Journal of Nuclear Medicine, published on September 22, 2022 as doi:10.2967/jnumed.122.264138

Clinical decision support for axillary lymph node staging in newly diagnosed breast cancer patients based on ^{18}F-FDG PET/MRI and machine-learning

Janna Morawitz[1], Benjamin Sigl[2], Christian Rubbert[1], Nils-Martin Bruckmann[1], Frederic Dietzel[1], Lena J. Häberle[3], Saskia Ting[4], Svjetlana Mohrmann[5], Eugen Ruckhäberle[5], Ann-Kathrin Bittner[6], Oliver Hoffmann[6], Pascal Baltzer[2], Panagiotis Kapetas[2], Thomas Helbich[2], Paola Clauser[2], Wolfgang P. Fendler[7], Christoph Rischpler[7], Ken Herrmann[7], Benedikt M. Schaarschmidt[8], Andreas Stang[9], Lale Umutlu[8], Gerald Antoch[1], Julian Caspers[1*] & Julian Kirchner[1*]

[1] Department of Diagnostic and Interventional Radiology, Medical Faculty, University Dussledorf, D-40225 Dusseldorf, Germany

[2] Department of Biomedical Imaging and Image-guided Therapy, Division of General Radiology, Medical University of Vienna, Austria

[3] Institute of Pathology, Medical Faculty, Heinrich-Heine-University and University Hospital Duesseldorf, Duesseldorf, Germany

[4] Institute of Pathology, University Hospital Essen, West German Cancer Center, University Duisburg-Essen and the German Cancer Consortium (DKTK) Essen, Germany

[5] Department of Gynecology, University Dusseldorf, Medical Faculty, D-40225 Dusseldorf, Germany

[6] Department Gynecology and Obstetrics, University Hospital Essen, University of Duisburg-Essen, D-45147 Essen, Germany

[7] Department of Nuclear Medicine, University of Duisburg-Essen and German Cancer Consortium (DKTK)-University Hospital Essen, Essen, Germany

[8] Department of Diagnostic and Interventional Radiology and Neuroradiology,
University Hospital Essen, University of Duisburg-Essen, D-45147 Essen, Germany

[9] Institute of Medical Informatics, Biometry and Epidemiology, Essen University Medical Center, Essen, Germany

*contributed equally

Corresponding Author:

Dr. Janna Morawitz
Resident
Department of Diagnostic and Interventional Radiology, Medical Faculty,
University Dusseldorf, Moorenstrasse 5, D-40225 Dusseldorf, Germany.

Tel: +49 211 8 11 75 52 / Fax: +49 211 8 11 61 45
Email: janna.morawitz@med.uni-duesseldorf.de

Word count: 5000

The study is funded by the Deutsche Forschungsgemeinschaft (DFG), the German Research Foundation [BU3075/2-1, KI2434/1-2]. The funding foundation was not involved in trial design, participant recruitment, data collection, analysis, interpretation or presentation, writing or editing of the reports, or the decision to submit for publication. The corresponding author had full access to all data in the study and had all responsibility for the decision to submit for publication.

Running title: Machine learning in PET/MRI

Abbreviations

^{18}F-FDG	^{18}F-Fluorodeoxyglucose
NPV	Negative predictive value
PET	Positron emission tomography
PPV	Positive predictive value
RADS	Reporting and Data System
ROC	Receiver operating characteristic
SUVmax	Standardized uptake value

ABSTRACT

Background: In addition to its high prognostic value, the involvement of axillary lymph nodes in breast cancer patients also plays an important role in therapy planning. Therefore, an imaging modality that can determine nodal status with high accuracy in primary breast cancer patients is desirable.

Purpose: To investigate if machine-learning prediction models based on simple assessable imaging features in MRI (magnetic resonance imaging) or PET (positron emission tomography)/MRI are able to determine nodal status in newly diagnosed breast cancer patients with comparable performance as experienced radiologists, if such models can be adjusted to achieve low rates of false negatives such that invasive procedures could potentially be omitted, and if a clinical framework for decision-support based on simple imaging features can be derived from these models.

Methods: 303 participants from three centres prospectively underwent dedicated whole-body ^{18}F-FDG (^{18}F-fluorodeoxyglucose) PET/MRI between August 2017 and September 2020. Imaging datasets were evaluated regarding axillary lymph node metastases based on morphologic and metabolic features. Predictive models were developed for MRI and PET/MRI separately using random forest classifiers on data of two centers and were tested on data of the third center.

Results: The diagnostic accuracy for MRI features was 87.5% both for radiologists and for machine learning algorithm. For PET/MRI the diagnostic accuracy was 89.3% for the radiologists and 91.2% for the machine learning algorithm with no significant differences in diagnostic performance of radiologists and the machine learning algorithm in MRI (p=0.671) and PET/MRI (p=0.683). Most important lymph node feature was tracer uptake, followed by lymph node size. With an adjusted threshold, a sensitivity of 96.2% was achieved by the random forest classifier, whereas specificity, positive predictive value, negative predictive value and accuracy were 68.2%, 78.1%, 93.8% and 83.3%. A decision tree based on three simple imaging features could be established for MRI and PET/MRI.

Conclusion: Applying a high sensitivity threshold to the random forest results could potentially avoid invasive procedures such as sentinel lymph node biopsy in 68.2% of the patients.

Key words: breast cancer, lymph node metastases, machine learning, PET/MRI

INTRODUCTION

With more than 2.3 million cases in 2020, breast cancer represents the world's most prevalent cancer (*1*). In primary breast cancer axillary lymph node involvement is the most important predictor for overall survival and recurrence in breast cancer patients (*2*) and has a decisive influence on the therapy regime. Whereas a few years ago mastectomy and extensive axillary dissection were performed in most clinical nodal positive patients, advances in imaging, among other factors, have helped to make therapeutic options for local control much less invasive (*3*,*4*). If imaging procedures like sonography and mammography do not reveal affected axillary lymph nodes, sentinel lymph node biopsy is now the gold standard for clinical node negative patients (*5*). With regard to the planned therapy, this is decisive, because depending on these findings, axillary dissection and axillary radiation are further therapy options (*6*). Nearly 60 % of breast carcinoma patients do not have lymph node metastases at the time of initial diagnosis (*7*). Particularly these patients would benefit from de-escalating invasive procedures. Although the recently introduced node-RADS (Reporting and Data System) classification tries to standardize reporting of possible lymph node metastases (*8*), no universal consensus exists regarding objective criteria for the evaluation of metastatic disease of axillary lymph nodes in breast cancer patients, and N-staging by imaging remains a challenge (*7*,*9*,*10*).

In recent years artificial intelligence and machine-learning have emerged strongly into the medical imaging field (*11*). Thus, incorporating machine-learning models into imaging-based decision-support tools has great potential to enhance diagnostic workup in breast cancer patients.

Therefore, the aim of this study was to investigate (1.) if machine-learning prediction models based on simple and easy assessable imaging features in MRI (magnetic resonance imaging) or PET (positron emission tomography)/MRI are able to detect lymph node metastases in newly diagnosed breast cancer patients with comparable performance as experienced radiologists, (2.) if such models can be adjusted to achieve low rates of false negatives such that invasive procedures could potentially be omitted, and (3.) if a clinical framework for decision support based on simple imaging features can be derived from these models.

MATERIAL AND METHODS

Due to the multifold aims of this study, the workflow of the study is structured in three consecutive steps involving different methods. All calculations are based on the assessment of predefined imaging features of axillary lymph nodes by radiologists. First, machine-learning based prediction models applying random forest classifiers were developed using the imaging features derived from the radiologist readers assessments and their predictive performance on an independent test sample was compared to that of radiologists. Secondly, an adjustment of the random forest classifiers was applied to minimize false negative results by ROC (receiver operating characteristic) -curve optimization. Third, in order to facilitate a simple decision framework for everyday clinical routine, a simple decision tree classifier was trained on the imaging features independent of the optimized random forest classifiers trained beforehand.

Participant Population, Inclusion Criteria and Imaging Protocol

The study sample consisted of two samples, i.e. a training sample that was derived from two centers (University Hospital Düsseldorf and University Hospital Essen) and a testing sample from a third center (Medical University of Vienna, General Hospital).

For the training sample, 255 participants were prospectively included (Fig. 1). All included patients had newly diagnosed, therapy-naive breast cancer with at least one of the following criteria for a worse prognosis: 1) newly diagnosed, therapy-naive T2 tumor or higher T-stage or 2) newly diagnosed, therapy-naive triple-negative tumor of any size or 3) newly diagnosed, therapy-naive tumor with a high-risk molecular profile (Ki67 > 14%, G3 or Her2neu-overexpression). All included participants underwent whole-body ^{18}F-FDG PET (^{18}F-fluorodeoxyglucose positron emission tomography)/MRI. Some participants have been reported before (*7*,*12*,*13*). This study was approved by the local ethics committees (study number: 6040R, 17-7396-BO + 510-2009). The test sample consisted of 48 participants. All PET/MRI examinations were performed on integrated hybrid 3.0 Tesla PET/MRI system (Biograph mMR, Siemens Healthcare, Erlangen, Germany) (*14*).

Image Analysis

Imaging data of the training and test samples were analyzed by one reader (J.M.), while data of the test sample was additionally rated by a second reader (B.S.). (PET/)MRI datasets were analyzed in random order utilizing an Osirix Workstation (Pixmeo SARL, Bernex, Switzerland). Readers were blinded to participants identity and all clinical information except for the diagnosis of breast cancer. For every participant, the presence or absence of axillary lymph node metastasis was evaluated in MRI and and subsequently PET/MRI separately. For this, predefined imaging features of axillary lymph nodes were assessed for the most suspicious lymph node in each participant. Morphologic features for the assessment of lymph node metastases were (Fig. 2): (a) short-axis diameter in mm, (b) irregular margin (yes/no), (c) inhomogeneous cortex (yes/no), (d) intact nodal border (yes/no), (e) perifocal edema (yes/no), (f) absent fatty hilum (yes/no) and (g) contrast media enhancement (yes/no). In PET/MRI, tracer-uptake in terms of SUVmax (Standardized uptake value) of the selected lymph node was assessed. For this, a manually drawn region of interest was placed around the respective lymph node. A lymph node SUVmax ratio was calculated with the bloodpool SUVmax of the ascending Aorta as the denominator. When all criteria were considered together, each reader then made a final evaluation of the lymph node status, although an absolute number of positive findings did not have to be present to evaluate the lymph node as benign or malignant.

Reference Standard

In all participants, histopathology of axillary lymph nodes served as reference standard. If available, sentinel lymph node biopsy or axillary dissection were used. Otherwise, histopathologic results were derived from pretherapeutic ultrasound-guided core needle biopsy of the suspicious lymph node. If no sufficient pretherapeutic sampling of lymph nodes was available, sentinel lymph node excision or axilla dissection after neoadjuvant systemic therapy were used as the reference standard. In these cases, additional histopathological preparations were evaluated, using focal fibrosis or focal necrosis as retrospective indicators for previously viable lymph node metastasis (*15*,*16*)

Model Development

Predictive models were developed for MRI and PET/MRI separately using random forest classifiers. For each modality, a random forest classifier was trained using the imaging features derived from the readers assessment as input features and the dichotomous reference standard ("benign" or "malignant") as output.

To further optimize the classification of the models for sensitivity and minimize false negatives (to identify a rule-out criterion), an adjusted random forest model was developed by applying adjustment of the classification threshold of a trained random forest model on an independent validation set that was split from the training sample beforehand (80:20 stratified split) so that sensitivities of >0.95 were achieved on this validation set.

To additionally create more clinically interpretable classifiers, simple decision tree classifiers with a maximum depth of 3 were additionally built using Gini impurity as optimization criterion.

Model development was conducted using the scikit-learn library (version 0.24.2) in python 3.9.

Statistics

For statistical analyses IBM SPSS Statistics (Version 21, IBM Deutschland GmbH) was used. Demographic participant data were reported using descriptive statistics. Cohen's Kappa was used to calculate interrater reliability between the two readers regarding prediction of lymph node status (metastatic vs. non-metastatic) in MRI and PET/MRI. Diagnostic performance of radiologists and machine-learning models for lymph node status in MRI and PET/MRI was assessed by determining sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and ROC-AUC (area under the curve). A McNemar test was used to compare the diagnostic performance of the radiologists with that of the machine-learning models. A Pearson's Chi-squared test was used to compare the tumor characteristics between training sample and validation sample. Statistical significance was defined as a *p*-value < 0.05.

RESULTS

Participant Demographics and Reference Standard

In this study a total of 255 female participants (mean age 51.2±11.9 years) from two centers were included for the training sample (Fig. 1). According to the reference standard 101/255 participants (39.6%) were nodal positive and 154/255 participants (60.4%) were nodal negative.

For the testing sample 48 participants (mean age 52.2±12.2 years) from a third center were evaluated. According to the reference standard 26/48 (45.8%) participants were nodal negative and 22/48 (54.2%) participants were nodal positive. For demographics and tumor characteristics of all participants see Table 1.

Radiologists' Performance

Based on MRI data, the radiologist was able to determine the correct lymph node status in 218/255 participants (85.5%) in the training set. This yielded a diagnostic performance indicated by sensitivity, specificity, PPV, NPV and accuracy of 74.3%, 92.9%, 87.2%, 84.6% and 85.5% for the training sample (Supplemental Table 1). Corresponding results for the radiologists performance (identical results for both readers) based on MRI in the testing sample were 84.6%, 90.9%, 91.7%, 83.3% and 87.5% (Table 2).

When taking PET/MRI into account, the radiologist was able to determine the correct lymph node status in 221/255 participants (86.7%) and sensitivity, specificity, PPV, NPV and accuracy were 84.0%, 88.4%, 82.4%, 89.5% and 86.7% for the training sample (Supplemental Table 1). In the testing sample radiologists performance on PET/MRI data were 92.3%, 86.4%, 88.9%, 90.5% and 89.6% (Table 2).

With regard to the individual features, there were isolated differences in the subjective evaluation of the lymph nodes by the raters (irregular margin κ=0.919, inhomogeneous cortex κ=0.879, perifocal edema κ=0.776, absent fatty hilum κ=0.865, contrast media enhancement κ=0.947, absent intact nodal border 0.957, all p<0.001), but together these lead to an equal evaluation of the lymph node status, so that the interrater reliability with regard to the lymph node status was excellent (κ=1.0, p<0.001).

Random Forest Algorithm Performance

The trained random forest classifiers yielded an accuracy of 88.3% for MRI and of 99.2% for PET/MRI on the training data, which is indicative for a very good model fit to the training data (Supplemental Table 1). When applied to the independent datasets of the testing sample, the respective random forest classifier was able to determine the correct lymph node status in 42/48 participants (87.5%) (23 true positive and 19 true negative) for MRI features, while 3 participants were rated false positive and 3 participants false negative (both readers, Table 2). The performance was unchanged when applying the PET/MRI-based random forest classifier to the testing sample, with 42/48 correct classifications (87.5%) (23 true positive and 19 true negative), while 3 participants were rated false positive and 3 participants false negative based on lymph node assessment of reader 1. Based on lymph node assessment of reader 2, there were 41/48 correct classifications (85.4%) (23 true positive and 18 true negative), while 4 participants were rated false positive and 3 participants false negative. Sensitivity, specificity, PPV and NPV for both classifiers were 88.5%, 86.4%, 88.5% and 86.4% (reader 1), and 88.5%, 81.8%, 85.2%, 85.7%, 85.4% for PET/MRI respectively (reader 2) (Table 2).

Comparison of Radiologist Performance and Random Forest Algorithm

In the testing sample highest ROC-AUC was achieved by the random forest classifier based on PET/MRI data with a value of 91.2% (95%-CI (confidence interval): 82.8-99.6%), followed by a ROC-AUC of 89.5% (95%-CI: 80.4–98.7%) by the random forest classifier based on MRI data (Fig. 3).

There were no significant differences in the assessment of lymph node status between the radiologists and random forest classifier, nor for MRI features (*p*=0.67) neither for PET/MRI features (*p*=0.68).

Feature Importance

The most important feature in MRI is size, followed by intact nodal border and irregular margin, whereas most important features for predicting the nodal status in PET/MRI were tracer uptake indicated by the ratio of SUVmax of the lymph node/SUVmax of the ascending Aorta, followed by size and intact nodal border (Fig. 4).

Decision Threshold Adjustment

To minimize the classifier's false negatives with regard to clinical need, we adjusted the decision threshold of the random forest classifier on PET/MRI data as a trade-off between precision (=PPV) and recall (=sensitivity). The default decision threshold in random forest was 0.5. Fig. 5 shows precision and recall as a function of decision values in the internal validation sample. The optimal decision threshold for this purpose was obtained at 0.19. A sensitivity (recall) of 96.2% was achieved with only one false negative in the test sample, whereas specificity, PPV, NPV and accuracy were 68.2%, 78.1%, 93.8% and 83.3% at this threshold. Applying these results to everyday routines in our cohort would mean that it would be possible to save 68.2% (15/22) of the women from an unnecessary biopsy, even though it has to be accepted that 3.8% (1/26) of the affected woman would be missed (Table 3).

Decision Tree For Clinical Decision Support

The decision tree classifier for distinguishing benign from malignant lymph nodes achieved an accuracy of 89.6% and a ROC-AUC of 87.6% (95%-CI: 77.6–97.5%) for MRI in the testing sample; for PET/MRI and an accuracy of 89.6% and ROC-AUC of 89.0% (95%-CI: 79.7-98.4%) for PET/MRI data in the testing sample.

These decision trees can support clinical decision making based on three simple imaging features, each (Fig. 6A). For MRI the root node indicative for the most important feature is “size”, which is consistent with the feature importances from random forests. Here, a short-axis diameter of ≥7.5mm serves as a cut-off value for highly suspicious lymph nodes. ROC-AUC evaluation of this feature alone shows a sensitivity of 71.6% and specificity of 86.4% (J=0.580) for this cut-off. A cut-off value of 12.5 mm led to a specificity of 100%, but a sensitivity of 34.3% (J=0.343) (Fig. 6B). The decision tree as well as these cut-off values were determined on the training data. The combination of an FDG-uptake above the 1.3-fold of the lymph node compared to the uptake of Aorta ascendens and a short axis diameter of 7.5mm is sufficient to characterize a lymph node as malignant.

The confusion matrices for the decision trees are shown in Table 4. For the performance of the decision trees on the training data see Supplemental Table 2. Supplemental Table 3 shows the detection rates of lymph nodes in ^{18}F-FDG PET/MRI per nodal stage (cN0 – cN3c).

DISCUSSION

In this study, we demonstrate that diagnosis of lymph node metastases in newly diagnosed breast cancer patients can be achieved by simple imaging features from MRI and PET/MRI, both, by radiologists and machine-learning-based prediction models with comparably high accuracies. However, our results indicate that a machine-learning-based prediction model can be advantageous in a clinical setting due to its opportunity to allow decision threshold adjustments. Based on the implemented random forest classifier on PET/MRI data, it would be possible - compared with the current gold standard, where every clinically node-negative patient would receive sentinel lymph node biopsy - to save 68.2 % of the women from an unnecessary biopsy, even though it has to be accepted that 3.8% of the affected woman would be missed. The latter is important for such model to be suitable for a clinical setting, where diagnostic imaging could potentially omit invasive diagnostic procedures such as lymph node biopsy when false negatives can reliably be reduced. Furthermore, we derived a decision tree for clinical decision support based on simple imaging features from MRI and PET/MRI, which can assist clinicians in the diagnostic workup in regard to lymph node involvement in breast cancer. Although the application of the model evaluated here does not result in time savings in the evaluation of lymph node criteria per se, the clear cascade of the three easily assessable imaging features can be helpful for the radiologist in classification of axillary lymph nodes in daily routine.

Different machine learning algorithms for the detection of axillary lymph node metastases have previously been shown to provide diagnostic performance comparable to or better than that of experienced physicians in other specialties (*17*), but only a few applications have been introduced into the everyday routine.

This study further rates the relevance of different imaging features of lymph nodes. While the size of a lymph node characterized by the short-axis diameter is a generally accepted criterion for assessing the metastatic status of a lymph node (*8*), it has been discussed in the past that diagnostic accuracy can be increased by adding factors such as contour and signal intensity.

Nevertheless, the feature importances of the random forest classifiers and the good performance of the simple decision tree classifier indicate that only a few features are necessary to predict lymph node malignancy with high accuracy. Our study is in line with a study by Ramirez-Galvan *et al.* (*18*), demonstrating lymph node size as the most important morphological feature. However, according to our investigation, a short axis diameter of ≥7.5mm seems to be most suitable for prediction of axillary lymph node involvement of breast cancer, while a diameter of ≥12.5 mm can even be seen as evidence for malignancy (Fig. 6B).

As with other cancer entities, there is no consensus about thresholds for tracer uptake in breast cancer to define a lymph node as benign or malignant (*19*), but a threshold of a SUVmax of 1.8-2.0 has reported to be a helpful criterion to diagnose malignancy (*20*,*21*). In our study, we demonstrated that a tracer uptake of the lymph node below the uptake of the mediastinal blood pool is a reliable feature of benignity, while tracer uptakes ≥ 1.3 times the uptake of the mediastinal blood pool should be considered malignant.

Using the adjusted threshold of the random forest classifier, the rate of false negatives could be substantially decreased to a range that would be acceptable for clinical purposes. The single participant missed by our machine learning algorithm after adjusting the threshold had a histopathologically proven micrometastasis (1 mm). The clinical impact of micrometastases does not appear to be comparable to that of macrometastases with an outcome of patients with micrometastases comparable to that of node-negative patients (*22*). Thus, machine learning algorithms may be expected to play a crucial role in reducing invasive procedures in the future.

This study has some limitations. Only therapy-naive patients were examined at baseline staging, so no general statements can be made regarding regressively altered lymph nodes after therapy or with regard to response to therapy. The reference standard is in part based on post-therapeutic specimens from axillary nodes and different ways of sample acquisition including axillary dissection and ultrasound-guided biopsy. This may have had an impact on definition of the reference standard. The imaging features used as input for the machine-learning-based prediction models still rely on subjective assessments of radiologists. Nevertheless, we could show that these imaging features are easy assessable and with a high interrater reliability. In addition, the size of the validation cohort is only moderate, so further studies with a larger study population are needed.

In conclusion, this study shows that 1. a random forest classifier based on simple imaging features provides comparable diagnostic performance compared with an experienced radiologist, 2. that ^{18}F-FDG-PET uptake and lymph node size assessed on MRI are the most informative features in determining the metastatic status of an axillary lymph node, 3. that a combination of three features can be helpful for the differentiation between malignant and benign axillary lymph nodes in newly diagnosed breast cancer in daily routine, and 4. that - accepting a low specificity - a sensitivity of >95% can be achieved with an adjusted random forest classifier on ^{18}F-FDG PET/MRI data, which can exclude lymph node involvement with high confidence and could play a central role in reducing invasive procedures in the future. Thus, especially the combination of the three imaging features may be used for daily use by the radiologist, as these can be determined and evaluated quickly and reliably, although the decision tree should not be taken as the only basis for therapy planning. For therapy decision making the adjusted random forest model is more reliable for the diffentiation between malignant and benign lymph nodes because of its higher sensitivity. Nevertheless, the adjusted random forest models need to be confirmed in large prospective studies to minimize the number of unnecessary invasive procedures and will then have great impact.

FUNDING

The study is funded by the Deutsche Forschungsgemeinschaft (DFG), the German Research Foundation [BU3075/2-1, KI2434/1-2]. The funding foundation was not involved in trial design, participant recruitment, data collection, analysis, interpretation or presentation, writing or editing of the reports, or the decision to submit for publication. The corresponding author had full access to all data in the study and had all responsibility for the decision to submit for publication.

DISCLOSURE

Ethical approval: All procedures performed were in accordance with the ethical standards of the institutional research committee and with the principles of the 1964 Declaration of Helsinki and its later amendments.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Conflict of interest: No potential conflicts of interest relevant to this article exist.

KEY POINTS

Question: Can machine learning prediction models determine nodal status in PET/MRI examinations from patients with newly diagnosed breast cancer with comparable performance as experienced radiologists?

Pertinent findings: Machine learning shows comparable diagnostic performane as experienced radiologists in identifying axillary lymph node metastases in PET/MRI in primary breast cancer patients**.** The most important lymph node feature is tracer uptake, followed by lymph node size. A combination of three features is helpful for the differentiation between malignant and benign axillary lymph nodes in newly diagnosed breast cancer, leading to an easlily applicable decision-tree in everyday clinical routine.

Implications for patient care: With the help of machine learning, axillary lymph node metastases can be reliably excluded in PET/MRI, sparing 68.2% of the patients an invasive procedure like a sentinel lymph node biopsy.

REFERENCES

1. Organization WH. WHO [Internet]. 2016. Available from: https://www.who.int/news-room/fact-sheets/detail/breast-cancer, Accessed on 16.10.2021

2. Chang JM, Leung JWT, Moy L, Ha SM, Moon WK. Axillary nodal evaluation in breast cancer: State of the Art. *Radiology*. 2020;295:500-515.

3. Giuliano AE, Ballman KV, McCall L, et al. Effect of axillary dissection vs no axillary dissection on 10-year overall survival among women with invasive breast cancer and sentinel node metastasis: The ACOSOG Z0011 (Alliance) randomized clinical trial. *Jama*. 2017;318:918–26.

4. Giuliano AE, Hunt KK, Ballman KV, et al. Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: A randomized clinical trial. *Jama*. 2011;305:569–75.

5. Duraes M, Guillot E, Seror J, Pouget N, Rouzier R. [Sentinel lymph node biopsy and neoadjuvant treatment in breast cancer]. *B Cancer*. 2017;104:892–901.

6. Yan M, Abdi MA, Falkson C. Axillary management in breast cancer patients: A comprehensive review of the key trials. *Clin Breast Cancer*. 2018;18:e1251–9.

7. Morawitz J, Bruckmann N-M, Dietzel F, et al. Determining the axillary nodal status with four current imaging modalities including 18 F-FDG PET/MRI in newly diagnosed breast cancer: A comparative study using histopathology as reference standard. *J Nucl Med*. 2021;62:1677-1683.

8. Elsholtz FHJ, Asbach P, Haas M, et al. Introducing the node reporting and data system 1.0 (Node-RADS): a concept for standardized assessment of lymph nodes in cancer. *Eur Radiol*. 2021;31:6116–24.

9. Zhao M, Wu Q, Guo L, Zhou L, Fu K. Magnetic resonance imaging features for predicting axillary lymph node metastasis in patients with breast cancer. *Eur J Radiol*. 2020;129:109093.

10. Atallah D, Moubarak M, Arab W, Kassis NE, Chahine G, Salem C. MRI-based predictive factors of axillary lymph node status in breast cancer. *Breast J*. 2020;26:2177–82.

11. Bejnordi BE, Veta M, Diest PJ van, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*. 2017;318:2199–210.

12. Bruckmann NM, Kirchner J, Umutlu L, et al. Prospective comparison of the diagnostic accuracy of 18F-FDG PET/MRI, MRI, CT, and bone scintigraphy for the detection of bone metastases in the initial staging of primary breast cancer patients. *Eur Radiol*. 2021;31:8714–24.

13. Morawitz J, Kirchner J, Martin O, et al. Prospective correlation of prognostic immunohistochemical markers with SUV and ADC derived from dedicated hybrid breast 18F-FDG PET/MRI in women with newly diagnosed breast cancer. *Clin Nucl Med*. 2020;46:201–5.

14. Kirchner J, Grueneisen J, Martin O, et al. Local and whole-body staging in patients with primary breast cancer: a comparison of one-step to two-step staging utilizing 18F-FDG-PET/MRI. *Eur J Nucl Med Mol* I. 2018;45:2328–37.

15. Newman LA, Pernick NL, Adsay V, et al. Histopathologic evidence of tumor regression in the axillary lymph nodes of patients treated with preoperative chemotherapy correlates with breast cancer outcome. *Ann Surg Oncol*. 2003;10:734–9.

16. Takashi Y, Soh J, Shien K, et al. Fibrosis or necrosis in resected lymph node indicate metastasis before chemoradiotherapy in lung cancer patients. *Anticancer Res*. 2020;40:4419–23.

17. Golden JA. Deep learning algorithms for detection of lymph node metastases from breast cancer: Helping Artificial Intelligence Be Seen. *Jama.* 2017;318:2184–6.

18. Ramírez-Galván YA, Cardona-Huerta S, Elizondo-Riojas G, Álvarez-Villalobos NA, Campos-Coy MA, Ferrara-Chapa CM. Does axillary lymph node size predict better metastatic involvement than apparent diffusion coefficient (ADC) value in women with newly diagnosed breast cancer? *Acta Radiol*. 2020;61:1494–504.

19. Yu C, Xia X, Qin C, Sun X, Zhang Y, Lan X. Is SUVmax helpful in the differential diagnosis of enlarged mediastinal lymph nodes? A Pilot Study. *Contrast Media Mol I*. 2018;2018:1–9.

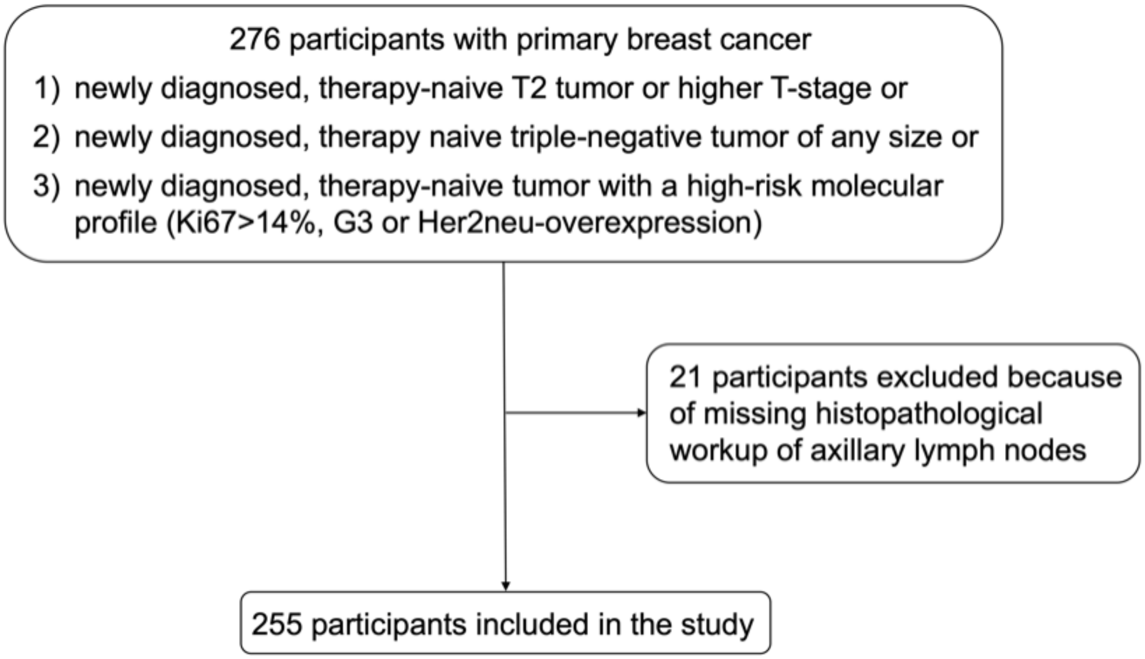
20. Rosen EL, Eubank WB, Mankoff DA. FDG PET, PET/CT, and breast cancer imaging. *Radiographics*. 2007;27(suppl_1):S215–29.

21. Carkaci S, Adrada BE, Rohren E, et al. Semiquantitative analysis of maximum standardized uptake values of regional lymph nodes in inflammatory breast cancer is there a reliable threshold for differentiating benign from malignant? *Acad Radiol*. 2012;19:535–41.

22. Wada N, Imoto S. Clinical evidence of breast cancer micrometastasis in the era of sentinel node biopsy. *Int J Clin Oncol*. 2008;13:24–32.

FIGURE LEDGENDS

Fig. 1. Flow chart diagram of included and excluded participants.



276 participants with primary breast cancer
1) newly diagnosed, therapy-naive T2 tumor or higher T-stage or
2) newly diagnosed, therapy naive triple-negative tumor of any size or
3) newly diagnosed, therapy-naive tumor with a high-risk molecular profile (Ki67>14%, G3 or Her2neu-overexpression)
21 participants excluded because of missing histopathological workup of axillary lymph nodes
255 participants included in the study

Fig. 2. Lymph node features. Example of morphologic and metabolic features for the assessment of axillary lymph nodes in axial T1vibe fat sat contrast enhanced images: unsuspicious lymph node with no feature of malignancy (left), enlarged lymph node with a short-axis diameter of 31 mm, lymph node with irregular margin, lymph node with an inhomogeneous cortex, lymph node with perifocal edema, lymph node with absense of fatty hilum, lymph node with contrast media enhancement, lymph node without intact nodal border, and lymph node with increased ^{18}F-FDG uptake (SUVmax 13.1).

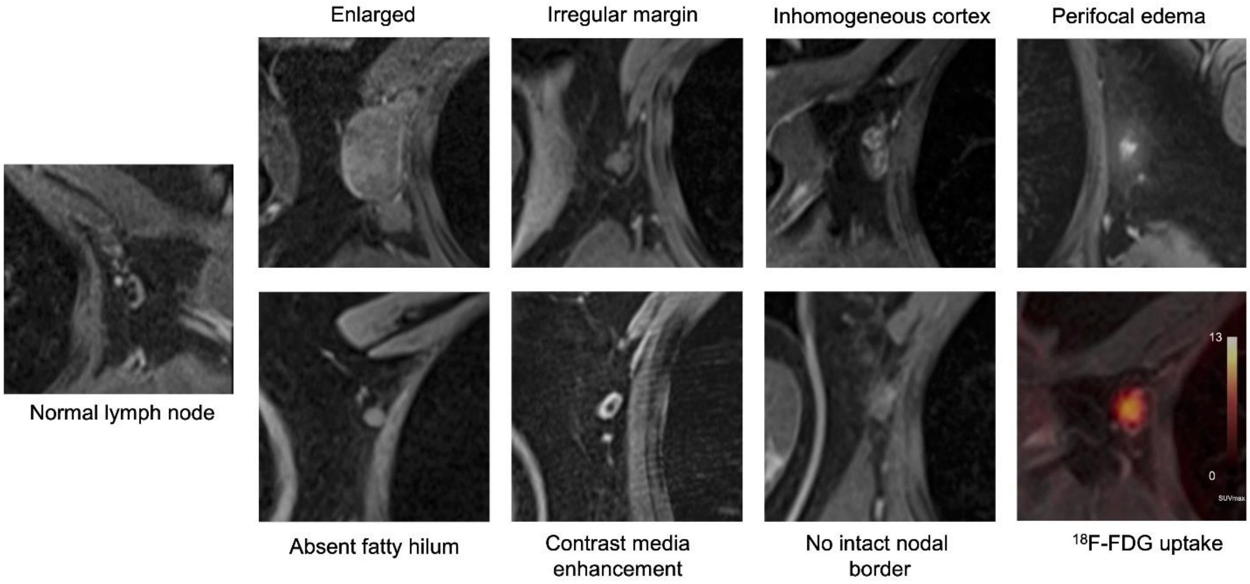


Fig. 3. ROC (receiver operating characteristic)-AUC (area under the curve) for Random Forest Model Performance on the testing Data and for prediction of lymph node status by radiologists in MRI and PET/MRI.

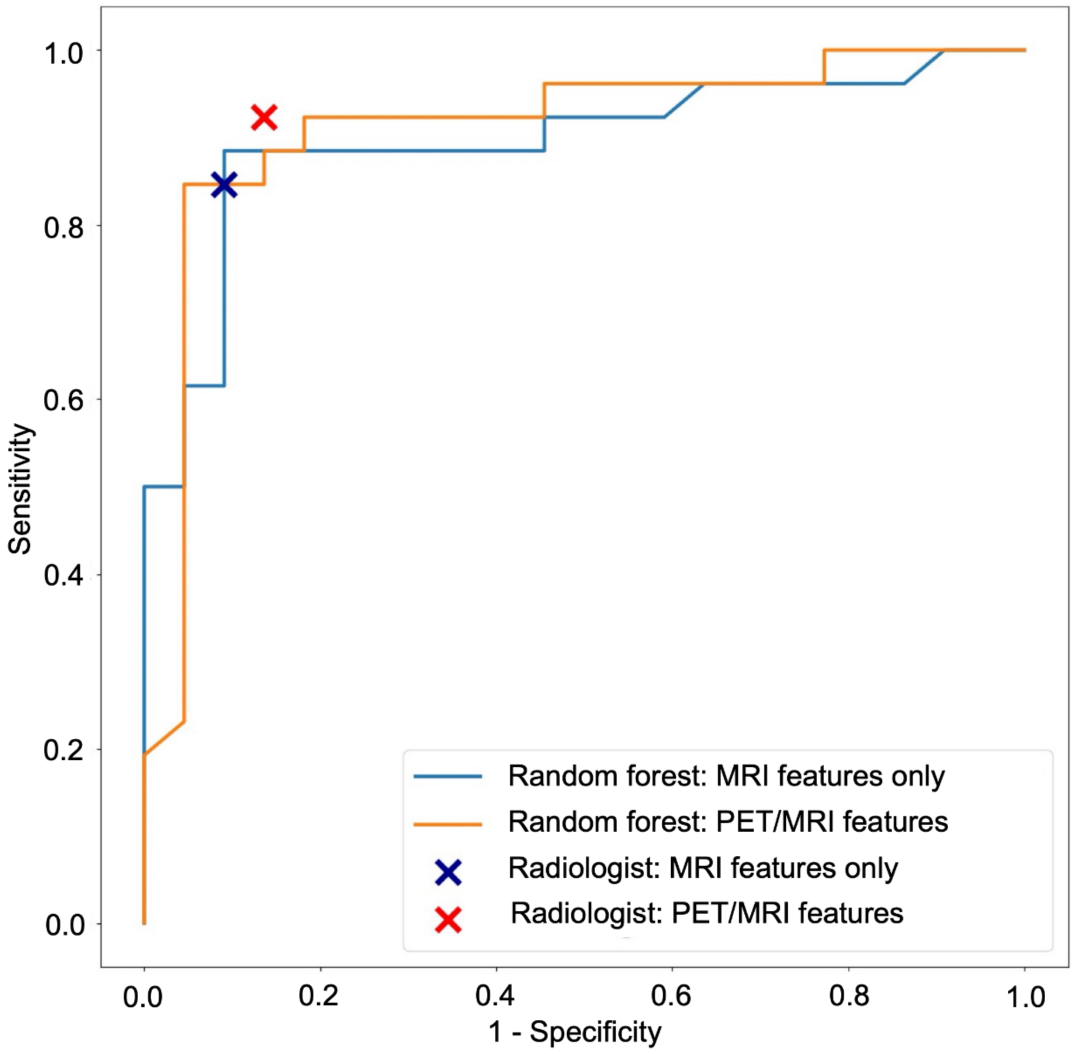


Fig. 4. Importance of different morphological and metabolical features of lymph nodes.

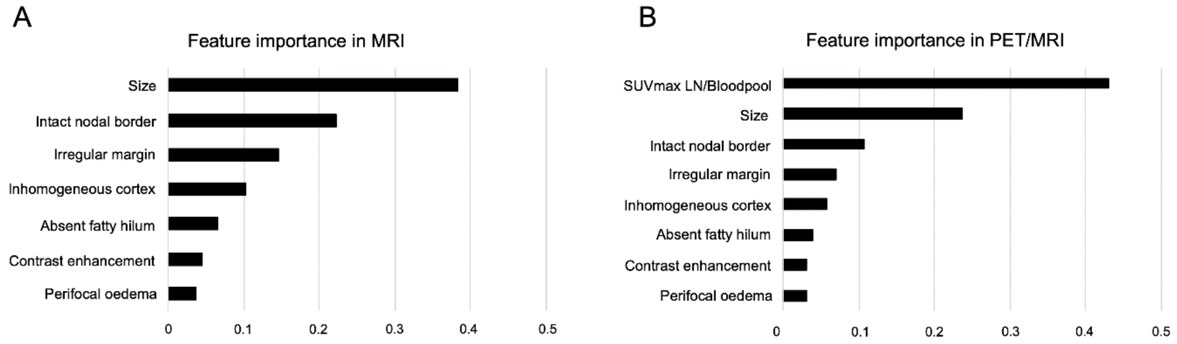


Fig. 5. Precision and Recall scores as a function of the decision threshold on the internal validation sample. X represents threshold values and y is the score of precision or recall. The adjusted decision threshold for optimized sensitivity is indicated by a dashed line.

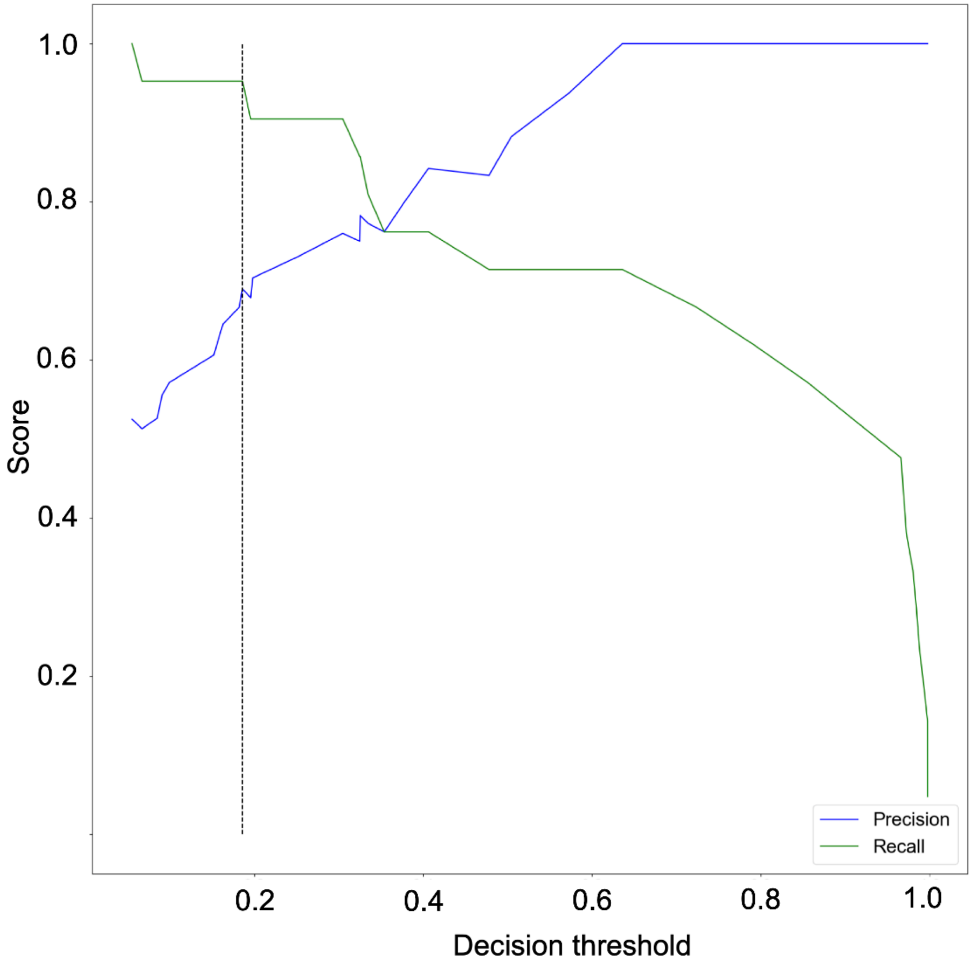
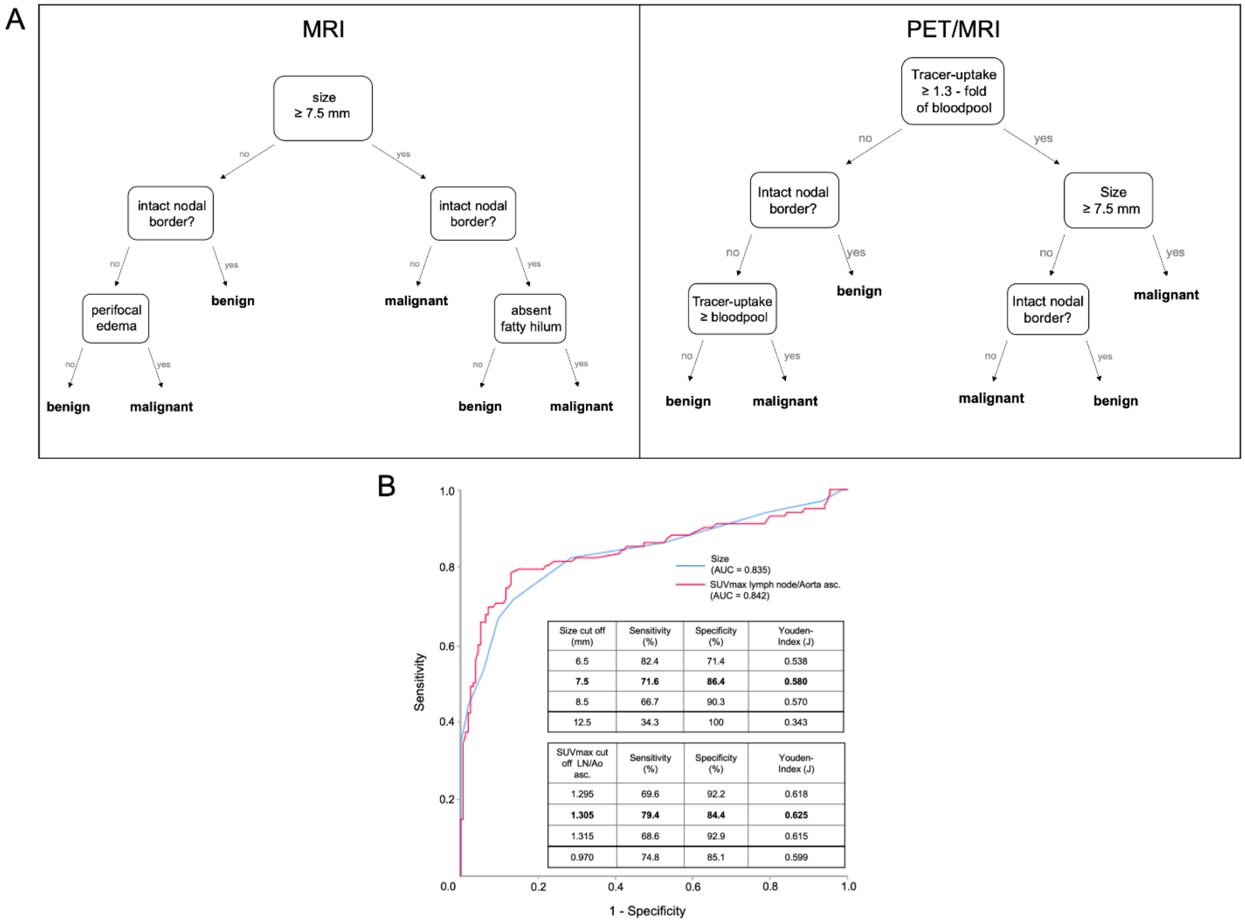


Fig. 6. Decision tree for predicting lymph node status in MRI and PET/MRI (A) and ROC (receiver operating characteristic)-AUC (area under the curve) for size and ratio of SUVmax of lymph node to mediastinal bloodpool for prediction of lymph node status (B)



Size cut off (mm)	Sensitivity (%)	Specificity (%)	Youden-Index (J)
6.5	82.4	71.4	0.538
7.5	**71.6**	**86.4**	**0.580**
8.5	66.7	90.3	0.570
12.5	34.3	100	0.343

SUVmax cut off LN/Ao asc.	Sensitivity (%)	Specificity (%)	Youden-Index (J)
1.295	69.6	92.2	0.618
1.305	**79.4**	**84.4**	**0.625**
1.315	68.6	92.9	0.615
0.970	74.8	85.1	0.599

TABLES

Table 1. Participant demographics and tumor characteristics

		Training Sample	Testing Sample	p-value
Total participants		255	48	
Mean age (± Standard deviation)		51.2 ± 11.9 years	52.2 ± 12.2 years	0.689
Lymph node status (reference standard)				
	negative	154 (60.4 %)	26 (54.2 %)	0.420
	positive	101 (39.6 %)	22 (45.8 %)	
Menopause status				
	pre	111 (43.5 %)	18 (37.5 %)	0.737
	peri	25 (9.8 %)	5 (10.4 %)	
	post	119 (46.7 %)	25 (52.1 %)	
Ki67				
	positive >14 %	226 (88.6 %)	41 (85.4 %)	0.528
	negative <14 %	29 (11.4 %)	7 (14.6 %)	
Progesterone status				
	positive	169 (66.3 %)	29 (60.4 %)	0.433
	negative	86 (33.7 %)	19 (39.6 %)	
Estrogen status				
	positive	187 (73.3 %)	28 (58.3 %)	<0.01
	negative	68 (26.7 %)	20 (41.7 %)	
HER2neu-expression				
	0	97 (38.0 %)	23 (47.9 %)	0.479
	1+	73 (28.6 %)	14 (29.2 %)	
	2+	34 (13.3 %)	5 (10.4 %)	
	3+	51 (20.0 %)	6 (12.5 %)	
Tumor grade				
	G1	10 (3.9 %)	4 (8.3 %)	0.025
	G2	137 (53.7 %)	16 (33.3 %)	
	G3	108 (42.4 %)	28 (58.3 %)	
Histology				
	NST	222 (87.1 %)	42 (87.5 %)	<0.01
	Lobular invasive	25 (9.8 %)	0 (0 %)	
	other	8 (3.1 %)	6 (12.5 %)	

NST = No Special Type

Table 2. Diagnostic performance of MRI and PET/MRI in assessment of lymph node status of radiologists and random forest classifier within the testing sample (values given in %).

Radiologists		
	MRI*	PET/MRI*
Sensitivity 95 % - CI	84.6 65.1 - 95.6	92.3 74.9 - 99.1
Specificity 95 % - CI	90.9 70.8 - 98.9	86.4 65.1 - 97.1
PPV 95 % - CI	91.7 74.4 - 97.7	88.9 73.5 - 96.8
NPV 95 % - CI	83.3 66.8 - 92.6	90.5 71.3 - 97.3
Accuracy 95 % - CI	87.5 74.8 - 95.3	89.6 77.3 - 96.5

Random Forest Algorithm			
	MRI*	PET/MRI	
		Reader 1	Reader 2
Sensitivity 95 % - CI	88.5 69.9 – 97.6	88.5 69.9 – 97.6	88.5 69.9 – 97.6
Specificity 95 % - CI	86.4 65.1 – 97.1	86.4 65.1 – 97.1	81.8 59.7 – 94.8
PPV 95 % - CI	88.5 72.6 – 95.7	88.5 72.6 – 95.7	85.2 70.1 – 93.4
NPV 95 % - CI	86.4 68.3 – 94.9	86.4 68.3 – 94.9	85.7 67.0 – 94.7
Accuracy 95 % - CI	87.5 74.8 – 96.3	87.5 74.8 – 96.3	85.4 72.2 – 93.9

PPV = positive predictive value; NPV = negative predictive value

Table 3. Confusion Matrix and performance metrices for adjusted threshold.

		Predicted	
		negative	**positive**
actual	**negative**	15	7
	positive	1	25
Sensitivity		96.2 % (80.4 - 99.9 %)	
Specificity		68.2 % (45.1 - 86.1 %)	
PPV		78.1 % (65.9 - 86.9 %)	
NPV		93.8 % (68.2 - 99.1 %)	
Accuracy		83.3 % (69.8 - 92.5 %)	

PPV = positive predictive value; NPV = negative predictive value

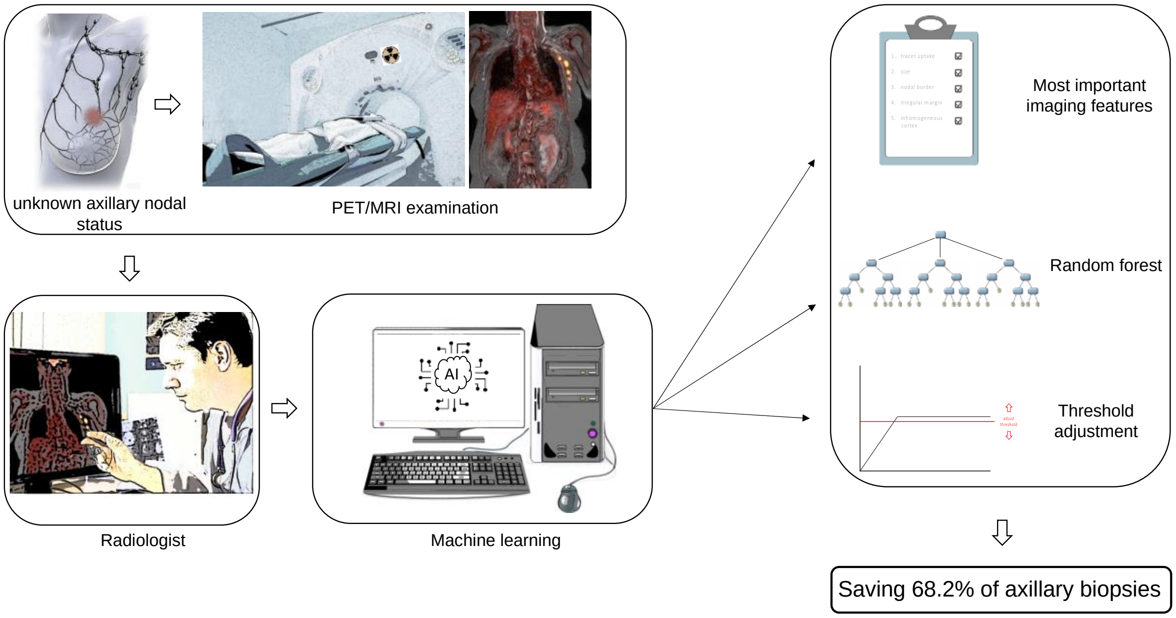
Table 4. Confusion matrices and performance metrtices for the decision trees.

MRI				PET/MRI			
		Predicted				Predicted	
		negative	positive			negative	positive
actual	negative	20	2	actual	negative	21	1
	positive	3	23		positive	4	22

MRI	
Sensitivity	88.5 % (69.9 - 97.6 %)
Specificity	90.9 % (70.8 - 98.9 %)
PPV	92.0 % (75.3 -97.8 %)
NPV	87.0 % (69.5 -95.1 %)
Accuracy	89.6 % (77.3 -96.5 %)
PET/MRI	
Sensitivity	84.6 % (65.1 - 95.6 %)
Specificity	95.5 % (77.2 - 99.9 %)
PPV	95.7 % (76.3 - 99.3 %)
NPV	84.0 % (68.0 - 92.9. %)
Accuracy	89.6 % (77.3 - 96.5 %)

PPV = positive predictive value; NPV = negative predictive value

GRAPHICAL ABSTRACT



unknown axillary nodal status

PET/MRI examination

Radiologist

Machine learning

Most important imaging features

Random forest

Threshold adjustment

Saving 68.2% of axillary biopsies

SUPPLEMENTS

Supplemental Table 1. Diagnostic performance of MRI and PET/MRI in assessment of lymph node status of radiologists and random forest classifier in training sample (values given in %).

	Radiologist		Random Forest Algorithm	
	MRI	PET/MRI	MRI	PET/MRI
Sensitivity (95% - CI)	74.3 64.6-82.4	84.0 75.3-90.6	77.5 68.1-85.1	98.0 93.1-99.8
Specificity (95% - CI)	92.9 87.6-96.4	88.4 82.3-93.0	95.5 90.9-98.2	100.0 97.6-100.0
PPV (95% - CI)	87.2 79.2-92.4	82.4 75.0-87.9	91.9 84.5-95.9	100.0
NPV (95% - CI)	84.6 79.8-88.5	89.5 84.5-93.1	86.5 81.7-90.2	98.7 95.1-99.7
Accuracy (95% - CI)	85.5 80.6-89.6	86.7 81.9-90.6	88.3 83.7-92.0	99.2 97.2-99.1

PPV = positive predictive value; NPV = negative predictive value

PPV = positive predictive value; NPV = negative predictive value

Supplemental Table 2. Performance of the decision trees for MRI (A) and PET/MRI (B) on the training data.

(A)

		Predicted	
		negative	**positive**
actual	**negative**	143	11
	positive	27	75
Sensitivity		73.5 % (63.9 – 81.8 %)	
Specificity		92.9 % (87.6 – 96.4 %)	
PPV		87.2 % (79.2 – 92.4 %)	
NPV		84.1 % (79.3 – 88.0 %)	
Accuracy		85.2 % (80.2 – 89.3 %)	

(B)

		Predicted	
		negative	**positive**
actual	**negative**	149	5
	positive	28	74
Sensitivity		72.6 % (62.8 – 80.9 %)	
Specificity		96.8 % (92.6 – 98.9 %)	
PPV		93.7 % (86.1 – 97.3 %)	
NPV		84.2 % (79.5 – 88.0 %)	
Accuracy		87.1 % (82.4 – 91.0 %)	

PPV = positive predictive value; NPV = negative predictive value

Supplemental Table 3. Exact, falsely low and falsely high detection of lymph nodes in PET/MRI per clinical lymph node stage.

		PET/MRI detection exact	**PET/MRI detection falsely low**	**PET/MRI detection falsely high**
Reference Standard				
Nodal stage	Absolute number of patients			
cN0	154	PET/MRI cN0 142 (92.3%)		PET/MRI cN1 or higher 12 (7.8%)
cN1	68	PET/MRI cN1 62 (91.2 %)	PET/MRI cN0 4 (5.9 %)	PET/MRI cN2 or higher 2 (2.9%)
cN2a	9	PET/MRI cN2a 7 (77.8 %)	PET/MRI cN1 or lower 0 (0 %)	PET/MRI cN2b or higher 2 (22.2%)
cN2b	2	PET/MRT cN2b 2 (100 %)	PET/MRI cN2a or lower 0 (0 %)	PET/MRI cN3a or higher 0 (0 %)
cN3a	11	PET/MRI cN3a 11 (100 %)	PET/MRI cN2b or lower 0 (0 %)	PET/MRI cN3b or higher 0 (0 %)
cN3b	8	PET/MRI cN3b 8 (100 %)	PET/MRI cN3a or lower 0 (0 %)	PET/MRI cN3c 0 (0 %)
cN3c	3	PET/MRI cN3c 2 (66.7 %)	PET/MRI cN3b or lower 1 (33.3 %)	

Imaging protocol

To ensure a blood glucose level of <150 mg/dl, patients fasted for six hours before the examinations. ^{18}F-FDG was administered intravenously with a weight adapted dose (4 MBq/kg body weight) one hour before the PET/MRI examination. The examination area covered a range from the head to the proximal thighs in the supine body position. To evaluate the axillary lymph nodes, the thoracic sections of the whole-body staging were evaluated.

The whole-body MRI protocol comprised the following sequences:

1) For center one and two: A transverse T2-w half Fourier acquisition single shot turbo spin echo (HASTE) sequence in breath-hold technique with a slice thickness of 7 mm (TE 97 ms; TR 1500 ms; Turbo factor (TF) 194; FOV 400 mm; phase FOV 75%; acquisition matrix 320 × 240 mm; in plane resolution 1.3 x 1.3 mm; TA 0:47 min / bed position). For center three: A coronal T2-w half Fourier acquisition single shot turbo spin echo (HASTE) sequence in breath-hold technique with a slice thickness of 6 mm (TE 121 ms; TR 1500 ms; Turbo factor (TF) 194; FOV 400 mm; phase FOV 75%; acquisition matrix 320 × 240 mm; in plane resolution 1.3 x 1.3 mm; TA 0:47 min / bed position)
2) A transversal diffusion-weighted (DWI) echo-planar imaging (EPI) sequence in free breathing with a slice thickness of 5.0 mm (TR 7400 ms; TE 72 ms; b-values: 0, 500 and 1000 s/mm2, matrix size 160 x 90; FOV 400 mm x 315 mm, phase FOV, 75 %; GRAPPA, acceleration factor 2; in plane resolution 2.6 x 2.6 mm; TA 2:06 min / bed position)
3) A fat-saturated post-contrast transverse 3-dimensional Volumetric Interpolated Breath-hold Examination (VIBE) sequence with a slice thickness of 3 mm (TE, 1.53 ms; TR, 3.64 ms; Flip angle 9°; FOV 400 x 280 mm; phase FOV 75%; acquisition matrix 512 × 384, in plane resolution 0.7 x 0.7 mm; TA 0:19 min / bed position)

PET images were reconstructed using the iterative ordered-subset expectation maximization (OSEM) algorithm, 3 iterations and 21 subsets, a Gaussian filter with 4-mm full width at half maximum (FWHM) and a 344 × 344 image matrix. For MR-based attenuation correction of the patient tissues a two-point (fat, water) coronal 3D-Dixon-VIBE sequence was acquired to generate a four-compartment model (background air, lungs, fat, muscle).