant to provide concepts related to reproducibility in *ML* research that are both easily interpretable and easily applicable in the computational healthcare domain, with a focus on medical imaging. While the topics are generally described across the sphere of radiomics with traditional *ML* and *DL* related to healthcare imaging, the ideas presented are generally relevant across problems and domains.

The inspiration for this work was to determine the path towards enabling clinical translation of computational research. Specifically, which considerations researchers need to keep in mind while designing their study, as well as which questions they need to ask when exploring the translational side of their research. This publication-based dissertation is comprised of three first-author journal and conference publications in Part II, and further supplemented by first and co-authored work in Part IV.

Emerging "zero-code" and "low-code" principles aim to broaden the landscape of software development by catering to diverse user groups. While "zero-code" empowers users to build solutions without writing a single line of code, "low-code" enables customization of existing solutions with minimal programming. It is imperative for open-source tools to embrace these principles, and thus target two distinct audiences: *i) computational researchers* should follow well-defined software standards and practices [3] to take advantage of the standardized input/output systems, common data loader interfaces, and well-defined application layers to focus on algorithmic development and thus push the boundaries of scientific research, and *ii) non-computational experts*, who can access the pre-built blocks provided by a tool to leverage their domain expertise to conduct analyses without requiring deep programming or technical knowledge, thus democratizing access to novel computational research for new domains. These concepts are critical to ensure the continued reproducibility, stability, and robustness of open-source tools. The work demonstrated in this thesis uses this idea as a cornerstone to ensure reproducibility.

7.2 Reproducibility Across Annotations & Radiomics

The exploration of inter-rater variability [78] was done to develop reliable image-based markers for glioblastoma tumors using radiomics and statistical modeling. The availability of large open-source repositories of data such as those from The Cancer Imaging Archive (*TCIA*)

[26] have enabled this research. Specifically, the availability of the entire Ivy Glioblastoma Atlas Project (*Ivy GAP*) cohort [79], which includes both preprocessed images and annotations from multiple clinical raters, as well as extracted radiomic features [27, 76], has enabled the increased use of radiomic analyses to further the goals of precision medicine (i.e., prognosis and predictive analysis) for glioblastoma tumors. Two main challenges were addressed with this work: *i)* the scarcity of consistent segmentation labels for different parts of the tumor, and *ii)* the variability of radiomic features across different segmentations. These challenges are addressed by quantifying the differences in the annotations generated by clinical radiology experts from two large health systems, which are then released to the community [79]. This is in contrast to previous related work [105], which only examined the robustness of radiomic features from the Cancer Genome Atlas Glioblastoma Multiforme Collection (*TCGA-GBM*) collection [26, 97]. Specifically, there are 4 key areas of difference between the two studies: *i)* the number and type of tumor regions, *ii)* the modalities that were considered, *iii)* the precise radiomic features used along with their parameterization, and *iv)* the statistical methods applied for analysis.

The analysis across the input images and segmentation labels obtained from the two institutions indicated a substantially high level of agreement between the two raters, which was to be expected since both are highly experienced clinical radiologists. This was especially true for the "tumor core" region of the tumor (comprising of the "enhancing" part of the tumor along with the "necrotic" tissue), as evidenced by the relatively high values of sensitivity (median value ≥ 0.85) and specificity (median value ≥ 0.95). This region is highly important from a clinical stand-point since it provides the definition of the region that is to be considered for surgical resection. Similarly, a high level of agreement was observed for even the "whole tumor" region (which comprises of the "tumor core" along with the edematous tissue), with median sensitivity ≥ 0.85. The analysis for radiomic feature reproducibility across the pair of annotations showcased 24.3% of $11,700$ of the computed radiomic features to be *robust* to annotation changes across the two sites. The majority of these features belonged to the morphology (which describes shape characteristics), intensity (which captures statistics across intensity profiles), and *COLLAGE* (which captures heterogeneity in local gradient orientations) families. This shows that even though the computational efficiency of *COLLAGE* features is high (see illustration in Figure 2.2), it is offset with a more than adequate level of *stability*. On the other hand, the computational expenditure for the majority of the texture feature families (i.e., the gray-level co-occurrence matrix (*GLCM*) family, gray-level run-length matrix (*GLRLM*) family, Gray-level size zone matrix (*GLSZM*) family, and Neighborhood gray tone difference matrix (*NGTDM*) family) was largely wasted, since they were not robust to variances across annotation differences [78].

The ***first study*** detailed in this thesis [78] has shown that although radiomics can measure tumor heterogeneity for glioblastomas (*GBM*) using non-invasive MRI scans [6], radiomic features can change depending on how the tumor is annotated at different sites. This work focused on performing a feasibility study to *(a)* assess how well two readers agreed on tumor annotations, and *(b)* find radiomic features that were *stable*, *robust*, and *reproducible* across different multi-institutional clinical experts for the same tumor region for the the cancer imaging archive's (*TCIA*) [26] Ivy Glioblastoma Atlas Project (*Ivy GAP*) dataset [85]. Firstly the most-commonly used metrics (*DSC*, Sensitivity, Specificity, and Hausdorff) were used to measure the inter-reader agreement. High values of the DSC, Sensitivity and Specificity and

low value of Hausdorff translated to better inter-reader agreement between the measured annotated regions. The results in this study showed that there was a high overall correlation between the two raters for all the annotated regions. Secondly, the radiomic variability analysis experiment indicated that *i)* some features and feature families such as intensity statistics (mean, median, standard deviation, kurtosis), morphologic (flatness, elongation, sphericity) and *COLLAGE* (statistics of local gradient entropy) might be more stable to variability in annotation labels from different readers, and *ii)* the more "complex" and relatively computationally expensive radiomic feature families such as histogram-based features, Gray-level co-occurrence matrix (*GLCM*), Gray-level run-length matrix (*GLRLM*), Gray-level size zone matrix (*GLSZM*), Neighborhood gray tone difference matrix (*NGTDM*) which rely on local intensity differences had lower correlation across features from the segmented tumor regions marked by two different raters. Although *GLCM* and *GLRLM* features have been shown to be prognostic of *GBM* [4, 16, 84], the results demonstrated in this work suggested that most of these features had large variations across the two annotations, and might need to be carefully examined for robustness across segmentations for prognostic modeling, specifically for *GBM* tumors. In contrast, however, most of morphology and intensity statistics-based features were seemingly resistant to differences in annotations between the two readers.

7.3 The Generally Nuanced Deep Learning Framework (GaNDLF)

The development of the Generally Nuanced Deep Learning Framework (*GaNDLF*) [77] could not have happened without all the public data that was used to benchmark its functionality. Benchmarking the performance of *GaNDLF* for various applications used public data repositories, such as: *i) segmentation of brain in magnetic resonance imaging (*MRI*)* used a combination of publicly available data [13, 15, 103] for training along with private collections from numerous institutions (such as Thomas Jefferson University and MD Anderson Cancer Center) for validation, *ii) segmentation of brain tumor sub-regions in* MRI was done using the data from the Brain Tumor Segmentation (BraTS) challenge of 2020 [12, 13, 14, 65], *iii) whole brain parcellation in* MRI used was from the Multi-Atlas Labelling challenge (*MALC*) of 2012 [54], *iv) segmentation of the structural tumor volume from breast* MRI was done using data obtained from the ACRIN 6657/I-SPY1 TRIAL [45, 68], and the annotations were obtained from a data repository from the cancer imaging archive [25], *v) segmentation of the retinal fundus* used the data from the Pathologic myopia (*PALM*) challenge [33], *vi) segmentation of colorectal cancer regions in whole slide images* used data from the DigestPath challenge [58], *vii) brain age prediction from* MRI used data from the United Kingdom Biobank [101] and a multi-site schizophrenia consortium [95], *viii) classification of diabetic foot ulcer images* [42] used data the Diabetic Foot Ulcer Grand Challenge (*DFUC*) of 2021 [118], and finally, the *ix) classification of tumor infiltrating lymphocyte density* [11] was done using data from the Cancer Genome Atlas (*TCGA*) [2]. *GaNDLF* has been proved to be a tool that demonstrates clinical viability by showcasing its capability to design end-to-end (starting with data curation, to preprocessing, training and post-processing) *DL* workflows across multiple imaging modalities, i.e., radiology (e.g., *MRI*, computed tomography (*CT*)), and histology (e.g., hematoxylin and eosin (*H&E*) stained slides). With *GaNDLF*, researchers can work with almost any type of

medical imaging data using the same framework, without writing any extra code. This makes it easier to conduct future research that depends on combining different diagnostics.

GaNDLF provides a "zero/low-code" solution enabling both computational and non-computational experts to train robust *DL* models to tackle a variety of workloads/tasks related to any type of healthcare imaging modality. It has been proved to be a tool that demonstrates clinical viability and the ability to quickly scale by showcasing its capability to design end-to-end (starting with data curation, to preprocessing, training and post-processing) *DL* workflows across multiple imaging modalities (radiographic or microscopic), without worrying about details such as appropriate data splitting for training to avoid leakage, validation & testing routines, tackling class imbalances, and implementing various training customization strategies for *DL* training (e.g., loss functions, optimizers). Specifically, *GaNDLF*'s abilities span across: *i)* processing images of various domains; *ii)* enabling work on various types of AI workloads (i.e., segmentation, regression, and classification); *iii)* offering built-in general-purpose functionality for augmentations and cross-validation; *iv)* integrating tools to promote the interpretability and explainability of *DL* topology outputs via M3D-CAM [41]; *v)* enabling built-in model optimization by leveraging the functionality provided via *OpenVINO* [10, 40, 104]; and *vi)* ensuring adherence of "good *ML*" practices, such as default cross-validation [29], to automatically generate optimized models after the training process is complete, allowing inference of these models on machines without requiring any specialized hardware, or large amounts of memory.

The ***second study*** detailed in this thesis [77] presented the *GaNDLF*, which was designed as an effort to standardize the definitions and initialization of all these processes for healthcare data science focusing in *DL*, while enabling reproducibility and clinical translation for research algorithms (see illustration in Figure 2.6). This coincided with the development of other healthcare-specific libraries, such as Project MONAI [23], however, in contrast to being a "toolbox" which computational researchers could leverage to create custom solutions, *GaNDLF* was designed to integrate the entire training and inference pipelines of a *DL* process in a human-readable text-based format by using the YAML syntax [20]. This allowed users to play around with and parameterize pipelines to their maximum extent while maintaining the reproducibility of the experiment for the open scientific community. This also allowed generation of baseline results using a large number of imaging modalities applied towards numerous types of workloads (segmentation, classification, and regression) in a relatively quick manner, as well as provided a vehicle for computational experts to package and ship their algorithms so that they could be applicable in multiple applications [77].

7.4 2D Pre-trained *DL* Models for Native 3D Medical Image Analysis

The ***third study*** detailed in this thesis [9] presented a novel *DL* network topology which harnesses the abilities of transfer learning to enable the training of 3 dimensional datasets in a native manner while using weights from models trained on large-scale 2 dimensional computer vision data [28] and has been an important contribution of this thesis [9]. This has been done using the concept of axial-coronal-sagittal convolutions (*ACSConv*) [117], where

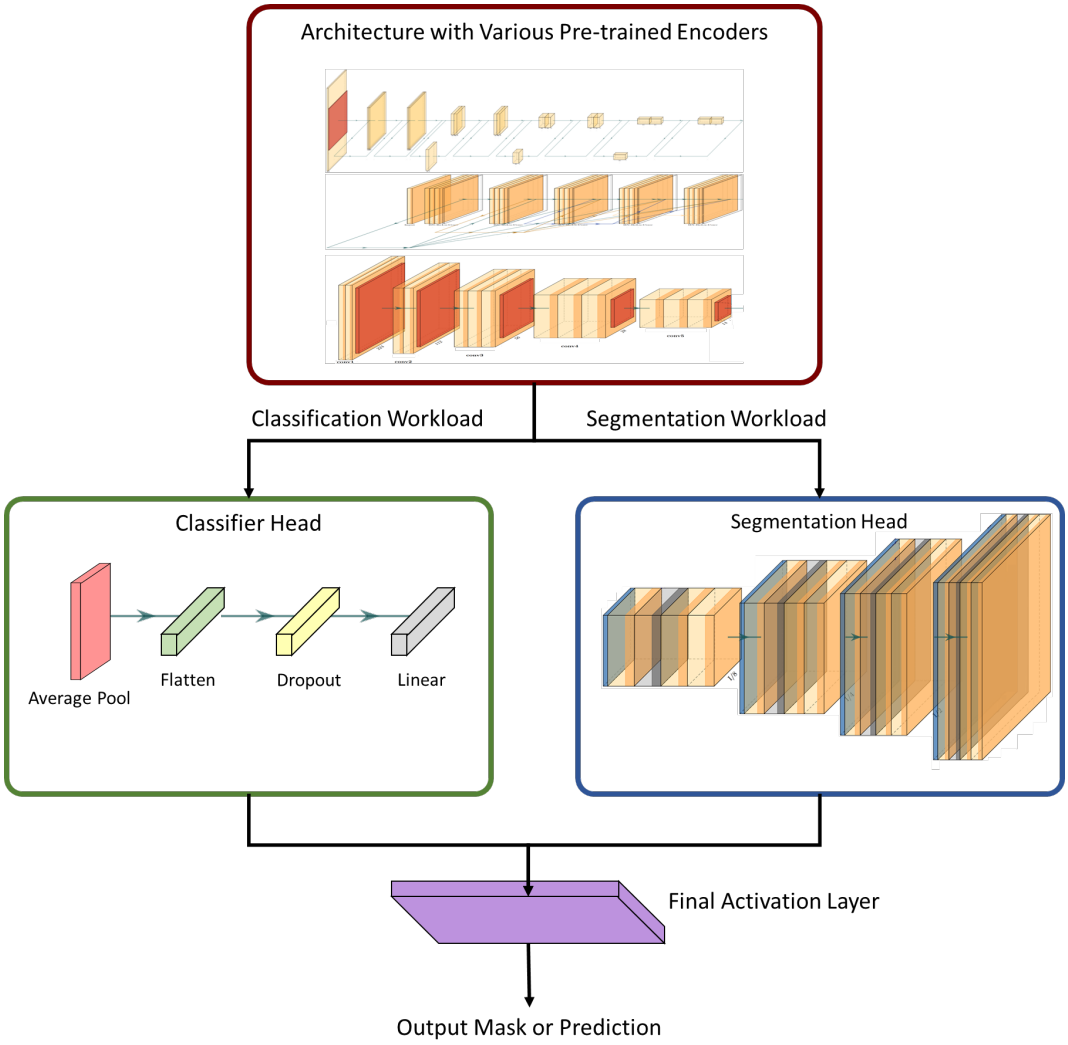


Fig. 7.1. Illustration of **FlexiNet**, a novel *DL* network topology that allows the usage of different pre-trained encoders using ImageNet [28] for either segmentation or classification workloads.

each 2D filter is sequentially applied across all the axes of a 3D image to produce the result. The network topology's first and last layers are flexible to be able to process input images of any size, with varying number of input channels (or modalities). Combining this capability with a two-pathway decoding mechanism (as illustrated in Figure 7.1), allows the topology to work for both semantic segmentation and classification tasks, thus making this applicable for a wide array of healthcare domains. The encoding layers are initialized with pre-trained weights from the models trained on ImageNet [28], and can be further fine-tuned per task. The topology represented significant performance improvements over the current state-of-the-art for topologies running on limited resources (not more than 11GB of dedicated *DL* accelerator card memory), both in terms of convergence speed, as well as in terms of utility metrics (both dice similarity coefficient for segmentation and accuracy for classification tasks). However, this topology has been explored only for fully-labeled datasets.

The performance comparisons were done by comparing an off-the-shelf 3D U-Net with residual connections [93] with a ResNet-50 encoder [43] with and without initialization using

weights derived from training a classification model on ImageNet [28]. The novel topology outperformed the standard U-Net when it was initialized with the ImageNet weights, and it was observed to reach convergence in $\leq 50\%$ of the number of epochs needed to train the topologies that were randomly initialized. The topology configured with a 4-layer ResNet-50 encoder has only 25.821 million parameters, which is around 22% less compared to 33.377 million parameters of the standard 4-layer 3D U-Net. This highlights the importance of this topology in the context of clinical translation, where fewer parameters translates to not only faster convergence but also to overall lower run-time requirements [104].

7.5 Good Reporting Practices for Computational Healthcare

In the realm of computational healthcare, adhering to established reporting guidelines is essential for accelerating clinical adoption, ensuring research reproducibility, and facilitating successful translation of findings into real-world applications. These guidelines include: *i)* Checklist for Artificial Intelligence in Medical Imaging (*CLAIM*) [66], which outlines reporting requirements for medical-imaging AI research; *ii)* AI-specific version of the Standards for Reporting of Diagnostic Accuracy Study (*STARD-AI*) [100], Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (*TRIPOD-AI*), and Prediction model Risk Of Bias ASsessment Tool (*PROBAST-AI*) checklists, which address challenges specific to AI applications in diagnostic accuracy studies; *iii)* AI extensions of Consolidated Standards of Reporting Trials (*CONSORT-AI*) and Standard Protocol Items: Recommendations for Interventional Trials (*SPIRIT-AI*), which provides guidance for reporting randomized clinical trials [59]; *v)* Minimum Information about Clinical Artificial Intelligence Modelling (*MI-CLAIM*) [69], which focuses on the clinical impact and the technical reproducibility of clinically relevant AI studies; *vi)* MINimum Information for Medical AI Reporting (*MINIMAR*) [44], which sets the reporting standards for medical AI applications by specifying the minimum information that AI manuscripts should include; and finally *vi)* Radiomics Quality Score (*RQS*) [53], which outlines 16 unique criteria by which to judge the quality of a publication whose analysis is based on radiomics [125].

8 Outlook

While *GaNDLF* has demonstrated its effectiveness across imaging modalities with both single inputs (e.g., radiology or histology images) and multi-channel support (e.g., combined MRI sequences), its applications have primarily been focused on segmentation, regression, and classification tasks. Exploring its potential in other computational domains, such as synthesis of "new data", semi-/self-supervised training, data fusion to integrate different imaging modalities (radiology/histology) with categorical healthcare data and genomic profiles, and physics-informed modeling could significantly expand its usefulness in terms of its clinical relevance. Furthermore, the framework's suitability for datasets involving images with a time-series component (e.g., dynamic sequences) or higher dimensional datasets (e.g., multi-spectral imaging) remains unexplored and deserves investigation. Furthermore, no mechanism is present to automatically enable the aggregation of various models (that is, run the training and inference of different models and then collect data from all), which have generally been shown to produce better results [14, 65]. Mechanisms that enable AutoML [114] and other network architecture search (*NAS*) techniques [30] are tremendously powerful tools that create robust models, but are currently not supported in *GaNDLF*. Finally, application of *GaNDLF* to other data types, such as genomics or electronic health records (*EHR*), which would allow *GaNDLF* to further inform and aid clinical decision making by training multi-modal models, has not been fully explored yet but it is considered as current work in progress.

While the novel *DL* network topology presented in [9] shows a great deal of promise, there is further work to be done towards investigating the applicability of this idea for weakly supervised or semi-supervised learning. There is also a need to explore this method towards fully unsupervised learning. This would allow datasets with noisy or absent annotations to be used to further improve model training. Since the core concept of this topology involves distilling (or *encoding*) the raw information from the input image(s) onto a common latent space, it could be explored towards the idea of data fusion, where different imaging and non-imaging sources of data are combined through a classifier to give a more holistic model output. Some of the application areas could be combining radiology imaging data with categorical healthcare reports, radiology and histology imaging data could be potentially trained together, and combining radiology and histology imaging data with categorical healthcare reports. All of these have

Finally, in the spirit of open science, **all** components of this thesis have been made available as the Comprehensive Federated Ecosystem (*COFE*) (see illustration in Figure 8.1). The overall idea behind *COFE* is to enable researchers (both clinically-focused and computationally-focused) to leverage existing libraries, tools, and software infrastructure to push the boundaries of science and our understanding of diseases. This includes enabling access to cutting-edge *ML* research within clinical workstations via an easy-to-use tool with a graphical interface [73], an algorithmic core that enables fast translation of computational research into different application domains (*GaNDLF* [77]), a security focused federated learning library to enable

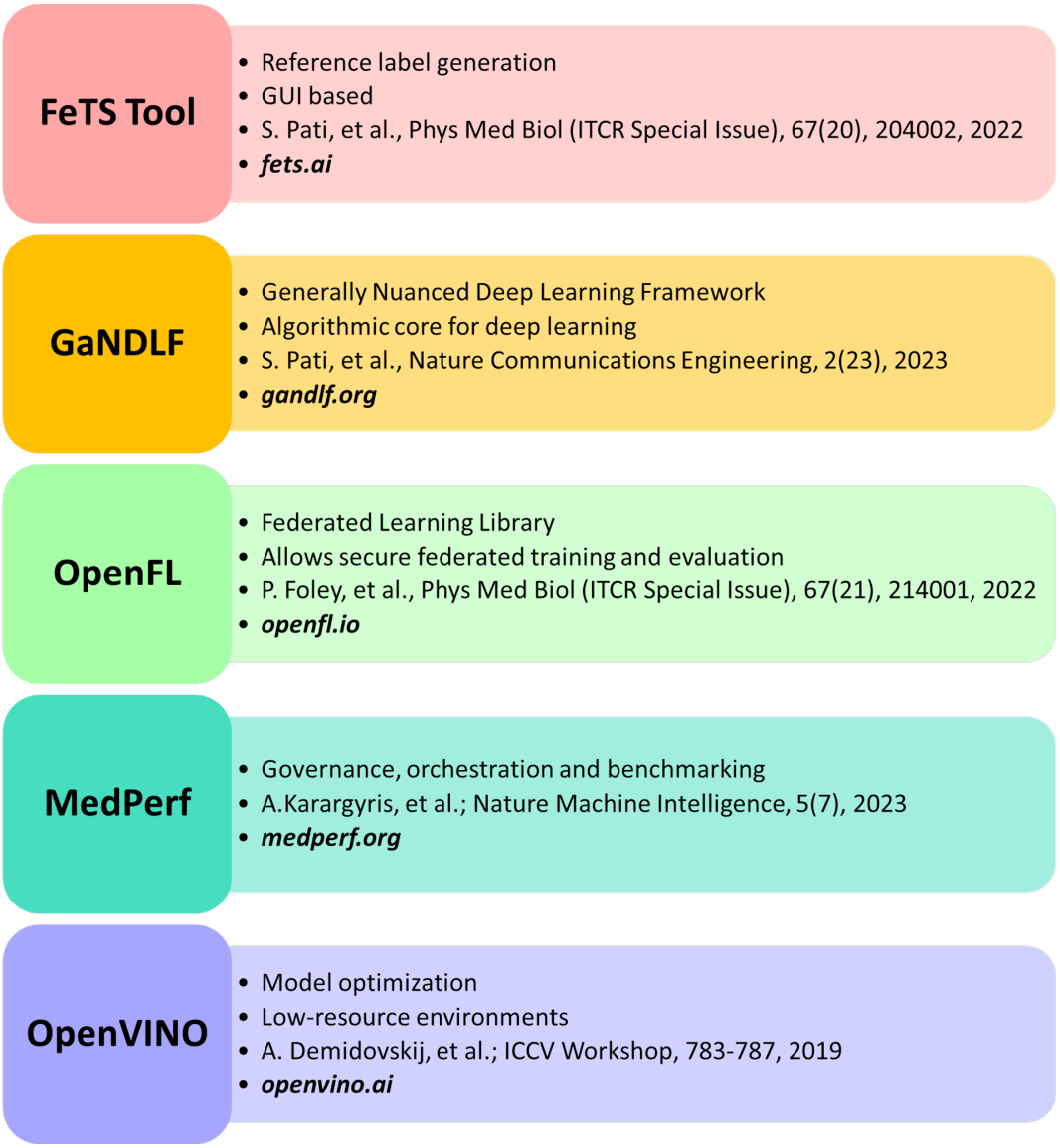


Fig. 8.1. Illustration of the Comprehensive Federated Ecosystem (*COFE*). Combining easy-to-use graphical user interfaces [73] with a powerful and robust *DL* algorithmic core [77] allows research to be propelled forward in a meaningful way by making labeling and training easy. Scaling this research to multiple clinical sites around the world requires their seamless integration with a secure federated learning library [32] along with a robust governance and orchestration application [50]. Finally, a reproducible and stable optimization toolkit [40] allows trained models to be inferred without requiring specialized hardware, thus democratizing precision medicine for under-served populations.

collaborative computing (*OpenFL* [32, 88]), a collaborative governance and orchestration solution for benchmarking (*MedPerf* [50]), and finally an application that can optimize models trained on specialized *DL* accelerator cards so that they can run on generic hardware (*OpenVINO* [40]).

Part IV

Appendix

A Abstracts of Publications not Discussed in this Thesis

A.1 The Cancer Imaging Phenomics Toolkit (CaPTk): Technical Overview

Sarthak Pati, Ashish Singh, Saima Rathore, Aimilia Gastounioti, Mark Bergman, Phuc Ngo, Sung Min Ha, Dimitrios Bounias, James Minock, Grayson Murphy, Hongming Li, Amit Bhattarai, Adam Wolf, Patmaa Sridaran, Ratheesh Kalarot, Hamed Akbari, Aristeidis Sotiras, Siddhesh P. Thakur, Ragini Verma, Russell T. Shinohara, Paul Yushkevich, Yong Fan, Despina Kontos, Christos Davatzikos, Spyridon Bakas

The purpose of this manuscript is to provide an overview of the technical specifications and architecture of the Cancer imaging Phenomics Toolkit (CaPTk - `www.cbica.upenn.edu/captk`), a cross-platform, open-source, easy-to-use, and extensible software platform for analyzing 2D and 3D images, currently focusing on radiographic scans of brain, breast, and lung cancer. The primary aim of this platform is to enable swift and efficient translation of cutting-edge academic research into clinically useful tools relating to clinical quantification, analysis, predictive modeling, decision-making, and reporting workflow. CaPTk builds upon established open-source software toolkits, such as the Insight Toolkit (ITK) and OpenCV, to bring together advanced computational functionality. This functionality describes specialized, as well as general-purpose, image analysis algorithms developed during active multi-disciplinary collaborative research studies to address real clinical requirements. The target audience of CaPTk consists of both computational scientists and clinical experts. For the former it provides i) an efficient image viewer offering the ability of integrating new algorithms, and ii) a library of readily-available clinically-relevant algorithms, allowing batch-processing of multiple subjects. For the latter it facilitates the use of complex algorithms for clinically-relevant studies through a user-friendly interface, eliminating the prerequisite of a substantial computational background. CaPTk's long-term goal is to provide widely-used technology to make use of advanced quantitative imaging analytics in cancer prediction, diagnosis and prognosis, leading toward a better understanding of the biological mechanisms of cancer development.

Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries 2019
Reprinted with permission from Springer.
DOI: `https://doi.org/10.1007/978-3-030-46643-5_38`

A.2 Estimating Glioblastoma Biophysical Growth Parameters Using Deep Learning Regression

Sarthak Pati, Vaibhav Sharma, Heena Aslam, Siddhesh P Thakur, Hamed Akbari, Andreas Mang, Shashank Subramanian, George Biros, Christos Davatzikos, Spyridon Bakas

Glioblastoma (GBM) is arguably the most aggressive, infiltrative, and heterogeneous type of adult brain tumor. Biophysical modeling of GBM growth has contributed to more informed clinical decision-making. However, deploying a biophysical model to a clinical environment is challenging since underlying computations are quite expensive and can take several hours using existing technologies. Here we present a scheme to accelerate the computation. In particular, we present a deep learning (DL)-based logistic regression model to estimate the GBM's biophysical growth in seconds. This growth is defined by three tumor-specific parameters: 1) a diffusion coefficient in white matter (Dw), which prescribes the rate of infiltration of tumor cells in white matter, 2) a mass-effect parameter (Mp), which defines the average tumor expansion, and 3) the estimated time (T) in number of days that the tumor has been growing. Preoperative structural multi-parametric MRI (mpMRI) scans from $n = 135$ subjects of the TCGA-GBM imaging collection are used to quantitatively evaluate our approach. We consider the mpMRI intensities within the region defined by the abnormal FLAIR signal envelope for training one DL model for each of the tumor-specific growth parameters. We train and validate the DL-based predictions against parameters derived from biophysical inversion models. The average Pearson correlation coefficients between our DL-based estimations and the biophysical parameters are 0.85 for Dw, 0.90 for Mp, and 0.94 for T, respectively. This study unlocks the power of tumor-specific parameters from biophysical tumor growth estimation. It paves the way towards their clinical translation and opens the door for leveraging advanced radiomic descriptors in future studies by means of a significantly faster parameter reconstruction compared to biophysical growth modeling approaches.

Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries 2020
Reprinted with permission from Springer.
DOI: https://doi.org/10.1007/978-3-030-72084-1_15

A.3 Classification of infection and ischemia in diabetic foot ulcers using vgg architectures

Orhun Güley, Sarthak Pati, Spyridon Bakas

Diabetic foot ulceration (DFU) is a serious complication of diabetes, and a major challenge for healthcare systems around the world. Further infection and ischemia in DFU can significantly prolong treatment and often result in limb amputation, with more severe cases resulting in terminal illness. Thus, early identification and regular monitoring is necessary to improve care, and reduce the burden on healthcare systems. With that in mind, this study attempts to address the problem of infection and ischemia classification in diabetic food ulcers, in four distinct classes. We have evaluated a series of VGG architectures with different layers, following numerous training strategies, including k-fold cross validation, data pre-processing options, augmentation techniques, and weighted loss calculations. In favor of transparency and reproducibility, we make all the implementations available through the Generally Nuanced Deep Learning Framework (GaNDLF, `github.com/CBICA/GaNDLF`. Our best model was evaluated during the DFU Challenge 2021, and was ranked 2th, 5th, and 7th based on the macro-averaged AUC (area under the curve), macro-averaged F1 score, and macro-averaged recall metrics, respectively. Our findings support that current state-of-the-art architectures provide good results for the DFU image classification task, and further experimentation is required to study the effects of pre-processing and augmentation strategies.

Diabetic Foot Ulcers Grand Challenge 2021
Reprinted with permission from Springer.
DOI: `https://doi.org/10.1007/978-3-030-94907-5_6`

A.4 Expert tumor annotations and radiomics for locally advanced breast cancer in DCE-MRI for ACRIN 6657/I-SPY1

Rhea Chitalia, Sarthak Pati, Megh Bhalerao, Siddhesh Pravin Thakur, Nariman Jahani, Vivian Belenky, Elizabeth S McDonald, Jessica Gibbs, David C Newitt, Nola M Hylton, Despina Kontos, Spyridon Bakas

Breast cancer is one of the most pervasive forms of cancer and its inherent intra- and inter-tumor heterogeneity contributes towards its poor prognosis. Multiple studies have reported results from either private institutional data or publicly available datasets. However, current public datasets are limited in terms of having consistency in: a) data quality, b) quality of expert annotation of pathology, and c) availability of baseline results from computational algorithms. To address these limitations, here we propose the enhancement of the I-SPY1 data collection, with uniformly curated data, tumor annotations, and quantitative imaging features. Specifically, the proposed dataset includes a) uniformly processed scans that are harmonized to match intensity and spatial characteristics, facilitating immediate use in computational studies, b) computationally-generated and manually-revised expert annotations of tumor regions, as well as c) a comprehensive set of quantitative imaging (also known as radiomic) features corresponding to the tumor regions. This collection describes our contribution towards repeatable, reproducible, and comparative quantitative studies leading to new predictive, prognostic, and diagnostic assessments.

Nature Scientific Data
Open Access Article.
DOI: https://doi.org/10.1038/s41597-022-01555-4

A.5 Federated learning enables big data for rare cancer boundary detection

Sarthak Pati, Ujjwal Baid, Brandon Edwards, Micah Sheller, Shih-Han Wang, G. Anthony Reina, Patrick Foley, Alexey Gruzdev, Deepthi Karkada, Christos Davatzikos, Chiharu Sako, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Philipp Vollmuth, Gianluca Brugnara, Chandrakanth J. Preetha, Felix Sahm, Klaus Maier-Hein, Maximilian Zenk, Martin Bendszus, Wolfgang Wick, Evan Calabrese, Jeffrey Rudie, Javier Villanueva-Meyer, Soonmee Cha, Madhura Ingalhalikar, Manali Jadhav, Umang Pandey, Jitender Saini, John Garrett, Matthew Larson, Robert Jeraj, Stuart Currie, Russell Frood, Kavi Fatania, Raymond Y. Huang, Ken Chang, Carmen Balaña, Jaume Capellades, Josep Puig, Johannes Trenkler, Josef Pichler, Georg Necker, Andreas Haunschmidt, Stephan Meckel, Gaurav Shukla, Spencer Liem, Gregory S. Alexander, Joseph Lombardo, Joshua D. Palmer, Adam E. Flanders, Adam P. Dicker, Haris I. Sair, Craig K. Jones, Archana Venkataraman, Meirui Jiang, Tiffany Y. So, Cheng Chen, Pheng Ann Heng, Qi Dou, Michal Kozubek, Filip Lux, Jan Michálek, Petr Matula, Miloš Keřkovský, Tereza Kopřivová, Marek Dostál, Václav Vybíhal, Michael A. Vogelbaum, J. Ross Mitchell, Joaquim Farinhas, Joseph A. Maldjian, Chandan Ganesh Bangalore Yogananda, Marco C. Pinho, Divya Reddy, James Holcomb, Benjamin C. Wagner, Benjamin M. Ellingson, Timothy F. Cloughesy, Catalina Raymond, Talia Oughourlian, Akifumi Hagiwara, Chencai Wang, Minh-Son To, Sargam Bhardwaj, Chee Chong, Marc Agzarian, Alexandre Xavier Falcão, Samuel B. Martins, Bernardo C. A. Teixeira, Flávia Sprenger, David Menotti, Diego R. Lucio, Pamela LaMontagne, Daniel Marcus, Benedikt Wiestler, Florian Kofler, Ivan Ezhov, Marie Metz, Rajan Jain, Matthew Lee, Yvonne W. Lui, Richard McKinley, Johannes Slotboom, Piotr Radojewski, Raphael Meier, Roland Wiest, Derrick Murcia, Eric Fu, Rourke Haas, John Thompson, David Ryan Ormond, Chaitra Badve, Andrew E. Sloan, Vachan Vadmal, Kristin Waite, Rivka R. Colen, Linmin Pei, Murat Ak, Ashok Srinivasan, J. Rajiv Bapuraj, Arvind Rao, Nicholas Wang, Ota Yoshiaki, Toshio Moritani, Sevcan Turk, Joonsang Lee, Snehal Prabhudesai, Fanny Morón, Jacob Mandel, Konstantinos Kamnitsas, Ben Glocker, Luke V. M. Dixon, Matthew Williams, Peter Zampakis, Vasileios Panagiotopoulos, Panagiotis Tsiganos, Sotiris Alexiou, Ilias Haliassos, Evangelia I. Zacharaki, Konstantinos Moustakas, Christina Kalogeropoulou, Dimitrios M. Kardamakis, Yoon Seong Choi, Seung-Koo Lee, Jong Hee Chang, Sung Soo Ahn, Bing Luo, Laila Poisson, Ning Wen, Pallavi Tiwari, Ruchika Verma, Rohan Bareja, Ipsa Yadav, Jonathan Chen, Neeraj Kumar, Marion Smits, Sebastian R. van der Voort, Ahmed Alafandi, Fatih Incekara, Maarten M. J. Wijnenga, Georgios Kapsas, Renske Gahrmann, Joost W. Schouten, Hendrikus J. Dubbink, Arnaud J. P. E. Vincent, Martin J. van den Bent, Pim J. French, Stefan Klein, Yading Yuan, Sonam Sharma, Tzu-Chi Tseng, Saba Adabi, Simone P. Niclou, Olivier Keunen, Ann-Christin Hau, Martin Vallières, David Fortin, Martin Lepage, Bennett Landman, Karthik Ramadass, Kaiwen Xu, Silky Chotai, Lola B. Chambless, Akshitkumar Mistry, Reid C. Thompson, Yuriy Gusev, Krithika Bhuvaneshwar, Anousheh Sayah, Camelia Bencheqroun, Anas Belouali, Subha Madhavan, Thomas C. Booth, Alysha Chelliah, Marc Modat, Haris Shuaib, Carmen Dragos, Aly Abayazeed, Kenneth Kolodziej, Michael Hill, Ahmed Abbassy, Shady Gamal, Mahmoud Mekhaimar, Mohamed Qayati, Mauricio Reyes, Ji Eun Park, Jihye Yun, Ho Sung Kim, Abhishek Mahajan, Mark Muzi, Sean Benson, Regina G. H. Beets-Tan, Jonas Teuwen,

Alejandro Herrera-Trujillo, Maria Trujillo, William Escobar, Ana Abello, Jose Bernal, Jhon Gómez, Joseph Choi, Stephen Baek, Yusung Kim, Heba Ismael, Bryan Allen, John M. Buatti, Aikaterini Kotrotsou, Hongwei Li, Tobias Weiss, Michael Weller, Andrea Bink, Bertrand Pouymayou, Hassan F. Shaykh, Joel Saltz, Prateek Prasanna, Sampurna Shrestha, Kartik M. Mani, David Payne, Tahsin Kurc, Enrique Pelaez, Heydy Franco-Maldonado, Francis Loayza, Sebastian Quevedo, Pamela Guevara, Esteban Torche, Cristobal Mendoza, Franco Vera, Elvis Ríos, Eduardo López, Sergio A. Velastin, Godwin Ogbole, Mayowa Soneye, Dotun Oyekunle, Olubunmi Odafe-Oyibotha, Babatunde Osobu, Mustapha Shu'aibu, Adeleye Dorcas, Farouk Dako, Amber L. Simpson, Mohammad Hamghalam, Jacob J. Peoples, Ricky Hu, Anh Tran, Danielle Cutler, Fabio Y. Moraes, Michael A. Boss, James Gimpel, Deepak Kattil Veettil, Kendall Schmidt, Brian Bialecki, Sailaja Marella, Cynthia Price, Lisa Cimino, Charles Apgar, Prashant Shah, Bjoern Menze, Jill S. Barnholtz-Sloan, Jason Martin, Spyridon Bakas

Although machine learning (ML) has shown promise across disciplines, out-of-sample generalizability is concerning. This is currently addressed by sharing multi-site data, but such centralization is challenging/infeasible to scale due to various limitations. Federated ML (FL) provides an alternative paradigm for accurate and generalizable ML, by only sharing numerical model updates. Here we present the largest FL study to-date, involving data from 71 sites across 6 continents, to generate an automatic tumor boundary detector for the rare disease of glioblastoma, reporting the largest such dataset in the literature ($n = 6,314$). We demonstrate a 33% delineation improvement for the surgically targetable tumor, and 23% for the complete tumor extent, over a publicly trained model. We anticipate our study to: 1) enable more healthcare studies informed by large diverse data, ensuring meaningful results for rare diseases and underrepresented populations, 2) facilitate further analyses for glioblastoma by releasing our consensus model, and 3) demonstrate the FL effectiveness at such scale and task-complexity as a paradigm shift for multi-site collaborations, alleviating the need for data-sharing.

Nature Communications
Open Access Article.
DOI: https://doi.org/10.1038/s41467-022-33407-5

A.6 Optimization of deep learning based brain extraction in mri for low resource environments

Siddhesh P Thakur, Sarthak Pati, Ravi Panchumarthy, Deepthi Karkada, Junwen Wu, Dmitry Kurtaev, Chiharu Sako, Prashant Shah, Spyridon Bakas

Brain extraction is an indispensable step in neuro-imaging with a direct impact on downstream analyses. Most such methods have been developed for non-pathologically affected brains, and hence tend to suffer in performance when applied on brains with pathologies, e.g., gliomas, multiple sclerosis, traumatic brain injuries. Deep Learning (DL) methodologies for healthcare have shown promising results, but their clinical translation has been limited, primarily due to these methods suffering from i) high computational cost, and ii) specific hardware requirements, e.g., DL acceleration cards. In this study, we explore the potential of mathematical optimizations, towards making DL methods amenable to application in low resource environments. We focus on both the qualitative and quantitative evaluation of such optimizations on an existing DL brain extraction method, designed for pathologically-affected brains and agnostic to the input modality. We conduct direct optimizations and quantization of the trained model (i.e., prior to inference on new data). Our results yield substantial gains, in terms of speedup, latency, throughput, and reduction in memory usage, while the segmentation performance of the initial and the optimized models remains stable, i.e., as quantified by both the Dice Similarity Coefficient and the Hausdorff Distance. These findings support post-training optimizations as a promising approach for enabling the execution of advanced DL methodologies on plain commercial-grade CPUs, and hence contributing to their translation in limited- and low- resource clinical environments.

Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries 2021
Reprinted with permission from Springer.
DOI: https://doi.org/10.1007/978-3-031-08999-2_12

A.7 Federated benchmarking of medical artificial intelligence with MedPerf

Alexandros Karargyris, Renato Umeton, Micah J Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, Prakash Narayana Moorthy, Alexander Chowdhury, Junyi Guo, Sahil Nalawade, Jacob Rosenthal, David Kanter, Maria Xenochristou, Daniel J Beutel, Verena Chung, Timothy Bergquist, James Eddy, Abubakar Abid, Lewis Tunstall, Omar Sanseviero, Dimitrios Dimitriadis, Yiming Qian, Xinxing Xu, Yong Liu, Rick Siow Mong Goh, Srini Bala, Victor Bittorf, Sreekar Reddy Puchala, Biagio Ricciuti, Soujanya Samineni, Eshna Sengupta, Akshay Chaudhari, Cody Coleman, Bala Desinghu, Gregory Diamos, Debo Dutta, Diane Feddema, Grigori Fursin, Xinyuan Huang, Satyananda Kashyap, Nicholas Lane, Indranil Mallick, FeTS Consortium, BraTS-2020 Consortium, AI4SafeChole Consortium, Pietro Mascagni, Virendra Mehta, Cassiano Ferro Moraes, Vivek Natarajan, Nikola Nikolov, Nicolas Padoy, Gennady Pekhimenko, Vijay Janapa Reddi, G Anthony Reina, Pablo Ribalta, Abhishek Singh, Jayaraman J Thiagarajan, Jacob Albrecht, Thomas Wolf, Geralyn Miller, Huazhu Fu, Prashant Shah, Daguang Xu, Poonam Yadav, David Talby, Mark M Awad, Jeremy P Howard, Michael Rosenthal, Luigi Marchionni, Massimo Loda, Jason M Johnson, Spyridon Bakas, Peter Mattson

Medical artificial intelligence (AI) has tremendous potential to advance healthcare by supporting and contributing to the evidence-based practice of medicine, personalizing patient treatment, reducing costs, and improving both healthcare provider and patient experience. Unlocking this potential requires systematic, quantitative evaluation of the performance of medical AI models on large-scale, heterogeneous data capturing diverse patient populations. Here, to meet this need, we introduce MedPerf, an open platform for benchmarking AI models in the medical domain. MedPerf focuses on enabling federated evaluation of AI models, by securely distributing them to different facilities, such as healthcare organizations. This process of bringing the model to the data empowers each facility to assess and verify the performance of AI models in an efficient and human-supervised process, while prioritizing privacy. We describe the current challenges healthcare and AI communities face, the need for an open platform, the design philosophy of MedPerf, its current implementation status and real-world deployment, our roadmap and, importantly, the use of MedPerf with multiple international institutions within cloud-based technology and on-premises scenarios. Finally, we welcome new contributions by researchers and organizations to further strengthen MedPerf as an open benchmarking platform.

Nature Machine Intelligence
Open Access Article.
DOI: https://doi.org/10.1038/s42256-023-00652-2

Bibliography

B

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. "Tensorflow: A system for large-scale machine learning". In: *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. 2016, pp. 265–283. DOI: 10.48550/arXiv.1605.08695 (cit. on pp. 16, 18).

[2] S. Abousamra, L. Hou, R. Gupta, C. Chen, D. Samaras, T. Kurc, R. Batiste, T. Zhao, S. Kenneth, and J. Saltz. "Learning from thresholds: fully automated classification of tumor infiltrating lymphocytes for multiple cancer types". In: *arXiv preprint arXiv:1907.03960* (2019). DOI: 10.48550/arXiv.1907.03960 (cit. on p. 103).

[3] A. Afiaz, A. Ivanov, J. Chamberlin, D. Hanauer, C. Savonen, M. J. Goldman, M. Morgan, M. Reich, A. Getka, A. Holmes, et al. "Evaluation of software impact designed for biomedical research: Are we measuring what's meaningful?" In: *arXiv preprint arXiv:2306.03255* (2023). DOI: 10.48550/arXiv.2306.03255 (cit. on p. 101).

[4] H. Akbari, S. Bakas, J. M. Pisapia, M. P. Nasrallah, M. Rozycki, M. Martinez-Lage, J. J. Morrissette, N. Dahmane, D. M. O'Rourke, and C. Davatzikos. "In vivo evaluation of EGFRvIII mutation in primary glioblastoma patients via complex multiparametric MRI signature". In: *Neuro-oncology* (2018). DOI: 10.1093/neuonc/noy033 (cit. on p. 103).

[5] H. Akbari, S. Rathore, S. Bakas, M. P. Nasrallah, G. Shukla, E. Mamourian, M. Rozycki, S. J. Bagley, J. D. Rudie, A. E. Flanders, et al. "Histopathology-validated machine learning radiographic biomarker for noninvasive discrimination between true progression and pseudo-progression in glioblastoma". In: *Cancer* 126.11 (2020), pp. 2625–2636. DOI: 10.1002/cncr.32790 (cit. on p. 9).

[6] L. Alic, W. J. Niessen, and J. F. Veenland. "Quantification of heterogeneity as a biomarker in tumor imaging: a systematic review". In: *PloS one* 9.10 (2014). DOI: 10.1371/journal.pone.0110300 (cit. on p. 102).

[7] D. M. Allen. "The relationship between variable selection and data agumentation and a method for prediction". In: *technometrics* 16.1 (1974), pp. 125–127. DOI: 10.1080/00401706.1974.10489157 (cit. on p. 22).

[8] G. J. Annas et al. "HIPAA regulations-a new era of medical-record privacy?" In: *New England Journal of Medicine* 348.15 (2003), pp. 1486–1490. DOI: 10.1056/NEJMlim035027 (cit. on pp. 10, 12).

[9] B. Baheti, S. Pati, B. Menze, and S. Bakas. "Leveraging 2D Deep Learning ImageNet-trained Models for Native 3D Medical Image Analysis". In: *International MICCAI Brainlesion Workshop*. Springer. 2022, pp. 68–79. DOI: 10.1007/978-3-031-33842-7_6 (cit. on pp. 27, 104, 107).

[10] B. Baheti, S. Thakur, S. Pati, D. Karkada, R. Panchumarthy, J. Wu, S. Mohan, M. Nasrallah, P. Shah, and S. Bakas. "NIMG-25. OPTIMIZATION OF ARTIFICIAL INTELLIGENCE ALGORITHMS FOR LOW-RESOURCE/CLINICAL ENVIRONMENTS: FOCUS ON CLINICALLY-RELEVANT GLIOMA REGION DELINEATION". In: *Neuro-Oncology* 24.Suppl 7 (2022), p. vii167. DOI: 10.1093/neuonc/noac209.643 (cit. on p. 104).

[11] U. Baid, S. Pati, T. M. Kurc, R. Gupta, E. Bremer, S. Abousamra, S. P. Thakur, J. H. Saltz, and S. Bakas. "Federated learning for the classification of tumor infiltrating lymphocytes". In: *arXiv preprint arXiv:2203.16622* (2022). DOI: 10.48550/arXiv.2203.16622 (cit. on p. 103).

[12] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos. "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection". In: *The cancer imaging archive* 286 (2017). DOI: 10.7937/K9/TCIA.2017.KLXWJJ1Q (cit. on pp. 9, 103).

[13] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4.1 (2017), pp. 1–13. DOI: 10.1038/sdata.2017.117 (cit. on pp. 9, 103).

[14] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* (2018). DOI: 10.48550/arXiv.1811.02629 (cit. on pp. 9, 16, 103, 107).

[15] S. Bakas, C. Sako, H. Akbari, M. Bilello, A. Sotiras, G. Shukla, J. D. Rudie, N. F. Santamaria, A. F. Kazerooni, S. Pati, et al. "The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics". In: *Scientific data* 9.1 (2022), p. 453. DOI: 10.1038/s41597-022-01560-7 (cit. on p. 103).

[16] S. Bakas, G. Shukla, H. Akbari, G. Erus, A. Sotiras, S. Rathore, C. Sako, S. M. Ha, M. Rozycki, A. Singh, et al. "Integrative radiomic analysis for pre-surgical prognostic stratification of glioblastoma patients: from advanced to basic MRI protocols". In: *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 11315. International Society for Optics and Photonics. 2020, 113151S. DOI: 10.1117/12.2566505 (cit. on p. 103).

[17] P. Baldi, P. Sadowski, and D. Whiteson. "Searching for exotic particles in high-energy physics with deep learning". In: *Nature communications* 5.1 (2014), pp. 1–9. DOI: 10.1038/ncomms5308 (cit. on p. 9).

[18] A. L. Beam, A. K. Manrai, and M. Ghassemi. "Challenges to the reproducibility of machine learning models in health care". In: *Jama* 323.4 (2020), pp. 305–306. DOI: 10.1001/jama.2019.20866 (cit. on p. 10).

[19] A. Beers, J. Brown, K. Chang, K. Hoebel, J. Patel, K. I. Ly, S. M. Tolaney, P. Brastianos, B. Rosen, E. R. Gerstner, et al. "DeepNeuro: an open-source deep learning toolbox for neuroimaging". In: *Neuroinformatics* (2020), pp. 1–14. DOI: 10.1007/s12021-020-09477-5 (cit. on p. 18).

[20] O. Ben-Kiki, C. Evans, and B. Ingerson. "Yaml ain't markup language (yaml™) version 1.1". In: *Working Draft 2008* 5.11 (2009) (cit. on p. 104).

[21] J. Borovec, J. Kybic, I. Arganda-Carreras, D. V. Sorokin, G. Bueno, A. V. Khvostikov, S. Bakas, I. Eric, C. Chang, S. Heldmann, et al. “ANHIR: automatic non-rigid histological image registration challenge”. In: *IEEE Transactions on Medical Imaging* (2020). DOI: 10.1109/TMI.2020.2986331 (cit. on p. 9).

[22] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. “The balanced accuracy and its posterior distribution”. In: *2010 20th international conference on pattern recognition*. IEEE. 2010, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764 (cit. on p. 24).

[23] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al. “MONAI: An open-source framework for deep learning in healthcare”. In: *arXiv preprint arXiv:2211.02701* (2022). DOI: 10.48550/arXiv.2211.02701 (cit. on pp. 18, 104).

[24] R. Castaldo, V. Brancato, C. Cavaliere, F. Trama, E. Illiano, E. Costantini, A. Ragozzino, M. Salvatore, E. Nicolai, and M. Franzese. “A framework of analysis to facilitate the harmonization of multicenter radiomic features in prostate cancer”. In: *Journal of Clinical Medicine* 12.1 (2022), p. 140. DOI: 10.3390/jcm12010140 (cit. on p. 9).

[25] R. Chitalia, S. Pati, M. Bhalerao, S. P. Thakur, N. Jahani, V. Belenky, E. S. McDonald, J. Gibbs, D. C. Newitt, N. M. Hylton, et al. “Expert tumor annotations and radiomics for locally advanced breast cancer in DCE-MRI for ACRIN 6657/I-SPY1”. In: *Scientific data* 9.1 (2022), p. 440. DOI: 10.1038/s41597-022-01555-4 (cit. on pp. 27, 103).

[26] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al. “The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository”. In: *Journal of digital imaging* 26.6 (2013), pp. 1045–1057. DOI: 10.1007/s10278-013-9622-7 (cit. on pp. 12, 102).

[27] C. Davatzikos, S. Rathore, S. Bakas, S. Pati, M. Bergman, R. Kalarot, P. Sridharan, A. Gastounioti, N. Jahani, E. Cohen, et al. “Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome”. In: *Journal of medical imaging* 5.1 (2018), pp. 011018–011018. DOI: 10.1117/1.JMI.5.1.011018 (cit. on pp. 14, 16, 18, 27, 102).

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cit. on pp. 16, 104–106).

[29] B. Efron and R. Tibshirani. “Improvements on cross-validation: the 632+ bootstrap method”. In: *Journal of the American Statistical Association* 92.438 (1997), pp. 548–560. DOI: 10.1080/01621459.1997.10474007 (cit. on p. 104).

[30] T. Elsken, J. H. Metzen, F. Hutter, et al. “Neural architecture search: A survey.” In: *J. Mach. Learn. Res.* 20.55 (2019), pp. 1–21. DOI: 10.48550/arXiv.1808.05377 (cit. on p. 107).

[31] M. A. Flower. *Webb’s physics of medical imaging*. CRC press, 2012. DOI: 10.1201/b12218 (cit. on p. 7).

[32] P. Foley, M. J. Sheller, B. Edwards, S. Pati, W. Riviera, M. Sharma, P. N. Moorthy, S.-h. Wang, J. Martin, P. Mirhaji, et al. “OpenFL: the open federated learning library”. In: *Physics in Medicine & Biology* 67.21 (2022), p. 214001. DOI: 10.1088/1361-6560/ac97d9 (cit. on pp. 108, 109).

[33] H. Fu, F. Li, J. I. Orlando, H. Bogunovic, X. Sun, J. Liao, Y. XU, S. ZHANG, and X. ZHANG. "PALM: Pathologic myopia challenge". In: *Proc. IEEE Dataport*. 2019, p. 1. DOI: 10.21227/55pk-8z03 (cit. on p. 103).

[34] W. Fu and T. Menzies. "Easy over hard: A case study on deep learning". In: *Proceedings of the 2017 11th joint meeting on foundations of software engineering*. 2017, pp. 49–60. DOI: 10.1145/3106237.3106256 (cit. on p. 9).

[35] Y. Fu, N. M. Brown, S. U. Saeed, A. Casamitjana, Z. Baum, R. Delaunay, Q. Yang, A. Grimwood, Z. Min, S. B. Blumberg, et al. "DeepReg: a deep learning toolkit for medical image registration". In: *arXiv preprint arXiv:2011.02580* (2020). DOI: 10.21105/joss.02705 (cit. on p. 18).

[36] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. "A survey on deep learning techniques for image and video semantic segmentation". In: *Applied Soft Computing* 70 (2018), pp. 41–65. DOI: 10.1016/j.asoc.2018.05.018 (cit. on p. 9).

[37] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu. "An artificial agent for anatomical landmark detection in medical images". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 229–237. DOI: 10.1007/978-3-319-46726-9_27 (cit. on p. 9).

[38] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, et al. "NiftyNet: a deep-learning platform for medical imaging". In: *Computer methods and programs in biomedicine* 158 (2018), pp. 113–122. DOI: 10.1016/j.cmpb.2018.01.025 (cit. on pp. 16, 18).

[39] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016 (cit. on p. 16).

[40] Y. Gorbachev, M. Fedorov, I. Slavutin, A. Tugarev, M. Fatekhov, and Y. Tarkan. *Openvino deep learning workbench: Comprehensive analysis and tuning of neural networks inference*. 2019. DOI: 10.1109/ICCVW.2019.00104 (cit. on pp. 104, 108, 109).

[41] K. Gotkowski, C. Gonzalez, A. Bucher, and A. Mukhopadhyay. "M3d-CAM". In: *Bildverarbeitung für die Medizin 2021*. Wiesbaden: Springer Fachmedien Wiesbaden, 2021, pp. 217–222. DOI: 10.1007/978-3-658-33198-6_52 (cit. on pp. 9, 104).

[42] O. Güley, S. Pati, and S. Bakas. "Classification of infection and ischemia in diabetic foot ulcers using vgg architectures". In: *Diabetic foot ulcers grand challenge*. Springer, 2021, pp. 76–89. DOI: 10.1007/978-3-030-94907-5_6 (cit. on pp. 27, 103).

[43] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90 (cit. on p. 105).

[44] T. Hernandez-Boussard, S. Bozkurt, J. P. Ioannidis, and N. H. Shah. "MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care". In: *Journal of the American Medical Informatics Association* 27.12 (2020), pp. 2011–2015. DOI: 10.1093/jamia/ocaa088 (cit. on p. 106).

[45] N. M. Hylton, C. A. Gatsonis, M. A. Rosen, C. D. Lehman, D. C. Newitt, S. C. Partridge, W. K. Bernreuter, E. D. Pisano, E. A. Morris, P. T. Weatherall, et al. "Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival—results from the ACRIN 6657/CALGB 150007 I-SPY 1 TRIAL". In: *Radiology* 279.1 (2016), pp. 44–55. DOI: 10.1148/radiol.2015150013 (cit. on p. 103).

[46] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* (2020), pp. 1–9. DOI: 10.1038/s41592-020-01008-z (cit. on p. 18).

[47] C. C. Jaffe. "Imaging and Genomics: Is There a Synergy?" In: *Radiology* 264.2 (2012). PMID: 22821693, pp. 329–331. DOI: 10.1148/radiol.12120871. eprint: https://doi.org/10.1148/radiol.12120871 (cit. on p. 8).

[48] A. Jungo, O. Scheidegger, M. Reyes, and F. Balsiger. "pymia: A Python package for data handling and evaluation in deep learning-based medical image analysis". In: *Computer methods and programs in biomedicine* 198 (2021), p. 105796. DOI: 10.1016/j.cmpb.2020.105796 (cit. on p. 18).

[49] T. G. Kannampallil, G. F. Schauer, T. Cohen, and V. L. Patel. "Considering complexity in healthcare systems". In: *Journal of biomedical informatics* 44.6 (2011), pp. 943–947. DOI: 10.1016/j.jbi.2011.06.006 (cit. on p. 7).

[50] A. Karargyris, R. Umeton, M. J. Sheller, A. Aristizabal, J. George, A. Wuest, S. Pati, H. Kassem, M. Zenk, U. Baid, et al. "Federated benchmarking of medical artificial intelligence with MedPerf". In: *Nature Machine Intelligence* 5.7 (2023), pp. 799–810. DOI: 10.1038/s42256-023-00652-2 (cit. on pp. 27, 108, 109).

[51] R. Kemker, C. Salvaggio, and C. Kanan. "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning". In: *ISPRS journal of photogrammetry and remote sensing* 145 (2018), pp. 60–77. DOI: 10.1016/j.isprsjprs.2018.04.014 (cit. on p. 9).

[52] R. Kikinis, S. D. Pieper, and K. G. Vosburgh. "3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support". In: *Intraoperative imaging and image-guided therapy*. Springer, 2014, pp. 277–289. DOI: 10.1007/978-1-4614-7657-3_19 (cit. on pp. 13, 18).

[53] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems, et al. "Radiomics: the bridge between medical imaging and personalized medicine". In: *Nature reviews Clinical oncology* 14.12 (2017), pp. 749–762. DOI: 10.1038/nrclinonc.2017.141 (cit. on p. 106).

[54] B. A. Landman and S. K. Warfield. *MICCAI 2012: Workshop on Multi-atlas Labeling*. CreateSpace Independent Publishing Platform, 2012, pp. 1–164 (cit. on p. 103).

[55] F. Lateef and Y. Ruichek. "Survey on semantic segmentation using deep learning techniques". In: *Neurocomputing* 338 (2019), pp. 321–348. DOI: 10.1016/j.neucom.2019.02.003 (cit. on p. 9).

[56] F. Q. Lauzon. "An introduction to deep learning". In: *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. IEEE. 2012, pp. 1438–1439. DOI: 10.1109/ISSPA.2012.6310529 (cit. on p. 9).

[57] H. Li and Y. Fan. “Non-rigid image registration using self-supervised fully convolutional networks without training data”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 1075–1078. DOI: 10.1109/ISBI.2018.8363757 (cit. on p. 9).

[58] J. Li, S. Yang, X. Huang, Q. Da, X. Yang, Z. Hu, Q. Duan, C. Wang, and H. Li. “Signet ring cell detection with a semi-supervised learning framework”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 842–854. DOI: 10.1007/978-3-030-20351-1_66 (cit. on p. 103).

[59] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, and A. K. Denniston. “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension”. In: *bmj* 370 (2020). DOI: 10.1038/s41591-020-1034-x (cit. on p. 106).

[60] O. H. Maghsoudi, A. Gastounioti, L. Pantalone, C. Davatzikos, S. Bakas, and D. Kontos. “O-Net: An Overall Convolutional Network for Segmentation Tasks”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2020, pp. 199–209. DOI: 10.1007/978-3-030-59861-7_21 (cit. on p. 9).

[61] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. Cook. “Introduction to radiomics”. In: *Journal of Nuclear Medicine* 61.4 (2020), pp. 488–495. DOI: 10.2967/jnumed.118.222893 (cit. on p. 9).

[62] C. McCague, S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, and R. Woitek. “Introduction to radiomics for a clinical audience”. In: *Clinical Radiology* 78.2 (2023), pp. 83–98. DOI: 10.1016/j.crad.2022.08.149 (cit. on p. 9).

[63] M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, and M. Ghassemi. “Reproducibility in machine learning for health research: Still a ways to go”. In: *Science Translational Medicine* 13.586 (2021), eabb1655. DOI: 10.1126/scitranslmed.abb1655 (cit. on p. 10).

[64] M. McNitt-Gray, S. Napel, A. Jaggi, S. Mattonen, L. Hadjiiski, M. Muzi, D. Goldgof, Y. Balagurunathan, L. Pierce, P. Kinahan, et al. “Standardization in quantitative imaging: a multicenter comparison of radiomic features from different software packages on digital reference objects and patient data sets”. In: *Tomography* 6.2 (2020), pp. 118–128. DOI: 10.18383/j.tom.2019.00031 (cit. on p. 14).

[65] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. “The multimodal brain tumor image segmentation benchmark (BRATS)”. In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024. DOI: 10.1109/TMI.2014.2377694 (cit. on pp. 9, 16, 103, 107).

[66] J. Mongan, L. Moy, and C. E. Kahn Jr. “Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers”. In: *Radiology. Artificial Intelligence* 2.2 (2020). DOI: 10.1148/ryai.2020200029 (cit. on p. 106).

[67] B. A. Mosqueda González, O. Hasan, W. Uriawan, Y. Badr, and L. Brunie. “Secure and efficient decentralized machine learning through group-based model aggregation”. In: *Cluster Computing* (2023), pp. 1–15. DOI: 10.1007/s10586-023-04174-9 (cit. on p. 10).

[68] D. Newitt, N. Hylton, et al. “Multi-center breast DCE-MRI data and segmentations from patients in the I-SPY 1/ACRIN 6657 trials”. In: *Cancer Imaging Arch; https://doi.org/10.7937/K9/TCIA.20* (2016). DOI: 10.7937/K9/TCIA.2016.HdHpgJLK (cit. on p. 103).

[69] B. Norgeot, G. Quer, B. K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I. S. Kohane, S. Saria, E. Topol, et al. “Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist”. In: *Nature medicine* 26.9 (2020), pp. 1320–1324. DOI: 10.1038/s41591-020-1041-y (cit. on p. 106).

[70] O. Oktay, J. Nanavati, A. Schwaighofer, D. Carter, M. Bristow, R. Tanno, R. Jena, G. Barnett, D. Noble, Y. Rimmer, et al. “Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers”. In: *JAMA network open* 3.11 (2020), e2027426–e2027426. DOI: 10.1001/jamanetworkopen.2020.27426 (cit. on p. 18).

[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems*. 2019, pp. 8026–8037. DOI: 10.48550/arXiv.1912.01703 (cit. on pp. 16, 18).

[72] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos, et al. “Federated learning enables big data for rare cancer boundary detection”. In: *Nature communications* 13.1 (2022), p. 7346. DOI: 10.1038/s41467-022-33407-5 (cit. on p. 11).

[73] S. Pati, U. Baid, B. Edwards, M. J. Sheller, P. Foley, G. A. Reina, S. Thakur, C. Sako, M. Bilello, C. Davatzikos, et al. “The federated tumor segmentation (FeTS) tool: an open-source solution to further solid tumor research”. In: *Physics in Medicine & Biology* 67.20 (2022), p. 204002. DOI: 10.1088/1361-6560/ac9449 (cit. on pp. 18, 107, 108).

[74] S. Pati, U. Baid, M. Zenk, B. Edwards, M. Sheller, G. A. Reina, P. Foley, A. Gruzdev, J. Martin, S. Albarqouni, et al. “The federated tumor segmentation (fets) challenge”. In: *arXiv preprint arXiv:2105.05874* (2021). DOI: 10.48550/arXiv.2105.05874 (cit. on p. 16).

[75] S. Pati, V. Sharma, H. Aslam, S. P. Thakur, H. Akbari, A. Mang, S. Subramanian, G. Biros, C. Davatzikos, and S. Bakas. “Estimating Glioblastoma Biophysical Growth Parameters Using Deep Learning Regression”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*. Springer. 2021, pp. 157–167. DOI: 10.1007/978-3-030-72084-1_15 (cit. on p. 27).

[76] S. Pati, A. Singh, S. Rathore, A. Gastounioti, M. Bergman, P. Ngo, S. M. Ha, D. Bounias, J. Minock, G. Murphy, et al. “The cancer imaging phenomics toolkit (CaPTk): technical overview”. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5*. Springer. 2020, pp. 380–394. DOI: 10.1007/978-3-030-46643-5_38 (cit. on pp. 14, 16, 18, 27, 102).

[77] S. Pati, S. P. Thakur, İ. E. Hamamcı, U. Baid, B. Baheti, M. Bhalerao, O. Güley, S. Mouchtaris, D. Lang, S. Thermos, et al. “GaNDLF: the generally nuanced deep learning framework for scalable end-to-end clinical workflows”. In: *Communications Engineering* 2.1 (2023), p. 23. DOI: 10.1038/s44172-023-00066-3 (cit. on pp. 23, 27, 103, 104, 107, 108).

[78] S. Pati, R. Verma, H. Akbari, M. Bilello, V. B. Hill, C. Sako, R. Correa, N. Beig, L. Venet, S. Thakur, et al. “Reproducibility analysis of multi-institutional paired expert annotations and radiomic features of the Ivy Glioblastoma Atlas Project (Ivy GAP) dataset”. In: *Medical physics* 47.12 (2020), pp. 6039–6052. DOI: 10.1002/mp.14556 (cit. on pp. 14, 27, 101, 102).

[79] S. Pati, R. Verma, H. Akbari, M. Bilello, V. B. Hill, C. Sako, R. Correa, N. Beig, L. Venet, S. Thakur, P. Serai, S. M. Ha, T. Shinohara, P. Tiwari, and S. Bakas. “Multi-Institutional Paired Expert Segmentations and Radiomic Features of the Ivy GAP Dataset”. In: *The Cancer Imaging Archive* (2020). DOI: 10.7937/9j41-7d44 (cit. on p. 102).

[80] N. Pawlowski, S. I. Ktena, M. C. Lee, B. Kainz, D. Rueckert, B. Glocker, and M. Rajchl. “DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images”. In: *arXiv preprint arXiv:1711.06853* (2017). DOI: 10.48550/arXiv.1711.06853 (cit. on p. 18).

[81] V. Piratla. “Robustness, Evaluation and Adaptation of Machine Learning Models in the Wild”. In: *arXiv preprint arXiv:2303.02781* (2023). DOI: 10.48550/arXiv.2303.02781 (cit. on p. 11).

[82] J. Pocock, S. Graham, Q. D. Vu, M. Jahanifar, S. Deshpande, G. Hadjigeorghiou, A. Shephard, R. M. S. Bashir, M. Bilal, W. Lu, et al. “TIAToolbox: An End-to-End Toolbox for Advanced Tissue Image Analytics”. In: *bioRxiv* (2021). DOI: 10.1101/2021.12.23.474029 (cit. on p. 19).

[83] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M. P. Reyes, M.-L. Shyu, S.-C. Chen, and S. Iyengar. “A survey on deep learning: Algorithms, techniques, and applications”. In: *ACM Computing Surveys (CSUR)* 51.5 (2018), pp. 1–36. DOI: 10.1145/3234150 (cit. on p. 9).

[84] P. Prasanna, P. Tiwari, and A. Madabhushi. “Co-occurrence of Local Anisotropic Gradient Orientations (CoLlAGe): distinguishing tumor confounders and molecular subtypes on MRI”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2014, pp. 73–80. DOI: 10.1007/978-3-319-10443-0_10 (cit. on pp. 15, 103).

[85] R. B. Puchalski, N. Shah, J. Miller, R. Dalley, S. R. Nomura, J.-G. Yoon, K. A. Smith, M. Lankerovich, D. Bertagnolli, K. Bickley, et al. “An anatomic transcriptional atlas of human glioblastoma”. In: *Science* 360.6389 (2018), pp. 660–663. DOI: 10.1126/science.aaf2666 (cit. on p. 102).

[86] Y. Qin, L.-H. Zhu, W. Zhao, J.-J. Wang, and H. Wang. “Review of radiomics-and dosiomics-based predicting models for rectal cancer”. In: *Frontiers in Oncology* 12 (2022), p. 913683. DOI: 10.3389/fonc.2022.913683 (cit. on p. 9).

[87] S. Ranjbar and J. R. Mitchell. “An introduction to radiomics: an evolving cornerstone of precision medicine”. In: *Biomedical Texture Analysis*. Elsevier, 2017, pp. 223–245. DOI: 10.1016/B978-0-12-812133-7.00008-9 (cit. on p. 9).

[88] G. A. Reina, A. Gruzdev, P. Foley, O. Perepelkina, M. Sharma, I. Davidyuk, I. Trushkin, M. Radionov, A. Mokrov, D. Agapov, et al. “OpenFL: An open-source framework for Federated Learning”. In: *arXiv preprint arXiv:2105.06413* (2021). DOI: 10.48550/arXiv.2105.06413 (cit. on p. 109).

[89] A. Reinke, M. Eisenmann, M. D. Tizabi, C. H. Sudre, T. Rädsch, M. Antonelli, T. Arbel, S. Bakas, M. J. Cardoso, V. Cheplygina, et al. “Common limitations of image processing metrics: A picture story”. In: *arXiv preprint arXiv:2104.05642* (2021). DOI: 10.48550/arXiv.2104.05642 (cit. on p. 22).

[90] A. Reinke and H. Müller. “Metrics reloaded: A new recommendation framework for biomedical image analysis validation”. In: *Proceedings of the Medical Imaging with Deep Learning (MIDL 2022)* (2022) (cit. on pp. 22, 24).

[91] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi. “Radiomics: the facts and the challenges of image analysis”. In: *European radiology experimental* 2.1 (2018), pp. 1–8. DOI: 10.1186/s41747-018-0068-z (cit. on p. 9).

[92] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2005. DOI: 10.1007/978-3-642-02431-3 (cit. on p. 23).

[93] O. Ronneberger, P. Fischer, and T. Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28 (cit. on p. 105).

[94] J. Rosenthal, R. Carelli, M. Omar, D. Brundage, E. Halbert, J. Nyman, S. N. Hari, E. M. Van Allen, L. Marchionni, R. Umeton, et al. “Building tools for machine learning and artificial intelligence in cancer research: best practices and a case study with the PathML toolkit for computational pathology”. In: *Molecular Cancer Research* 20.2 (2022), pp. 202–206. DOI: 10.1158/1541-7786.MCR-21-0665 (cit. on p. 18).

[95] M. Rozycki, T. D. Satterthwaite, N. Koutsouleris, G. Erus, J. Doshi, D. H. Wolf, Y. Fan, R. E. Gur, R. C. Gur, E. M. Meisenzahl, et al. “Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals”. In: *Schizophrenia bulletin* 44.5 (2018), pp. 1035–1044. DOI: 10.1093/schbul/sbx137 (cit. on p. 103).

[96] J. D. Rudie, D. A. Weiss, R. Saluja, A. M. Rauschecker, J. Wang, L. Sugrue, S. Bakas, and J. B. Colby. “Multi-Disease Segmentation of Gliomas and White Matter Hyperintensities in the BraTS Data Using a 3D Convolutional Neural Network”. In: *Frontiers in Computational Neuroscience* 13 (2019), p. 84. DOI: 10.3389/fncom.2019.00084 (cit. on p. 9).

[97] L. Scarpace, T. Mikkelsen, S. Cha, S. Rao, S. Tekchandani, D. Gutman, J. Saltz, B. Erickson, N. Pedano, A. Flanders, et al. “The Cancer Genome Atlas Glioblastoma Multiforme Collection (TCGA-GBM)(Version 4)[Data set]”. In: *Cancer Imaging Arch. Published online* (2016). DOI: 10.7937/K9/TCIA.2016.RNYFUYE9 (cit. on p. 102).

[98] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen, et al. “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data”. In: *Scientific reports* 10.1 (2020), pp. 1–12. DOI: 10.1038/s41598-020-69250-1 (cit. on p. 9).

[99] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas. “Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation”. In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 92–104. DOI: 10.1007/978-3-030-11723-8_9 (cit. on p. 9).

[100] V. Sounderajah, H. Ashrafian, R. M. Golub, S. Shetty, J. De Fauw, L. Hooft, K. Moons, G. Collins, D. Moher, P. M. Bossuyt, et al. “Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol”. In: *BMJ open* 11.6 (2021), e047709. DOI: 10.1136/bmjopen-2020-047709 (cit. on p. 106).

[101] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *Plos med* 12.3 (2015), e1001779. DOI: 10.1371/journal.pmed.1001779 (cit. on p. 103).

[102] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594 (cit. on p. 9).

[103] S. Thakur, J. Doshi, S. Pati, S. Rathore, C. Sako, M. Bilello, S. M. Ha, G. Shukla, A. Flanders, A. Kotrotsou, et al. “Brain extraction on MRI scans in presence of diffuse glioma: Multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training”. In: *NeuroImage* 220 (2020), p. 117081. DOI: 10.1016/j.neuroimage.2020.117081 (cit. on pp. 9, 18, 103).

[104] S. P. Thakur, S. Pati, R. Panchumarthy, D. Karkada, J. Wu, D. Kurtaev, C. Sako, P. Shah, and S. Bakas. “Optimization of Deep Learning Based Brain Extraction in MRI for Low Resource Environments”. In: *International MICCAI Brainlesion Workshop*. Springer. 2022, pp. 151–167. DOI: 10.1007/978-3-031-08999-2_12 (cit. on pp. 27, 104, 106).

[105] F. Tixier, H. Um, R. J. Young, and H. Veeraraghavan. “Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features”. In: *Medical Physics* 46.8 (2019), pp. 3582–3591. DOI: 10.1002/mp.13624 (cit. on p. 102).

[106] A. Tocchetti, L. Corti, A. Balayn, M. Yurrita, P. Lippmann, M. Brambilla, and J. Yang. “AI robustness: a human-centered perspective on technological challenges and opportunities”. In: *arXiv preprint arXiv:2210.08906* (2022). DOI: 10.48550/arXiv.2210.08906 (cit. on p. 11).

[107] P. Turney. “Bias and the quantification of stability”. In: *Machine Learning* 20 (1995), pp. 23–33. DOI: 10.1023/A:1022682001417 (cit. on p. 11).

[108] N. J. Tustison, P. A. Cook, A. J. Holbrook, H. J. Johnson, J. Muschelli, G. A. Devanyi, J. T. Duda, S. R. Das, N. C. Cullen, D. L. Gillen, et al. “ANTsX: A dynamic ecosystem for quantitative biological and medical imaging”. In: *medRxiv* (2020). DOI: 10.1101/2020.10.19.20215392 (cit. on p. 18).

[109] P. F. Uhlir and P. Schröder. “Open data for global science”. In: *Data Science Journal* 6 (2007), OD36–OD53. DOI: 10.2481/dsj.6.OD36 (cit. on p. 12).

[110] M. Vallières, C. R. Freeman, S. R. Skamene, and I. El Naqa. “A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities”. In: *Physics in Medicine & Biology* 60.14 (2015), p. 5471. DOI: 10.1088/0031-9155/60/14/5471 (cit. on p. 14).

[111] J. J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts. “Computational radiomics system to decode the radiographic phenotype”. In: *Cancer research* 77.21 (2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339 (cit. on p. 14).

[112] G. Varoquaux and V. Cheplygina. “Machine learning for medical imaging: methodological failures and recommendations for the future”. In: *NPJ digital medicine* 5.1 (2022), pp. 1–8. DOI: 10.1038/s41746-022-00592-y (cit. on p. 16).

[113] P. Voigt and A. Von dem Bussche. “The eu general data protection regulation (gdpr)”. In: *A Practical Guide, 1st Ed., Cham: Springer International Publishing* (2017). DOI: 10.1007/978-3-319-57959-7 (cit. on pp. 10, 12).

[114] J. Waring, C. Lindvall, and R. Umeton. “Automated machine learning: Review of the state-of-the-art and opportunities for healthcare”. In: *Artificial Intelligence in Medicine* 104 (2020), p. 101822. DOI: 10.1016/j.artmed.2020.101822 (cit. on p. 107).

[115] I. Wolf, M. Vetter, I. Wegner, T. Böttger, M. Nolden, M. Schöbinger, M. Hastenteufel, T. Kunert, and H.-P. Meinzer. “The medical imaging interaction toolkit”. In: *Medical image analysis* 9.6 (2005), pp. 594–604. DOI: 10.1016/j.media.2005.04.005 (cit. on pp. 13, 18).

[116] T. Xiao, W. Hua, C. Li, and S. Wang. “Glioma Grading Prediction by Exploring Radiomics and Deep Learning Features”. In: *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*. 2019, pp. 208–213. DOI: 10.1145/3364836.3364877 (cit. on p. 8).

[117] J. Yang, X. Huang, Y. He, J. Xu, C. Yang, G. Xu, and B. Ni. “Reinventing 2d convolutions for 3d images”. In: *IEEE Journal of Biomedical and Health Informatics* 25.8 (2021), pp. 3009–3018. DOI: 10.1109/JBHI.2021.3049452 (cit. on pp. 16, 19, 104).

[118] M. H. Yap, B. Cassidy, J. M. Pappachan, C. O’Shea, D. Gillespie, and N. Reeves. “Analysis Towards Classification of Infection and Ischaemia of Diabetic Foot Ulcers”. In: *arXiv preprint arXiv:2104.03068* (2021). DOI: 10.1109/BHI50953.2021.9508563 (cit. on p. 103).

[119] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. “How transferable are features in deep neural networks?” In: *Advances in neural information processing systems* 27 (2014) (cit. on p. 16).

[120] P. A. Yushkevich, Y. Gao, and G. Gerig. “ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2016, pp. 3342–3345. DOI: 10.1109/EMBC.2016.7591443 (cit. on p. 18).

[121] J. Zhang, M. Liu, and D. Shen. “Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks”. In: *IEEE Transactions on Image Processing* 26.10 (2017), pp. 4753–4764. DOI: 10.1109/TIP.2017.2721106 (cit. on p. 9).

[122] S. K. Zhou, D. Rueckert, and G. Fichtinger. *Handbook of medical image computing and computer assisted intervention*. Academic Press, 2019. DOI: 10.1016/C2017-0-04608-6 (cit. on p. 7).

[123] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. “Morphometric analysis of white matter lesions in MR images: method and validation”. In: *IEEE transactions on medical imaging* 13.4 (1994), pp. 716–724. DOI: 10.1109/42.363096 (cit. on p. 22).

[124] A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E. G. Troost, C. Richter, and S. Löck. “Assessing robustness of radiomic features by image perturbation”. In: *Scientific reports* 9.1 (2019), p. 614. DOI: 10.1038/s41598-018-36938-4 (cit. on p. 14).

[125] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, et al. “The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping”. In: *Radiology* 295.2 (2020), pp. 328–338. DOI: 10.1148/radiol.2020191145 (cit. on pp. 9, 14, 15, 106).

List of Figures

List of Tables