Dataset Size Sensitivity Analysis of Machine Learning Classifiers to Differentiate Molecular Markers of Pediatric Low-Grade Gliomas Based on MRI

Matthias W. Wagner (m.w.wagner@me.com)
The Hospital for Sick Children, University of Toronto

Khashayar Namdar
The Hospital for Sick Children, University of Toronto

Abdullah Alqabbani
The Hospital for Sick Children, University of Toronto

Nicolin Hainc
The Hospital for Sick Children, University of Toronto

Liana Nobre Figuereido
The Hospital for Sick Children, University of Toronto

Min Sheng
The Hospital for Sick Children, University of Toronto

Manohar M Shroff
The Hospital for Sick Children, University of Toronto

Eric Bouffet
The Hospital for Sick Children, University of Toronto

Uri Tabori
The Hospital for Sick Children, University of Toronto

Cynthia Hawkins
The Hospital for Sick Children, University of Toronto

Michael Zhang
Lucile Packard Children's Hospital

Kristen W. Yeom
Lucile Packard Children's Hospital

Farzad Khalvati
The Hospital for Sick Children, University of Toronto

Birgit B. Ertl-Wagner
The Hospital for Sick Children, University of Toronto

Research Article

Keywords: Random Forest, Neural Net, XGBoost, Radiomics, Glioma, Low Grade, BRAF, Children

Posted Date: September 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-883606/v1

Abstract

Machine learning (ML) approaches can predict BRAF status of pediatric low-grade gliomas (pLGG) on pre-therapeutic brain MRI. The impact of training data sample size and type of ML model is not established. In this bi-institutional retrospective study, 251 pLGG FLAIR MRI datasets from 2 children's hospitals were included. Radiomics features were extracted from tumor segmentations and five models (Random Forest, XGBoost, Neural Network (NN) 1 (100:20:2), NN2 (50:10:2), NN3 (50:20:10:2)) were tested to classify them. Classifiers were cross-validated on data from institution 1 and validated on data from institution 2. Starting with 10% of the training data, models were cross-validated using a 4-fold approach at every step with an additional 2.25% increase in sample size. Two-hundred-twenty patients (mean age 8.53 ± 4.94 years, 114 males, 67% BRAF fusion) were included in the training dataset, and 31 patients (mean age 7.97±6.20 years, 18 males, 77% BRAF fusion) in the independent test dataset. NN1 (100:20:2) yielded the highest area under the receiver operating characteristic curve (AUC). It predicted BRAF status with a mean AUC of 0.85, 95% CI [0.83, 0.87] using 60% of the training data and with mean AUC of 0.83, 95% CI [0.82, 0.84] on the independent validation data set.

Introduction

Pediatric low-grade gliomas (pLGG) comprise a heterogeneous variety of tumors classified by the World Health Organization as grades I or II (1, 2). They are the most common brain tumors in children, accounting for approximately 40% of tumors of the central nervous system (CNS) in childhood (3). If total resection is not possible, pLGG become a chronic disease with protracted reduction in quality of life (1, 4) with a 10-year progression-free survival (PFS) of less than 50% (5, 6). Molecular characterization of sporadic pLGG has identified frequent alterations in the RAS-MAPK pathway, most commonly fusions or mutations in the BRAF gene (7, 8). Lassaletta et al. recently showed that patient prognosis differs based on the underlying molecular alteration: pLGG with BRAF fusion have a favorable outcome, while those with BRAF V600E mutation are at increased risk of progression and transformation (9, 10). This has led to clinical trials using RAS-MAPK pathway targeted agents such as MEK inhibitors and BRAF V600E inhibitors for patients with molecular evidence of BRAF alterations. These new therapies are promising and many pLGG that were refractory to traditional chemotherapy have had significant responses (11, 12).

In the past decade, radiomics has emerged as an imaging-based method to link quantitative features extracted from medical images to outcomes, such as cancer genotype or survival (13, 14). Radiomic signatures have been extensively investigated for different cancer sites including liver cancer (15), bone tumors (16), glioblastoma (17), medulloblastoma (21), and midline high-grade glioma (18, 19). Recently, we applied Random Forest (RF) to differentiate BRAF fused from BRAF V600E mutated pLGG and yielded an area under the receiver operating characteristic curve (AUC) of 0.85 on an independent validation set (20). It has not been well established to what extent different classification models and the size of the training data affect diagnostic performance. This may also serve as a model for classification algorithms in other tumors.

We therefore aimed to assess the performance of five commonly used machine learning (ML) models to predict BRAF fusion or BRAF V600E mutation on an independent validation set with systematic step-wise increase of training data.

Material And Methods

Patients: This retrospective study was approved by the institutional review board or research ethics board of the two participating academic institutions: The Hospital for Sick Children (Toronto, Ontario, Canada) and The Lucile Packard Children's Hospital (Stanford University, Palo Alto, California). This study was performed in accordance with the relevant guidelines and regulations. Informed consent was waived by the local institutional review or research ethics boards due to the retrospective nature of the study. An inter-institutional data transfer agreement was obtained for data-sharing. Patients were identified from the electronic health record data bases at Toronto from January 2000 to December 2018 and at Stanford from January 2009 to January 2016. Patient inclusion criteria were: 1) age 0–18 years, 2) availability of molecular information on BRAF status in histopathologically confirmed pLGG, and 3) availability of preoperative brain MRI with a non-motion degraded FLAIR sequence. Patients with histone H3 K27M mutation and neurofibromatosis 1 were excluded. Spinal cord tumors were also excluded.

The datasets of 94 patients from The Hospital for Sick Children, Toronto, and 21 patients from The Lucile Packard Children's Hospital, Stanford, used in this study have been previously published (20). The previous study applied an RF model without variations in sample size to differentiate BRAF fused from BRAF V600E mutated pLGG. Our current study investigates the performance of five commonly used ML models and various sample sizes to predict BRAF fusion or BRAF V600E mutation on an independent validation set using a systematic step-wise increase of training data.

Molecular Analysis: BRAF fusion status was determined using a nanoString panel or fluorescence in situ hybridisation (FISH) while BRAF p.V600E mutation was determined using immunohistochemistry or droplet digital PCR as previously described (21).

MRI Acquisition, Data Retrieval, Image Segmentation: All patients from The Hospital for Sick Children, Toronto, underwent brain MRI at 1.5T or 3T across various vendors (Signa, GE Healthcare; Achieva, Philips Healthcare; Magnetom Skyra, Siemens Healthineers). Sequences were acquired according to the institutional tumor protocol and included a 2D axial T2 FLAIR sequence (TR/TE, 7000–10000/140–170 ms; 3–6 mm slice thickness; 3-7.5 mm gap). Patients from Lucile Packard Children's Hospital, Stanford, underwent brain MR imaging at 1.5T or 3T scanners from a single vendor (Signa or Discovery 750; GE Healthcare, Milwaukee, Wisconsin). Sequences were acquired using the institutional brain tumor protocol, which included a 2D axial T2 FLAIR sequence (TR/TE, 7000–10000/140–170 ms; 4–5 mm slice thickness; 1-1.5 mm gap). All MRI data were extracted from the respective PACS and were de-identified for further analyses. Tumor segmentation was performed by a 4th year radiology resident with neuroradiology research experience (AA) using 3D Slicer (ver. 4.10.2) (22) (http://www.slicer.org). The scripted loadable module SlicerRadiomics extension was used to obtain access to the radiomics feature

calculation classes implemented in the pyradiomics library (http://pyradiomics.readthedocs.io/). This extension offers to select all available feature classes and ensures isotropic resampling under “Resampled voxel size” when extracting 3D features. Semi-automated tumor segmentation on FLAIR images was performed with the Level-Tracing-Effect tool. This semi-automatic approach had been found superior to multi-user manual delineation with regard to reproducibility and robustness of results (23). The final and proper placement of ROIs was confirmed by a pediatric neuroradiology trained and board-certified radiologist (MWW, 7 years of neuroradiology research experience).

Radiomic Feature-Extraction Methodology: A total of 851 MRI-based radiomic features were extracted from the ROIs on FLAIR images. Radiomic features included histogram, shape, and texture features with and without wavelet-based filters. Features of Laplacian of Gaussian filters were not extracted. All features are summarized in Supplementable Table. Bias field correction prior to z-score normalization were used to standardize the range of all image features (24, 25). Once the features were extracted, we applied z-score normalization again followed by L2 normalization to the features of cohort 1 and used the distribution of the features in cohort 1 (training data) to normalize cohort 2 (validation data). Details of pre-processing and radiomic feature extraction in 3D Slicer have been described elsewhere (13, 17, 26).

Statistical and ML analysis: We used t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize our dataset. RF, XGBoost, NN1 (100:20:2), NN2 (50:10:2), NN3 (50:20:10:2) were utilized as classification models (27–29).

Data visualization

In dimensionality reduction, t-SNE can be applied to different types of data including radiomics features. It applies Principal Components Analysis to map data points from an original high dimensional space to an intermediate lower dimensional space (default dimensions = 30). Subsequently, pair-wise distances are calculated and probability distributions are fit to examples so that data points in closer proximity are assigned with higher probabilities. Initialization for embeddings (i.e., representation of data points in latent space, usually a 2D or 3D space) is realized. The same procedure is repeated for the points in the latent space. Iteratively and applying a gradient descent approach, Kullback-Leibler divergences between the two sets of probability distributions are minimized. Ultimately, if two data points are similar in high dimensional space, their embeddings will be close to one another in 2D/3D space.

Random Forest

RF is a learning method consisting of several decision trees that can be used for classification and regression. Decision trees are multi-step thresholders, which can overfit to any data, if there is no controlling mechanism such as maximum depth. In order to enhance their generalization capacity, RF ensemble decision trees. Similar to any other tree-based algorithm, RF are suitable for classification of tabular data which makes them a high potential option for radiomics pipelines. The most critical hyperparameters of RF are number of trees (number of estimators), maximum depth of each tree, and the minimum and maximum number of examples at leaf nodes.

XGBoost

eXtreme Gradient Boost (XGBoost) is a popular gradient boosted trees (GBT) algorithm. Similar to RF, GBTs ensemble a collection of decision trees. However, GBTs add trees in a sequential manner such that errors of the previous tree are revised by the next tree. Compared to other GBTs, XGBoost utilizes a customized loss function and implements multiple regularization techniques to enhance the model's generalization and computational efficiency. Learning rate of the tree booster (default value is 0.3) and maximum depth of trees are the most important hyperparameters of the algorithm.

Neural Networks

Neural Networks (NNs) are highly nonlinear classifiers that were initially built based on ensembling of perceptron blocks. To date, there are multiple well-established categories of NNs including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks. Each of these categories is suitable for specific sets of problems. For example, CNNs have a high potential for image and video processing tasks such as object detection or segmentation. Given the type of data, we used feedforward NNs in this study. Feedforward NNs are considered a conventional type of NN, where individual perceptrons form layers and a stack of layers creates the architecture without recurrent paths in the network. We designed an initial NN (NN1) and derived two other architectures (NN2 and NN3, respectively) by changing its width and depth. Figure 1 illustrates the three architectures, NN1, NN2, and NN3. Rectified linear unit was used as activation function of the linear layers. In order to enhance generalization of the models, we implemented dropout layers in our architectures. The dropout mechanism arbitrarily excludes some nodes of the network from the weight updating process during training.

Internal Cross Validation

Starting with 10% of the training data, all models were cross-validated using a 4-fold approach with a systematic step wise 2.25% increase in sample size. At each step, experiments were repeated 10 times using randomized versions of the respective percentage of the training data, resulting in 10 classifiers per step.

External Validation

At each step, the 10 classifiers were validated on the entire independent external data set.

Classification performance metrics

Mean AUC and 95% confidence intervals (CI) were calculated for every step for both training and validation data sets, and the process was repeated for all five models. The external validation data set was never used in any stage of the training of the models and was dedicated to external validation. To examine whether the difference between performance of the models were significant, we conducted a

two-sided two-sample Kolmogorov-Smirnov (KS) test on mean AUCs across training sample sizes for each pair of our models.

Results

Patients: A total of 251 children (132 males (53%), mean age 8.5 years, standard deviation (SD) 5.1 years) were included. The internal cohort consisted of 220 patients (114 males (52%), mean age 8.5 years, SD 4.9 years) from The Hospital for Sick Children. The external cohort consisted of 31 patients (18 males (58%), mean age 8.0 years, SD 6.2 years) from The Lucile Packard Children's Hospital, were analyzed. BRAF fusion was found in 172 of 251 patients (69%), in 148 of 220 patients from the Toronto cohort (67%), and in 24 of 31 patients from the Stanford cohort (77%). Patient demographic information and pathologic information including age at diagnosis, sex, histologic diagnosis, and molecular diagnosis regarding BRAF status are provided in Table 1.

Table 1
Patient demographics

		Institutional Cohort	
		Toronto	Stanford
No. of patients		220	31
Age (mean) (yr)		8.53	7.97
Male sex (No.) (%)		114 (52)	18 (58)
Histological diagnosis (No.)			
	JPA	122	21
	LGA	32	-
	GG	30	7
	DA	12	-
	PMA	9	3
	PXA	6	-
	ODG	2	-
	NC	2	-
	DNET	2	-
	GC	1	-
	GNT	1	-
	Mixed	1	-
Molecular subgroup (No.) (%)			
	BRAF fusion	148 (67)	24 (77)
	BRAF mutation	72 (33)	7 (23)

Table 1
Patient demographics. JPA = Juvenile Pilocytic Astrocytoma, LGA = Low Grade Astrocytoma, GG = Ganglioglioma, DA = Diffuse Astrocytoma, PMA = Pilomyxoid Astrocytoma, PXA = Pleomorphic Xanthoastrocytoma, ODG = Oligodendroglioma, NC = Neurocytoma, DNET = Dysembryoplastic Neuroepithelial Tumor, GC = Gangliocytoma, GNT = Glioneuronal tumor, Mixed = Mixed histology

Table 2. Comparison of model performance

Internal						
		NN1	NN2	NN3	XGBoost	RF
	NN1		0.42	< 0.0001	< 0.0001	< 0.0001
	NN2	0.42		0.016	< 0.0001	< 0.0001
	NN3	< 0.0001	0.016		< 0.0001	< 0.0001
	XGBoost	< 0.0001	< 0.0001	< 0.0001		0.99
	RF	< 0.0001	< 0.0001	< 0.0001	0.99	
External						
		NN1	NN2	NN3	XGBoost	RF
	NN1		0.59	0.1	< 0.0001	< 0.0001
	NN2	0.59		0.1	< 0.0001	< 0.0001
	NN3	0.1	0.1		< 0.0001	< 0.0001
	XGBoost	< 0.0001	< 0.0001	< 0.0001		< 0.0001
	RF	< 0.0001	< 0.0001	< 0.0001	< 0.0001	

Data visualization: t-SNE was used for data visualization (Fig. 2). No apparent separation was found between the internal and external data set, neither for the whole data nor for the two classes. However, locality of BRAF V600E mutation examples suggests separability of the two classes.

Classification model evaluation and comparison of performance: AUC values over the entire training or external data were averaged for each model and performances were compared using two sided two-sample KS tests (Table 2). The performance of the five models over a 2.25% step wise increase of training data is shown in Fig. 3. All classifiers showed a decreasing variance on the internal data with a continued increase in sample size. While the performance on the internal data was significantly higher for RF and XGBoost (p-values: all < 0.0001 for RF vs NN1 / NN2 / NN3 and all < 0.0001 for XGBoost vs NN1 / NN2 / NN3), their performance was significantly lower and their variation was higher on the external data compared to the three NNs (Fig. 4) (p-values: all $p < 0.0001$ for RF vs NN1 / NN2 / NN3 and all $p < 0.0001$ for XGBoost vs NN1 / NN2 / NN3). XGBoost had a significantly different performance compared to RF on the external data (p-value: <0.0001) while demonstrating a higher variance. On the internal data, NN1 and

NN2 showed similar performance, but they were significantly different from NN3 (p-value: <0.0001 for NN1 vs NN3 and p-value: 0.016 for NN2 vs NN3). All three NNs demonstrated similar high performance and low variance on the external cohort up until 70% of the training data, where performance dropped to the level of XGBoost and their variance increased. Mean, upper and lower confidence interval sensitivity, specificity, accuracy, and F1 score across the entire training and external validation dataset are summarized in Table 3. At 60% of the training data (132 patients), NN1 and NN2 yielded the best results on the validation data (NN1 and NN2: mean AUC with [95% Confidence Interval]: 0.83 [0.82–0.84]). NN3 performed slightly below NN1 and NN2: 0.82 [0.81–0.84]. RF and XGBoost AUC values at 60% were 0.72 [0.7–0.74] and 0.75 [0.72–0.78], respectively. On the training data set, RF had the highest AUC at 60%: 0.87 [0.86–0.9]. XGBoost performed slightly below RF at 0.87 [0.85–0.89]. NN1 and NN2 were performing at the same level: 0.85 [0.83 0.87] and NN3 was slightly below: 0.84 [0.81 0.86].

Table 3
Mean performance metrics over entire training and validation data set

Metric	Model	Training data	Validation data
		Mean [CI]	Mean [CI]
Sensitivity	RF	0.85 [0.82–0.89]	0.71 [0.71–0.71]
Sensitivity	XGB	0.85 [0.81–0.88]	0.69 [0.63–0.74]
Sensitivity	NN1	0.82 [0.79–0.85]	0.71 [0.71–0.71]
Sensitivity	NN2	0.81 [0.77–0.84]	0.71 [0.71–0.71]
Sensitivity	NN3	0.81 [0.77–0.84]	0.73 [0.70–0.76]
Specificity	RF	0.87 [0.84–0.90]	0.82 [0.79–0.85]
Specificity	XGB	0.85 [0.82–0.88]	0.87 [0.84–0.90]
Specificity	NN1	0.85 [0.81–0.88]	0.92 [0.91–0.93]
Specificity	NN2	0.85 [0.82–0.89]	0.93 [0.92–0.94]
Specificity	NN3	0.85 [0.81–0.89]	0.91 [0.89–0.93]
Accuracy	RF	0.86 [0.84–0.88]	0.80 [0.77–0.82]
Accuracy	XGB	0.85 [0.83–0.87]	0.83 [0.81–0.85]
Accuracy	NN1	0.84 [0.82–0.86]	0.87 [0.87–0.88]
Accuracy	NN2	0.84 [0.82–0.86]	0.88 [0.87–0.89]
Accuracy	NN3	0.83 [0.81–0.86]	0.87 [0.86–0.88]
F1 Score	RF	0.81 [0.78–0.83]	0.62 [0.59–0.64]
F1 Score	XGB	0.79 [0.77–0.81]	0.65 [0.62–0.67]
F1 Score	NN1	0.78 [0.75–0.80]	0.72 [0.71–0.73]
F1 Score	NN2	0.77 [0.75–0.79]	0.73 [0.72–0.75]
F1 Score	NN3	0.77 [0.74–0.79]	0.72 [0.70–0.74]

Discussion

In this bi-institutional study, we assessed five commonly used ML classifiers to predict BRAF fusion or BRAF V600E mutation on independent data using a systematic step-wise increase of training data. Our results indicate that although classifier performance is generally high, certain classifiers can perform better than others. We found that NNs1-3 outperformed XGBoost and RF on the external data, while they demonstrated lower AUCs on the internal data. This was visible starting from > 20% of the training data or

42 patients. NNs1-3 achieved a high level of performance on the external data with only a limited amount of training data. This remained at a similar level with further increase of training data. At 70% of the training data, performance levels of NNs1-3 dropped to the level of XGBoost. This effect was attributed to overfitting. NN1 and 2 (100:20:2 and 50:10:2) yielded the best AUC on the external data at 60% of the training data (AUC: 0.83). Though differences in performance in NN1-3 were not statistically significant on the external data, we observed the least variation in performance with NN1.

RF has been a popular data mining and statistical tool in radiomics research due to its transparency and success in classification and regression tasks (20, 30–33). It generates a large number of decision trees using random subsamples of the training data while also randomly varying the features used in the trees (34). GBT differ from RF in that they add decision trees sequentially so that errors of the previous tree are revised by the next tree. XGBoost further enhances this process by correlating the new tree with the negative gradient of the loss function associated with the whole tree assembly (35). Accordingly and in line with a prior radiomics study on response assessment of rare cancers, XGBoost significantly increased our model's performance compared to RF on the external data set (36).

The rationale to applying NN with different architectures was that it employs several layers possibly facilitating a higher dimensional feature selection algorithm. Similar to our results, Yun et al (37) and Bae et al (38) found that a NN approach using radiomics features as input outperformed other ML classifiers on external data sets differentiating brain tumor types. In the study of Yun et al, NNs fed with radiomics features significantly outperformed support vector machine, RF, generalized linear model, human readers, and CNN on external validation data (37). The relatively low performance of CNN was attributed to the small training data set (n = 123) and heterogeneity in image acquisition (37). Using a training cohort of 166 patients with glioblastoma or brain metastasis, Bae et al could show that deep neural networks outperform human readers and traditional ML classifiers including adaptive boosting, support vector machine, and linear discriminant analysis (38).

Our study also investigated the role of the training data sample size on model performance on the external data set. Using an incremental 2.25% increase in training data, we found the best performance of NN1 and NN2 on the external data at 60% or 132 patients. Notably, performance was high on the external data starting already at 20% of the training data, however, there was marked variation on internal cross validation. This may be explained by the heterogeneity of the training data diagnoses and the relative homogeneity of the validation set.

NN demonstrated increased performance over conventional ML classifiers (37) and CNN (37) when MRI data is limited. We therefore recommend a combination of radiomic features and NN classification as a ML classifier when data are limited.

Our study has several limitations. With the small samples, large feature sets, and low signal-to-noise that are characteristic of neuroimaging data, prediction models built using neuroimaging data are at a high risk of overfitting (39). We experienced overfitting of NN1-3 at 70% of the training data. Due to the retrospective and bi-institutional nature of our study, there were heterogeneous FLAIR sequence

acquisitions, various scanner vendors, and different field strengths in our sample. Given that this heterogeneity reflects clinical practice, a reliable model should incorporate these technical variations. For our study, we only used FLAIR images. Incorporating additional MR imaging sequences such as T2-weighted images, DWI, and contrast-enhanced T1-weighted sequences could further increase model performance.

Conclusion

A combination of radiomic features and NN led to a high performing and reliable model for the challenging classification task of differentiating the molecular status of pediatric low grade glioma based on MRI data. The model was superior to RF and XGB for small datasets. This may have implications for the classification of other tumors with limited sample sizes as well.

Declarations

Funding statement:

MWW, KN, AA, NH, FK, LN, MS, MMS, EB, UT, MZ, KWW, FK, BEW received no grant for this study from any funding agency in the public, commercial or not-for-profit sectors.

References

1. Sturm D, Pfister SM, Jones DT. Pediatric gliomas: current concepts on diagnosis, biology, and clinical management. Journal of Clinical Oncology. 2017;35(21):2370-7.
2. Goebel A-M, Gnekow AK, Kandels D, Witt O, Schmidt R, Hernáiz Driever P. Natural History of Pediatric Low-Grade Glioma Disease - First Multi-State Model Analysis. J Cancer. 2019;10(25):6314-26.
3. Ostrom QT, Gittleman H, Liao P, Vecchione-Koval T, Wolinsky Y, Kruchko C, et al. CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2010–2014. Neuro-oncology. 2017;19(suppl_5):v1-v88.
4. Armstrong GT, Conklin HM, Huang S, Srivastava D, Sanford R, Ellison DW, et al. Survival and long-term health and cognitive outcomes after low-grade glioma. Neuro-oncology. 2011;13(2):223-34.
5. Lassaletta A, Scheinemann K, Zelcer SM, Hukin J, Wilson BA, Jabado N, et al. Phase II weekly vinblastine for chemotherapy-naive children with progressive low-grade glioma: a Canadian Pediatric Brain Tumor Consortium Study. Journal of Clinical Oncology. 2016;34(29):3537-43.
6. Krishnatry R, Zhukova N, Guerreiro Stucklin AS, Pole JD, Mistry M, Fried I, et al. Clinical and treatment factors determining long-term outcomes for adult survivors of childhood low-grade glioma: A population-based study. Cancer. 2016;122(8):1261-9.
7. AlRayahi J, Zapotocky M, Ramaswamy V, Hanagandi P, Branson H, Mubarak W, et al. Pediatric Brain Tumor Genetics: What Radiologists Need to Know. Radiographics. 2018;38(7):2102-22.

8. Ryall S, Tabori U, Hawkins C. Pediatric low-grade glioma in the era of molecular diagnostics. Acta Neuropathologica Communications. 2020;8(1):30.
9. Lassaletta A, Zapotocky M, Mistry M, Ramaswamy V, Honnorat M, Krishnatry R, et al. Therapeutic and prognostic implications of BRAF V600E in pediatric low-grade gliomas. Journal of Clinical Oncology. 2017;35(25):2934.
10. Mistry M, Zhukova N, Merico D, Rakopoulos P, Krishnatry R, Shago M, et al. BRAF mutation and CDKN2A deletion define a clinically distinct subgroup of childhood secondary high-grade glioma. Journal of clinical oncology. 2015;33(9):1015.
11. Fangusaro J, Onar-Thomas A, Poussaint TY, Wu S, Ligon AH, Lindeman N, et al. LGG-08. A PHASE II PROSPECTIVE STUDY OF SELUMETINIB IN CHILDREN WITH RECURRENT OR REFRACTORY LOW-GRADE GLIOMA (LGG): A PEDIATRIC BRAIN TUMOR CONSORTIUM (PBTC) STUDY. Neuro-oncology. 2017;19(suppl_4):iv34-iv5.
12. Hargrave DR, Bouffet E, Tabori U, Broniscer A, Cohen KJ, Hansford JR, et al. Efficacy and Safety of Dabrafenib in Pediatric Patients with BRAF V600 Mutation–Positive Relapsed or Refractory Low-Grade Glioma: Results from a Phase I/IIa Study. Clinical Cancer Research. 2019;25(24):7303-11.
13. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
14. Khalvati F, Zhang Y, Wong A, Haider MA. Radiomics. 2019.
15. Stocker D, Marquez HP, Wagner MW, Raptis DA, Clavien P-A, Boss A, et al. MRI texture analysis for differentiation of malignant and benign hepatocellular tumors in the non-cirrhotic liver. Heliyon. 2018;4(11):e00987.
16. Fritz B, Müller DA, Sutter R, Wurnig MC, Wagner MW, Pfirrmann CW, et al. Magnetic Resonance Imaging–Based Grading of Cartilaginous Bone Tumors: Added Value of Quantitative Texture Analysis. Investigative radiology. 2018;53(11):663-72.
17. Chaddad A, Kucharczyk MJ, Daniel P, Sabri S, Jean-Claude BJ, Niazi T, et al. Radiomics in glioblastoma: current status and challenges facing clinical implementation. Frontiers in oncology. 2019;9.
18. Iv M, Zhou M, Shpanskaya K, Perreault S, Wang Z, Tranvinh E, et al. MR Imaging-Based Radiomic Signatures of Distinct Molecular Subgroups of Medulloblastoma. AJNR Am J Neuroradiol. 2019;40(1):154-61.
19. Goya-Outi J, Calmon R, Orlhac F, Philippe C, Boddaert N, Puget S, et al., editors. Can Structural MRI Radiomics Predict DIPG Histone H3 Mutation and Patient Overall Survival at Diagnosis Time? 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); 2019: IEEE.
20. Wagner MW, Hainc N, Khalvati F, Namdar K, Figueiredo L, Sheng M, et al. Radiomics of Pediatric Low-Grade Gliomas: Toward a Pretherapeutic Differentiation of BRAF-Mutated and BRAF-Fused Tumors. AJNR Am J Neuroradiol. 2021.

21. Ryall S, Zapotocky M, Fukuoka K, Nobre L, Guerreiro Stucklin A, Bennett J, et al. Integrated Molecular and Clinical Analysis of 1,000 Pediatric Low-Grade Gliomas. Cancer Cell. 2020;37(4):569-83.e5.
22. Pieper S, Halle M, Kikinis R, editors. 3D Slicer. 2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No 04EX821); 2004: IEEE.
23. Parmar C, Velazquez ER, Leijenaar R, Jermoumi M, Carvalho S, Mak RH, et al. Robust radiomics feature quantification using semiautomatic volumetric segmentation. PloS one. 2014;9(7).
24. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. IEEE Trans Med Imaging. 2010;29(6):1310-20.
25. Li J, Liu S, Qin Y, Zhang Y, Wang N, Liu H. High-order radiomics features based on T2 FLAIR MRI predict multiple glioma immunohistochemical features: A more precise and personalized gliomas management. PLoS One. 2020;15(1):e0227703.
26. Park JE, Kim HS. Radiomics as a Quantitative Imaging Biomarker: Practical Considerations and the Current Standpoint in Neuro-oncologic Studies. Nucl Med Mol Imaging. 2018;52(2):99-108.
27. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123-40.
28. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001:1189-232.
29. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A. 1982;79(8):2554-8.
30. He B, Zhao W, Pi J-Y, Han D, Jiang Y-M, Zhang Z-G. A biomarker basing on radiomics for the prediction of overall survival in non–small cell lung cancer patients. Respiratory research. 2018;19(1):1-8.
31. Shinde S, Prasad S, Saboo Y, Kaushick R, Saini J, Pal PK, et al. Predictive markers for Parkinson's disease using deep neural nets on neuromelanin sensitive MRI. NeuroImage: Clinical. 2019;22:101748.
32. Bernatz S, Ackermann J, Mandel P, Kaltenbach B, Zhdanovich Y, Harter PN, et al. Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features. European radiology. 2020;30(12):6757-69.
33. Cao H, Bernard S, Sabourin R, Heutte L. Random forest dissimilarity based multi-view learning for Radiomics application. Pattern Recognition. 2019;88:185-97.
34. Deist TM, Dankers FJWM, Valdes G, Wijsman R, Hsu IC, Oberije C, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers. Medical physics. 2018;45(7):3449-59.
35. Torlay L, Perrone-Bertolotti M, Thomas E, Baciu M. Machine learning-XGBoost analysis of language networks to classify patients with epilepsy. Brain Inform. 2017;4(3):159-69.
36. Colen RR, Rolfo C, Ak M, Ayoub M, Ahmed S, Elshafeey N, et al. Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers. J Immunother Cancer. 2021;9(4).

37. Yun J, Park JE, Lee H, Ham S, Kim N, Kim HS. Radiomic features and multilayer perceptron network classifier: a robust MRI classification strategy for distinguishing glioblastoma from primary central nervous system lymphoma. Scientific reports. 2019;9(1):1-10.
38. Bae S, An C, Ahn SS, Kim H, Han K, Kim SW, et al. Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. Scientific reports. 2020;10(1):1-10.
39. Jollans L, Boyle R, Artiges E, Banaschewski T, Desrivières S, Grigis A, et al. Quantifying performance of machine learning methods for neuroimaging data. NeuroImage. 2019;199:351-65.

Figures

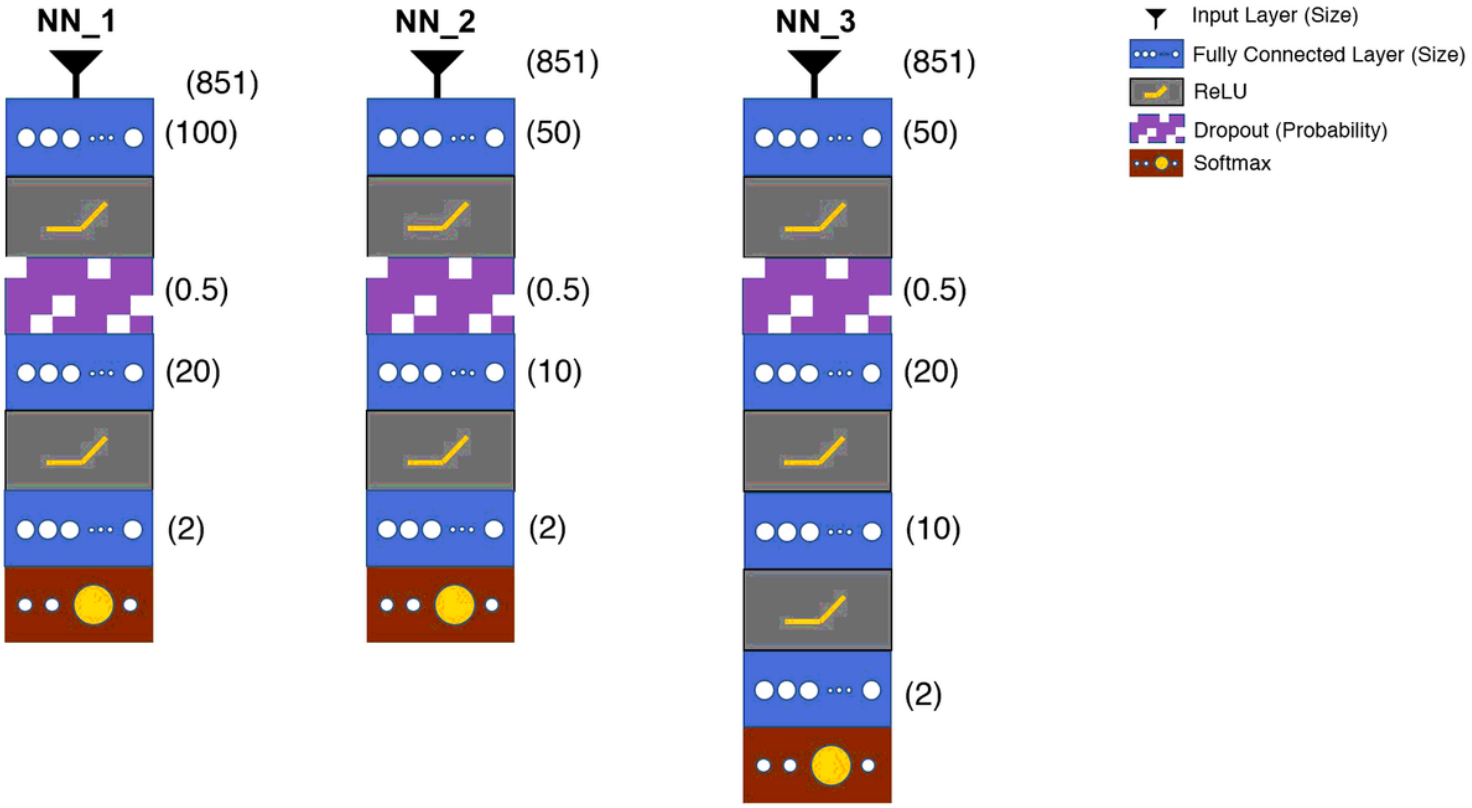

Figure 1

Neural Network architectures used for the experiments

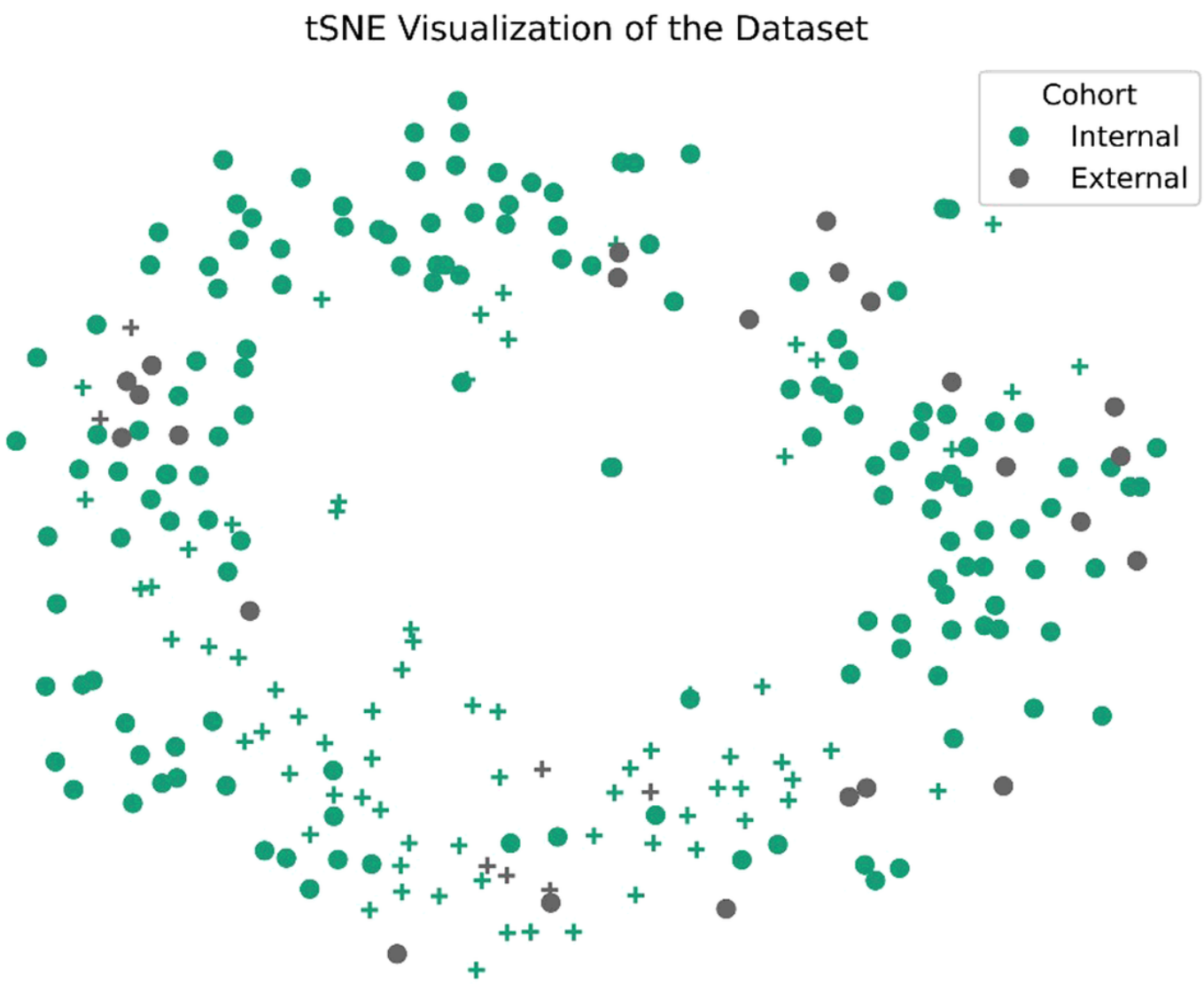

Figure 2

tSNE Visualization of the Dataset: Circles denote BRAF fusion, while plus markers represent BRAF V600E mutation. Green and grey colors highlight internal and external data, respectively.

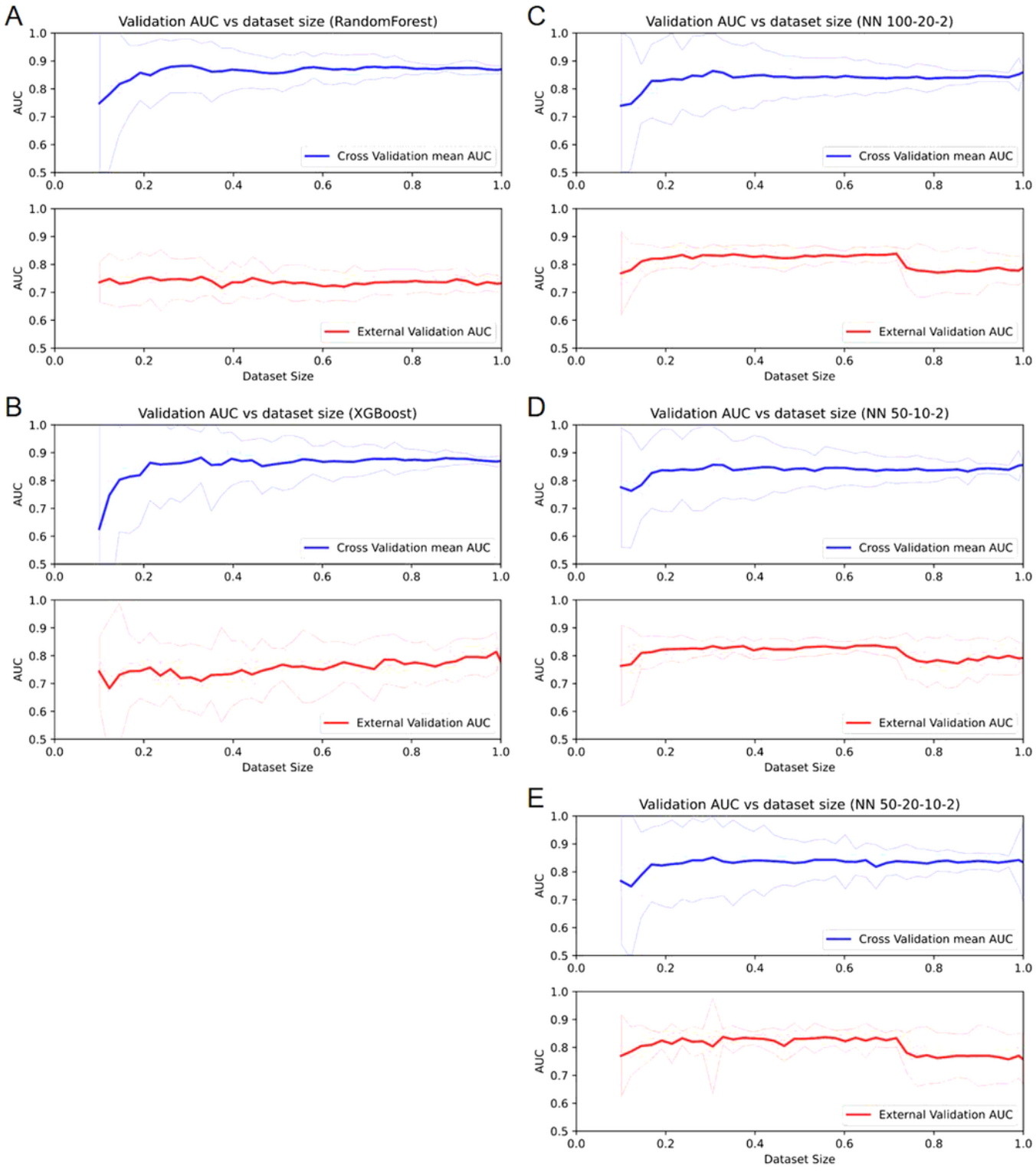

Figure 3

Area under the curve performance of different models on internal and external cohorts across training dataset sizes. A: Random Forest; B: XGBoost; C: Neural Network 1; D: Neural Network 2, E: Neural Network 3

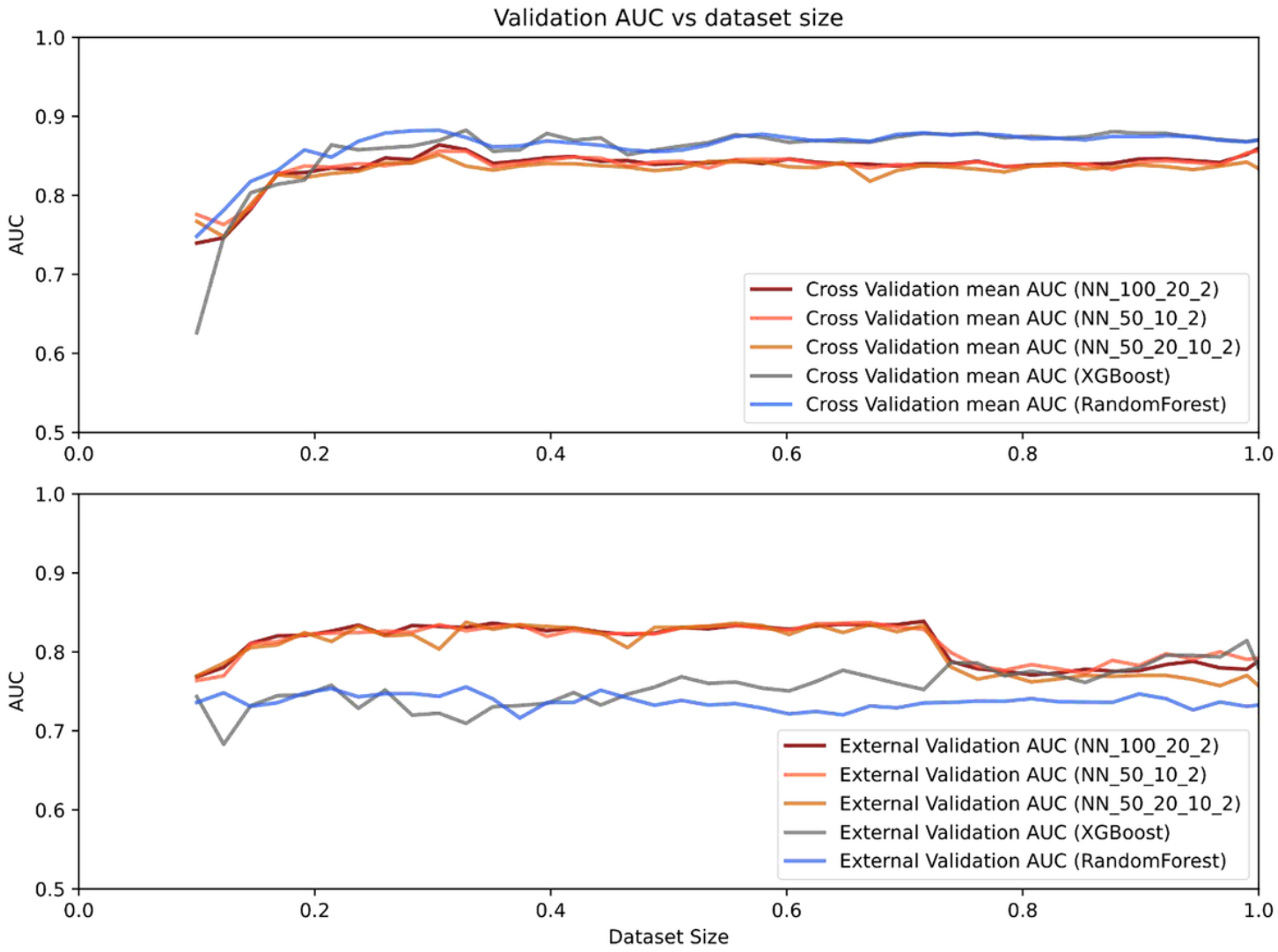

Figure 4

Comparison of area under the curve performance of different models across training dataset sizes

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalTable.docx