Radiology: Artificial Intelligence

Impact of Scanner Manufacturer, Endorectal Coil Use, and Clinical Variables on Deep Learning-assisted Prostate Cancer Classification Using Multiparametric MRI

José Guilherme de Almeida[1]

Nuno M. Rodrigues[1,2]

Ana Sofia Castro Verde[1]

Ana Mascarenhas Gaivão[3]

Carlos Bilreiro[3]

Inês Santiago[3]

Joana Ip[3]

Sara Belião[3]

Celso Matos[1]

Sara Silva[2]

Manolis Tsiknakis[4,5]

Kostantinos Marias[5,6]

Daniele Regge[7,8]

Nikolaos Papanikolaou[1,9]

On behalf of the ProCAncer-I Consortium[10]

[1] Champalimaud Research, Champalimaud Foundation, Lisbon, Portugal.

[2] Department of Informatics, LASIGE, Faculty of Sciences, University of Lisbon, Portugal.

[3] Department of Radiology, Champalimaud Clinical Center, Champalimaud Foundation, Lisbon, Portugal.

[4] Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, Greece.

[5] Department of Electrical and Computer Engineering, Hellenic Mediterranean University, Heraklion, Greece.

[6] Computational BioMedicine Laboratory (CBML), Institute of Computer Science, Foundation for Research and Technology–Hellas (FORTH), Heraklion, Greece.

[7] Department of Radiology, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Turin, Italy.

[8] Department of Surgical Sciences, University of Turin, Turin, Italy.

[9] Department of Radiology, Royal Marsden Hospital, Sutton, United Kingdom.

[10] https://www.procancer-i.eu/.

Received XXX; revision requested XXX; revision received XXX; accepted XXX.

Address correspondence to J.G.d.A. (email: jose.almeida@research.fchampalimaud.pt).

https://doi.org/10.1148/ryai.230555

Purpose: To assess the impact of scanner manufacturer and scan protocol on the performance of deep learning models to classify prostate cancer (PCa) aggressiveness on biparametric MRI (bpMRI).

Materials and Methods: In this retrospective study, 5,478 cases from ProstateNet, a PCa bpMRI dataset with examinations from 13 centers, were used to develop five deep learning (DL) models to predict PCa aggressiveness with minimal lesion information and test how using data from different subgroups—scanner manufacturers and endorectal coil (ERC) use (Siemens, Philips, GE with and without ERC and the full dataset)—impacts model performance. Performance was assessed using the area under the receiver operating characteristic curve (AUC). The impact of clinical features (age, prostate-specific antigen level, Prostate Imaging Reporting and Data System [PI-RADS] score) on model performance was also evaluated.

Results: DL models were trained on 4,328 bpMRI cases, and the best model achieved AUC = 0.73 when trained and tested using data from all manufacturers. Hold-out test set performance was higher when models trained with data from a manufacturer were tested on the same manufacturer (within-and between-manufacturer AUC differences of 0.05 on average, $P < .001$). The addition of clinical features did not improve performance ($P = .24$). Learning curve analyses showed that performance remained stable as training data increased. Analysis of DL features showed that scanner manufacturer and scan protocol heavily influenced feature distributions.

Conclusion: In automated classification of PCa aggressiveness using bpMRI data, scanner manufacturer and endorectal coil use had a major impact on DL model performance and features.

Prostate cancer aggressiveness could be predicted using biparametric MRI and deep learning

with negligible expert input, but performance was impacted by scanner manufacturer and scan protocol.

Key Points

Deep learning models predicted prostate cancer aggressiveness using only biparametric MRI with no lesion annotations or lesion location information (area under the receiver operating characteristic curve [AUC]=0.73).

Scanner manufacturer and endorectal coil use affected predictive performance of models (AUC improved by 0.05 when models were tested on data similar to the training data, $P < .001$).

Inclusion of clinical variables (age, prostate specific antigen level, and PI-RADS score) led to no performance improvements ($P = .24$).

Author contributions:
Guarantors of integrity of entire study, **S.S., M.T., N.P.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **J.G.d.A., A.M.G., K.M.**; clinical studies, **A.S.C.V., A.M.G., C.B., S.B.**; experimental studies, **N.M.R., A.S.C.V., A.M.G., C.B.**; statistical analysis, **J.G.d.A., A.M.G.**; and manuscript editing, **J.G.d.A., A.S.C.V., A.M.G., C.M., S.S., M.T., K.M., D.R., N.P.**

Conflicts of interest are listed at the end of this article.

Prostate cancer (PCa) staging constitutes an important part of disease management (1,2). The most recent International Society of Urological Pathology (ISUP) grading system is associated with biochemical relapse-free survival, metastasis-free survival and PCa-specific survival (3). This grading requires a biopsy (2), an invasive procedure associated with complications such as pain and infection (4); additionally, biopsy sampling errors may lead to false negative results (5). Noninvasive techniques, such as biparametric MRI (bpMRI), are usually performed before the biopsy and as a part of PCa staging (6).

Previous works have shown that it is possible to predict PCa aggressiveness using both radiomics (7,8) and deep learning (DL) methods (9–11). Other work has focused on predicting PI-RADS score (12,13), but that task is not considered here. Radiomics methods require the segmentation of lesions (or prostate gland (8)), and previously suggested DL methods typically require a nonnegligible amount of information regarding the location of the lesion (bounding box or lesion segmentation masks) and/or make use of relatively small datasets to train and validate the models (Table C.1).

To simplify the task of PCa aggressiveness classification using DL models, ISUP1 lesions should be distinguished from ISUP2–5 lesions. While some models group ISUP1 and ISUP2 together (14), ISUP2 lesions can still have high recurrence (22.6%) (15). As such, classi-

fying ISUP2 as low PCa aggressiveness can lead to patients missing necessary care. Similar to other ISUP-based tasks, this categorization could be used to reduce the number of unnecessary biopsies by identifying cases that are likely to be clinically insignificant among patients referred for biopsy.

The aim of this retrospective study was to show the impact of scanner manufacturer and endorectal coil use on PCa aggressiveness (ISUP1 versus ISUP2–5) prediction in the absence of lesion location information with one of the largest PCa bpMRI datasets to date, ProstateNet (16). Additionally, deep features were inspected to better understand how the presence of scanner manufacturer and scan protocol diversity can lead to changes in performance.

Materials and Methods

Due to the retrospective nature of this study, informed consent was waived by the independent review board of each participating institution (Table C.2).

Data

The ProstateNet dataset.—

This study incorporated 8,891 bpMRI studies from the ProstateNet dataset in Digital Imaging and Communications in Medicine (DICOM) format, accessed on March 13th, 2023. The cases, available at https://prostatenet.eu/, were collected from 12 distinct European clinical centers as a part of the ProCAncer-I project (Table C.2; https://www.procancer-i.eu/). Given the diverse acquisition protocols implemented across centers, a unified view was not feasible; however, while not formally analyzed in the current study, the composition of magnetic field strength (between 1.5T and 3.0T) was variable when stratifying by manufacturer (Figure B.1). All examinations were interpreted by through histopathology and annotated using Gleason scores the available histopathologist at each center, allowing for the derivation of ISUP scores (1). Prostate-specific antigen (PSA) level, age at baseline and Prostate Imaging Reporting and Data System (PI-RADS) data were also available for most studies. As inclusion criteria, only patients referenced for biopsy were included, leading to a subset of 5,478 bpMRI studies with available Gleason grades and ISUP grades (Fig 1; demographic parameters in Table C.3). To avoid issues of data leakage, a single study corresponding to the MRI examination that preceded a confirmatory biopsy was selected for each patient. Prior to uploading to the ProstateNet server, all data were de-identified and anonymized using a protocol identical to that used by the RSNA (17). Ethics committee approval and patient consent were obtained by each clinical partner (Table C.2). Finally, we note that while using an endorectal coil (ERC) in MRI examinations is controversial and may not be beneficial (18), we have included studies using ERC as they allowed us to determine how this impacts the predictive performance of potential artificial intelligence solutions.

Dataset preparation and composition.—

Table 1 shows the dataset composition in terms of manufacturer and PCa aggressiveness. The dataset comprised 2,418 Siemens studies, 1,712 Philips studies and 1,575 GE studies (539 with ERC and 1,036 with no ERC), with a small proportion of MRI negative studies (7.1% studies with PI-RADS ≤ 2; Table C.4). All sequences were center cropped using a 128 × 128 window

and cropped/padded such that a total of 24 slices were present, similar to a previous work (11). T2-weighted (T2 W) and high b-value diffusion weighted images (DWI) sequences were individually scaled between 0 and 1, whereas apparent diffusion coefficient (ADC) sequences were first converted to mm^2/s (if in $\mu\, m^2/s$) and multiplied by $\frac{1}{3}$ to maintain their dynamic range. Whenever models with T2 W, ADC, and DWI sequences were trained, all three sequences for a given study were concatenated. To determine whether the used crop size could negatively impact performance, sequences were also cropped to size 192 × 192 × 24 for a crop size sensitivity analysis.

Model Specifications and Training

Models were trained on T2 W sequences alone or T2 W together with DWI and ADC sequences to predict whether studies had ISUP grade 1 or ISUP grades 2–5. In clinical practice, T2 W sequences are used for anatomic resolution, whereas DWI and ADC are used for lesion localization and assessment of lesion morphology and intensity (6). Five distinct 3D DL models were trained: 1) a simple 3D VGG-based model composed only of convolutions, batch normalization and activation layers, and maximum pooling (19); 2) a 3D residual network model; 3) a 3D ConvNeXt model (20); 4) a 3D vision transformer (ViT) model (21); and 5) a factorized ViT model. The factorized ViT architecture is similar to that proposed in (21) but separates the processing of within and between slice information by applying a ViT to each slice and a second ViT to the slice information in each sequence. Networks were implemented in PyTorch (22) and trained for a total of 100 epochs using AdamW (23). Details on hyperparameters, DL models, training, initialization and augmentation are available in the Supplementary Methods.

Inclusion of Clinical Features in Deep Learning Models

To test how age at baseline and total PSA level may contribute toward prediction, "hybrid models" combining an image (MRI sequence) and tabular network were trained and compared with "sequence-only models." The image network is the same as those described above, whereas the tabular network is a linear model (Supplementary Methods). Additionally, cross-validated elastic net-regularized linear classification models (24) were trained. These models used glmnet (25) with probability predictions from sequence-only models, age, total PSA level and PI-RADS score as predictors, such that $p_{final} = sigmoid(p_{image} + \beta_1\, PSA + \beta_2\, age + \beta_3\, PI\text{–}RADS)$, where β_1, β_2 and β_3 are linear coefficients.

Model Training and Evaluation

Each model was trained five times using five-fold stratified cross-validation at the patient level and further evaluated on an independent hold-out test set. All models were trained using ISUP grade and manufacturer stratification. For training and testing, we consider the full training data (Full) and 4 subsets derived from these data: i) GE (ERC), ii) GE (no ERC), iii) Philips, and iv) Siemens. Each validation fold was constructed from the subset of all studies containing all image sequences and all clinical data (age and PSA level; Fig 1), guaranteeing that validation and test-

ing are performed on the same studies. Training was performed on the set of available sequences not belonging to the validation fold or hold-out test set.

To understand the relationship between the amount of available training data and performance a learning curve analysis was performed. This involved training models with 10%, 30%, 50% and 70% of the available training data and calculating the performance for the same cross-validation and hold-out test sets.

Analysis of Differences between Dataset Subsets

To assess the difference between feature representations for different manufacturers and classifications, the features from the last feature extraction layer of the best-performing fold of the best-performing model were extracted. Then, *t*-distributed stochastic neighbor embeddings (*t*-SNE) (26) (implemented in the Rtsne package for R (26,27)) were calculated. Given that *t*-SNE is appropriate for only qualitative analysis, optimal transport dataset distances were used (28) as a quantitative measure of feature differences. To do this, distances between subsets constructed from the original testing set (stratified by manufacturer and classification) were calculated using the formulation of the Wasserstein 2-distance noted by Alvarez-Melis and Fusi (28) (Supplementary Methods).

Alternative Target Categorization

To better understand if there could be an impact from the ISUP grade binarization, we trained four sequence-only (T2 W and T2 W+ADC+DWI) models (VGG, ConvNext, ViT, factorized ViT) stratified by scanner manufacturer using a different categorization—ISUP grades 1–2 versus ISUP grades 3–5; this enables us to make our work compatible with others using this alternative categorization. Additionally, and as described above for the methodology using ISUP grade 1 versus ISUP grades 2–5, we trained elastic net-regularized linear classification models to predict the target (ISUP grades 1–2 versus ISUP grades 3–5) using classification probabilities from deep-learning models and clinical features.

Statistical Analysis

Reported cross-validation and hold-out test set performances were obtained using the model with the highest observed area under the receiver operating characteristic curve (AUC) during training. For external testing, we used the average of the five folds from the best performing model on the Prostate Imaging: Cancer AI (PI-CAI) studies with available ISUP scores (n = 653, 425 with ISUP > 1) (29). Statistical analysis was conducted using R version 4.2.2 (27).

To better account for the relatively large number of comparisons, four multivariate linear models were constructed: (i) AUC_{CV} = Manufacturer + Architecture + Sequences (Model I), (ii) AUC_{CV} = Manufacturer + Convolutional Architecture + Sequences (Model II), (iii) AUC_{Test} = Train Manufacturer + Test Manufacturer + Is Same + Test Manufacturer: Is Same + Clinical + Architecture + Sequences (Model III) and (iv) AUC_{Test} = Train Manufacturer + Test Manufacturer + Architecture + Sequences + Size (Model IV), where:

- AUC_{CV} and AUC_{Test} refer to the CV and hold-out test set AUC, respectively

- Manufacturer, Train Manufacturer and Test Manufacturer are factors (Full, Philips, Siemens, GE (no ERC) and GE (ERC)) referring to manufacturer data used during CV, training and testing
- Architecture is a factor (VGG, ResNet, ConvNeXt, ViT, Factorized ViT) referring to model architecture
- Convolutional Architecture is a factor (No, Yes) referring to whether the model architecture is convolutional
- Sequences is a factor (T2W and T2W+DWI+ADC) referring to the sequences used
- Is Same is a factor (No, Yes) corresponding to whether Train Manufacturer and Test Manufacturer are identical
- Test Manufacturer: Is Same is the interaction term between Test Manufacturer and Is Same
- Clinical is a factor (No, Yes) corresponding to whether clinical information was used
- Size is a factor (128x128, 192x192) corresponding to input in-plane size

Analysis of variance (ANOVA) was used to determine if differences within factors were statistically significant according to an F-test. If so, posthoc pairwise comparisons were performed using Tukey Honest Significant Differences tests (THSD), which automatically adjusted for multiple comparisons within each ANOVA. To further control for multiple testing, we adjusted *P* values using the Benjamini-Hochberg correction (30), considering the total number of posthoc pairwise comparisons ($n = 25$) and a significance threshold and false discovery rate of 0.05.

Differences in performance for the cross-validated elastic net-regularized linear classification models with and without clinical features were tested using paired Student *t* tests.

Implementation and Code Availability

Training, testing and inference were performed using an internal library developed in PyTorch (22). Lightning (31) and MONAI (32) are available in https://github.com/CCIG-Champalimaud/adell-mri. Training, testing, and inference routines are present in lib/entrypoints/classification, while neural network architectures are provided in lib/modules/classification.

Results

Performance Analyses

While performance in classification of PCa aggressiveness across different manufacturers and ERC use was widely variable (Fig 2A), a few trends were identified during cross-validation.

Although T2 W sequences provided greater anatomic resolution, using these alone led to lower performance (0.04 cross-validation AUC improvement for bpMRI over T2 W, $P < .001$ for THSD in Model I; Fig 2A). Across different data subsets, VGG models outperformed more complicated and modern models such as ResNet, ViT or ConvNext; in all cases, the average AUC for VGG models was higher than that of other models. VGG models were associated with higher AUC (at least 4.0% higher AUC compared with other models; Table 2). ViT-based models performed worse than convolutional neural network-based models (cross-validation AUC reduction of 0.03 for ViT-based models, $P < .001$ for posthoc THSD in Model II).

When considering manufacturer-specific results, models trained on GE (ERC) scanners had lower performance than other models (Figs 2A, B.2). Models trained on data from Philips scanners and from GE scanners with no ERC performed better than those trained on Siemens data, which are twice as prevalent (Fig B.2; Table 3). Performance for models trained and tested on all manufacturers was consistent between cross-validation (AUC = 0.71) and the hold-out test set (AUC = 0.73). However, an inconsistent drop in performance was observed when models were tested on data different from the train data (test set AUC dropped by 0.05, $P < .001$ for THSD; Figs 2B and B.3). This performance drop was more evident across T2-only models, models trained on GE data and ViT models trained on Siemens data. Models trained on specific scanners generally performed better when tested on data from the same scanner (Figs B.4 and 2C). However, it should be highlighted that models trained on all manufacturers (Full) recapitulated the within-manufacturer performance and performed similarly across different scanners (excluding GE ERC and Siemens data; Table 4). AUC was not independent of architecture during cross-validation, and this trend is consistent in the hold-out test set–VGG outperformed all other models (test set AUC increase of 0.016 when compared with ResNet, the second best-performing model $P = .014$; Table 2).

These findings were consistent with an alternative ISUP categorization (ISUP = 1–2 versus ISUP = 3–5), as shown in Figure B.5.

Sensitivity and specificity of the VGG model on the hold-out test set were 90 ± 2.5% (437/486 cases) and 30.4 ± 6% (56/184 cases), respectively, across the five folds. When testing the ensemble of the best-performing model (the average of all folds of VGG trained on T2 W+DWI+ADC from all manufacturers) on an external test set, there was a small drop in performance (AUC (bootstrapped 95% CI) = 0.66 (0.62–0.70) for the external test set versus 0.74 (95% CI: 0.70–0.78) for the hold-out test set.

Comparison of cross-validation performance between hybrid models and sequence-only models revealed that these variables did not lead to improvements in predictive performance (Figure B.6, Figure B.7, Figure B.8), with results for the hold-out test set confirming this observation ($P = .24$ for F-test for clinical features (Clinical) in Model III ANOVA). Training linear classification models with sequence-only probabilities, age, PSA and PI-RADS values as predictors did not improve results ($P = .22$ for a paired Student's t test; Figure B.9). However, when using the alternative ISUP = 1–2 versus ISUP = 3–5 categorization, the inclusion of clinical variables led to a performance improvement of 0.04 AUC ($P < .001$ for a paired Student's t test; Figure B.10), indicating that clinical variables were important in limited instances.

Sensitivity Analysis and Learning Curves

To better understand the relationship between the amount of data and performance, sequence-only VGG models with all sequences were trained with different amounts of training data. For cross-validated performance, there was an expected relation between training data size and performance for all manufacturers, excluding GE data with ERC (Fig 3A). This trend, however, was not as clear for the hold-out test set; while an upward trend was observable in most cases when training and testing on the same data, this was unpredictable when testing on data from other

manufacturers (Fig 3B). For instance, increasing the amount of Philips data led to improved performance on GE data with no ERC but was detrimental when testing on Siemens data. On the other hand, when training with data from all manufacturers, performance improved for all manufacturers except GE with ERC, which plateaued at approximately 50% (no better than chance).

The effect of a larger crop size (192 × 192 × 24) on performance was tested, showing that the performance in both the cross-validation and hold-out test sets did not change (Figure B.11; $P = .92$ for F-test for input size (Size) in Model IV ANOVA).

Association between High-dimensional Features and Scanner Manufacturer, Target and Protocol

DL features from the best performing sequence-only T2 W+DWI+ADC VGG fold were first visually inspected using *t*-SNE, showing how samples from the same manufacturer cluster together. Particularly, GE data with ERC, mostly stemming from data provider B (data providers were pseudonymized to preserve confidentiality), shared very few neighbors with samples from other manufacturers (Fig 4A).

Quantitative analysis of feature differences confirmed qualitative findings (Fig 4): data derived from the same manufacturer and with the same ERC protocol were, in general, more similar than data with the same classification (aggressiveness).

Discussion

This work assessed the impact of several factors on the performance of DL models to predict PCa aggressiveness on biparametric MRI. Such factors included the use of different model architectures, sequence combinations (T2 W and T2 W+DWI+ADC), scanner manufacturer, scan protocols, and the inclusion of clinical variables (age at baseline, total PSA level, PI-RADS score). Models performed better when a diffusion weighted sequence was included (cross-validation AUC improved by 0.04 when bpMRI was used instead of T2 W alone, $P < .001$).Performance was also better when tested on data acquired using the same manufacturer/protocol as the training data (test set AUC improved by 0.05 when models were tested on data similar to the train data instead of data from other manufacturers, $P < .001$). Improvements in performance were possible with increases in data, but such improvements do not appear to be associated with cross-manufacturer generalization; indeed, the effect of training and testing on similar data appears to affect how data are represented in the DL high-dimensional feature space.

The main objective of this work was to elucidate how data variability, using data from different manufacturers and centers, could provide more robust and clinically applicable models. The results presented for the Full models highlighted this. Indeed, the performance of these models was consistent between training and testing, and this effect did not depend on model architecture. Transformer-based architectures tended to underperform, explained by their lack of inductive biases and larger data requirements. A comprehensive hyperparameter optimization could alter some of these observations, as well as improve the performance of other convolutional methods. Additionally, models trained using all bpMRI sequences matched the performance of models trained using data from specific manufacturers. In contrast, performance dropped when

models trained on data acquired using specific manufacturers were tested on different manufacturers. This highlights an important aspect of these models—generalizability to different manufacturers can only be guaranteed when similar examples are included in the training data. Analyzing the high-dimensional structure of the hold-out test data confirmed that training models with all manufacturers still led to a clear manufacturer and protocol shift in DL features, partly associated with ERC use. This may be caused by study-specific aspects, as the relevant signal may be weaker in studies using ERC, or by the relatively small dataset size, as GE studies with ERC were half as prevalent as GE studies without ERC. Other factors, such as center-specific protocols, can also play a role in these differences. Previous work analyzing PI-RADS results across 26 centers showed that the positive predictive value for PI-RADS between centers had high variability (33). It is possible that some of this variability stems from different scanners across different centers. While these manufacturer-specific hurdles are prevalent, we note that performance is fairly similar to that exhibited by radiologists. Specifically, our results showed a sensitivity and specificity of 90 ± 2.5% and 30.4 ± 6%, respectively, while the PROMIS trial showed that radiologists had a sensitivity and specificity of 93% (95% CI = 88%–96%) and 41% (95% CI = 36%–46%) (34). A later meta-analysis showed that pooled estimates for the sensitivity and specificity were 91% (95% CI = 83%–95%) and 37% (95% CI = 29%–46%) (35).

Models trained using Philips data performed significantly better than models trained using other data from other scanner manufacturers. This is quite remarkable considering that data obtained using Siemens scanners existed at nearly twice the proportion as data obtained using Philips scanners in our dataset. This difference in performance was likely not explained by the available training data; indeed, through the learning curve analysis, predictive performance for models trained and tested using Philips scanners was already significantly higher at relatively low fractions of training data. Indeed, while it may be possible to improve performance by including more data, it is likely that the numbers required to do so are quite high. A better approach may be the inclusion of other types of data. A study by Hosseinzadeh and colleagues (36) demonstrated that lesion segmentation and detection methods using bpMRI prostate examinations benefit from more data and anatomic (ie, prostate zone segmentation) information.

This study had important limitations. This work should act as a proof-of-concept showing that i) DL methods without anatomic information can predict PCa and ii) important sources of variability captured only by large datasets led to striking differences in performance. Despite its less demanding requirements, a potential caveat of this research was its performance when compared with other models using high quality lesion annotations (9,10,37,38). Furthermore, the concrete definition of aggressiveness in this study can be contested. Some works indicate larger differences in outcome between ISUP grades 1 and 2 (39), while others indicate larger differences between grades 2 and 3 (3). However, by repeating our analysis using ISUP grades 1–2 versus ISUP grades 3–5, we showed that our conclusions hold regardless of the characterization. Additionally, even though an informal exploration of hyperparameters (learning rate and weight decay) was performed, it is imperative for future endeavors to substantially enhance this search process to improve the performance of these models (ie, determining the best patch size for ViT-based models or include a better study of the initial convolutions for convolutional methods). Moreover, DL research moves at a rapid pace, and better models and optimization techniques are

likely to appear as time progresses; therefore, future analyses should strive to expand this aspect to a more complete exploration. Similarly, while an assessment of the impact of clinical features was performed, this was incomplete and showed potential only in limited scenarios. Other clinical and molecular variables—particularly PSA density or the genetic and molecular characterization of PCa—may further help in classifying Pca aggressiveness, as hinted by previous work (40–42). Additionally, a prospective evaluation of these models is key to better understanding how alterations to clinical practice may lead to model drift, a pervasive issue in machine learning models known to cause performance deterioration (43). Finally, given that the dataset was sourced from European centers and that no information on patient race/ethnicity was available, models developed using these data could suffer from biases that cause them to underperform in underrepresented populations (44,45).

In conclusion, this study demonstrated a substantial impact of scanner manufacturers and scan protocol on the performance of classification models to classify PCa aggressiveness on parametric MRI. This effect was reduced when models were trained on data similar to that used during testing, and feature distribution was largely affected by scanner manufacturer and scan protocol. The addition of clinical features did not lead to consistent improvements in model predictive performance. This work can be further improved by replicating its findings in terms of lesion segmentation and detection with segmentation annotations, by including other relevant clinical information such as molecular features, by prospectively validating these results, or with a more consistent hyperparameter search method during training. Additionally, increasing the fairness of the model by including fairness analyses in terms of sensitive attributes such as age, race and ethnicity can further improve the impact of this work.

Funding: Horizon 2020; Fundação para a Ciência e a Tecnologia.

Acknowledgments: The authors would like to acknowledge the funding bodies.

Disclosures of conflicts of interest: J.G.d.A. Support, Champalimaud Foundation Horizon 2020; support for attending meeting, European Multidisciplinary Congress in Urological Cancers 2023. **N.M.R.** PhD scholarship, Fundação para a Ciência e Técnologia (FCT). **A.S.C.V.** Grants and support for meetings/travel, ProCAncer-I project (European Union's Horizon 2020 research and innovation programme, grant number 952159). **A.M.G.** No relevant relationships. **C.B.** No relevant relationships. **I.S.** No relevant relationships. **J.I.** No relevant relationships. **S.B.** No relevant relationships. **C.M.** No relevant relationships. **S.S.** No relevant relationships. **M.T.** Funding from a research project funded by the European Commission. **K.M.** No relevant relationships. **D.R.** No relevant relationships. **N.P.** Stock or stock options, MRIcons LTD.

References

1. Egevad L, Delahunt B, Srigley JR, Samaratunga H. International Society of Urological Pathology (ISUP) grading of prostate cancer - An ISUP consensus on contemporary grading. Acta Pathol Microbiol Scand Suppl 2016;124(6):433–435.

2. van Leenders GJLH, van der Kwast TH, Grignon DJ, et al; ISUP Grading Workshop Panel Members. The 2019 International Society of Urological Pathology (ISUP) consensus conference on grading of prostatic carcinoma. Am J Surg Pathol 2020;44(8):e87–e99.

3. Spratt DE, Jackson WC, Abugharib A, et al. Independent validation of the prognostic capacity of the ISUP prostate cancer grade grouping system for radiation treated patients with long-term follow-up. Prostate Cancer Prostatic Dis 2016;19(3):292–297.

4. Loeb S, Vellekoop A, Ahmed HU, et al. Systematic review of complications of prostate biopsy. Eur Urol 2013;64(6):876–892.

5. King CR, Long JP. Prostate biopsy grading errors: a sampling problem? Int J Cancer 2000;90(6):326–330.

6. Scott R, Misser SK, Cioni D, Neri E. PI-RADS v2.1: What has changed and how to report. SA J Radiol 2021;25(1):2062.

7. Cuocolo R, Cipullo MB, Stanzione A, et al. Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. Eur Radiol 2020;30(12):6877–6887.

8. Rodrigues A, Santinha J, Galvão B, Matos C, Couto FM, Papanikolaou N. Prediction of prostate cancer disease aggressiveness using bi-parametric MRI radiomics. Cancers (Basel) 2021;13(23):6065.

9. Hiremath A, Shiradkar R, Fu P, et al. An integrated nomogram combining deep learning, Prostate Imaging-Reporting and Data System (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. Lancet Digit Health 2021;3(7):e445–e454.

10. Liu S, Zheng H, Feng Y, Li W. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. Medical Imaging 2017: Computer-Aided Diagnosis. SPIE 2017;581–584:1013428. https://www.semanticscholar.org/reader/29bfbdf16d4b59c67085ab928cad384a44476cd5.

11. Pachetti E, Colantonio S, Pascali MA. On the Effectiveness of 3D Vision Transformers for the Prediction of Prostate Cancer Aggressiveness. Image Analysis and Processing ICIAP 2022 Workshops. Springer International Publishing, 2022; 317–328.

12. Yildirim K, Yildirim M, Eryesil H, et al. Deep learning-based PI-RADS score estimation to detect prostate cancer using multiparametric magnetic resonance imaging. Comput Electr Eng 2022;102(108275):108275.

13. Chen S, Yang Y, Peng T, Yu X, Deng H, Guo Z. The prediction value of PI-RADS v2 score in high-grade Prostate Cancer: a multicenter retrospective study. Int J Med Sci 2020;17(10):1366–1374.

14. Bertelli E, Mercatelli L, Marzi C, et al. Machine and Deep Learning Prediction Of Prostate Cancer Aggressiveness Using Multiparametric MRI. Front Oncol 2022;11:802964.

15. Hollemans E, Verhoef EI, Bangma CH, et al. Large cribriform growth pattern identifies ISUP grade 2 prostate cancer at high risk for recurrence and metastasis. Mod Pathol 2019;32(1):139–146.

16. Prostate NET. https://prostatenet.eu/. Accessed August 5, 2024.

17. The DicomAnonymizerTool. https://mircwiki.rsna.org/index.php?title=The_DicomAnonymizerTool. Accessed June 3, 2024.

18. Gawlitza J, Reiss-Zimmermann M, Thörmer G, et al. Impact of the use of an endorectal coil for 3 T prostate MRI on image quality and cancer detection rate. Sci Rep 2017;7(1):40640.

9. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556 [preprint] https://arxiv.org/abs/1409.1556. Posted September 4, 2014. Updated April 10, 2015. Accessed DATE.

20. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, 2022; 11966–11976. https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

21. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. 2021. https://openreview.net/forum?id=YicbFdNTTy.

22. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Fox E, Garnett R, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc., 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

23. Loshchilov I, Hutter F. Fixing Weight Decay Regularization in Adam. https://openreview.net/forum?id=rk6qdGgCZ. Published February 15, 2018. Updated October 14, 2024. Accessed DATE.

24. Zou H, Hastie T. Regularization and Variable Selection Via the Elastic Net. J R Stat Soc Series B Stat Methodol 2005;67(2):301–320.

25. Tay JK, Narasimhan B, Hastie T. Elastic net regularization paths for all generalized linear models. J Stat Softw 2023;106(1):1.

26. van der Maaten L, Hinton G. Visualizing Data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl. Published November 2008. Accessed July 13, 2023.

27. R Foundation for Statistical Computing R. R. A language and environment for statistical computing. RA Lang Environ Stat Comput, 2018.

28. Alvarez-Melis D, Fusi N. Geometric dataset distances via optimal transport. Adv Neural Inf Process Syst 2020;33:21428–21439. https://proceedings.neurips.cc/paper_files/paper/2020/hash/f52a7b2610fb4d3f74b4106fb80b233d-Abstract.html.

29. Saha A, Bosma J, Twilt J, et al. Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI — The PI-CAI Challenge. https://openreview.net/pdf?id=XfXcA9-0XxR. Published 2023. Accessed February 23, 2024.

30. Jafari M, Ansari-Pour N. Why, when and how to adjust your P values? Cell J 2019;20(4):604–607.

31. Falcon W. The PyTorch Lightning team. Published 2019.

32. MONAI Consortium. MONAI: Medical Open Network for AI. Published 2023.

33. Westphalen AC, McCulloch CE, Anaokar JM, et al. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: Experience of the Society of Abdominal Radiology Prostate Cancer Disease-focused Panel. Radiology 2020;296(1):76–84.

34. Ahmed HU, El-Shater Bosaily A, Brown LC, et al; PROMIS study group. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet 2017;389(10071):815–822.

35. Drost FH, Osses DF, Nieboer D, et al. Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. Cochrane Libr 2019;4(4):CD012663.

36. Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. Eur Radiol 2022;32(4):2224–2234.

37. Khosravi P, Lysandrou M, Eljalby M, et al. Biopsy-free prediction of prostate cancer aggressiveness using deep learning and radiology imaging. medRxiv 2019.12.16.19015057 [preprint] https://www.medrxiv.org/content/10.1101/2019.12.16.19015057. Posted December1 19, 2019. Accessed DATE.

38. Bosma JS, Saha A, Hosseinzadeh M, Slootweg I, de Rooij M, Huisman H. Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning-based Prostate Cancer Detection Using Biparametric MRI. Radiol Artif Intell 2023;5(5):e230031.

39. Sun GX, Shen PF, Zhang XM, et al. Predictive efficacy of the 2014 International Society of Urological Pathology Gleason grading system in initially diagnosed metastatic prostate cancer. Asian J Androl 2017;19(5):573–578.

40. Yusim I, Krenawi M, Mazor E, Novack V, Mabjeesh NJ. The use of prostate specific antigen density to predict clinically significant prostate cancer. Sci Rep 2020;10(1):20015.

41. Green HD, Merriel SWD, Oram RA, et al. Applying a genetic risk score for prostate cancer to men with lower urinary tract symptoms in primary care to predict prostate cancer diagnosis: a cohort study in the UK Biobank. Br J Cancer 2022;127(8):1534–1539.

42. Rebello RJ, Oing C, Knudsen KE, et al. Prostate cancer. Nat Rev Dis Primers 2021;7(1):9.

43. Carter RE, Anand V, Harmon DM Jr, Pellikka PA. Model drift: When it can be a sign of success and when it can be an occult problem. Intell Based Med. 2022;6(100058):100058.

44. Perez-Downes JC, Tseng AS, McConn KA, et al. Mitigating bias in clinical machine learning models. Curr Treat Options Cardiovasc Med 2024;26(3):29–45.

45. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. J Med Imaging (Bellingham) 2023;10(6):061104.

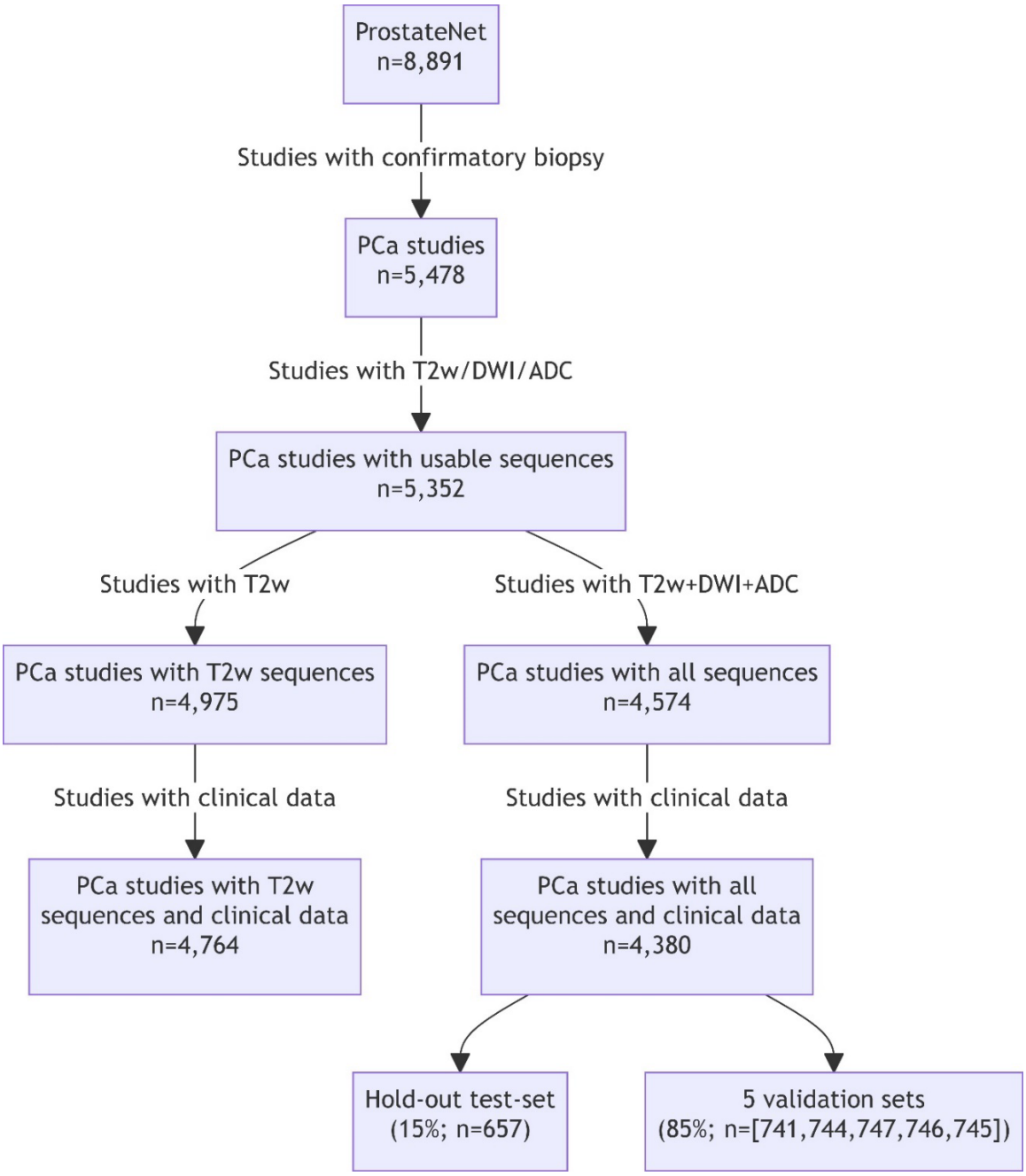
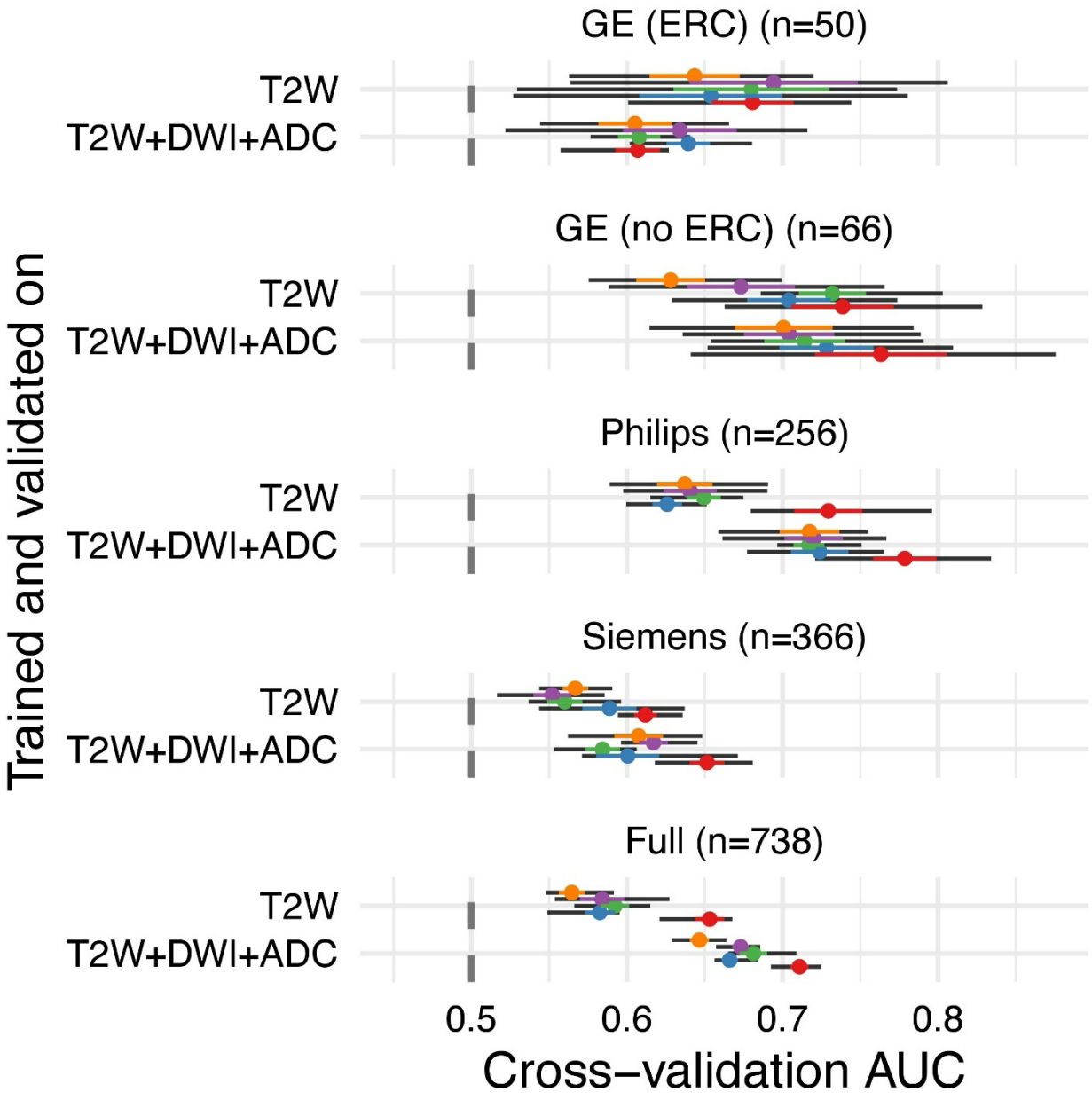
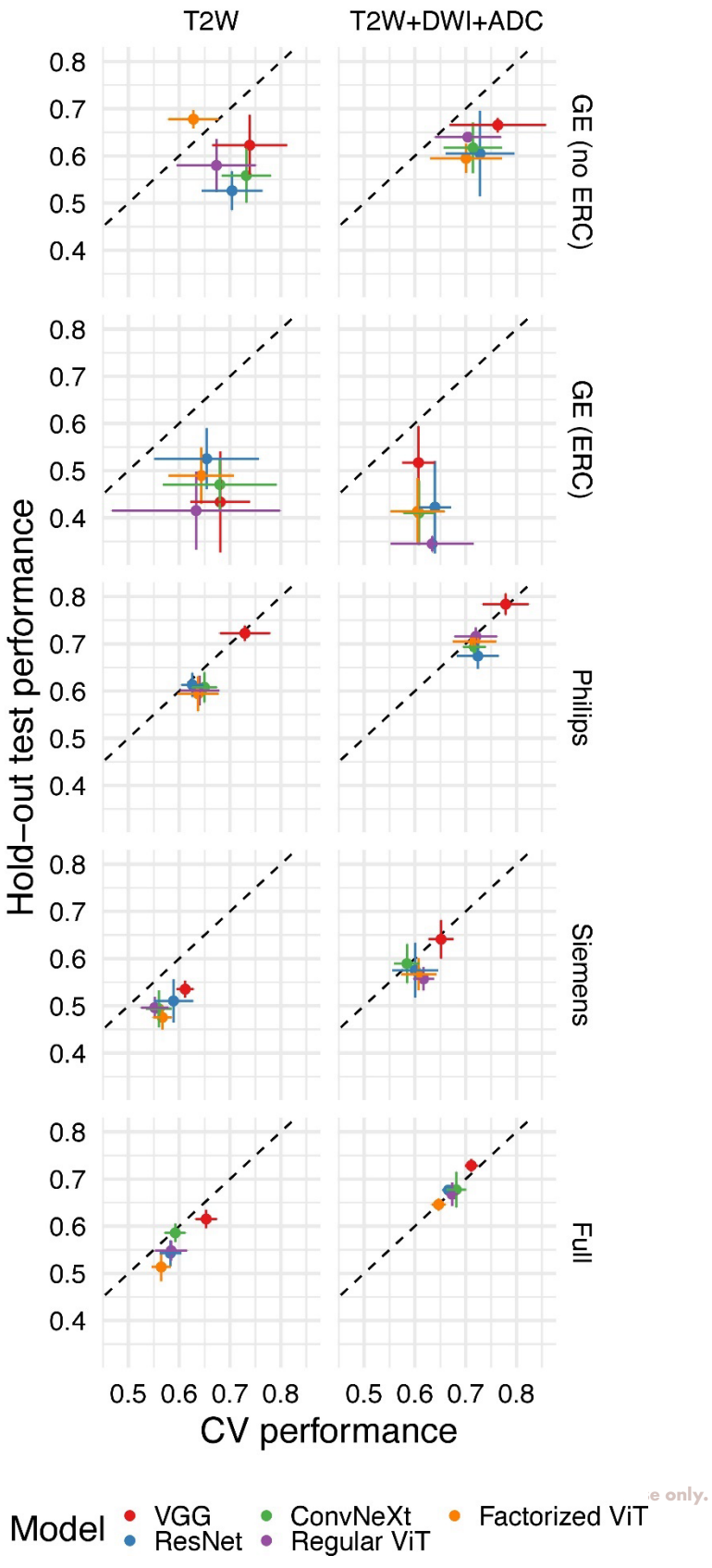


Figure 1: Flow diagram of studies according to selection criteria and bottlenecks. The five-fold cross-validation set and hold-out test set were selected using the least complete set of data so that models can be compared. Sequences that were not usable included sequences with missing slices or relevant metadata, such as image position, or wrong sequences, such as fat-suppressed T2-weighted images. T2 W = T2-weighted.

DWI = diffusion weighted imaging. ADC = apparent diffusion coefficient. PCa = prostate cancer.



Radiology: Artificial Intelligence



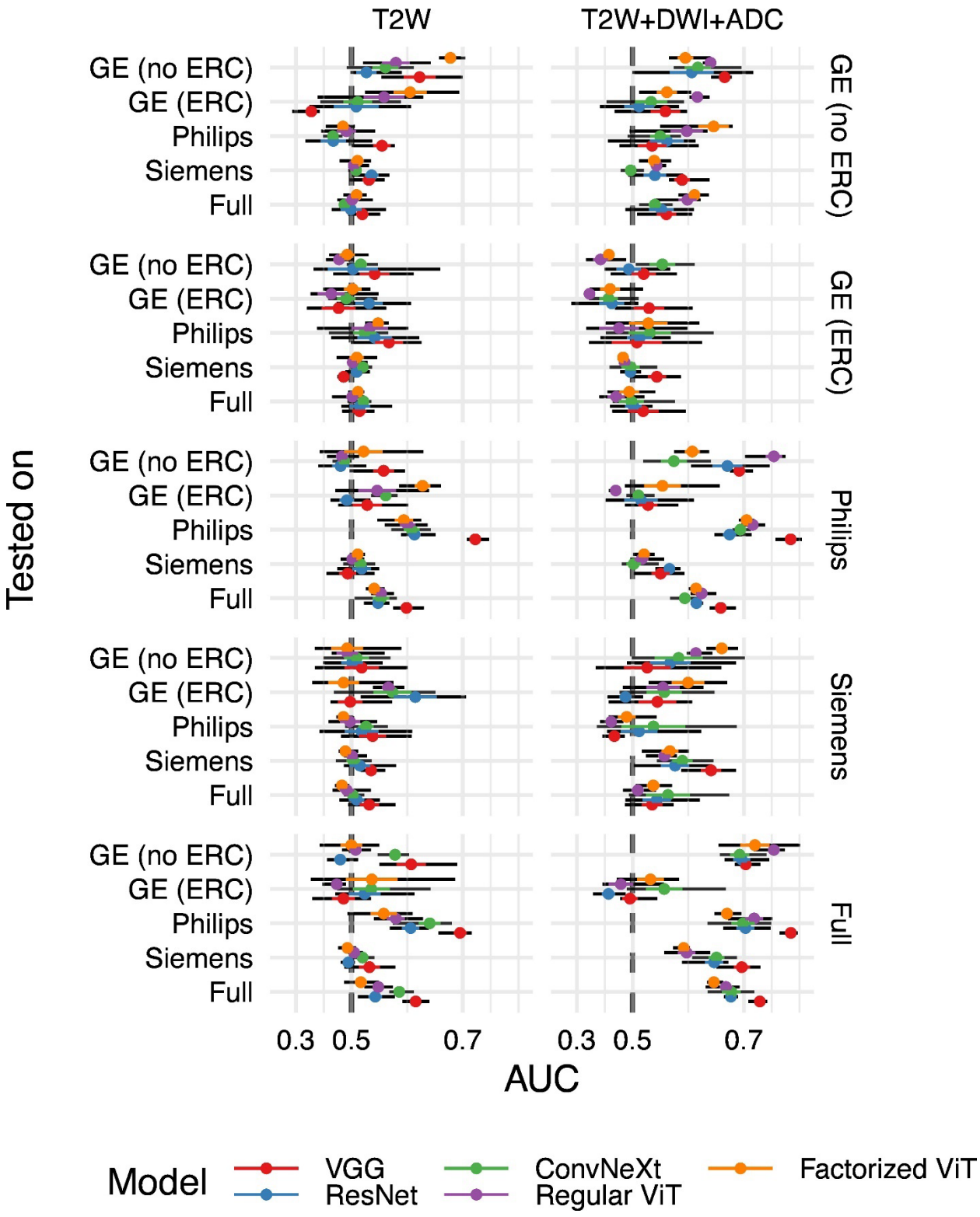


Figure 2: Model performance in cross-validation and hold-out test sets. **(A)** Cross-validation area under the receiver operating characteristic curve (AUC) values for different models on different manufacturer datasets. **(B)** Comparison of cross-validated (CV) and test AUCs. **(C)** Test AUCs of models trained and tested on different scanners. In all panels, colored points represent the average, and the vertical/horizontal lines represent standard errors. In a and c, the black horizontal lines represent the range of AUC scores. Facets in c represent the training and testing data subset. T2 W = T2-weighted. DWI = diffusion weighted imaging. ADC = apparent diffusion coefficient. ERC = endorectal coil.

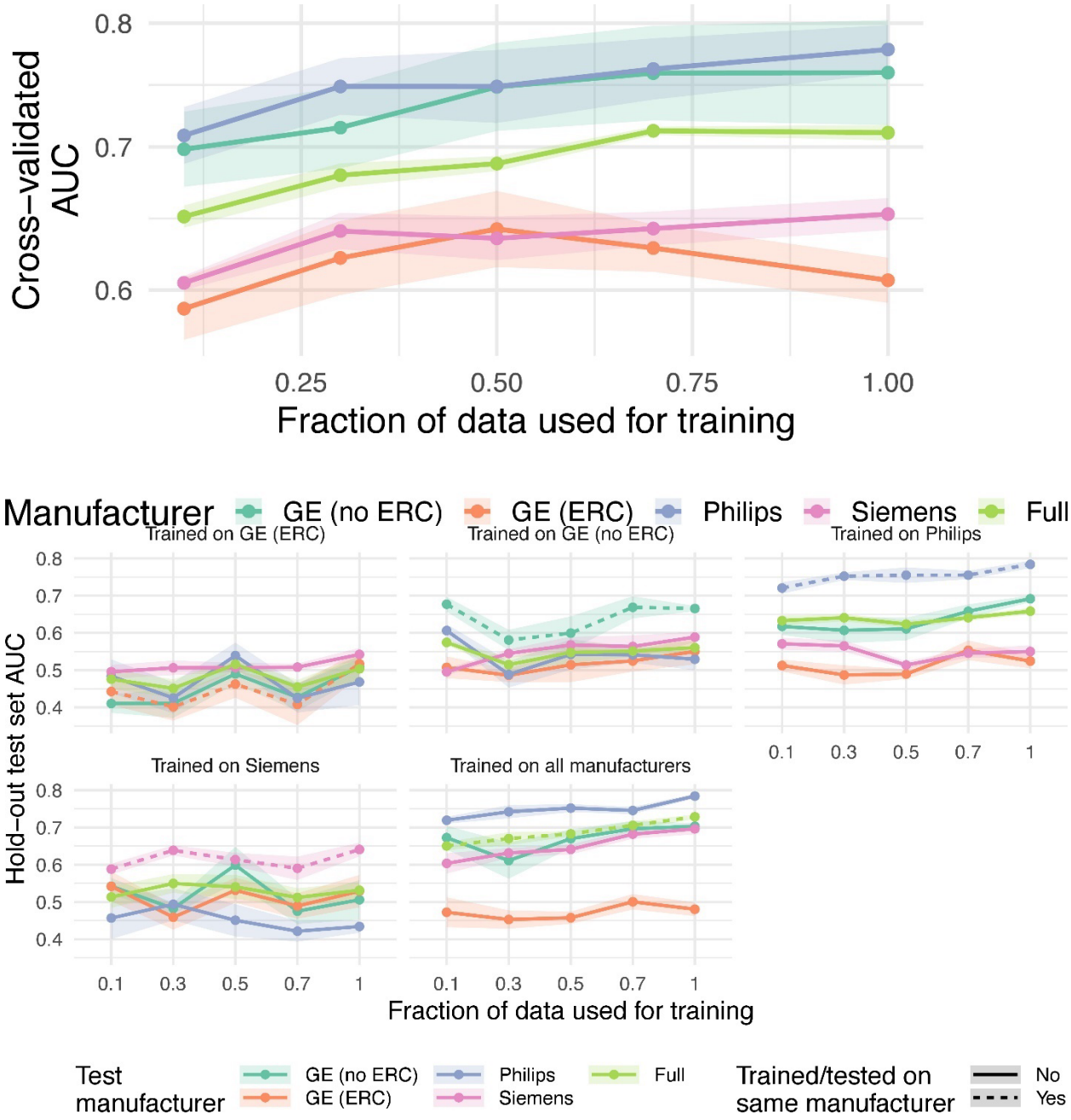
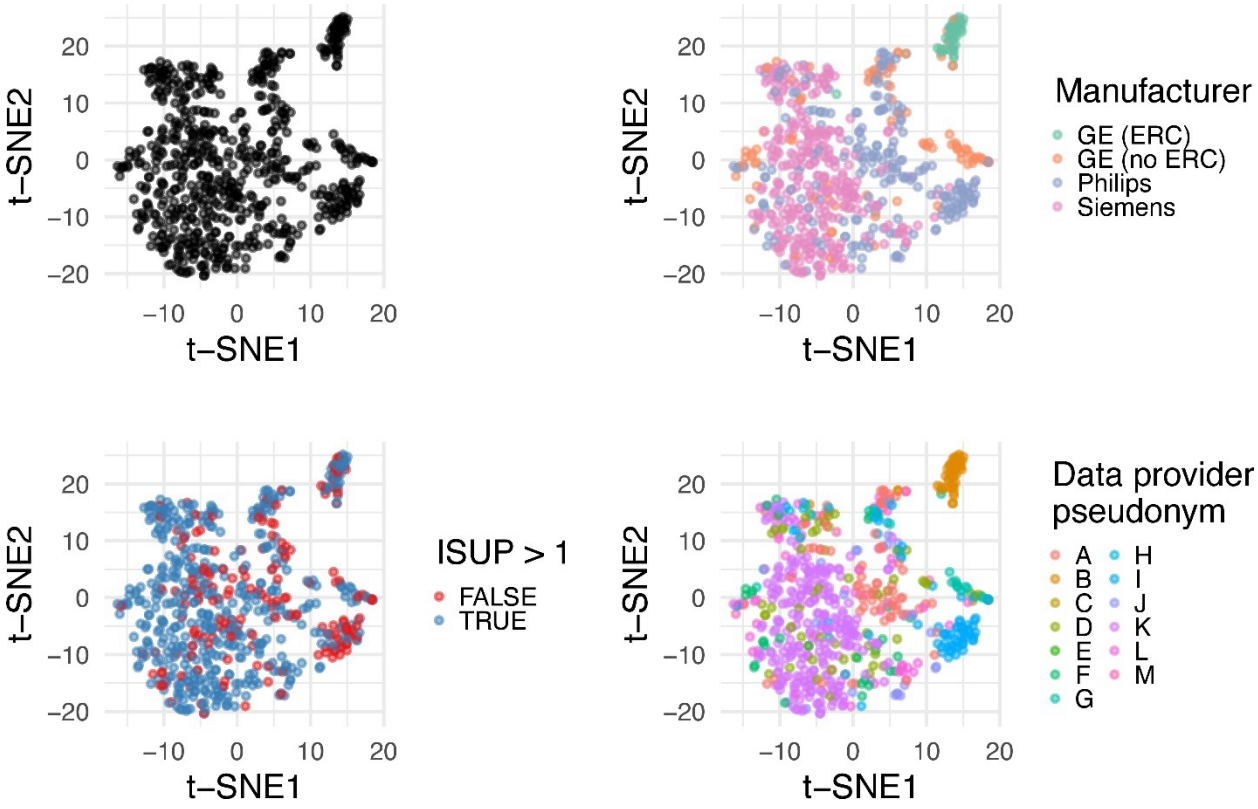


Figure 3: Learning curve analysis demonstrating the relationship between the amount of training data and cross-validation (CV) and hold-out test set AUCs. **(A)** Learning curve for CV AUC. **(B)** Learning curve for hold-out test set AUC stratified by test set manufacturer. Points represent the average estimates for all folds, and the shaded area represents the standard error for each estimate (each estimate is the average performance across 5 folds). Colors represent the testing manufacturer. Shaded areas repre-

sent the standard error. T2 W = T2-weighted. DWI = diffusion weighted imaging. ADC = apparent diffusion coefficient. ERC = endorectal coil.



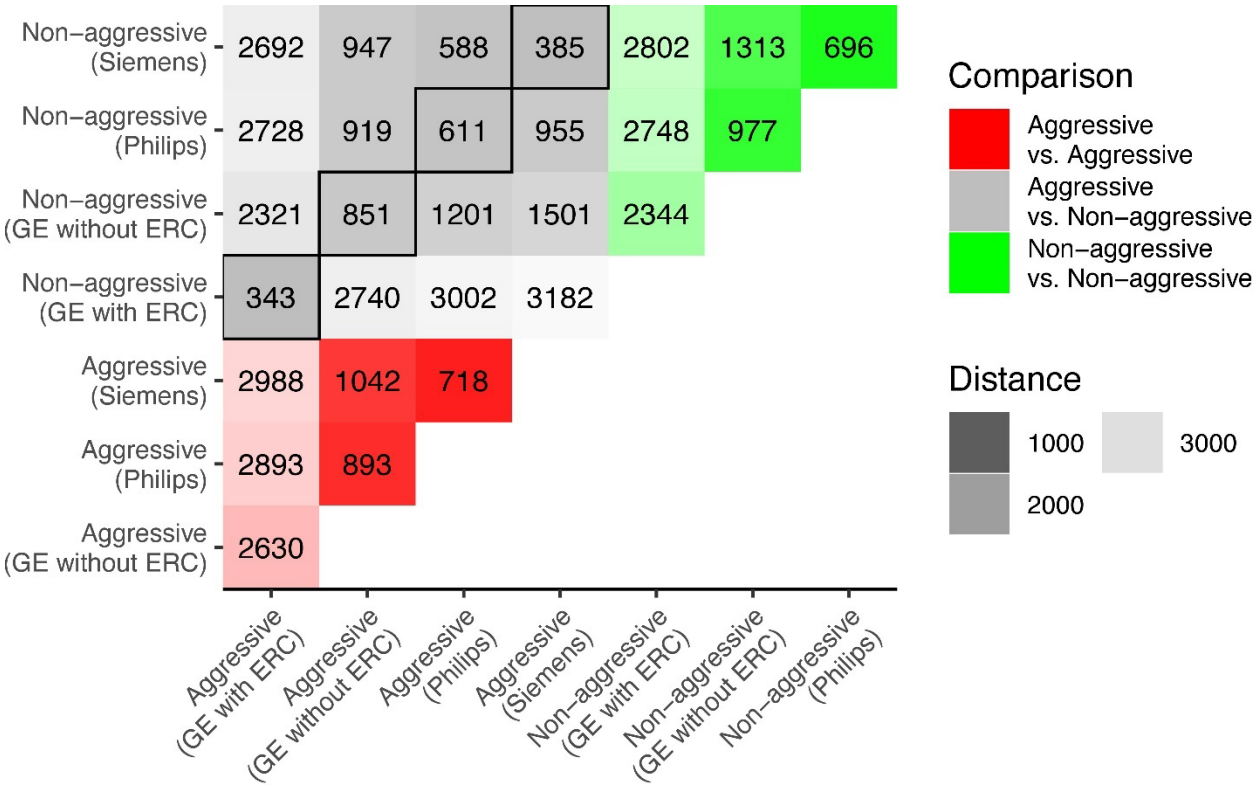


Figure 4: Analysis of deep feature distribution. **(A)** *t*-distributed stochastic neighbor embedding (*t*-SNE) visualization of all data (*n* = 669 studies; first column, first row) and stratified by manufacturer (second column, first row), by aggressiveness (first column, bottom row) and by data provider pseudonym (second column, bottom row). The embedding is the same across panels, and *t*-SNE1 and *t*-SNE2 represent the *t*-SNE dimensions. **(B)** Optimal transport dataset distance between different data subsets. The colors correspond to different aggressiveness comparisons, and the saturation of each grid cell corresponds to the distance between data subsets (higher saturation values, ie, cell grids that are “more green” or “more red,” imply greater dissimilarity). Cells surrounded by black lines correspond to between-class (aggressive, nonaggressive) and within-manufacturer/protocol (GE with ERC, GE without ERC, Philips, Siemens) comparisons. ERC = endorectal coil.

Table 1

Number of Studies Used in Training and Internal Validation (Hold-out Test Set), Stratified by Manufacturer

Manufacturer	ISUP = 1	ISUP = 2	ISUP = 3	ISUP = 4	ISUP = 5	Total
Train set						
GE (ERC)	143 (28.9%)	191 (38.6%)	88 (17.8%)	51 (10.3%)	22 (4.4%)	495
GE (no ERC)	216 (22.7%)	417 (43.9%)	170 (17.9%)	55 (5.8%)	92 (9.7%)	950
Philips	550 (37.0%)	525 (35.3%)	251 (16.9%)	87 (5.8%)	75 (5.0%)	1488
Siemens	515 (24.5%)	804 (38.2%)	342 (16.3%)	185 (8.8%)	256 (12.2%)	2102
Train set (with T2 W+DWI+ADC, total PSA level, age at baseline)						
GE (ERC)	86 (29.9%)	111 (38.5%)	51 (17.7%)	30 (10.4%)	10 (3.5%)	288
GE (no ERC)	74 (17.6%)	179 (42.6%)	77 (18.3%)	40 (9.5%)	50 (11.9%)	420
Philips	558 (37.6%)	534 (36.0%)	241 (16.2%)	83 (5.6%)	69 (4.6%)	1485
Siemens	516 (24.2%)	791 (37.0%)	356 (16.7%)	192 (9.0%)	280 (13.1%)	2135
Hold-out test set						
GE (ERC)	14 (31.8%)	16 (36.4%)	7 (15.9%)	5 (11.4%)	2 (4.5%)	44
GE (no ERC)	17 (19.8%)	37 (43.0%)	17 (19.8%)	6 (7.0%)	9 (10.5%)	86
Philips	84 (37.5%)	81 (36.2%)	36 (16.1%)	13 (5.8%)	10 (4.5%)	224
Siemens	69 (21.8%)	124 (39.2%)	55 (17.4%)	25 (7.9%)	43 (13.6%)	316
Total						
GE (ERC)	157 (29.1%)	207 (38.4%)	95 (17.6%)	56 (10.4%)	24 (4.5%)	539
GE (no ERC)	233 (22.5%)	454 (43.8%)	187 (18.1%)	61 (5.9%)	101 (9.7%)	1036
Philips	634 (37.0%)	606 (35.4%)	287 (16.8%)	100 (5.8%)	85 (5.0%)	1712
Siemens	584 (24.2%)	928 (38.4%)	397 (16.4%)	210 (8.7%)	299 (12.4%)	2418

Note.—Data reported as number (percentage), unless otherwise indicated. ADC = apparent diffusion coefficient, DWI = diffusion-weighted imaging, ERC = endorectal coil, ISUP = International Society of Urological Pathology, PSA = prostate-specific antigen, T2 W = T2-weighted.

Table 2

Cross-validation and Hold-out Test Set AUC Comparison between VGG and Other Models

Model	Difference to VGG	95% Confidence Interval	Adj. *P* Value
Cross-validation			
ResNet	−0.041	[-0.07,-0.01]	0.004
ConvNeXt	−0.041	[-0.07,-0.01]	0.005
Regular ViT	−0.049	[-0.08,-0.02]	<0.001
Factorized ViT	−0.061	[-0.09,-0.03]	<0.001
Hold-out test set			
ResNet	−0.016	[-0.030,-0.002]	0.014
ConvNeXt	−0.017	[-0.031,-0.004]	0.006
Regular ViT	−0.017	[-0.031,-0.004]	0.007
Factorized ViT	−0.024	[-0.037,-0.010]	<0.001

Note.—*P* values were calculated using Tukey Honest Significant Differences tests following a significant analysis of variance for linear Model I. ViT–vision transformer.

Table 3

Difference in Cross-validation AUC between Training and Validation Manufacturers (Manufacturer 1–Manufacturer 2)

Manufacturer 1	Manufacturer 2	Difference in AUC	95% Confidence Interval	Adj. *P* Value
GE (ERC)	Full	0.003	[-0.028, 0.034]	>0.99
GE (no ERC)	Full	0.073	[0.042, 0.104]	<0.001
GE (no ERC)	GE (ERC)	0.070	[0.039, 0.101]	<0.001
Philips	Full	0.058	[0.028, 0.089]	<0.001
Philips	GE (ERC)	0.055	[0.025, 0.086]	<0.001
Philips	GE (no ERC)	−0.015	[-0.046, 0.016]	0.715
Siemens	Full	−0.041	[-0.072,-0.011]	0.004
Siemens	GE (ERC)	−0.044	[-0.075,-0.014]	0.002
Siemens	GE (no ERC)	−0.115	[-0.146,-0.084]	<0.001
Siemens	Philips	−0.100	[-0.131,-0.069]	<0.001

Note.—Differences in performance, their respective intervals and the associated *P* values were calculated using the Tukey Honest Significant Differences test following a significant analysis of variance for manufacturer in Model I. ERC = endorectal coil.

Table 4

Hold-out Test Set AUC Differences between Models Trained on All Data and Models Trained and Tested on Specific Scanners

Training/Testing Manufacturer	Difference in AUC	95% Confidence Interval	Adj. *P* Value
GE (ERC)	−0.125	[-0.160,-0.090]	<0.001
GE (no ERC)	−0.027	[-0.062, 0.008]	0.363
Philips	0.026	[-0.009, 0.062]	0.374
Siemens	−0.045	[-0.080,-0.010]	0.004

Note.—Differences in performance, their respective intervals and the associated *P* values were calculated using the Tukey Honest Significant Differences test following a significant analysis of variance for training manufacturer in Model III. ERC = endorectal coil. Negative values imply better performance from models trained on all data, positive values imply otherwise.

Supplementary Methods

Deep-learning Model

Selection

These models were selected as they represented a small yet comprehensive selection of models between convolutional-and transformer-based. Additionally, the inclusion of ResNet and ConvNeXt models also allowed us to assess how residual connection-based models would perform on this task. Finally, we note that the hyperparameter selection was performed such that these models would be sufficiently similar to those originally published. Before training, a few ad hoc experimental runs were performed to ensure each model would converge when trained on all of the data (this was done without assessing the performance on any hold-out test set).

Hyperparameters

ViT and Factorized ViT

- batch size = 64 (32 per GPU)
- patch size = 16 × 16 × 4
- learning rate = $5 * 10^{-5}$
- weight decay = 0.1
- dropout rate = 0.1
- Image embedding was performed using a convolutional layer with kernel size and stride identical to patch size
- Learnable position embeddings
- Architecture: 8 transformer blocks, 512 tokens split across 8 heads (64 features *per* head)

ConvNeXt and VGG

- batch size = 128 (32 per GPU) and 64 (16 per GPU)
- warmup steps = 10
- learning rate = $5 * 10^{-4}$
- weight decay = 0.005
- dropout rate = 0.1
- VGG Architecture: 3 standard VGG blocks (details in Table A.1)
- ConvNeXt Architecture: 4 standard ConvNeXt blocks (details in Table A.1)

ResNet

- batch size = 64 (32 per GPU)
- warmup steps = 10
- learning rate = 0.001
- weight decay = 0.005
- dropout rate = 0.1
- GeLU activation function and batch normalisation
- Structure: four residual blocks, each composed of 3 residual layers with 32, 64, 128 and 256 features

Hybrid Model

A hybrid model is composed of two distinct parts—a standard image classification model, identical to those described in the Methods section, a tabular network, which takes tabular (a vector of features) features as input and produces a probability score. Both the image and tabular networks produce a nonnormalized probability score (logit)—p_{image} and $p_{clinical}$ for image and clinical predictions, respectively—and these are combined using a trainable weight

$w_{comb} = sigmoid\left(w'_{comb}\right)$ to obtain a final prediction $p_{final} = sigmoid(w_{comb} \times p_{image} + (1-w_{comb}) \times p_{clinical})$. In other words, a hybrid model simply combines the nonnormalized probability produced by the image network with the nonnormalized probability produced by the tabular network using a learnable weight. The sigmoid of this linear combination is then used to obtain a value between 0 and 1.

Details Regarding Factorized ViT and Convolutional Embedding

The factorized ViT architecture was highly similar to that of a 3D ViT with the separation of within-and between-slice feature extraction. To clarify details regarding ViT-based architectures, we first introduce the forward pass for ViT architectures. Then, we introduce the forward pass in factorized ViT architectures. Finally, we explain how the embedding can be replaced by a convolutional operation in both regular and factorized ViT architectures.

Embedding and transformer application in ViT architectures.—

Given a three-dimensional input with c channels $I \in R^{c \times h \times w \times d}$, a patch size $P = [x, y, z]$ and number of patches $n_p = h/x \times w/y \times d/z \times c$:

1. The input I is "embedded" such that $I_{linear} \in R^{n_p \times (x \times y \times z \times c)}$. This ensures that the input is adequate for a transformer (1) — each patch is a "token" and the number of values in this token is its embedding size
2. (optional) a positional embedding can be added to the input to incorporate spatial information
3. (optional) the embedding size can be altered using a linear projection

A Sequence of *t* Transformers Is Applied to This Input

Embedding and transformer application in factorized ViT architectures.—

Given, a three-dimensional input with c channels $I \in R^{c \times h \times w \times d}$, a patch size $P = [x, y]$ and number of patches $n_p = \frac{h}{x} \times \frac{w}{y} \times c$:

1. The input I is "embedded" such that $I_{withinslice} \in R^{z \times n_p \times (x \times y \times c)}$. Here, the true number of tokens is np × z, while the embedding size is x × y × c
2. (optional) same as for regular ViT
3. (optional) same as for regular ViT
4. Afterwards, a sequence of $t_{with\ in\ slice}$ transformer blocks is applied to $I_{with\ in\ slice}$. It should be noted that, at this stage, transformer blocks operate only on within slice information
5. An aggregation operation turns $I_{with\ in\ slice}$ into $I_{betweenslice} \in R^{n_p \times (x \times y \times c)}$. This aggregation operation can be, for instance, the average over the z axis or a classification token
6. A set of $t_{with\ in\ slice}$ transformer blocks is applied to $I_{between\ slice}$

To ensure that a similar number of parameters is used in regular ViT and factorized ViT architectures, we project both I_{linear} and $I_{with\ in\ slice}$ to have the same positional embedding and embedding sizes (this is achieved as specified in 3., using a linear projection) and set $t_{with\ in\ slice}$ and

$t_{between\ slice}$ such that $t = t_{with\ in\ slice} + t_{between\ slice}$. For simplicity, and since t is a pair number, $\frac{t}{2} = t_{withinslice} = t_{betweenslice}$.

Convolutional embedding.—

Replacing the image embedding operation described above with a 3D convolutional layer is relatively simple and can be used to replace steps 1 and 3 by an operation that calculates an embedding of size E for every patch. Indeed, this can be achieved for regular ViT architectures by i) applying a 3D convolutional layer with stride (patch size) $P = [x, y, z]$ to the input $I \in R^{c \times h \times w \times d} \rightarrow I_E \in R^{E \times \frac{h}{x} \times \frac{w}{y} \times \frac{d}{z}}$ and ii) reshaping it such that $I_E \in R^{E \times \frac{h}{x} \times \frac{w}{y} \times \frac{d}{z}} \rightarrow I_{E_{reshaped}} \in R^{\left(\frac{h}{x} \times \frac{w}{y} \times \frac{d}{z}\right) \times E}$. For factorized ViT architectures, a 2D convolutional layer with stride (patch size) $P = [x, y]$ is applied to the input $I \in R^{c \times h \times w \times d} \rightarrow I_E \in R^{E \times \frac{h}{x} \times \frac{w}{y} \times d}$, which is then reshaped to $I_E \in R^{E \times \frac{h}{x} \times \frac{w}{y} \times d} \rightarrow I_{E_{reshaped}} \in R^{d \times \left(\frac{h}{x} \times \frac{w}{y}\right) \times E}$.

Details Regarding Initialization and Training

Model Implementation and Initialization

All models are implemented using PyTorch (2) in an in-house library (https://github.com/CCIG-Champalimaud/adell-mri) using the default initialisation in PyTorch for all layers excluding positional embeddings (initialized with a truncated normal distribution with mean 0, standard deviation 0.02 and lower and upper bounds set to-2 and 2, respectively).

Augmentations

We used a wide array of augmentations from MONAI (3), namely:

- Identity (no transform)
- Random contrast adjustment ($\gamma = [0.5, 1.5]$)
- Random standard shift in intensity (range = $[-0.1, 0.1]$)
- Random shift in intensity (range = $[-0.1, 0.1]$)
- Random Rician noise (std = 0.02)
- Random bias field (degree = 3 (T2W-only))
- Affine transforms (translation range = [4, 4, 1], rotation range = $\left[\frac{\pi}{16}, \frac{\pi}{16}, \frac{\pi}{16}\right]$)
- Horizontal flip

Each study is augmented with one of the aforementioned transforms, which is picked at random with uniform probability.

Training

Each model is trained for 100 epochs and the best performing checkpoint on the validation set is picked for further evaluation (this avoided the optimization of an early stopping criteria). For training, we used an AdamW optimizer (4) with the learning rate/weight decay parameters specified above; a binary cross-entropy loss was used.

Dataset Distances

To calculate dataset distances, we use the formulation of the Wasserstein 2-distance noted by Alvarez-Melis and Fusi (5). In short, for a given dataset of feature vectors D with n samples and f features, we calculate the Wasserstein 2-distance between any two nonoverlapping subsets $D_A \in R^{n_A \times f}$ and $D_B \in R^{n_B \times f}$. To do this, we first assume that the features in each subset follow a multivariate normal distribution and calculate the mean (μ_A and μ_B) and covariance (Σ_A and Σ_B) of each subset. Finally, using the fact that the Wasserstein 2-distance of two multivariate Gaussian distributions can be calculated as $W_2^2(A,B) = \vee \mu_A - \mu_B \vee_2^2 + tr\left(\sqrt{\Sigma_A - \Sigma_B + \sqrt{\Sigma_A}\Sigma_B\sqrt{\Sigma_A}}\right)$, we can feasibly calculate a simple and closed-form estimate of the optimal transport (here calculated as the Wasserstein 2-distance) between two datasets.

References

1. Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. In: Guyon I, Luxburg UV, Bengio S, et al, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc: 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

2. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, eds. Advances in Neural Information Processing Systems. Curran Associates, Inc: 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

3. MONAI Consortium. MONAI: Medical Open Network for AI. Published 2023.

4. Loshchilov I, Hutter F. Fixing Weight Decay Regularization in Adam. https://openreview.net/forum?id=rk6qdGgCZ. Published February 15, 2018. Updated October 14, 2024. Accessed DATE.

5. Alvarez-Melis D, Fusi N. Geometric dataset distances via optimal transport. Adv Neural Inf Process Syst 2020;33:21428–21439. https://proceedings.neurips.cc/paper_files/paper/2020/file/f52a7b2610fb4d3f74b4106fb80b233d-Paper.pdf.

6. Hiremath A, Shiradkar R, Fu P, et al. An integrated nomogram combining deep learning, Prostate Imaging-Reporting and Data System (PI-RADS) scoring, and clinical variables for identifi-

cation of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. Lancet Digit Health 2021;3(7):e445–e454.

7. Liu S, Zheng H, Feng Y, Li W. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. Medical Imaging 2017: Computer-Aided Diagnosis. SPIE, 2017; 581–584.

8. Pachetti E, Colantonio S, Pascali MA. On the Effectiveness of 3D Vision Transformers for the Prediction of Prostate Cancer Aggressiveness. Image Analysis and Processing ICIAP 2022 Workshops. Springer International Publishing, 2022; 317–328.

9. Bertelli E, Mercatelli L, Marzi C, et al. Machine and Deep Learning Prediction Of Prostate Cancer Aggressiveness Using Multiparametric MRI. Front Oncol 2022;11:802964.

10. Khosravi P, Lysandrou M, Eljalby M, et al. Biopsy-free prediction of prostate cancer aggressiveness using deep learning and radiology imaging. medRxiv 2019.12.16.19015057 [preprint] https://www.medrxiv.org/content/10.1101/2019.12.16.19015057. Posted December 19, 2019. Accessed DATE.

11. Bosma JS, Saha A, Hosseinzadeh M, Slootweg I, de Rooij M, Huisman H. Semisupervised Learning with Report-guided Pseudo Labels for Deep Learning-based Prostate Cancer Detection Using Biparametric MRI. Radiol Artif Intell 2023;5(5):e230031.

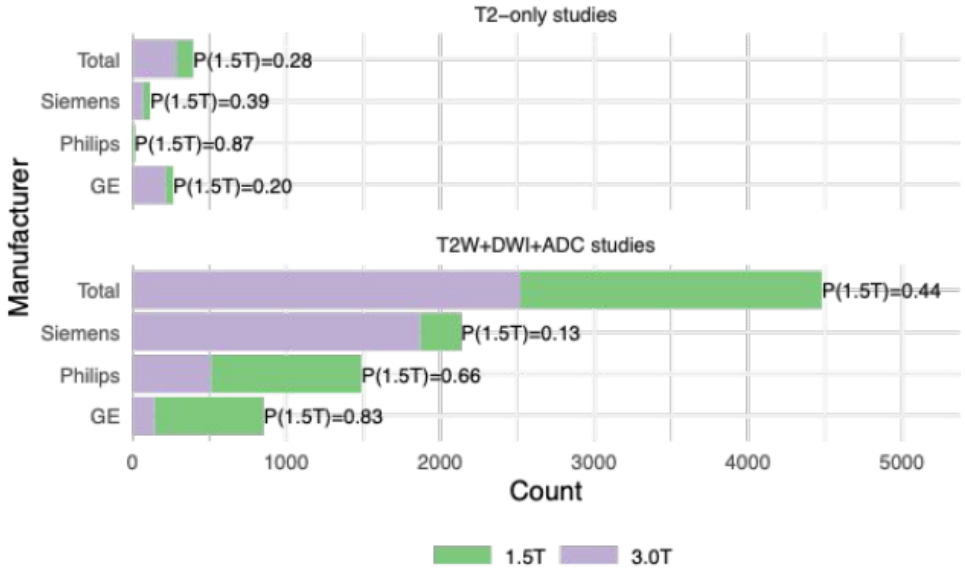


Figure S1: Magnetic field strength by manufacturer, stratified by studies with only T2 W sequences and by studies with all three (T2 W+DWI+ADC) sequences.

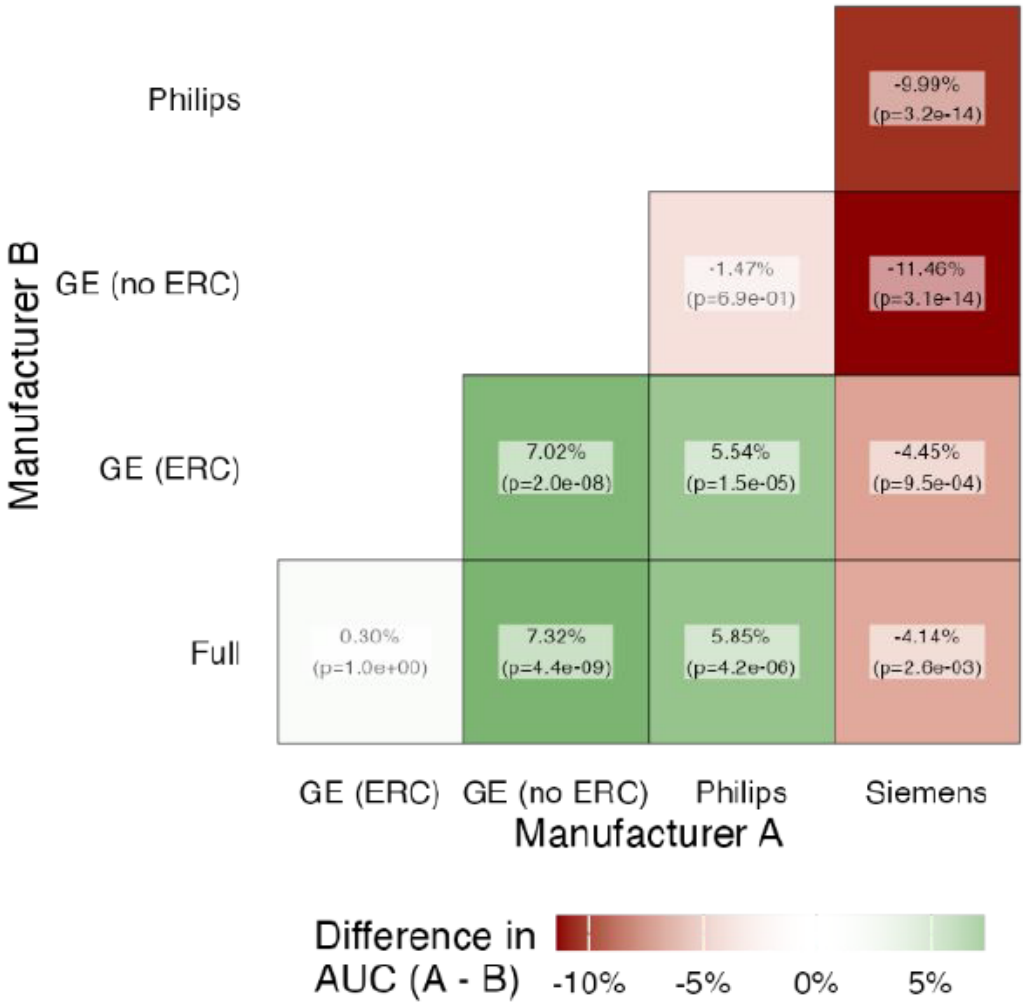


Figure S2: Difference between the average cross-validation AUC of models trained on different manufacturers. The color corresponds to the difference, whereas the text on each cell corresponds to the difference and to the *P* value of a sum of ranks test comparing both. Black text implies statistical significance ($P < .05$ according to a Tukey Honest Statistical Differences test following a statistically significant analysis of variance for manufacturer in Model I), whereas gray text signals the absence of statistical significance.

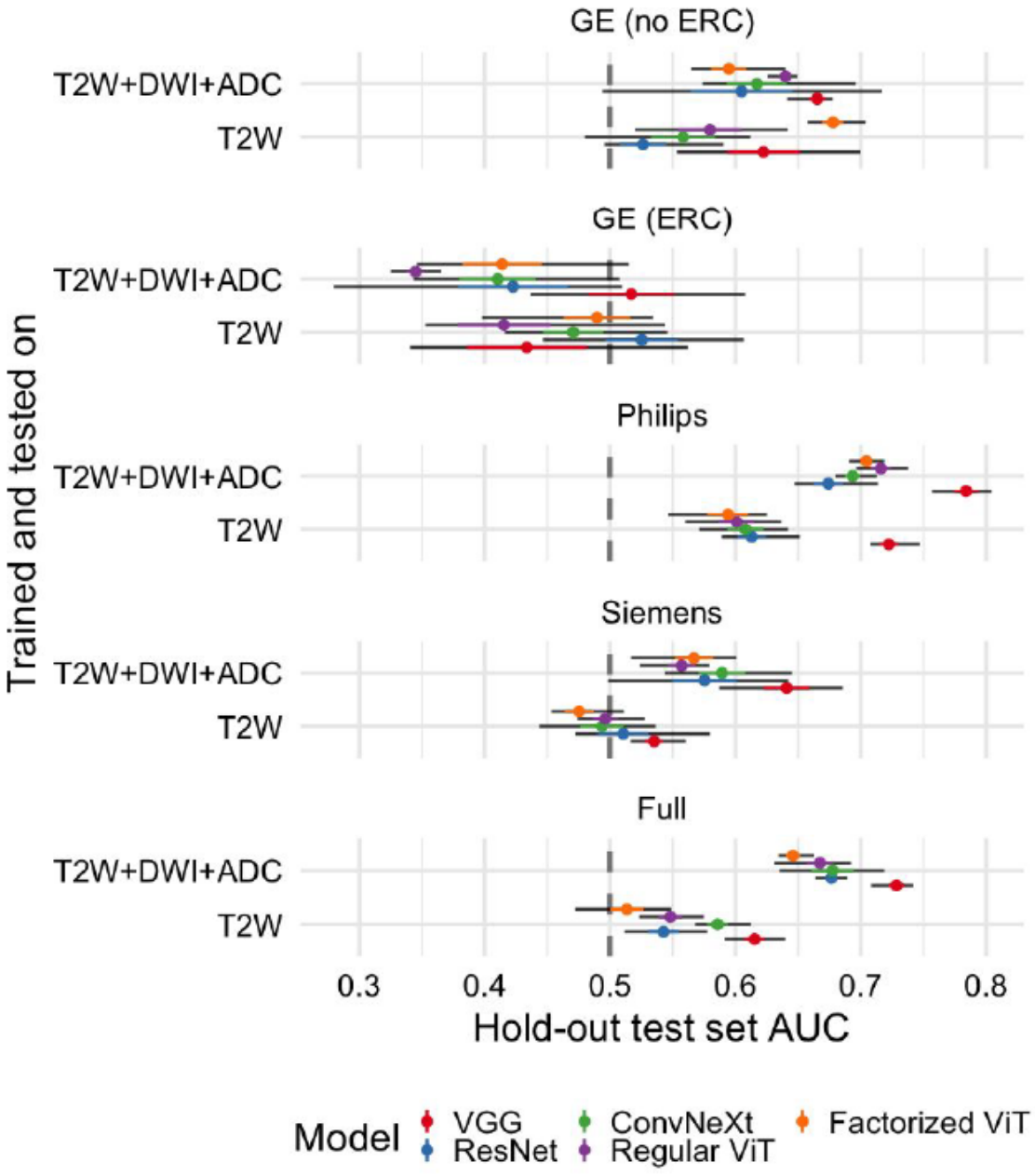


Figure S3: Test area under the curve (AUC) of different models on different manufacturer testing datasets.

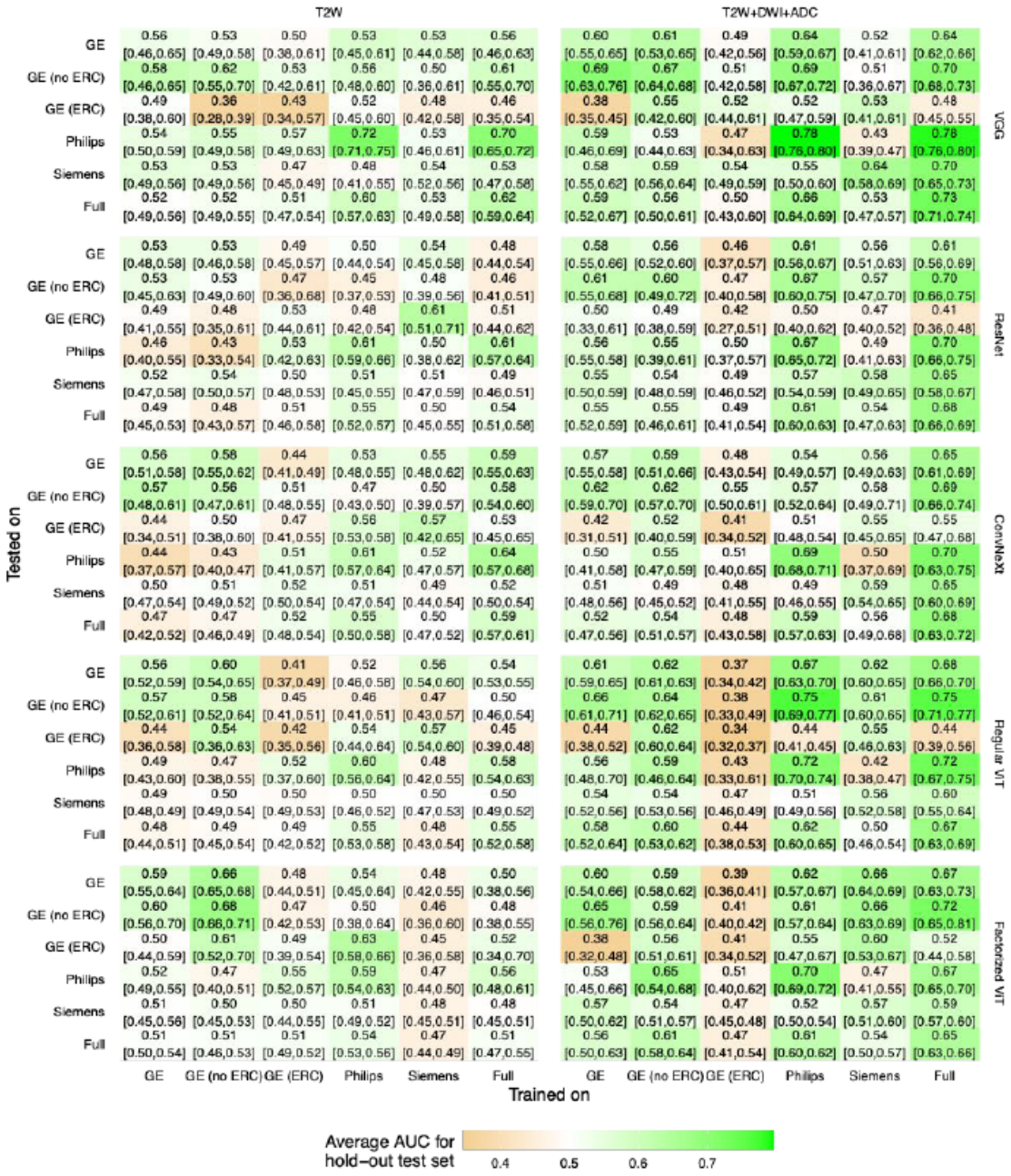


Figure S4: Test AUC of models trained and tested on different scanners. The text corresponds to the average, minimum and maximum AUC values (minimum and maximum values are between brackets) and the color corresponds to the average AUC value.

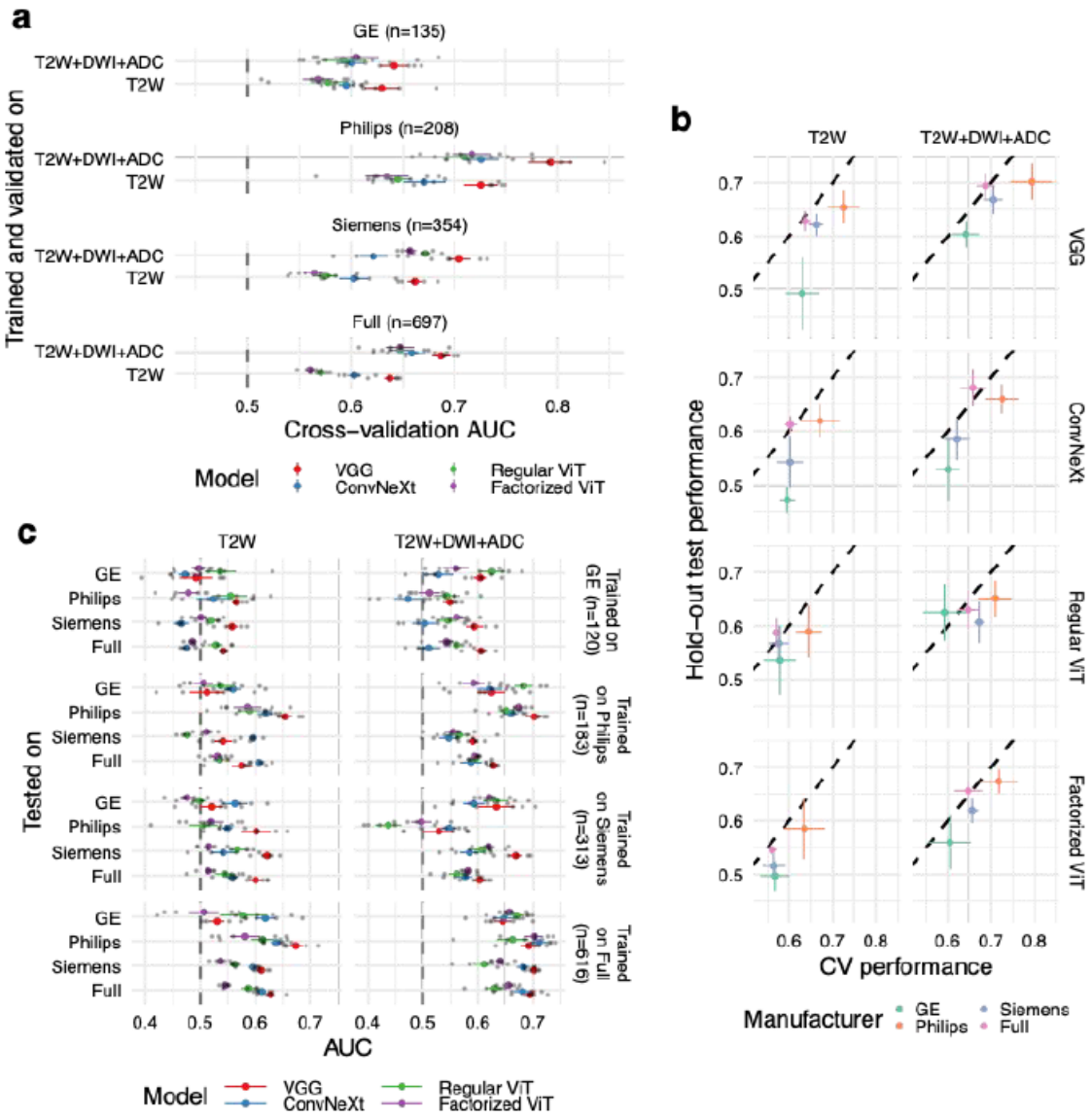


Figure S5: Model performance in CV and hold-out test sets for alternative ISUP categorization (ISUP = 1–2 versus ISUP = 3–5). **(A)** Cross validation area under the curve (AUC) for different models on different manufacturer datasets. **(B)** Comparison of cross-validated (CV) and test area under the curve (AUC). **(C)** Test AUC of models trained and tested on different scanners. In all panels colored points represent the average and the vertical/horizontal lines represent standard errors. In a and c the black horizontal lines represent the range of AUC scores. Facets in c represent the training and testing data subset.

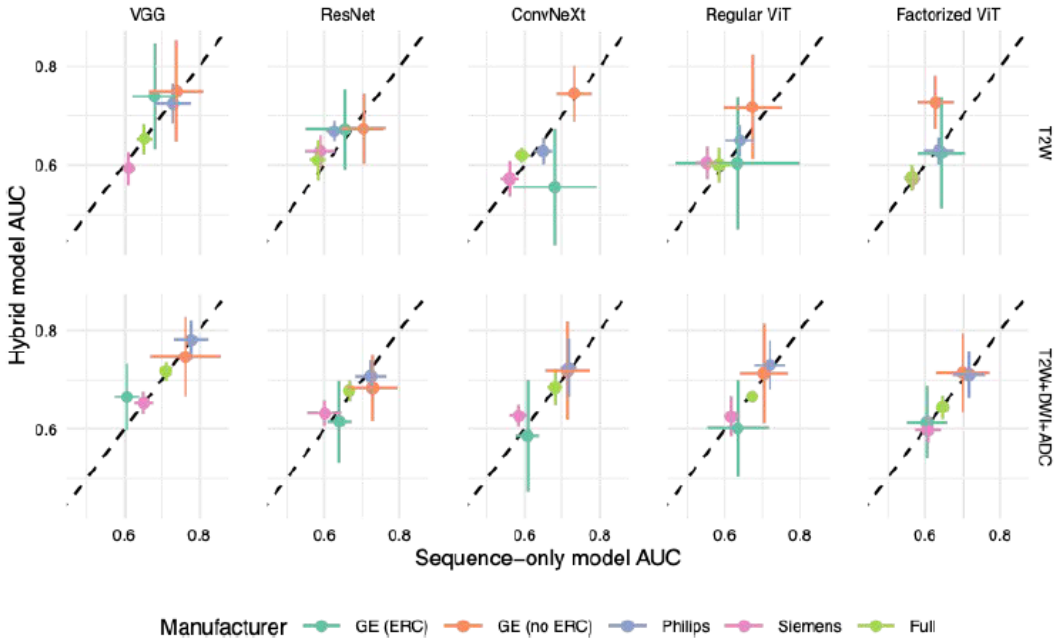


Figure S6: Comparison of AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axes.

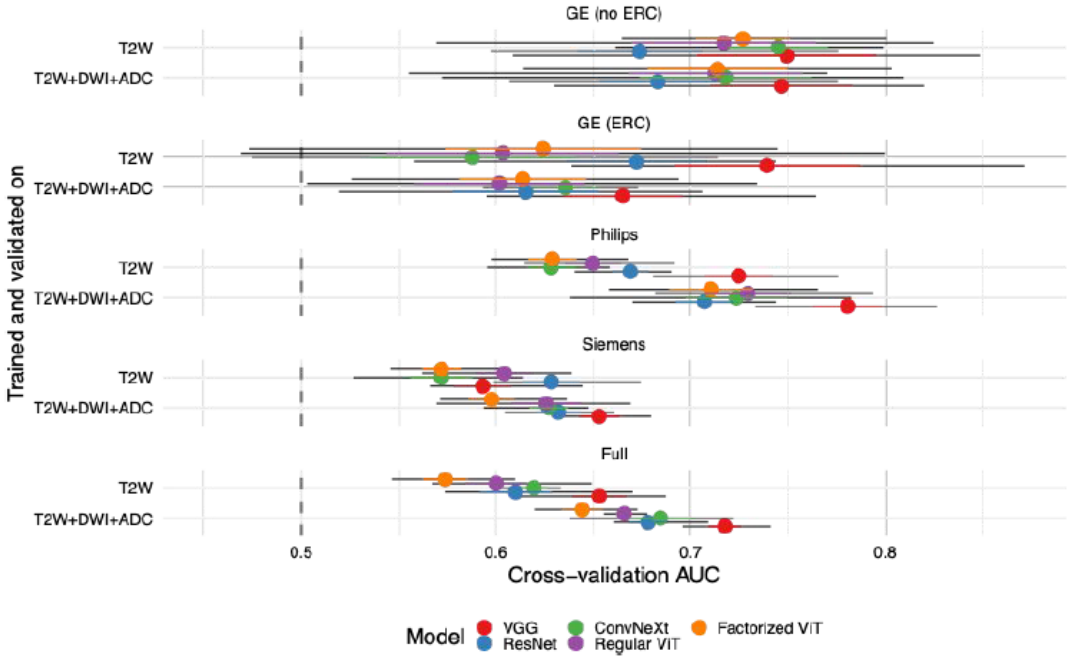


Figure S7: Cross validation area under the curve (AUC) of different hybrid models (bpMRI + clinical) on different manufacturer datasets.

Radiology: Artificial Intelligence

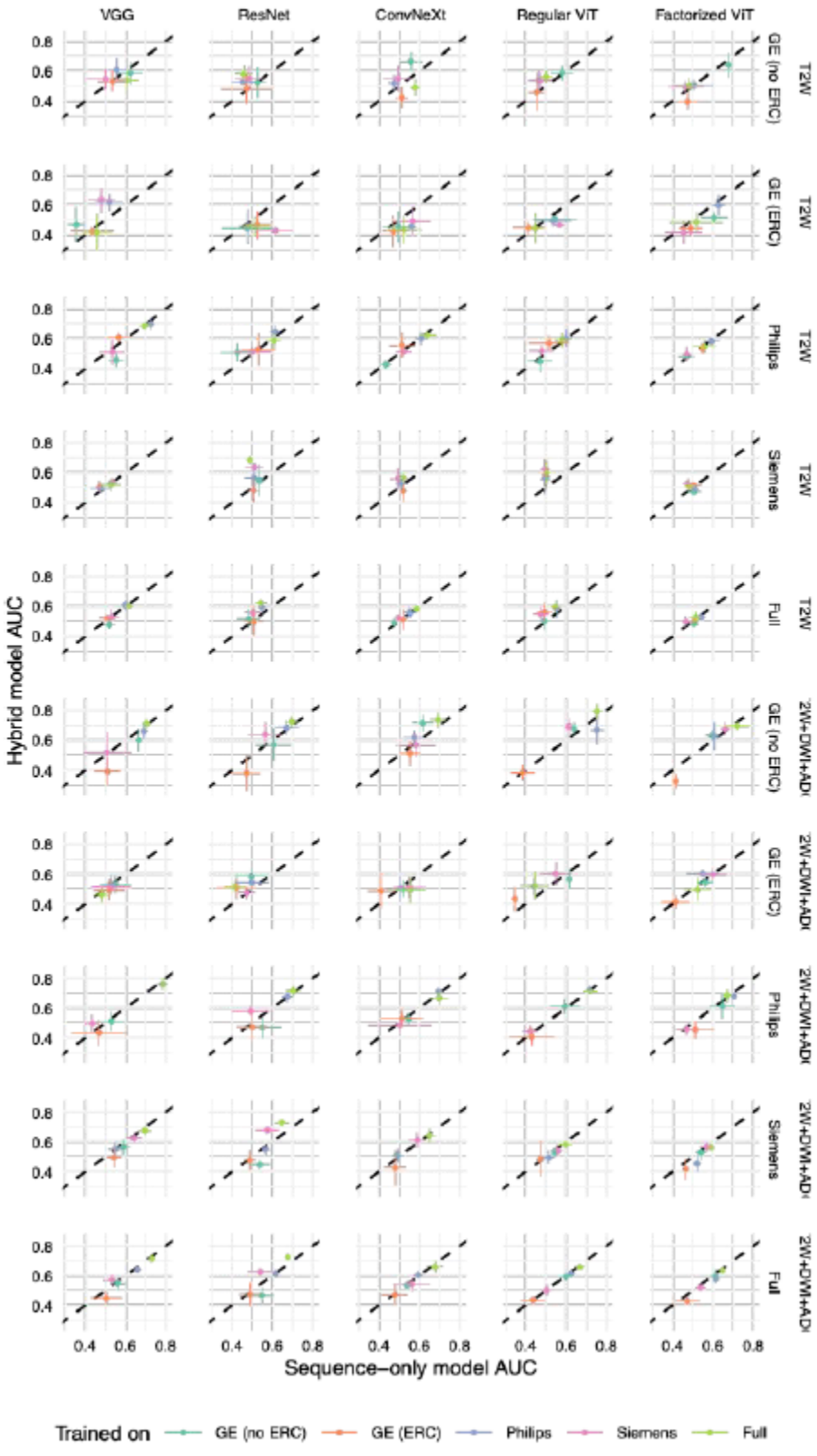


Figure S8: Comparison of hold-out test set AUC for sequence-only and hybrid models. Each point represents the average AUC, whereas the vertical and horizontal error bars represent the mean with the addition and subtraction of the standard error, respectively. The diagonal dashed line represents equality between both axis.

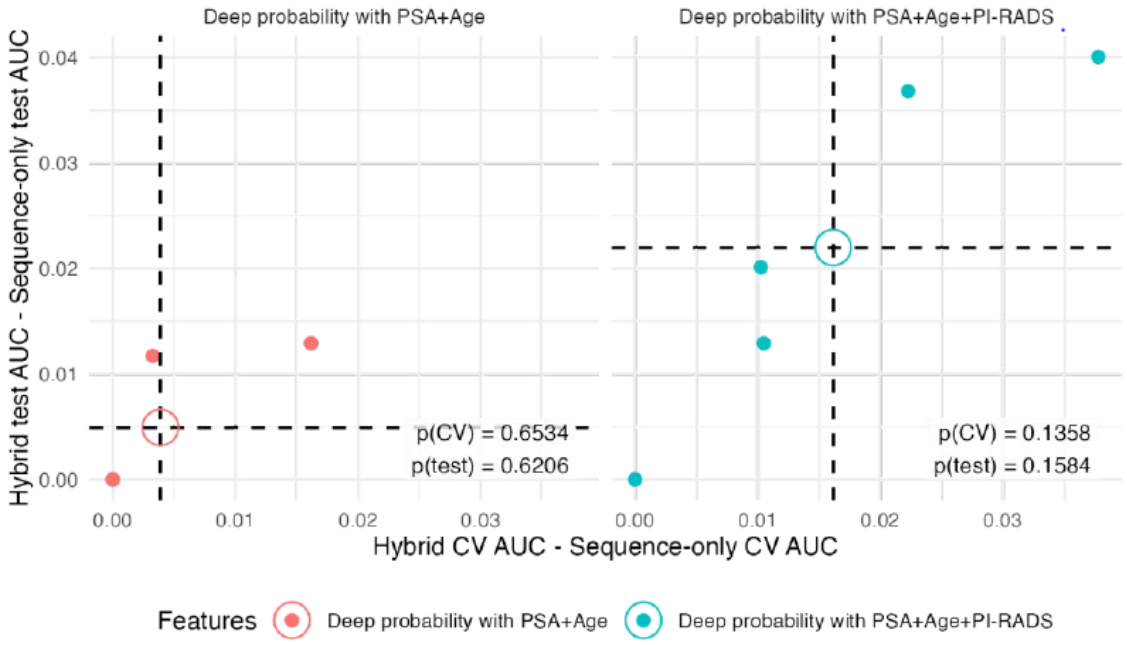


Figure S9: Difference between sequence-only and elastic net-regularized linear classification model AUC. Both CV (x axis) and test (y axis) AUC is represented, with the average value noted as circle at the intersection of the dashed lines. The *P* values shown in the figure were obtained using a paired Student *t* test.

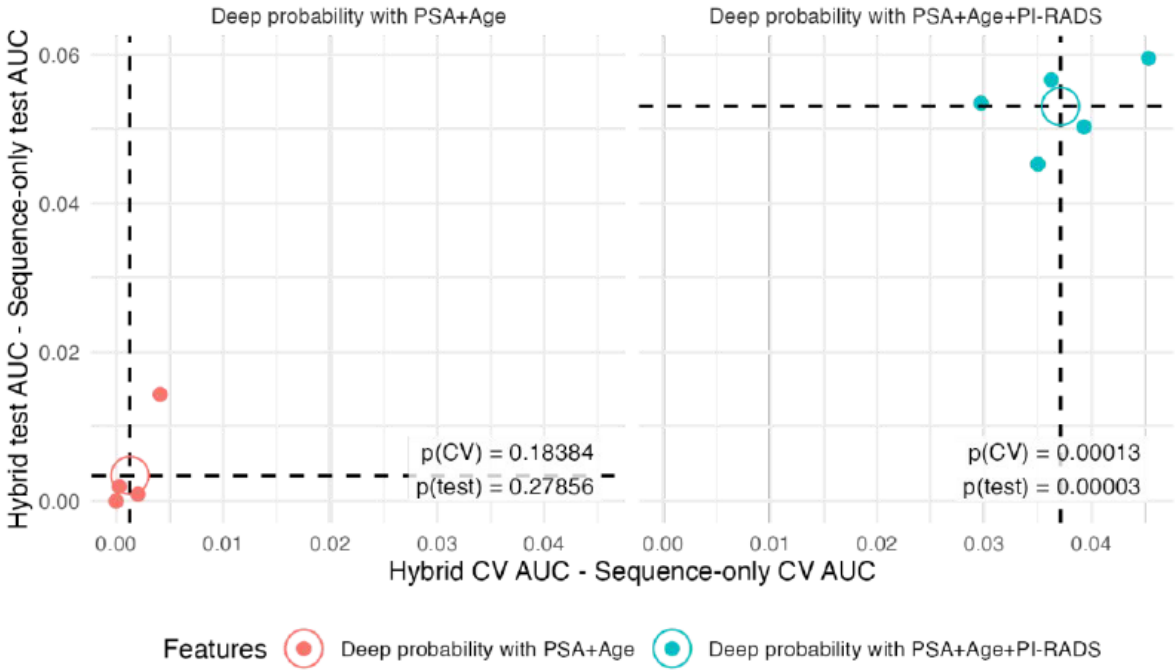


Figure S10: Difference between sequence-only and elastic net-regularized linear classification model AUC for alternative ISUP categorization (ISUP = 1–2 versus ISUP = 3–5). Both CV (x axis) and test (y axis) AUC is represented, with the average value noted as circle at the intersection of the dashed lines. The *P* values shown in the figure were obtained using a paired Student *t* test.

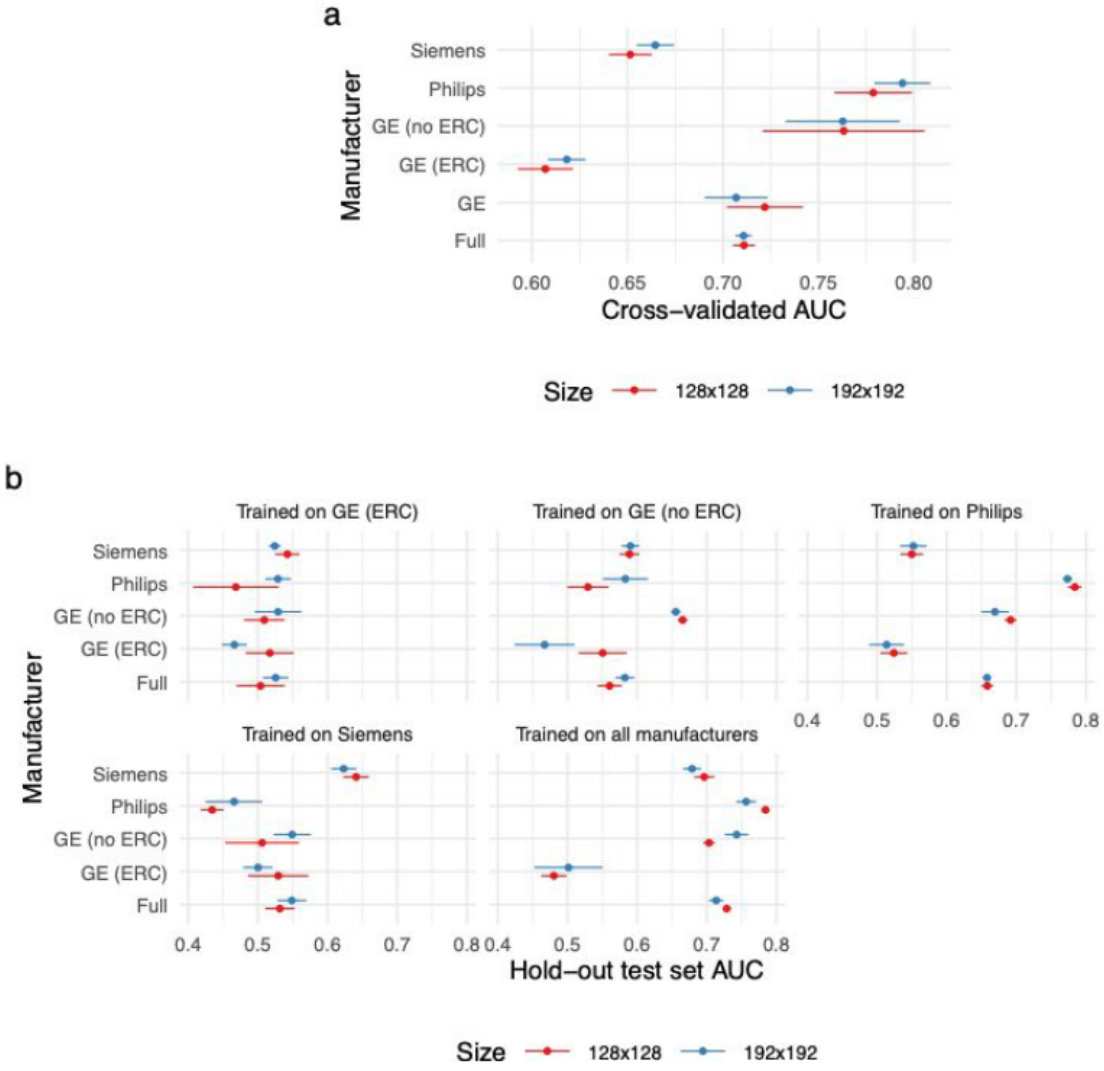


Figure S11: Impact of crop size on cross-validation **(A)** and hold-out test set AUC **(B)**.

Table A1

Detailed Architecture Specifications for the VGG and ConvNeXt Models

	VGG	ConvNeXt
block 1	Conv (64, 3 × 3x3) Conv (128, 3 × 3x3) Pool (max, 2 × 2x2)	Conv (32, 7 × 7x5) × 6 Pool (max, 2 × 2x2)
block 2	Conv (128, 3 × 3x3) Conv (256, 3 × 3x3) Pool (max, 2 × 2x2)	Conv (64, 7 × 7x5) × 6 Pool (max, 2 × 2x2)
block 3	Conv (256, 3 × 3x3) Conv (512, 3 × 3x3) Pool (max, 2 × 2x2)	Conv (128, 5 × 5x3) × 6 Pool (max, 2 × 2x2)
block 4		Conv (256, 5 × 5x3) × 6 Pool (max, 2 × 2x2)

Table C.1

Nonextensive Summary of Other Deep-learning Studies on Prostate Cancer Aggressiveness Prediction

Number Examinations	No. Centers	Lesion Location Requirements	Best AUC	Target	Ref.
592	5	bounding box	0.81	ISUP = 1 versus ISUP > 1	(6)
341	1	bounding box	0.84	ISUP = 1 versus ISUP > 1	(7)
99	1	none	0.78	ISUP = 1 versus ISUP > 1	(8)
112	1	lesion location	0.88	ISUP = 1,2 versus ISUP > 2	(9)
376	4		0.86	ISUP = 1,2 versus ISUP > 2	(10)
8056	2	segmentation masks	0.86	ISUP = 1 versus ISUP > 1	(11)

Table C.2

Centres Participating in Data Provision for Prostatenet

Centres	Country
Champalimaud Foundation	Portugal
Candiolo Cancer Institute	Italy
General Anti-Cancer and Oncological Hospital of Athens	Greece
Hacettepe University, School of Medicine, Department of Radiology	Turkey
Fundacion Para La Investigacion Del Hospital Universitario La Fe De La Comunidad Valenciana	Spain
Fundacio Institut D'Investigacio Biomedica De Girona Doctor Josep Trueta	Spain
Institut Paoli-Calmettes	France
JCC Diagnostic Imaging	Portugal
National Cancer Institute	Lithuania
QS Instituto de Investigacion e Innovacion SL	Spain
RadboudUMC	Netherlands
Royal Marsden National Health Service Trust	United Kingdom
University of Pisa	Italy

Table C.3

Mean and Standard Deviation (Std.) For Age and Prostate Specific Antigen (PSA) Stratified by Validation Folds and Hold-out Test Set

Fold	Age	Age Std.	PSA	PSA Std.
CV1	66.60	8.41	11.49	14.98
CV2	66.13	7.21	12.55	18.69
CV3	66.62	8.33	12.44	29.52
CV4	66.19	7.83	15.47	56.69
CV5	65.85	7.90	14.31	46.15
Hold-out test set	66.47	8.39	13.17	24.99

Table C.4

PI-RADS Frequency for the Data Used During This Study

PI-RADS	Frequency
0	4.1%
1	0.9%
2	2.1%
3	7.8%
4	44.1%
5	40.9%

RSNA

Impact of Scanner Manufacturer, Endorectal Coil Use, and Clinical Variables on Deep Learning-Assisted Prostate Cancer Classification Using Multiparametric MRI

Key Result

Prostate cancer (PCa) aggressiveness could be predicted using biparametric MRI (bpMRI) and deep learning with negligible expert input, but performance was impacted by scanner manufacturer and scan protocol.

Dataset:

- 5,478 cases from a PCa bpMRI multicenter dataset (ProstateNet)

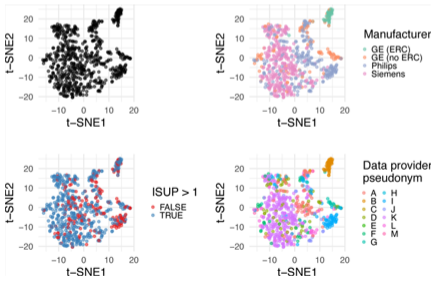
Methods:

- Five different models were trained on 4,328 bpMRI cases.
- The impact of different factors—scanner manufacturer, endorectal coil use, and addition of clinical features—on model performance was assessed.

Results:

- The models predicted PCa aggressiveness using only bpMRI with no lesion annotations or lesion location information (AUC = 0.73).
- Scanner manufacturer and endorectal coil use affected predictive performance of models (AUC improved by 0.05 when models were tested on data similar to the training data, $P < .001$). Inclusion of clinical variables led to no performance improvements ($P = .24$).

Analysis of Deep Feature Distribution



de Almeida JG et al. Published Online: January 22, 2025
https://doi.org/10.1148/ryai.230555