RESEARCH ARTICLE

Developing an ensemble machine learning study: Insights from a multi-center proof-of-concept study

Annarita Fanizzi[1], Federico Fadda[1], Michele Maddalo[2]*, Sara Saponaro[3], Leda Lorenzon[4], Leonardo Ubaldi[5,6], Nicola Lambri[7,8], Alessia Giuliano[9], Emiliano Loi[10], Michele Signoriello[11], Marco Branchini[12], Gina Belmonte[3], Marco Giannelli[9], Pietro Mancosu[7], Cinzia Talamonti[5,6], Mauro Iori[13], Sabina Tangaro[14], Michele Avanzo[15]*, Raffaella Massafra[1]

1 Laboratorio Biostatistica e Bioinformatica, I.R.C.C.S. Istituto Tumori 'Giovanni Paolo II', Bari, Italy, **2** Servizio di Fisica Sanitaria, Azienda Ospedaliero-Universitaria di Parma, Parma, Italy, **3** Fisica Sanitaria, Azienda Usl Toscana Nord Ovest, Lucca, Italy, **4** Fisica Sanitaria, Azienda Sanitaria dell'Alto Adige, Bolzano, Italy, **5** Dip. Scienze Biomediche Sperimentali e Cliniche "Mario Serio", Università degli Studi di Firenze,Viale Morgagni, Firenze, **6** Istituto Nazionale di Fisica Nucleare, Sez. Firenze, Via Sansone 1, Sesto Fiorentino, Firenze, **7** IRCCS Humanitas Research Hospital, Medical Physics Unit of Radiotherapy and Radiosurgery Department, via Manzoni, Rozzano, Milan, Italy, **8** Department of Biomedical Sciences, Humanitas University, via Rita Levi Montalcini, Pieve Emanuele, Milan, Italy, **9** U.O.C. Fisica Sanitaria, Azienda Ospedaliero-Universitaria Pisana, Pisa, Italy, **10** SC Fisica Sanitaria, IRCCS Istituto Romagnolo per lo Studio dei Tumori (IRST) "Dino Amadori", Meldola, Italy, **11** Fisica Sanitaria, Azienda sanitaria universitaria Giuliano Isontina, Trieste, Italy, **12** Fisica Sanitaria, Azienda Socio Sanitaria Territoriale della Valtellina e dell'Alto Lario, Sondrio, Italy, **13** Medical Physics Unit, Azienda USL-IRCCS di Reggio Emilia, Reggio Emilia, Italy, **14** Dipartimento di Fisica Applicata, Università degli Studi di Bari Aldo Moro, Bari, Italy, **15** Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, Via F. Gallini, Aviano, Italy

* michele.maddalo@ymail.com (MM); mavanzo@cro.it (MA)

OPEN ACCESS

Citation: Fanizzi A, Fadda F, Maddalo M, Saponaro S, Lorenzon L, Ubaldi L, et al. (2024) Developing an ensemble machine learning study: Insights from a multi-center proof-of-concept study. PLoS ONE 19(9): e0303217. https://doi.org/10.1371/journal.pone.0303217

Editor: Muhammad Umer, Islamia University of Bahawalpur: The Islamia University of Bahawalpur Pakistan, PAKISTAN

Received: November 15, 2023

Accepted: April 21, 2024

Published: September 10, 2024

Data Availability Statement: The data was obtained from the open-access database of M. Kirienko et al. (2020). Relevant data is also within the paper.

Funding: This work was supported by funding from the Italian Ministry of Health, Ricerca Corrente 2023/2024. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Background

To address the numerous unmeet clinical needs, in recent years several Machine Learning models applied to medical images and clinical data have been introduced and developed. Even when they achieve encouraging results, they lack evolutionary progression, thus perpetuating their status as autonomous entities. We postulated that different algorithms which have been proposed in the literature to address the same diagnostic task, can be aggregated to enhance classification performance. We suggested a proof of concept to define an ensemble approach useful for integrating different algorithms proposed to solve the same clinical task.

Methods

The proposed approach was developed starting from a public database consisting of radiomic features extracted from CT images relating to 535 patients suffering from lung cancer. Seven algorithms were trained independently by participants in the AI4MP working group on Artificial Intelligence of the Italian Association of Physics in Medicine to discriminate metastatic from non-metastatic patients. The classification scores generated by these algorithms are used to train SVM classifier. The Explainable Artificial Intelligence approach is applied to

Competing interests: The authors have declared that no competing interests exist.

the final model. The ensemble model was validated following an 80–20 hold-out and leave-one-out scheme on the training set.

Results

Compared to individual algorithms, a more accurate result was achieved. On the independent test the ensemble model achieved an accuracy of 0.78, a F1-score of 0.57 and a log-loss of 0.49. Shapley values representing the contribution of each algorithm to the final classification result of the ensemble model were calculated. This information represents an added value for the end user useful for evaluating the appropriateness of the classification result on a particular case. It also allows us to evaluate on a global level which methodological approaches of the individual algorithms are likely to have the most impact.

Conclusion

Our proposal represents an innovative approach useful for integrating different algorithms that populate the literature and which lays the foundations for future evaluations in broader application scenarios.

1. Introduction

In recent years, several kinds of machine learning-based algorithms have been used to address the problems brought on by -omics analyses of clinical data and medical images [1–3]. These models integrate various strategies and methodologies at different stages of their development, from feature selection to data pre-processing until classifier training.

Although encouraging results have been achieved, they have limited application in clinical practice. Frequently in the literature, the models proposed lack evolutionary progression, thus perpetuating their status as self-contained entities. A way to overcome this kind of limitation, is with the development of ensemble predictive models [4–11], a class of techniques that aggregate multiple basic models, thus outperforming the individual constituents.

However, to the best of our knowledge, the ensemble approaches proposed in the literature concern the integration of single algorithms or classifiers and not complex models ranging from data pre-processing to final classification.

In this scenario, we postulated that different algorithms which have been proposed in the literature to address the same diagnostic task, can be aggregated to enhance classification performance.

In this paper, we present the result of a multicenter experience in which each participant independently developed a classification algorithm using the techniques deemed most appropriate by each research group. The trained models were then used to develop an ensemble model to solve a specific diagnostic task. Moreover, we exploited the potentialities of the emerging Explainable Artificial Intelligence (XAI), and we applied a known algorithm to extract additional information for the final user allowing a more reliable use of the instrument [12–16].

Explainable artificial intelligence (XAI) could be closed the gap between the medical professionals and the artificial intelligence algorithms. In the oncological fields, several works have been proposed to explain machine learning and deep learning diagnostic and prognostic outcome [17–23]. In this paper, we applied that approach to explain the result of end-to-end ensemble machine learning model trained to solve a specific medical task independently.

Starting from a public dataset of radiomic features extracted from CT images, each research group implemented its own Machine Learning classification algorithm to predict the lung metastatic state [24].

To further improve the predicted power of the individual models, they were aggregated together via suitable Ensemble Methods like the Classifier algorithm. Afterwards, an algorithm of XAI has been implemented to add new information to the result of the classification generated from the ensemble algorithm to explain which algorithms impacted more on the final result. The innovative approach proposed in this work allows the integration of different algorithms aimed at solving the same search task on the same set of training data. The resulting ensemble model in association with emerging XAI techniques has the potential to define a more robust prediction model on the one hand, and on the other the identification of the individual algorithms that contribute most to achieving the final result. Indeed, XAI algorithm could be useful both to evaluate a result with a borderline classification score more consciously and to evaluate the characteristics of the algorithms that have contributed most to the achievement of the output, generated by the ensemble classifier.

2. Materials and methods

2.1 Experimental dataset

A public dataset from M. Kirienko et al., of patients who underwent FDG PET/CT in the period 1 January 2011 and 27 June 2017 for the detection of lung nodule [24], was used in the present study. The database includes radiomic features of both CT and PET images of patients with lung lesions, but we chose to focus on the CT ones for this study. The CT cohort consists of 535 patients with lung cancer. Following the histological data, the lung cancers can be divided in metastatic (131 cases, 24.49%) and primary (404 cases, 75.51%). All the clinical data are listed in the Table 1 below.

The choice of this specific dataset was dictated by the consistency of the cases and the number of features. Furthermore, the dataset was numerically unbalanced in the two classes of interest (metastatic/non-metastatic). A particular factor that allowed us to make appropriate considerations during the study.

Seven different algorithms were developed independently by the research groups related to centres involved in this study to differentiate between primary and metastatic lung lesions from CT-based radiomic features. Fig 1 shows the common workflow followed by all the centres, consisting of a split of the starting database in two sub-datasets: one used for training validation and a second one used as independent final test. In our specific case: the entire dataset has been initially split in stratified way between a validation subset of 428 patients (105 metastatic cases), 80% of the total, and a subset for independent test of 107 patients (26 metastatic cases), 20% of the total.

In each model, the validation sub-subset of 428 patients was used for training via a 10-fold cross-validation in 100 external rounds; the small dataset of 107 patients is, instead, used as independent final test.

Performances were evaluated for both training validation and independent test, in terms of the Area Under the Curve (AUC), Accuracy, Sensitivity, Specificity, Precision and F1 score. Moreover, to evaluate how well the classifier ensemble model forecasts the likelihood of each observation as the actual value with respect to the individual algorithms we also calculated the log-loss metric. Log Loss is similar to accuracy, but it will favor models that distinguish more strongly among classes.

Table 1. Table of the clinical features of the CT dataset. For each feature the absolute value and its frequency is shown.

Clinical Feature	Distribution
Age (years)	
(Median; 1st-3th quantile)	(67.3; 69.6; 26–90)
Gender	
Male (abs; %) Female (abs; %)	(356; 66.7%) (178; 33.3%)
Scanner	
Biograph (Siemens) (abs; %)	(234; 43.8%)
Discovery 690 (GE) (abs; %)	(300; 56.2%)
Histological type	
Primary lung tumor (abs; %)	(404; 75.5%)
Metastatic (abs; %)	(131; 24.5%)
Histological type of primary lung cancer	
Adenocarcinoma (abs; %)	(190; 35.6%)
Squamous cell carcinoma (abs; %)	(134; 25.1%)
Other primary lung cancer (abs; %)	(80; 19.8%)
Metastasis site (abs; %)	
Colorectal (abs; %)	(65; 49.6%)
Lymphoma (abs; %)	(20; 15.3%)
Breast (abs; %)	(14; 10.7%)
Other (abs; %)	(32; 22.9%)

https://doi.org/10.1371/journal.pone.0303217.t001

Each model was built independently by the developers following the same workflow scheme (see Fig 1). Users were free of using different softwares, platforms, libraries and steps in the machine learning pipeline including the balancing methods, feature selection mechanisms, and classifiers. The implementation code to develop Ensemble model was re-adapted from the Python code at the following link: https://keras.io/exples/vision/and then run by using Colab-Pro Notebook [25]. A brief discussion of the technical details of the specific algorithm is presented in the Supplementary Material.

2.2 Ensemble methods

In the machine learning field, ensemble learning refers to a series of methods that use multiple models or algorithms to achieve better predictive performance than that obtained from the same models applied individually [26, 27]. The approaches known to the state of the art are divided into three fundamental typologies [28]: (a) Bagging: this technique aims to create a set of classifiers each of which will provide a classification result; the final output will be the class that has received the highest number of votes; (b) Boosting: unlike bagging, each classifier influences the final vote with a different weight; the final result will be weighted with respect to the classification accuracy of the individual classifier in the learning phase; (c) Stacking: a further classifier is introduced (called a meta-classifier) which uses the predictions of other sub-models to carry out further learning. Bagging and boosting methods are used to combine algorithms of the same kind, whereas stacking can be used to aggregate algorithms of different types. Therefore, in this paper we used a stacking approach. The classification score generated by each previously independently trained machine learning algorithm became a “feature” for this classifier. This methodology is also called Meta-learning because it is a process of learning from learners in which a higher-level classifier is trained to improve the predictions [29]. The

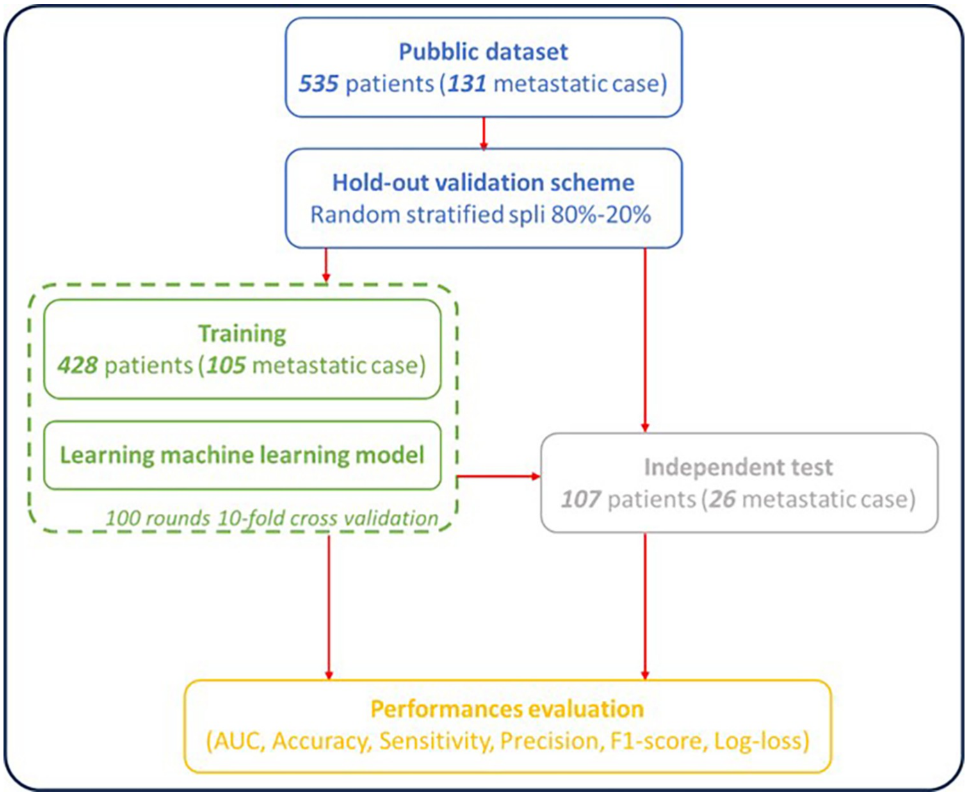

Pubblic dataset
535 patients (131 metastatic case)
Hold-out validation scheme
Random stratified spli 80%-20%
Training
428 patients (105 metastatic case)
Learning machine learning model
100 rounds 10-fold cross validation
Independent test
107 patients (26 metastatic case)
Performances evaluation
(AUC, Accuracy, Sensitivity, Precision, F1-score, Log-loss)

Fig 1. Workflow for implementing machine learning algorithms in the challenge. A same 80:20 hod-out validation scheme was used for all initially trained machine learning algorithms. Additionally, each machine learning algorithm was trained on the training sample and validated in 100 10-fold cross-validation rounds. The algorithms thus defined were validated on the independent dataset. Performances were evaluated for both training validation and independent test, in terms of the Area Under the Curve (AUC), Accuracy, Sensitivity, Specificity, Precision and F1 score.

https://doi.org/10.1371/journal.pone.0303217.g001

kay principle of meta-learning is that learning proceeds faster with more experience, via the acquisition of inductive biases or knowledge that allows for more efficient learning in the future [30].

[29]. As summarized In Fig 2, the general procedure is as follows:

- Each score of the various involved algorithms was averaged and concatenated obtaining final arrays consisting of seven features
- SVM classifier with rbf kernel was used to calculate the training score turned binary via Youden test index, and the final training performances are evaluated [31]. Leave-one-out validation was used to evaluate the model on training set. We also checked that the choice of other classifiers, such as Random Forest, XGBoost, Logistic Regression, did not significantly improve the model;
- The same training validation Youden threshold was then used for the independent test on the subset (107,7) and the final test performances were computed;
- A XAI approach was then adopted to unveil which feature-algorithm impacted more on the final prediction for positive (metastatic patient) and negative cases (no metastatic patient), as described in subsection 2.3.

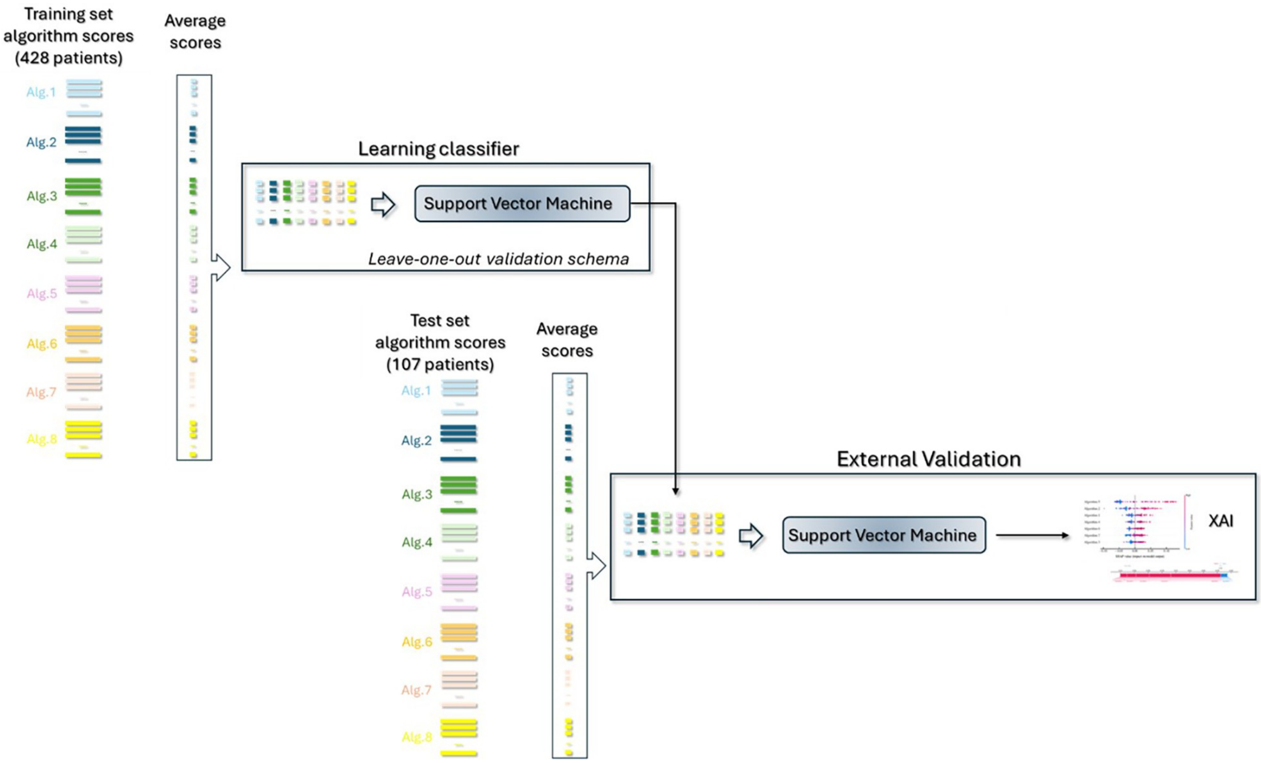


Fig 2. Workflow of the classifier Ensemble method. The scores of the various algorithms were averaged and aggregated; they became "features" of a ensemble machine learning model. Final performances for train and test were evaluated and an XAI approach was implemented to explain which feature-algorithm impacted more on the final predictions.

https://doi.org/10.1371/journal.pone.0303217.g002

The other two approaches, such as bagging and boosting, were also evaluated in the experimental phase. However, not having reported better performances also given the preliminary nature of the proposed study, we preferred not to burden the discussion with the exposition of these results.

2.3 The explainable algorithm

The XAI mechanism to explain the final classification on a specific patient was implemented via the cutting-edge local explanation SHAP Python package. The SHAP package consists of a local model-agnostic approach using only a classifier's input and output [32, 33]. In practice, it evaluates the Shapley values, corresponding to the contribution of each algorithm-feature to the final prediction referred to each individual test sample by evaluating each marginal contribution with respect to all the features considered together. For a classification model, the range of the values will be -1 to 1, then because the range of the model output is a probability between 0 to 1. The absolute SHAP value of a feature increases when its weight is larger in defining the classification score.

For the present study we adopted two evaluation tools built on SHAP values. The first tool was the Bee Swarm Plot used to display an information-rich summary of how the features of a dataset impact the model output. It provides a general picture of the SHAP values in the entire reference sample. For each predictor, the distribution of the Shapley values calculated on the reference sample is reported. The colour of each instance is used to display the value of the reference feature.

The second tool was the force plot. For a specific case, the force plot is another way to view the effect each feature has on the prediction. In this plot the positive SHAP values are displayed on the left side and the negative on the right side, as if competing against each other.

3. Results

Each algorithm produced validation-training scores matrix with dimension (428,100) and independent test scores matrix (107, 1000) that were then used for the Ensemble method.

Fig 3 shows the pie-charts of the usage of various softwares (a), balancing techniques (b), adopted classifiers (c), and feature selection (d) among the participants in the challenge. Regarding the choice of the software in Fig 3A, 5 algorithms have been developed with Python, 1 algorithm with Matlab and 1 with R software. 5 out of 7 research groups have implemented balancing techniques in their algorithm, such as SMOTE, ADASYN, the hybrid SMOTEEN or downsampling (Fig 3B). In 2 algorithms feature reduction mechanism was not implemented (Fig 5D), and finally, different classifiers were used because they were evaluated as more performing in the analysis phase for the specific algorithm implemented, such as Random Forest, XGBoost, neural network, Linear Discriminant Analysis (Fig 5C).

Fig 4 shows the heatmaps of the correlation coefficient for both training (Fig 4A) and test (Fig 4B). It shows high correlation coefficients among the 7 involved research groups demonstrating that there is agreement between all the algorithms despite the different methods used. All the correlations were statistically significant with p-value < 0.05.

As described in the methods, the average of the scores generated in the different cross-validation rounds of each algorithm became a 'feature' of the ensemble model. Fig 5 shows the boxplots of the distributions of the scores thus obtained on the training and test set with respect to the label. The distribution of the classification scores generated by the ensemble model is also reported. It should be noted that in our particular case, the median of the score distribution of positive cases generated by the ensemble classifier was significantly lower than that of all the algorithms (p-value <0.05) except for algorithms 3 and 5.

Fig 6 shows the comparison of ROC curves and the resulting AUC values on training and test sets. On the training set we found that the optimal threshold was 0.41. The AUC, Sensitivity, Specificity, Accuracy, Precision and F1 score of the Classifier Ensemble SVM Method were 0.74, 0.59, 0.79, 0.74, 0.48, and 0.53 for training, 0.75, 0.62, 0.83, 0.78, 0.53, and 0.57 for the independent test.

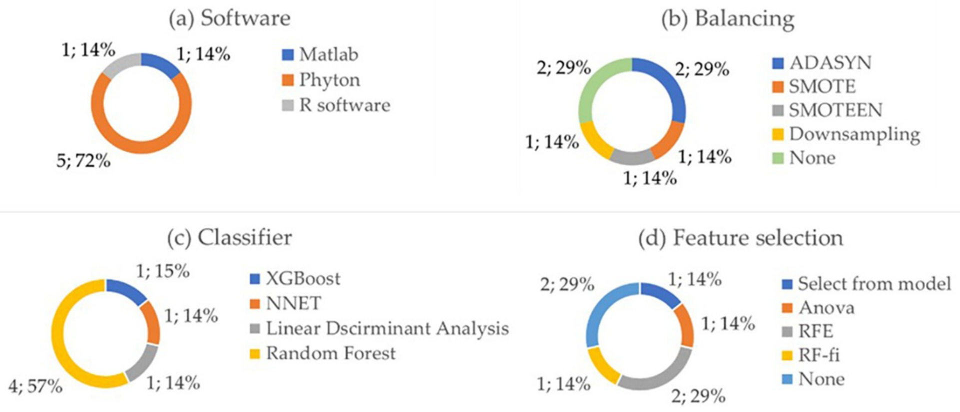


Fig 3. Pie charts of the adopted software (a), balancing technique (b), adopted classifier (c) and feature selection technique (d) by the various algorithms.

https://doi.org/10.1371/journal.pone.0303217.g003

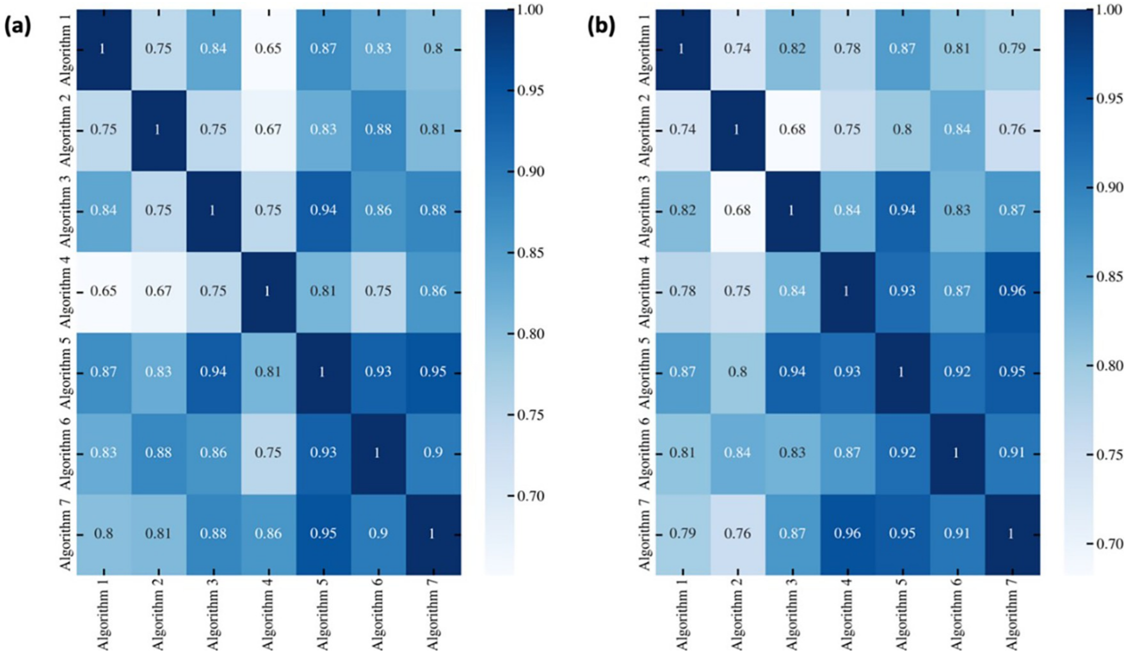


Fig 4. Heatmaps of the correlation coefficients among the classification score of all the seven algorithms for training (a) and test (b).

https://doi.org/10.1371/journal.pone.0303217.g004

Fig 7 shows the radar plots of the performances of all the seven algorithms (dashed lines) and the Classifier Ensemble model both for training (a) and test (b). Compared to the single algorithms, the ensemble algorithm showed an inverted behaviours with respect to the sensitivity and specificity metrics. Probably, this is due to the distributions of classification scores generated by the SVM classifier.

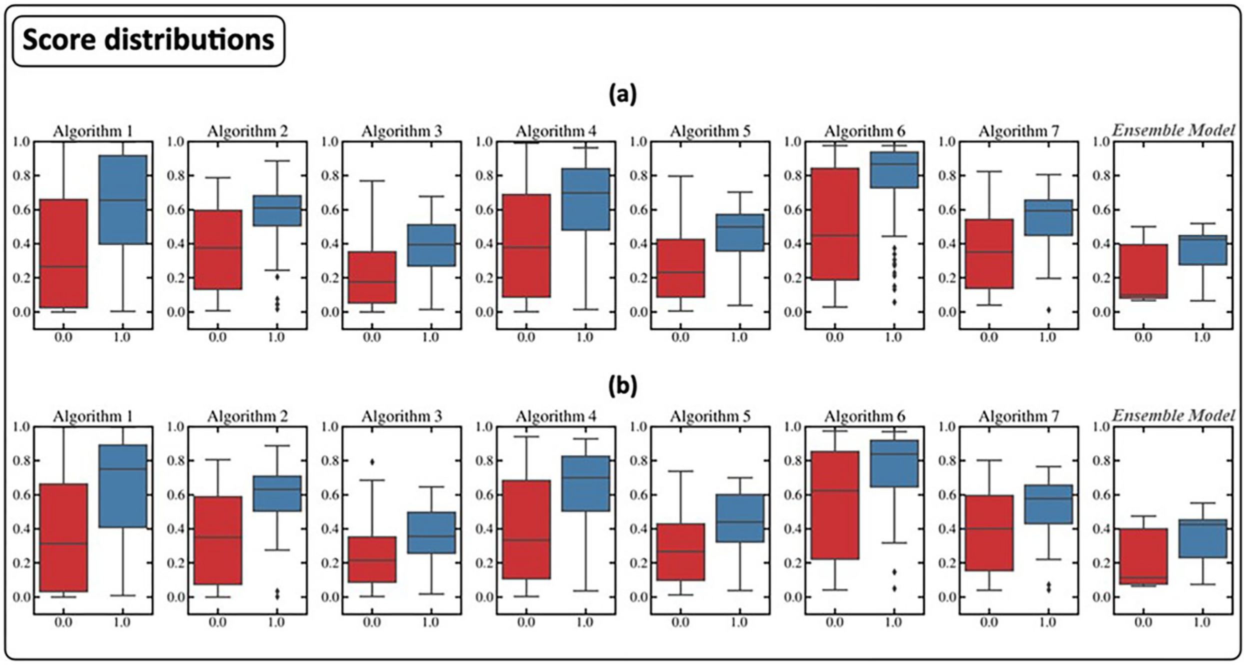


Fig 5. Score distributions for training (a) and test (b) of the various algorithms and the Classifier Ensemble model.

https://doi.org/10.1371/journal.pone.0303217.g005

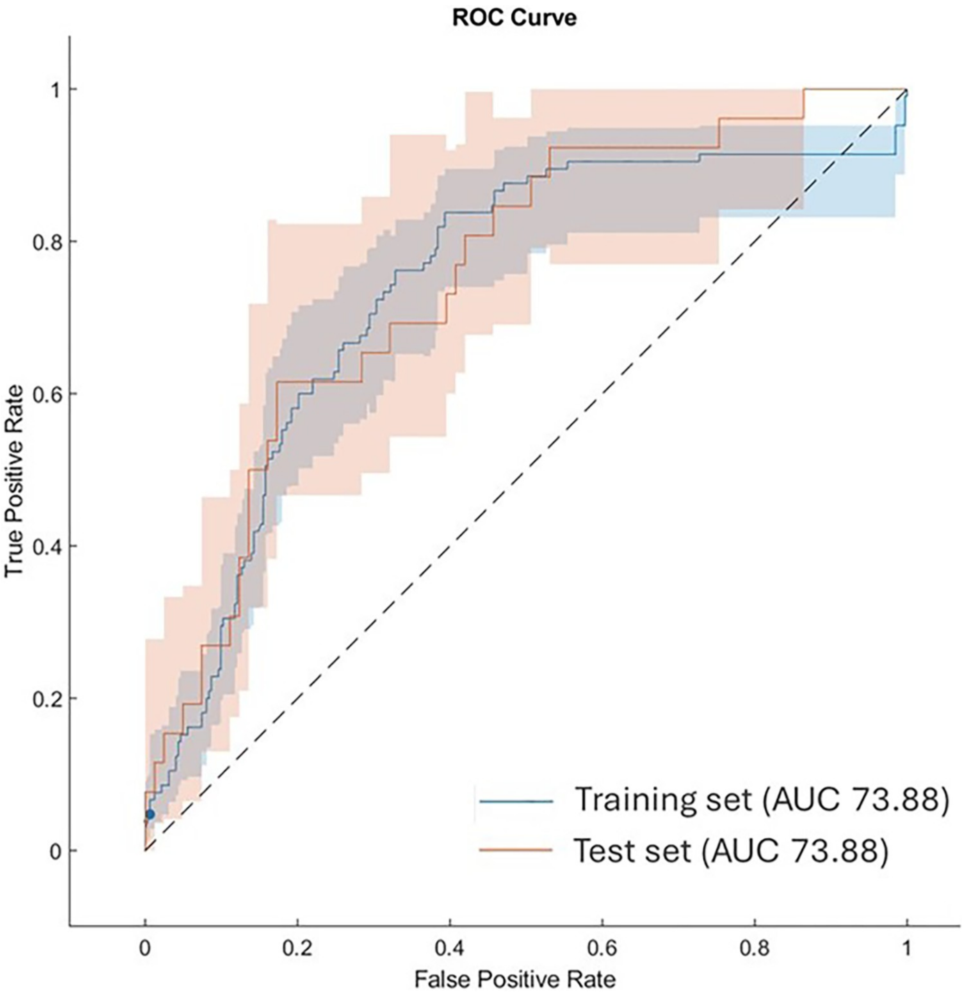


Fig 6. Comparison of ROC curves and the resulting AUC values. Blues curve: Ensemble model in Leave-one-out validation scheme over the training set; Red curve: Ensemble model over the test set. The shaded area around each curve indicates the confidence intervals at 95% level.

https://doi.org/10.1371/journal.pone.0303217.g006

However, accuracy is not indicative of how close the prediction probability is to the corresponding actual/true value. We therefore also calculated the log-loss metric, which is indicative of how close the prediction probability is to the corresponding actual/true value. The log-loss metric shows that the ensemble model outperformed the single algorithms on the test set (Fig 8).

To identify the algorithms which most impacted on the classification result of the ensemble model, after aggregating the results of all the seven algorithms, we implemented a XAI approach. Fig 9 shows the bee-swarm plot of all the algorithms that contributed more on the positive and negative prediction. Higher (lower) values for each algorithm, indicates a prediction more oriented towards the positive (negative) class. The right bar indicates the value of the feature correspondent to the classification score of the single algorithm. Algorithms 5 and 2 impacted more than the others on the final classification of the Ensemble model. In particular, high classification scores of these two algorithms were more oriented towards the positive (1) class. It should be noted that the strong impact of algorithm 5 justified the lowering of the median classification score of positive cases, which then led to an inversion of the behavior of the sensitivity and specificity classification metrics.

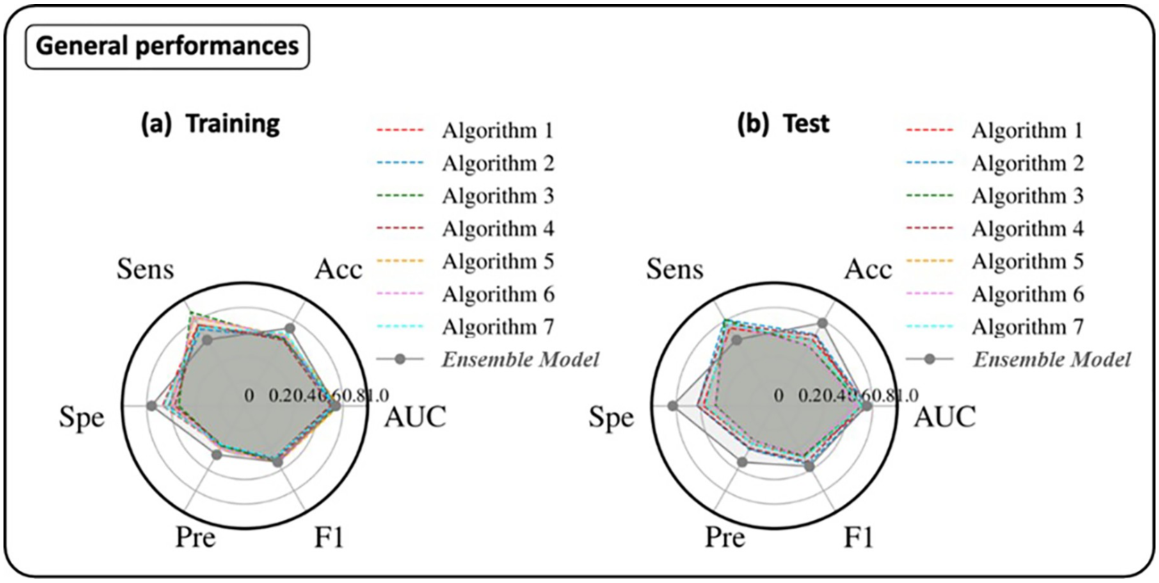


Fig 7. Radar plots of the performances of the various algorithms (dashed lines) and the Classifier Ensemble model for training (a) and test (b). The performance metrics were AUC, Accuracy (ACC), Sensitivity (Sens), Specificity (Spe), Precision (Pre) and F1 score.

https://doi.org/10.1371/journal.pone.0303217.g007

Algorithm 3 is the only one that did not apply any outlier identification, feature selection and dataset balancing techniques. His contribution to the definition of the final score was lower on average.

Purely by way of example in Fig 10 is shown the Force-plots of a particular patients. It is a graphical representation of how the individual algorithm-features contributed to achieving the classification score for a specific patient. Specifically, the algorithm features associated with a blue bar reduced (lower) the classification score, i.e. the probability that the patient is metastatic. On the other hand, the algorithm features associated with a red bar increased (higher) the classification score, i.e. the probability that the patient is metastatic. Considering that the cut-off used to binarize the classification result identified by index Youden was 0.40, this case is borderline. This tool would allow for a more accurate assessment of the classification result. Algorithm 5 was the one that pushed the classification score towards the non- metastatic class, however this algorithm had a low specificity (0.58). This tool could provide additional information to the simple classification result, allowing the final user of the model to pay attention to particular cases.

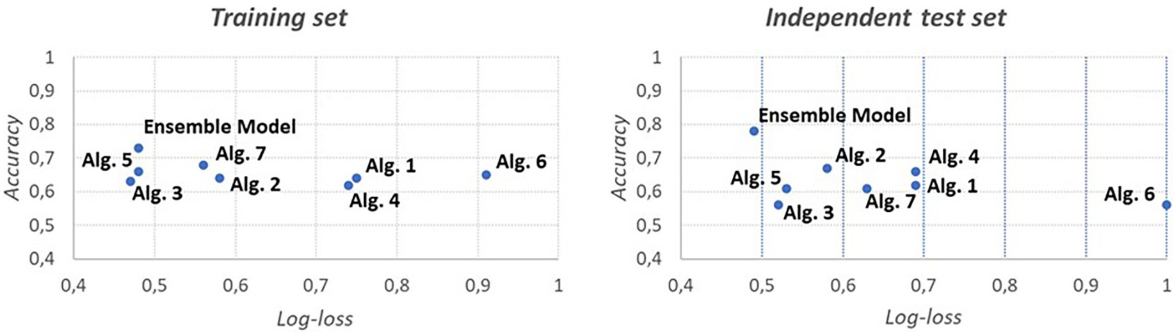
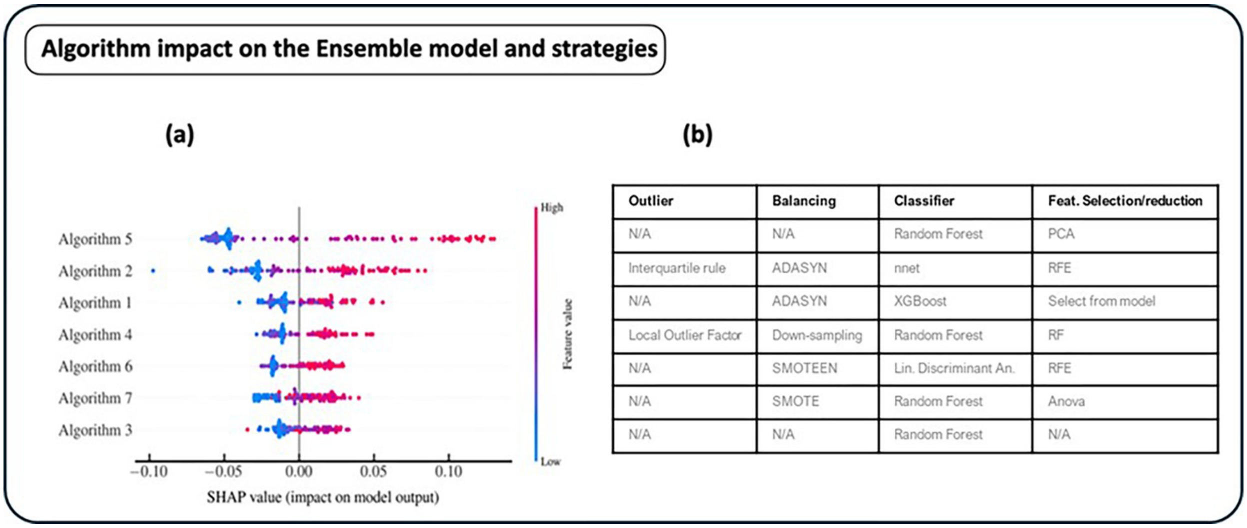


Fig 8. Charts of log-loss metrics for the various algorithms for training and test. Each algorithm has been averaged first and then used for the comparison.

https://doi.org/10.1371/journal.pone.0303217.g008



Outlier	Balancing	Classifier	Feat. Selection/reduction
N/A	N/A	Random Forest	PCA
Interquartile rule	ADASYN	nnet	RFE
N/A	ADASYN	XGBoost	Select from model
Local Outlier Factor	Down-sampling	Random Forest	RF
N/A	SMOTEEN	Lin. Discriminant An.	RFE
N/A	SMOTE	Random Forest	Anova
N/A	N/A	Random Forest	N/A

Fig 9. Bee-swarm of the global model (a) and table of the correspondent strategies (b) adopted by the specific algorithm (outlier mechanism, balancing technique, used classifier, and feature selection algorithm).

https://doi.org/10.1371/journal.pone.0303217.g009

4. Discussion and conclusion

Although the last twenty years have witnessed an exponential growth in applications of machine learning models in the oncology field to respond to the main clinical tasks, very few have seen real use in clinical care [34, 35]. Furthermore, the different models proposed to solve the same clinical problem are often heterogeneous and represent neither an evolution of previous models nor an integration of the previous ones. In this paper, we propose an approach that could be used to combine the different relatively accurate models to obtain a better performing and robust model.

The proposed study represented the result of a national challenge promoted by the Artificial Intelligence working group of the Italian Association of Medical Physics. The working group was created with the aim of investigating new methodological ideas for the analysis of biomedical data using artificial intelligence models [36].

Seven research groups have independently developed machine learning algorithms aimed at differentiating between primary and metastatic lung lesions in oncological patients. These algorithms adopted different analysis schemes in terms of data pre-processing, feature selection, classifier. Different data analysis softwarwere used.

We therefore trained a classification model on the scores generated by these algorithms to obtain more accurate and robust performance.

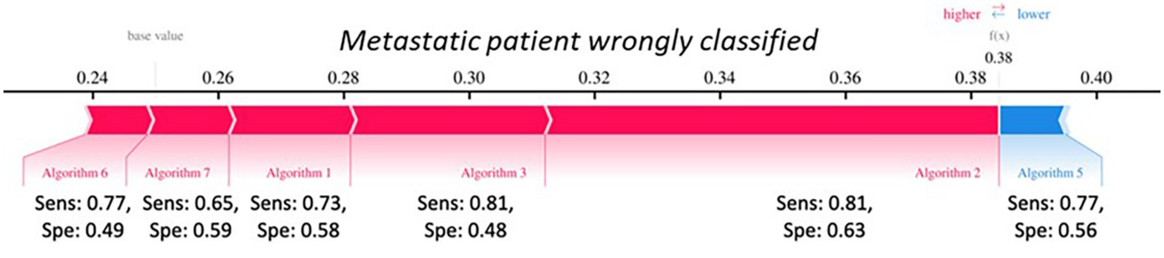


Fig 10. Force-plots of no metastatic sample wrongly classified.

https://doi.org/10.1371/journal.pone.0303217.g010

Ensemble models offer several advantages over single models in terms of performance, especially for complex problems and noise [6, 37–39]. By combining multiple trained models to make predictions and using different subsets and predictors of the data, the distribute errors among sub-models of different complexities. Therefore, they can reduce the risk of overfitting and underfitting. In the oncology field, there are various applications of ensemble approaches of machine learning algorithms proposed in the literature [40–44]. However, as far as we know, these proposals only concerned the combination of different machine learning classifiers (SVM, Random Forest, Logistic Regression, etc.) and not of algorithms, which are therefore more heterogeneous under other aspects other than the classifier used, such as the feature selection technique, the balancing method, and the pre-processing operations.

In this paper, we have aggregated the classification scores of the individual algorithms developed by the different research groups using a state-of-the-art classifier. The ensemble model achieved in terms of accuracy, F1-score, and log-loss. On the independent test the ensemble model achieved an accuracy of 0.78, a F1-score of 0.57 and a log-loss of 0.49.

The performance of the ensemble model was significantly better, especially in terms of specificity, also compared to the same work from which the dataset was extrapolated, as reported in Fig 7. The performance of the ensemble model was significantly better, also compared to the same work from which the dataset was extrapolated. In [24], on the validation set, the authors declared an AUC, accuracy, sensitivity, and specificity of 70%, 78%, 95%, and 26%, respectively. Our ensemble model outperformed the proposed model in [24] showing more balanced performances in terms of sensitivity and specificity. Indeed, sensitivity and specificity were equal to 62% and 83%, respectively.

In consideration of the classification results of the proposed individual algorithms and the model proposed in [24], it emerges that the problem of classification from the radiomic features extracted from the same [24] is a hostile one. Although the main objective of the study is not the prediction of metastatic vs. nonmetastatic state, but to define a new approach for the integration of different end-to-end algorithms that is more robust than the individual algorithms, the proposed ensemble model achieves appreciable and superior performance than the individual algorithms.

We have not reported other literature works, because to the best of our knowledge no there are other works on the same reference dataset.Log Loss is useful to compare models not only on their output but on their probabilistic outcome. It should be underlined that in general terms for the evaluation of a classification model the log-loss metric is not an appropriate indicator for evaluating performance on unbalanced datasets. As it is constructed, predicting many low probabilities (associated with negative cases) results in a low log loss value.

Then, we applied the XAI paradigm. The different machine learning applications have already widely shown their potential in different aspects of medicine in the literature. However, especially in the medical field, the lack of transparency of such applications has become increasingly problematic even outside [45, 46]. XAI approach allowed us an added value for the clinical usability of a diagnostic system since it could provide to the clinician a precise evaluation of how the automated system arrived at a given decision [47–49].

For each patient, Shapley values representing the contribution of each algorithm to the final classification result generated by the ensemble model were calculated. This information represented an added value useful for evaluating the correctness/consistency of the classification result on a particular case (patient), for example by looking at the classification performance of the single algorithm which contributed more than others to achieving the final classification result. Furthermore, it also allowed to evaluate on a global level which methodological approaches of the individual algorithms are likely to have the most impact.

This evaluation allows the identification of algorithms with a lower contribution. Reading the distribution of Shapley values associated with the detail of the salient characteristics of the individual algorithms could bring out important considerations. For example, algorithm 3 was the simplest algorithm: no outlier detection, feature selection and balancing techniques had been adopted.

The tools provided by the application of the XAI approach, which were described in a purely illustrative manner in this paper, could have considerable potential in a broader application perspective. The proposed work has limitations related to the nature of the study: what is proposed is a proof of concept conceived in a small working group, but which could be re-proposed by considering different algorithms proposed in the literature. Another limitation is related to the data that have been used: they are retrospective, collected by a single institution at one time point only. Future development will concern the evaluation of the ensemble model generalizability and its comparison with the generalizability of individual algorithms. Nonetheless, we believe that what has been proposed in this paper represents an innovative approach useful for integrating different algorithms that populate the literature and which lays the foundations for future evaluations in broader application scenarios. It is important to clarify that now the integration of multiple algorithms that comes from literature cannot systematically implemented due to the heterogeneity of surrounding conditions not related to data processing and analysis. They could be identified in discrepancy on data acquisition methodologies, patient inclusion criteria, endpoint definition and more. However, the availability of approaches for the integration of multiple machine learning algorithms from literature could spontaneously place a new requirement for a tighter definition of study designs, their sharing and compliance.

Supporting information

S1 File.
(DOCX)

Acknowledgments

This work has been carried out within the working group on Artificial Intelligence for Medical Physics (AI4MP) of Associazione Italiana di Fisica Medica e Sanitaria (AIFM).

Author Contributions

Conceptualization: Annarita Fanizzi, Michele Maddalo, Sabina Tangaro, Michele Avanzo, Raffaella Massafra.

Data curation: Michele Maddalo, Sara Saponaro, Leda Lorenzon, Leonardo Ubaldi, Nicola Lambri, Alessia Giuliano, Emiliano Loi, Michele Signoriello, Marco Branchini, Gina Belmonte, Marco Giannelli, Pietro Mancosu, Cinzia Talamonti, Mauro Iori, Sabina Tangaro, Michele Avanzo, Raffaella Massafra.

Formal analysis: Annarita Fanizzi, Federico Fadda, Michele Maddalo, Sara Saponaro, Leda Lorenzon, Leonardo Ubaldi, Nicola Lambri, Alessia Giuliano, Emiliano Loi, Michele Signoriello, Marco Branchini, Gina Belmonte, Marco Giannelli, Pietro Mancosu, Cinzia Talamonti, Mauro Iori, Sabina Tangaro, Michele Avanzo, Raffaella Massafra.

Methodology: Annarita Fanizzi, Federico Fadda, Michele Maddalo.

Resources: Raffaella Massafra.

Software: Annarita Fanizzi, Federico Fadda.

Supervision: Sabina Tangaro, Michele Avanzo, Raffaella Massafra.

Validation: Annarita Fanizzi, Federico Fadda.

Writing – original draft: Annarita Fanizzi, Federico Fadda, Raffaella Massafra.

Writing – review & editing: Annarita Fanizzi, Federico Fadda, Michele Maddalo, Sara Saponaro, Leda Lorenzon, Leonardo Ubaldi, Nicola Lambri, Alessia Giuliano, Emiliano Loi, Michele Signoriello, Marco Branchini, Gina Belmonte, Marco Giannelli, Pietro Mancosu, Cinzia Talamonti, Mauro Iori, Sabina Tangaro, Michele Avanzo, Raffaella Massafra.

References

1. Shimizu H., & Nakayama K. I. (2020). Artificial intelligence in oncology. *Cancer science*, 111(5), 1452–1460. https://doi.org/10.1111/cas.14377 PMID: 32133724
2. Massafra R., Comes M. C., Bove S., Didonna V., Diotaiuti S., Giotta F.,. . . & Fanizzi A. (2022). A machine learning ensemble approach for 5-and 10-year breast cancer invasive disease event classification. *Plos one*, 17(9), e0274691. https://doi.org/10.1371/journal.pone.0274691 PMID: 36121822
3. Massafra R., Latorre A., Fanizzi A., Bellotti R., Didonna V., Giotta F.,. . . & Lorusso V. (2021). A clinical decision support system for predicting invasive breast cancer recurrence: preliminary results. *Frontiers in Oncology*, 11, 576007. https://doi.org/10.3389/fonc.2021.576007 PMID: 33777733
4. Zhang W., Chien J., Yong J., & Kuang R. (2017). Network-based machine learning and graph theory algorithms for precision oncology. *NPJ precision oncology*, 1(1), 25. https://doi.org/10.1038/s41698-017-0029-7 PMID: 29872707
5. Delgado R. (2022). A semi-hard voting combiner scheme to ensemble multi-class probabilistic classifiers. *Applied Intelligence*, 52(4), 3653–3677.
6. Dong X., Yu Z., Cao W., Shi Y., & Ma Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14, 241–258.
7. Assiri A. S., Nazir S., & Velastin S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6), 39. https://doi.org/10.3390/jimaging6060039 PMID: 34460585
8. Hsieh S. L., Hsieh S. H., Cheng P. H., Chen C. H., Hsu K. P., Lee I. S.,. . . & Lai F. (2012). Design ensemble machine learning model for breast cancer diagnosis. *Journal of medical systems*, 36, 2841–2847. https://doi.org/10.1007/s10916-011-9762-6 PMID: 21811801
9. Kumar M., Singhal S., Shekhar S., Sharma B., & Srivastava G. (2022). Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability*, 14 (21), 13998.
10. Baradaran Rezaei H., Amjadian A., Sebt M. V., Askari R., & Gharaei A. (2023). An ensemble method of the machine learning to prognosticate the gastric cancer. *Annals of Operations Research*, 328(1), 151–192.
11. Cai Z., Xu D., Zhang Q., Zhang J., Ngai S. M., & Shao J. (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*, 11(3), 791–800. https://doi.org/10.1039/c4mb00659c PMID: 25512221
12. Amoroso N., Pomarico D., Fanizzi A., Didonna V., Giotta F., La Forgia D.,. . . & Massafra R. (2021). A roadmap towards breast cancer therapies supported by explainable artificial intelligence. *Applied Sciences*, 11(11), 4881.
13. de Vries B. M., Zwezerijnen G. J., Burchell G. L., van Velden F. H., & Boellaard R. (2023). Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. *Frontiers in medicine*, 10, 1180773. https://doi.org/10.3389/fmed.2023.1180773 PMID: 37250654
14. Strumbelj E., & Kononenko I. (2010). An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11, 1–18.
15. Slack D., Hilgard S., Jia E., Singh S., & Lakkaraju H. (2020, February). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI*, *Ethics*, *and Society* (pp. 180–186).
16. Gunning D., Stefik M., Choi J., Miller T., Stumpf S., & Yang G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, 4(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120 PMID: 33137719

17. Al-Najjar H. A., Pradhan B., Beydoun G., Sarkar R., Park H. J., & Alamri A. (2023). A novel method using explainable artificial intelligence (XAI)-based Shapley Additive Explanations for spatial landslide prediction using Time-Series SAR dataset. *Gondwana Research*, 123, 107–124.

18. Silva-Aravena F., Núñez Delafuente H., Gutiérrez-Bahamondes J. H., & Morales J. (2023). A hybrid algorithm of ML and XAI to prevent breast cancer: a strategy to support decision making. *Cancers*, 15 (9), 2443. https://doi.org/10.3390/cancers15092443 PMID: 37173910

19. Hasanpour Zaryabi E., Moradi L., Kalantar B., Ueda N., & Halin A. A. (2022). Unboxing the black box of attention mechanisms in remote sensing big data using xai. *Remote Sensing*, 14(24), 6254.

20. de Vries B. M., Zwezerijnen G. J., Burchell G. L., van Velden F. H., & Boellaard R. (2023). Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. *Frontiers in medicine*, 10, 1180773. https://doi.org/10.3389/fmed.2023.1180773 PMID: 37250654

21. Massafra R., Fanizzi A., Amoroso N., Bove S., Comes M. C., Pomarico D.,. . . & Lorusso V. (2023). Analyzing breast cancer invasive disease event classification through explainable artificial intelligence. *Frontiers in Medicine*, 10, 1116354. https://doi.org/10.3389/fmed.2023.1116354 PMID: 36817766

22. Yagin F. H., Alkhateeb A., Raza A., Samee N. A., Mahmoud N. F., Colak C., & Yagin B. (2023). An Explainable Artificial Intelligence Model Proposed for the Prediction of Myalgic Encephalomyelitis/ Chronic Fatigue Syndrome and the Identification of Distinctive Metabolites. *Diagnostics*, 13(23), 3495. https://doi.org/10.3390/diagnostics13233495 PMID: 38066735

23. Elkarami B., Alkhateeb A., & Rueda L. (2016, May). Cost-sensitive classification on class-balanced ensembles for imbalanced non-coding RNA data. In *2016 IEEE EMBS International Student Conference (ISC)* (pp. 1–4). IEEE.

24. Kirienko M., Cozzi L., Rossi A., Voulaz E., Antunovic L., Fogliata A.,. . . & Sollini M. (2018). Ability of FDG PET and CT radiomics features to differentiate between primary and metastatic lung lesions. *European Journal of Nuclear Medicine and Molecular Imaging*, 45, 1649–1660. https://doi.org/10.1007/s00259-018-3987-2 PMID: 29623375

25. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O.,. . . & Duchesnay É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825–2830.

26. Polikar R. (2012). Ensemble learning. *Ensemble machine learning*: *Methods and applications*, 1–34.

27. Sagi O., & Rokach L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, 8(4), e1249.

28. Kunapuli G. (2023). *Ensemble Methods for Machine Learning*. Simon and Schuster.

29. Wang J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, 38, 90–95.

30. Thrun S., & Pratt L. (1998). Learning to learn: Introduction and overview. In *Learning to learn* (pp. 3–17). Boston, MA: Springer US.

31. Youden W.J. "Index for rating diagnostic tests", Cancer 3, 32–35 (1950).

32. Slack D., Hilgard S., Jia E.; Singh S., Lakkaraju H. Fooling LIME and SHAP. In Proceedings of the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; ACM: New York, NY, USA, February 7 2020; pp. 180–186.

33. Strumbelj E., Kononenko I. An Efficient Explanation of Individual Classifications Using Game Theory. *Journal of Machine Learning Research* 2010, 11, 1–18.

34. Miller D. D., & Brown E. W. (2018). Artificial intelligence in medical practice: the question to the answer?. *The American journal of medicine*, 131(2), 129–133. https://doi.org/10.1016/j.amjmed.2017.10.035 PMID: 29126825

35. Corti C., Cobanaj M., Dee E. C., Criscitiello C., Tolaney S. M., Celi L. A., & Curigliano G. (2022). Artificial intelligence in cancer research and precision medicine: Applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care. *Cancer Treatment Reviews*, 102498. https://doi.org/10.1016/j.ctrv.2022.102498 PMID: 36527795

36. Avanzo M., Porzio M., Lorenzon L., Milan L., Sghedoni R., Russo G.,. . . & Mettivier G. (2021). Artificial intelligence applications in medical imaging: A review of the medical physics research in Italy. *Physica Medica*, 83, 221–241. https://doi.org/10.1016/j.ejmp.2021.04.010 PMID: 33951590

37. Akinbo R. S., & Daramola O. A. (2021). Ensemble machine learning algorithms for prediction and classification of medical images. *Machine Learning-Algorithms*, *Models and Applications*.

38. Losurdo L., Fanizzi A., Basile T. M., Bellotti R., Bottigli U., Dentamaro R.,. . . & La Forgia D. (2018, March). A combined approach of multiscale texture analysis and interest point/corner detectors for microcalcifications diagnosis. In International Conference on Bioinformatics and Biomedical Engineering (pp. 302–313). Cham: Springer International Publishing.

39. Bergstra J. and Bengio Y., Random search for hyper-parameter optimization, The Journal of Machine Learning Research (2012).

40. Bellotti R., De Carlo F., Massafra R., de Tommaso M., & Sciruicchio V. (2004). Topographic classification of EEG patterns in Huntington's disease. *Neurology & clinical neurophysiology: NCN*, 2004, 37–37. PMID: 16012598

41. Assiri A. S., Nazir S., & Velastin S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(6), 39. https://doi.org/10.3390/jimaging6060039 PMID: 34460585

42. Hsieh S. L., Hsieh S. H., Cheng P. H., Chen C. H., Hsu K. P., Lee I. S.,. . . & Lai F. (2012). Design ensemble machine learning model for breast cancer diagnosis. *Journal of medical systems*, 36, 2841–2847. https://doi.org/10.1007/s10916-011-9762-6 PMID: 21811801

43. Kumar M., Singhal S., Shekhar S., Sharma B., & Srivastava G. (2022). Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning. *Sustainability*, 14 (21), 13998.

44. Baradaran Rezaei H., Amjadian A., Sebt M. V., Askari R., & Gharaei A. (2023). An ensemble method of the machine learning to prognosticate the gastric cancer. *Annals of Operations Research*, 328(1), 151–192.

45. ZCai Z., Xu D., Zhang Q., Zhang J., Ngai S. M., & Shao J. (2015). Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular BioSystems*, 11(3), 791–800. https://doi.org/10.1039/c4mb00659c PMID: 25512221

46. Vellido A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24), 18069–18083.

47. Adadi A., & Berrada M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138–52160.

48. Antoniadi A. M., Du Y., Guendouz Y., Wei L., Mazo C., Becker B. A., & Mooney C. (2021). Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11), 5088.

49. Dauda O. I., Awotunde J. B., AbdulRaheem M., & Salihu S. A. (2022). Basic issues and challenges on Explainable Artificial Intelligence (XAI) in healthcare systems. *Principles and methods of explainable artificial intelligence in healthcare*, 248–271.

50. Petrauskas V., Damuleviciene G., Dobrovolskis A., Dovydaitis J., Janaviciute A., Jasinevicius R.,. . . & Meskauskas Z. (2020). XAI-based medical decision support system model. *IJSRP*, 10, 598–607.

51. Bergstra J. and Bengio Y., Random search for hyper-parameter optimization, The Journal of Machine Learning Research (2012).

52. Kuhn M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28, 1–26.

53. Cordón I., García S., Fernández A., & Herrera F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. *Knowledge-Based Systems*, 161, 329–341.

54. Kramer O., & Kramer O. (2016). Scikit-learn. *Machine learning for evolution strategies*, 45–53.

55. Akiba T., Sano S., Yanase T., Ohta T., & Koyama M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2623–2631).