Journal of Translational Medicine

RESEARCH Open Access

An artificial intelligence-based model exploiting H&E images to predict recurrence in negative sentinel lymph-node melanoma patients

Maria Colomba Comes[1†], Livia Fucci[2†], Sabino Strippoli[3†], Samantha Bove[1], Gerardo Cazzato[4], Carmen Colangiuli[1], Ivana De Risi[3], Ileana De Roma[3], Annarita Fanizzi[1], Fabio Mele[2], Maurizio Ressa[5], Concetta Saponaro[2], Clara Soranno[6], Rosita Tinelli[2], Michele Guida[3*], Alfredo Zito[2†] and Raffaella Massafra[1†]

Abstract

Background Risk stratification and treatment benefit prediction models are urgent to improve negative sentinel lymph node (SLN-) melanoma patient selection, thus avoiding costly and toxic treatments in patients at low risk of recurrence. To this end, the application of artificial intelligence (AI) could help clinicians to better calculate the recurrence risk and choose whether to perform adjuvant therapy.

Methods We made use of AI to predict recurrence-free status (RFS) within 2-years from diagnosis in 94 SLN- melanoma patients. In detail, we detected quantitative imaging information from H&E slides of a cohort of 71 SLN- melanoma patients, who registered at Istituto Tumori "Giovanni Paolo II" in Bari, Italy (investigational cohort, IC). For each slide, two expert pathologists firstly annotated two Regions of Interest (ROIs) containing tumor cells alone (TUMOR ROI) or with infiltrating cells (TUMOR + INF ROI). In correspondence of the two kinds of ROIs, two AI-based models were developed to extract information directly from the tiles in which each ROI was automatically divided. This information was then used to predict RFS. Performances of the models were computed according to a 5-fold cross validation scheme. We further validated the prediction power of the two models on an independent external validation cohort of 23 SLN- melanoma patients (validation cohort, VC).

Results The TUMOR ROIs have revealed more informative than the TUMOR + INF ROIs. An Area Under the Curve (AUC) value of 79.1% and 62.3%, a sensitivity value of 81.2% and 76.9%, a specificity value of 70.0% and 43.3%, an accuracy value of 73.2% and 53.4%, were achieved on the TUMOR and TUMOR + INF ROIs extracted for the IC cohort, respectively. An AUC value of 76.5% and 65.2%, a sensitivity value of 66.7% and 41.6%, a specificity value of 70.0% and

[†] Maria Colomba Comes, Livia Fucci and Sabino Strippoli equally contributed to this work.

[†] Alfredo Zito and Raffaella Massafra equally contributed to this work.

*Correspondence:
Michele Guida
micguida57@gmail.com; m.guida@oncologico.bari.it

Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

55.9%, an accuracy value of 70.0% and 56.5%, were achieved on the TUMOR and TUMOR + INF ROIs extracted for the VC cohort, respectively.

Conclusions Our approach represents a first effort to develop a non-invasive prognostic method to better define the recurrence risk and improve the management of SLN- melanoma patients.

Keywords Melanoma, Artificial intelligence, Recurrence risk prediction, Digital pathology

Introduction

Malignant melanoma is one of the most aggressive skin cancers [1, 2]. Characteristics of the primary tumor, such as location, stage, ulceration, mitotic index as well as loco-regional lymph node involvement play a key role in the prediction of the risk of recurrence in melanoma patients [3–5].

Since the early 1990s, a crucial advance in the management of melanoma patients has involved the sentinel lymph node (SLN) biopsy technique, that is now routinely used as a staging procedure for patients with pT1b, pT2, pT3, and pT4 melanoma, in agreement with the 8th edition of the American Joint Committee on Cancer (AJCC) Cancer Staging Manual [6]. The frequency of SLN metastasis is growing with the increase in thickness of primary lesion and other clinic-pathologic prognostic factors, such as ulceration and number of mitoses. Although the new AJCC classification has improved the assessment of recurrence risk stratification, negative SLN (SLN-) stage IB-IIC melanoma patients represent a heterogeneous population in terms of recurrence risk. Despite the absence of lymph node involvement, stage IB-IIC patients have a very high risk of melanoma recurrence and death. Patients with stages IIB and IIC melanoma are even associated with a higher risk than stage IIIA, and a similar risk if compared to stage IIIB [7].

Anti-CTLA-4 and anti-PD1 immune checkpoint inhibitors and combination BRAF/MEK inhibitors for patients with BRAF V600 mutations has significantly improve recurrence-free survival and distant metastasis-free survival in patients with stage III melanoma after surgical resection, and they are standards of care in this setting [8–10]. Recently, two pivotal adjuvant trials of pembrolizumab and nivolumab have also reported a significant improvement in both RFS and DMFS in stage IIB-IIC disease. However, the absolute benefit is low and there are concerns about the risk of severe toxicities [11, 12].

To improve the assessment of recurrence risk, less invasive and more accurate techniques are being under investigation. Application of artificial intelligence (AI)-based models has gaining increasing interest in many fields of medicine and oncology, particularly in the image analysis branch, including the emerging digital pathology [13], which utilizes specialized scanners to digitize histological specimens, namely, glass slides, thus generating digital images. AI techniques have shown value in the automatic identification of quantitative imaging information from the raw digitalized slides directly, to be subsequently used as potential diagnostic or prognostic biomarkers. In other words, such systems are able to automatically extract not only information that are usually evaluated manually and visually by pathologists, but also *"unperceivable-to-humans"* insights hidden to naked eyes [14]. To date, the efforts to develop AI models based on digital slide analysis have shown great potential to provide accurate predictions with respect to clinical outcomes (e.g., prediction of disease recurrence or response to therapy) [15]. This kind of models is still not so popular in the field of melanoma prognosis and therapy. Only recently, some examples of AI models exploring these issues have been developed [16]; however, they do not reach performance accurate enough to be applied in clinical practice [17–19].

In this study, we wanted to contribute to this debated topic by proposing an AI model, which exploited digital slides referred to primary lesion with the purpose of predicting 2-year recurrence-free status (RFS) in SLN- melanoma patients. To the aim, Regions of Interest (ROIs) containing tumor cells alone or with infiltrating cells have been selected by expert pathologists and then automatically analysed by our AI model to classify each patient as a recurrence case or a non-recurrence case. Finally, to give an interpretation of the decision-making process, which could be usable for clinical practitioners, the regions of the image mostly contributing to the model prediction have been highlighted through the implementation of an explainable AI technique [20, 21].

Materials and methods

Data collection

This retrospective study was approved by the Scientific Board of the Istituto Tumori "Giovanni Paolo II" in Bari, Italy Prot. 1177/CE. A cohort of 71 melanoma patients, who were cared for at the same Institute, were enrolled (investigational cohort, IC). The following criteria were required for inclusion: (i) melanoma patients from stage IB-IIC with completely resected primary tumor (T) and negative SLN; (ii) with available primary tumor specimen; (iii) with available clinical information of 2-year RFS (a follow-up of at least 24 months for disease-free patients, or who presented disease recurrence within 24 months). A validation cohort of 23 melanoma patients selected according to the previous eligibility criteria was provided by Azienda Ospedaliero Universitaria Consorziale Policlinico in Bari, Italy (validation cohort, VC).

Table 1 summarize the main clinical characteristics of the IC and VC cohorts, respectively. The association between each clinical feature and the classification label (recurrence cases vs. non-recurrence cases) was evaluated by means of suitable statistical tests, i.e., Wilcoxon–Mann– Whitney test [22] for continuous features, Chi-Square Test [23] for ordinal features. A result was considered statistically significant when the p-value was less than 0.05.

For each patient, 3-µm thickness were cut from formalin-fixed and paraffin-embedded histological blocks and for each sample, staining procedures were performed on HE 600 system automated immunostainer (Ventana Medical Systems, Tucson, AZ, USA), from deparaffinization to counterstaining with hematoxylin and mounting with VENTANA HE 600 Coverslipping Activator (Ventana Medical Systems, Tucson, AZ, USA). One slide stained with hematoxylin/eosin (H&E) was selected for the digitalization process. Digital slides were finally obtained by using a high-performing slide scanner at 40×magnifcation (Aperio AT2, Leica Biosystems).

Table 1 Clinical characteristics of the investigational and validation cohorts

Characteristic, *n* (%)	**Investigational cohort**	**Validation cohort**
Outcome		
non-Recurrence	51 (71.8)	17 (73.9)
Recurrence	20 (28.2)	6 (26.1)
Gender, n (%)		
Male	39 (54.9)	15 (65.2)
Female	32 (45.1)	8 (34.8)
Age		
Median [q1; q3]	58 [52; 71]	60 [51; 70]
Tumor site, n (%)		
Trunk	37 (52.1)	15 (65.2)
Extremities	28 (39.4)	6 (26.1)
Head and neck	6 (8.5)	2 (8.7)
pT, n (%)		
T2a	21 (29.6)	3 (13.0)
T2b	8 (11.3)	3 (13.0)
T3a	17 (23.9)	4 (17.4)
T3b	9 (12.7)	5 (21.7)
T4a	7 (9.8)	1 (4.5)
T4b	9 (12.7)	7 (30.4)
Stage, n (%)		
IB	21 (29.6)	0 (0.0)
IIA	25 (35.2)	10 (43.5)
IIB	16 (22.5)	6 (26.1)
IIC	9 (12.7)	7 (30.4)

For categorical variables, percentage (%) counts are reported. For continuous values, the median and 1rst and 3rd quartiles values are indicated

ROI identification and image pre-processing

The digitalized H&E slides were gigapixel images constructed into a multi-layered "pyramid", enabling optimized real-time viewing across multiple resolutions [24]. Thereby, to be handled by AI models, a pre-processing phase was required to decrease their computational burden. First of all, two ROIs, one containing tumor cells alone (TUMOR ROI) and the other one with tumor-infiltrating cells (TUMOR+INF ROI) were manually identified on each gigapixel H&E image by two expert pathologists of our Institute (left panel of Fig. 1A). Areas including hyperpigmentation and necrosis as well as regions containing artifacts due to staining or cutting procedures were not included. Each ROI was automatically divided into sub-regions (right panel of Fig. 1A), i.e., tiles with 200×200 pixels at 40×magnifcation using QuPath open-source software [25].

An automated cell detection to identify both tumor and infiltrating cells was implemented and performed on each tile by using QuPath software. Only squared tiles containing high cell density, i.e., with a number of cells occupying at least 25% of the entire tile area were retained for further analysis. For the IC cohort, whereas a total of 3811 tiles were extracted from TUMOR ROIs (TUMOR tiles), an amount of 2732 tiles were collected from TUMOR+INF ROIs (TUMOR+INF tiles). The VC cohort counted 950 TUMOR tiles and 593 TUMOR+INF tiles. As final step of pre-processing analysis, a colour-normalization based on Macenko's method [26] was implemented to overcome possible inconsistencies during the staining process. The retained tiles of the two kinds of ROIs were later given in input to AI-based models to finally classify the corresponding patients into two classes, namely, recurrence cases vs. non-recurrence cases.

Transfer learning-based approach

Transfer learning-based approaches have gained increasing attention in the research field focused on biomedical image analysis, especially for reduced size datasets [27]. Basically, they exploit the features learned on one task by pre-trained neural networks, which have been previously trained on a huge number (millions) of natural nonmedical images to learn how to automatically extract features of different level of abstraction, i.e., from low-level features, e.g., edge and dots, to high-level features, e.g., shapes and objects, from a raw image. These features are re-used for the task of interest (in our case, RFS prediction). The most common transfer learning-based approach consists of freezing layers from a pre-trained network, and then adding and training some trainable layers on top of the frozen layers to turn the old features into predictions on the dataset under analysis. In this work, we involved some pre-trained architectures,

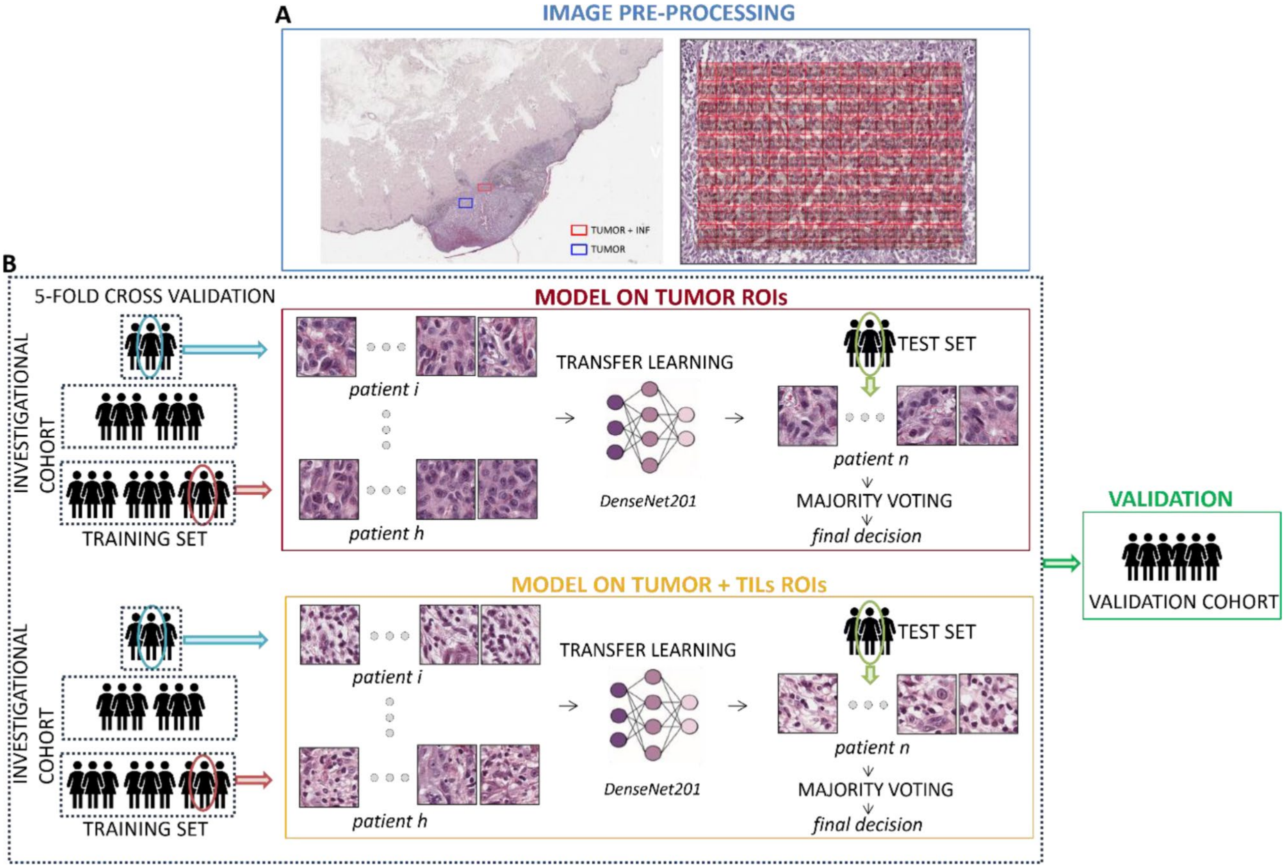

Fig. 1 Pipeline of image data analysis. (**A**) ROI identification and image pre-processing. Two ROIs, one containing tumor cells alone (TUMOR ROI) and the other one with tumor-infiltrating cells (TUMOR+INF ROI) were manually identified. Each ROI was automatically divided into sub-regions, i.e., tiles with 200×200 pixels at 40×magnifcation. (**B**) Transfer learning-based approach. Two predictive models based on DenseNet101 architecture were trained and validated according to a five-fold cross validation procedure on images related to the Investigational Cohort, i.e., taking in input the TUMOR tiles and to TUMOR+INF tiles, respectively. The two models were further validated on an external cohort of patients, the validation cohort

both CNNs, such as Xception [28], InceptionV3 [29], ResNet50 [30], Densenet201 [31] and, and Vision Transformers, ViT16 [32], and, then, we defined a criterion to choose only one of these networks to develop the learning model we proposed here. Xception is a 71 layers deep architecture, whose functioning is to apply the filters on each of the depth map and then compress the input space using 1×1 convolution across the depth. InceptionV3 has an architectural design with repeated components called inception modules. Both architectures receive 299×299 size images as input. ResNet50 architecture is a 50 layer-net belonging to the class of residual CNNs, which makes use of stacking residual blocks to train much deeper networks with the aim of maintaining compelling performances. DenseNet201 model is composed by layers receiving additional inputs from all preceding layers and passing their feature-maps to all subsequent layers. Both architectures receive 224×224 size images as input. ViT16 splits an input image in 16×16 size patches and encodes each patch as a token using self-attention map mechanism. It receives 224×224 size images as input. To decide which pre-trained network to use as training module of our model, for all the pre-trained networks taken into accounts, we extracted features from both TUMOR and TUMOR+INF tiles by using the last frozen layer, i.e., the layer immediately preceding the classification layer. A total number of 2048 features was computed by means of Xception, InceptionV3 and ResNet50; 1920 features were extracted through DensenNet201; 768 features were obtained by using ViT16. The statistical significance of each feature f_i extracted by the above mentioned pre-trained networks from both TUMOR ROIs and TUMOR+INF ROIs was assessed through the computation of an individual discriminant power (DP) [33], which was obtained as

$$DP\left(f_i\right) = max\left(AUC\left(f_i\right), \, 1 - AUC\left(f_i\right)\right)$$

where AUC stands for the area under the Receiver Operating Characteristic (ROC) curve, which expresses the general capability of the feature to discern samples with respect to a binary classification task, that, in our case, is recurrence cases vs. non-recurrence cases. The possible DP values belong to the range [0.5, 1]. A score of 0.5 indicates random guessing, while a value of 1 indicates perfect separability ability. A score slightly above 0.5 shows that a feature has at least some (albeit small) predictive power. The network showing features with the maximum DP, that was DenseNet201 (*see* Results), was then used as building block of the predictive models, which were developed starting from the tiles identified from the two ROIs type of each patient.

Pipeline of image data analysis

We developed two predictive models on images related to the IC cohort, i.e., taking in input the 3811 TUMOR tiles and 2732 TUMOR+INF tiles, respectively. The dataset of patients at disposal was divided in turn into training and test sets, in agreement with a five-fold cross validation procedure (Fig. 1B). The splitting was the same for the two models. All the tiles associated to the ROIs (either TUMOR ROIs or TUMOR+INF ROIs) of one patient were part either of the training set or the test set depending on whether the patient was assigned to the training set or the test set, respectively. The two predictive models shared the same backbone structure, that was a transfer learning module based on DenseNet201 architecture (Fig. 1B). To apply transfer learning, the last layer of the network was replaced with some consequent trainable layers: a flattening layer, a batch normalization layer, a dense layer with ReLu (Rectified Linear Unit) activation, another batch normalization layer, and a final dense layer as classifier with a sigmoid activation function. The trainable layers of the transfer learning module were trained by choosing 8 as batch size and 30 as epochs. Focal loss rather than the binary cross entropy error was defined as the loss function, with the aim of addressing class imbalance [34]. As optimizer for the weights of the network, Adam optimization algorithm with a starting learning rate of 10^{-4} was performed [35]. Data augmentation including random flip horizontally and vertically, random rotation with angles in the range [-20, 20] degrees with a step of 5 degree, and randomly contrast adjustment with a factor of 0.2, was implemented in the training phase to overcome overfitting. The implementation code was written in Python 3.6 and run using ColabPro Notebook.

The predictive models returned classification scores for each tile corresponding to one patient. To obtain a unique classification score per patient in correspondence of each kind of ROI, a majority voting rationale was applied: the final class assignment returned by a model corresponds to the class that was most frequently assigned for the tiles related to that patient (Fig. 1B). The corresponding classification score was computed as the maximum/minimum score of the models labelling the patient into the recurrence/non-recurrence class, if the class assigned by the majority voting was the recurrence/non-recurrence class, respectively. Finally, the models were validated on an external cohort of patients, the VC cohort (Fig. 1B), which involved 950 TUMOR tiles and 593 TUMOR+INF tiles.

The developed models were evaluated after majority voting by computing AUC of ROC curve and other standard metrics such as accuracy, sensitivity, specificity, and precision. Moreover, G-mean, that is the square-root of the product between sensitivity and specificity, thus allowing a balance between sensitivity and specificity, is also calculated, since it is an explanatory metric to assess model performance over imbalanced datasets, as in our case (recurrence cases represent the 28.2% and 26.1% of the IC and VC cohorts). All the metrics except for AUC are threshold metrics since their value depends on a threshold, which is usually set a priori pr defined by well-known techniques such as Youden's index. In this work, we imposed this threshold as the ratio of the number of patients belonging to the recurrence class over the total number of patients composing the training sets under study (approximated to the first decimal digit) [36].

When the models were validated on the independent cohort, the weights of the models saved for the five training sets of the cross-validation scheme of the IC cohort were used. Hence, the median and interquartile ranges (IRGs) of all the standard metrics were computed as evaluation metrics.

Explanation of the decision-making process

The decision-making process underlying the transfer learning module of the proposed model after training the trainable top layers was visually explained by applying the Local Interpretable Model-agnostic Explanations (LIME) [20, 21], whose functioning arises from the construction of a new dataset of "perturbed" samples with the corresponding predictions of the network. Then, an interpretable model is trained on the new dataset, taking into account 'local' approximations, i.e., the proximity of the sampled instances to the instance for which we want to have an explanation. When samples are images, variations of the images are generated through "superpixels" segmentation and turning superpixels off or on. The regions mainly contributing to the decision-making process, whose number is chosen a priori by users (in our case 20), are highlighted as superpixels on heatmaps overlapped to the raw images. With respect to the label predicted by the network (recurrence vs. non-recurrence), the positive contributing superpixels are colored green, whereas the regions which contribute to the assignment of that image into the predicted class are colored green.

Results

To decide which pre-trained network to use as the transfer learning module, we firstly evaluated the discriminant power of the features extracted from the last frozen layer of pre-trained networks mentioned in the Methods section. Figure 2A-B depict the percentage of the extracted features, whose DP value was higher than 0.6 (left panels), and the maximum DP score among the extracted features (right panels) for the TUMOR tiles and TUMOR+INF tiles, respectively.

From the comparison between the two kinds of tiles emerges how the percentage of features with DP score higher than 0.6 is greater for all the pre-trained networks in the case of TUMOR tiles (left panels). Such a result suggests how the TUMOR ROIs contain potential more informative content than TUMOR+INF ROIs. The best DP values were achieved by DenseNet201 architecture in correspondence of both kinds of tiles: 0.78 for TUMOR tiles and 0.66 for TUMOR+INF tiles. Hence, this architecture was then used to develop the transfer learning module of the predictive models.

The prominence of the TUMOR ROIs over the TUMOR+INF ROIs was also quantitatively confirmed in terms of performance reached by the predictive modes on the two types of ROIs, as shown in Table 2. For TUMOR+INF ROIs and TUMOR ROIs, an AUC value of 62.3% and 79.1%, a specificity value of 43.3% and 70.0%, a sensitivity value of 76.9% and 81.2%, an accuracy value of 53.4% and 73.2%, a precision value of 37.1% and 52.0%, a G-mean value of 57.7% and 75.4% were achieved, respectively. The corresponding ROC curves are depicted in Fig. 3A.

The models developed in this paper were finally tested on the external validation VC cohort of patients, reaching AUC value of 76.5% and 65.2%, a median accuracy value of 70.0% and 56.5%, a median sensitivity value of 66.7% and 41.6%, a median specificity value of 70.0% and 55.9%, a precision value of, 31.7% and 44.4%, and a median G-mean value of 68.6% and 52.4% for the TUMOR ROIs and TUMOR+INF ROIs, respectively. The resulting ROC curves are represented in Fig. 3B. These results are consistent with those reached on the IC cohort.

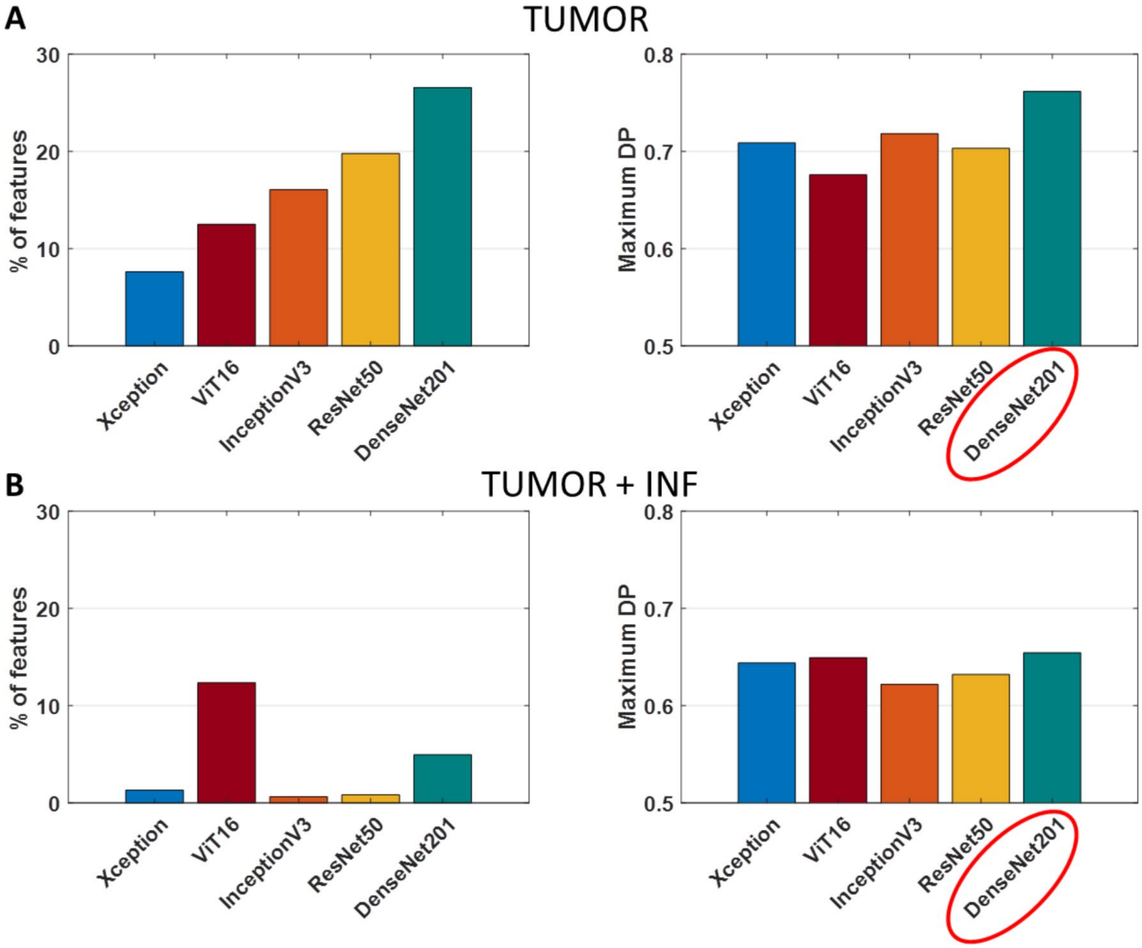

Fig. 2 Discriminant power (DP) evaluation. (**A**) The percentage of the extracted features, whose DP value was higher than 0.6 (left panels), and (**B**) the maximum DP score among the extracted features (right panels) for the TUMOR tiles and TUMOR + INF tiles, respectively

Table 2 Performance evaluation of the two AI-based models on investigational and validation cohorts, respectively. The metric values for the IC cohort are represented by a unique number according to the evaluation using a cross-validation scheme. The metric values for the VC cohort are expressed as median values and IRGs according to the evaluation using the training sets of the cross-validation scheme

Cohort	Metric (%)	ROI type	
		TUMOR+INF	TUMOR
Investigational cohort	AUC	62.3	79.1
	Specificity	43.3	70.0
	Sensitivity	76.9	81.2
	Accuracy	53.4	73.2
	Precision	37.1	52.0
	G-mean	57.7	75.4
Validation cohort	AUC	65.2 [58.3; 67.2]	76.5 [71.1; 80.2]
	Specificity	55.9 [52.9; 67.7]	70.6 [69.1; 80.9]
	Sensitivity	41.6 [33.3; 58.3]	66.7 [50.0; 66.7]
	Accuracy	56.5 [52.2; 60.9]	70.0 [68.5; 72.8]
	Precision	31.7 [25.0; 33.3]	44.4 [42.1; 52.1]
	G-mean	52.4 [46.3; 56.8]	68.6 [64.7; 68.6]

Finally, the statistical tests mentioned in the Methods Section were performed between RFS and each clinical factor to understand whether there was a relationship between the two variables: no association between RFS and the clinical factors emerged for the IC cohort ($p > 0.05$); a significant association of the RFS with stage was highlighted for the VC cohort ($p = 0.0042$). This result can be justified from the distribution of the recurrence cases in the diverse categories of the stage variable. The recurrence cases included in the IC cohort were distributed across all the categories of the stage variable: among the 20 recurrence cases, three belong to stage IB, nine to stage IIA, four to stage IIB, and four to stage IIC. Conversely, the six recurrence cases referred to the VC cohort are divided between stage IIA (one case) and stage IIC (five cases).

According to these last results, we investigated the behavior of the best model, i.e., that based on TUMOR ROIs, with respect to sub-cohorts of patients of the IC cohort divided by stage, by combining the stages with better and worse recurrence-free curves in the same category [37]. Specifically, we joined the patients with stage IB-IIA melanoma and those with stage IIB-IIC melanoma into two different categories. Then, we computed the metrics already being used: stage IB-IIA (AUC=80.3%, sensitivity=88.9%, specificity=61.5%, accuracy=68.5%), stage IIB-IIC (AUC=80.6%, sensitivity=57.1%, specificity=85.7%, accuracy=76.2%). The corresponding ROC curves are depicted in Fig. 4.

In the vein of explainable artificial intelligence, Fig 5 A-B depict a TUMOR tile and a TUMOR+INF tile, respectively, which are related to a correctly classified recurrence case (left panels) alongside to the same tiles overlaid by the areas which mostly contribute to the decision of the AI model (right panels). The red color highlights the negatively contributing areas to the assignment to recurrence class (i.e., against the occurrence of recurrence), whereas the green represents the positively contributing areas to the assignment to recurrence class (i.e., in favour of the occurrence of recurrence). This explanation was then used by our pathologists to make an informed decision on the reliability of the predictions in agreement with their expectations. Afterwards, a naked-eye analysis of the images was performed by our pathologists, who highlighted some findings. In the

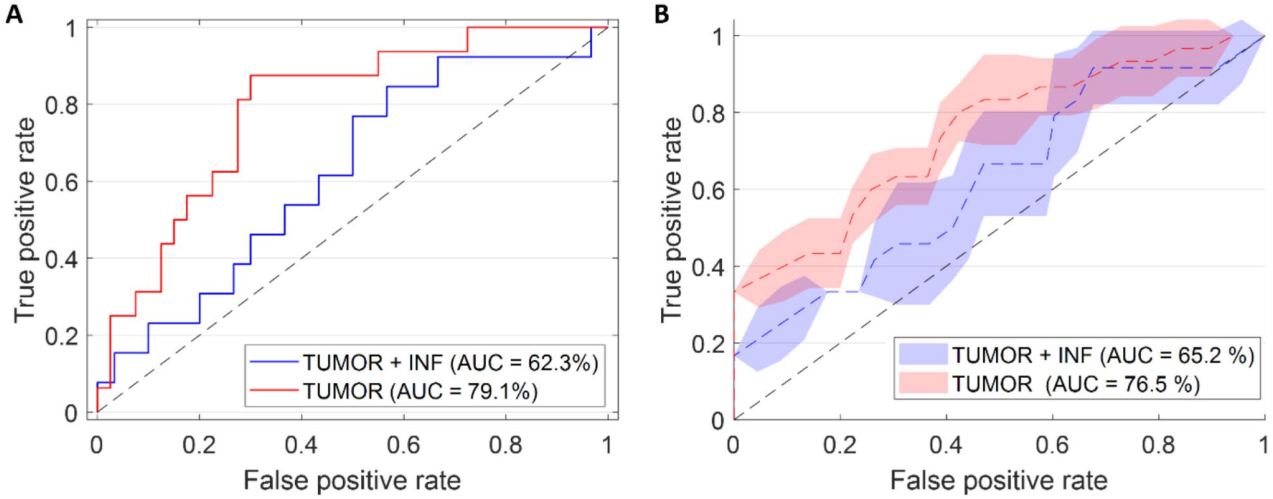

Fig. 3 ROC curves. (**A**) ROC curves related to the two AI models, analyzing tiles from TUMOR ROIs and TUMOR+INF ROIs on the IC cohort, respectively (**B**) ROC curves related to the two AI models, analyzing tiles from TUMOR ROIs and TUMOR+INF ROIs on the VC cohort, respectively. The curves are obtained by averaging the y-values over the values obtained on the VC cohort by using the diverse training sets of the cross-validation scheme in turn. The corresponding IRGs are represented as shaded areas

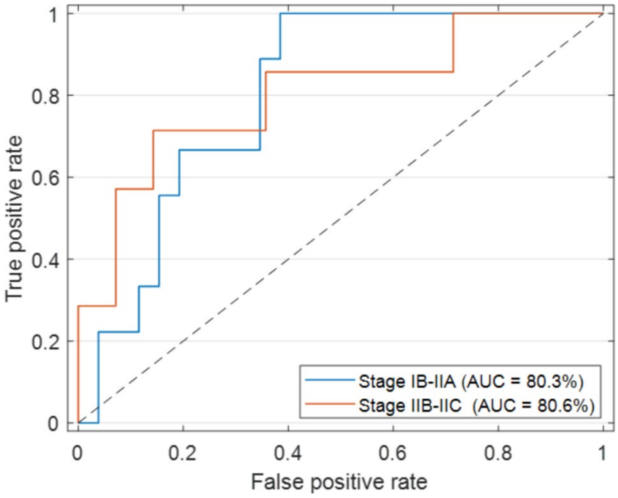

Fig. 4 ROC curves related to the best AI model by dividing the melanoma population under analysis by stage

regions including both tumor and infiltrating cells, the AI algorithm was able to automatically detect and morphologically divide tumor cells from the infiltrating ones. In addition, it assigned them in green and red areas, respectively. It confirmed pathologists' expectations since tumor cells contribute to recurrence, while infiltrating cells could have a protective effect on the spread of the disease. Conversely, the pathologists did not notice perceivable-to-humans morphological differences among more and less aggressive tumor cells involved on TUMOR tiles.

Discussion

In this study, we proposed an AI-based model exploiting ROIs annotated from H&E slides to predict recurrence risk in SLN- melanoma patients. We decided to analyze stage IB-IIC patients due to their high recurrence rate. Indeed, roughly 20% of stage IB patients and up to about 60% of stage IIC patients show recurrence within 10 years after radical surgery [37, 38]. Therefore, all these patients would deserve an adjuvant therapy, if available. Recent research works demonstrated a significantly improvement of the adjuvant therapy efficacy in terms of recurrence-free survival and distant metastasis-free survival in stage IIB-IIC and III melanoma after radical surgical resection, including immune checkpoint inhibitors for stage II/III, and a combination of BRAF/MEK inhibitors for BRAFV600 mutated stage III patients [8–12]. Notoriously, stage I and stage IIA patients were not included in those studies, and, to date, there are not available adjuvant therapies for these patients.

Despite the undoubted advantages, the absolute benefit of adjuvant therapy is low, and there are concerns about the risk of severe toxicity [8–12]. Thus, the validation of patient risk-stratification and treatment-benefit prediction models are needed to improve patient selection and limit exposure to toxicity in the large population of patients with low recurrence risk [39].

Here, we proposed an AI-model to address this challenging task. The proposed model was able to automatically extract some quantitative imaging information directly from H&E images, that are usually evaluated manually and visually by pathologists. Some interesting observation could be discussed. First, the achieved performances highlighted how the regions containing tumor cells alone have revealed as more informative than those containing both tumor and infiltrating cells. An AUC value of 79.1% and 62.3%, a sensitivity value of 81.2% and 76.9%, a specificity value of 70.0% and 43.3%, an accuracy value of 73.2% and 53.4%, were achieved on the TUMOR and TUMOR+INF ROIs extracted for the IC cohort, respectively. An AUC value of 76.5% and 65.2%, a sensitivity value of 66.7% and 41.6%, a specificity value of 70.0% and 55.9%, an accuracy value of 70.0% and 56.5%, were achieved on the TUMOR and TUMOR+INF ROIs extracted for the VC cohort, respectively. These results could be justified since the heterogeneity of the immune cell populations. Then, we deeper investigated the decision made by the algorithm by applying an explainable AI technique, which enabled us to visually interpret the regions of the input images that the algorithm as deemed as the most informative in its decision-making process. A variety of tumor microenvironment (TME) cells, namely T and B lymphocytes, natural killer cells, macrophages and dendritic cells, as well as neutrophils and fibroblasts, support the growth and invasiveness of melanoma cells, thanks to a plethora of mechanisms including secretion of pro-inflammatory molecules, induction of receptor inhibitors expression, or depletion of essential nutrients. The connection between tumor cells and TME determines the architecture and cellular plasticity of TME. These continuous changes affect tumor growth and therapeutic response with significant impact on clinical outcomes [40–43]. Understanding the interactions between tumor cells and TME is essential for the development of new algorithms to analyze different TME patterns. However, since an overall assessment of the TME is not sufficient, a detailed study of individual cellular components, their functional state, and spatial distribution is required. Therefore, our next step will be to verify if the analysis of specific cellular components of the TME and their spatial distribution is able to improve the sensitivity and specificity of our model in predicting the RFS of SLN- melanoma patients.

To the best of our knowledge, few research works investigated the potential of deep learning and transfer learning to fulfill the task of predicting recurrence in melanoma patients [17, 19]. Specifically, in our previous work [17], we developed a transfer learning model

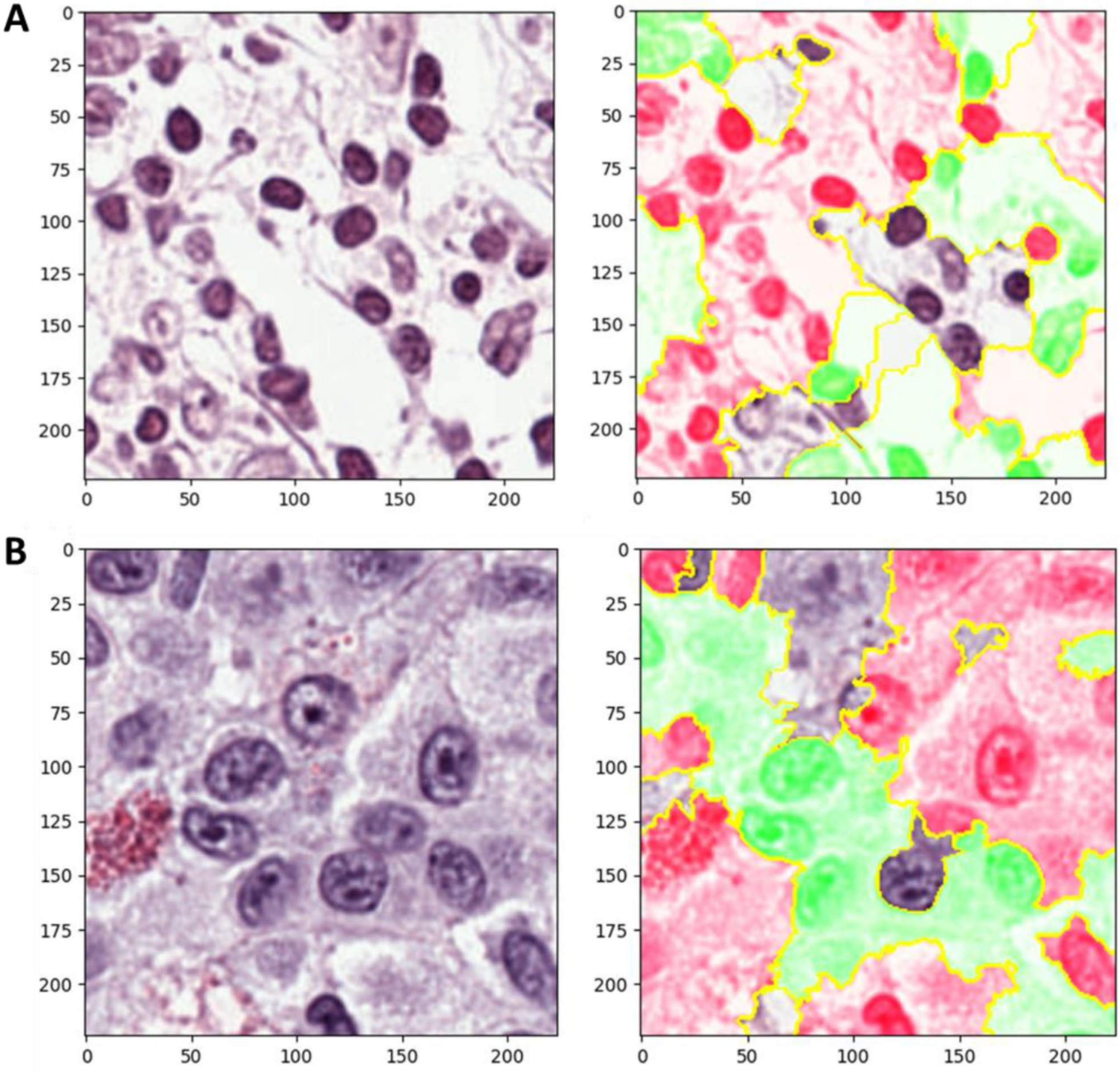

Fig. 5 Explainability results. (**A**) A TUMOR tile and a (**B**) TUMOR + INF tile related to a correctly classified recurrence case (left panels) alongside to the same tiles overlaid by the areas which mostly contribute to the decision of the AI model (right panels). The red color highlights the negatively contributing areas to the assignment to recurrence class, whereas the green represents otherwise

analyzing H&E images referred to the primary melanoma to predict 2-year RFS in stage I-III melanoma patients. An investigational cohort of 43 patients from Clinical Proteomic Tumor Analysis Consortium Cutaneous Melanoma (CPTAC-CM) public database was firstly used to train the model. A validation cohort of 11 cutaneous melanoma patients referred to our Institute was then used to test the model. Areas containing both tumor and infiltrating cells were taken into account to extract imaging information. As a result, AUC values of 69.5% and 66.7% were achieved on public and private databases, respectively. Recently, Kulkarni and colleagues [19] developed a deep learning model using H&E images of primary melanoma to predict visceral recurrence and death in stage I-III melanoma patients, reaching an AUC value of 88.0% on a validation cohort of 51 patients. Compared to these studies, we analyzed a more homogeneous population, thus achieving less biased results. In addition, we probed the differences in model performances by analyzing TUMOR ROIs and TUMOR+INF ROIs, also visualizing the areas on the image that the model judged as the most informative to the prediction. As far as we know, this is the first work addressing RFS prediction in melanoma patients in which explainable algorithms was used, thus adding value to make clearer the decision-making process of the algorithm.

Beyond the promising results, the main limitation of the study consists in the relatively small size of the dataset under study; therefore, a larger dataset is needed for a more comprehensive evaluation of the performance of the proposed model. However, the current analysis is a hypothesis-generating study which aims to answer an interesting unmet clinical need in this setting. As future work, to make the entire image analysis workflow as fully operator-independent, an automatic cell detection and classification (tumor cells and infiltrating cells), and, hence, an automatic ROI identification will be performed. Therefore, we are currently collecting data from multiple centres across Italy to re-train the optimized model on a wider cohort of patients. In conclusion, the added value of this work is represented by the automatic identification of quantitative imaging information from the raw H&E slides directly: fine tumor and infiltrating cells' characteristics, such as morphology of tumor nuclei as well as density distribution of infiltrating cells, are identified and then used as prognostic factors in SLN- melanoma patients. The promising results make this study as a valuable basis for future research investigation on wider cohorts of patients enrolled with a multi-centric study.

Acknowledgements
Not applicable.

Author contributions
Conceptualization, M.G., R.M., M.C.C., A.Z., L.F., S.S.; methodology, M.G., R.M., M.C.C., A.Z., L.F., S.S; software, M.C.C., validation, R.M., A.F., S.B.; formal analysis, M.C.C., M.G., R.M.; resources, R.M.; data curation, I.D.RI., C.C., C.SO., G.C., R.T.; writing—original draft preparation, M.G., R.M., M.C.C., L.F., C.SA., A.F., S.B.; writing—review and editing, all the authors; supervision, M.G., R.M., A.Z. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by funding from the Italian Ministry of Health "Ricerca Corrente 2022–2024".

Data availability
The raw data supporting the conclusions of this article will be made available by the authors upon request.

Declarations

Ethics approval and consent to participate
The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Scientific Board of Istituto Tumori 'Giovanni Paolo II', Bari, Italy- prot 1177/CE. The authors affiliated to Istituto Tumori "Giovanni Paolo II", IRCCS, Bari are responsible for the views expressed in this article, which do not necessarily represent the ones of the Institute.

Consent for publication
'Informed consent' for publication was collected for all the patients involved in the study, except for patients who are dead or not reachable, as it is a retrospective study (Garante della Privacy n. 9/2016 in data 15 dicembre 2016).

Competing interests
The authors declare that there are no competing interests.

Author details
[1]Laboratorio di Biostatistica e Bioinformatica, I.R.C.C.S. Istituto Tumori 'Giovanni Paolo II', Bari, Italy
[2]Unità Operativa Complessa di Anatomia Patologica, I.R.C.C.S. Istituto Tumori "Giovanni Paolo II", Bari, Italy
[3]Unità Operativa Tumori Rari e Melanoma, I.R.C.C.S. Istituto Tumori "Giovanni Paolo II", Bari, Italy
[4]Dipartimento di Medicina di Precisione e Rigenerativa e Area Jonica, Università degli Studi di Bari Aldo Moro, Bari, Italy
[5]Unità Operativa Complessa di Chirurgica Plastica e Ricostruttiva, I.R.C.C.S. Istituto Tumori "Giovanni Paolo II", Bari, Italy
[6]ASST dei Sette Laghi, Varese, Italy

Received: 7 March 2024 / Accepted: 18 August 2024
Published online: 12 September 2024

References

1. Ali Z, Yousaf N, Larkin J. Melanoma epidemiology, biology and prognosis. Eur J Cancer Suppl. 2013;11:81–91. https://doi.org/10.1016/j.ejcsup.2013.07.012.
2. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html#:~:text=Cancer%20of%20the%20skin%20is,majority%20of%20skin%20cancer%20deaths.
3. Gershenwald JE, Thompson W, Mansfield PF, et al. Multi-institutional Melanoma Lymphatic Mapping experience: the Prognostic Value of Sentinel Lymph Node Status in 612 stage I or II melanoma patients. J Clin Oncol. 1999;17:976–976. https://doi.org/10.1200/JCO.1999.17.3.976.
4. Hyams DM, Cook RW, Buzaid AC. Identification of risk in cutaneous melanoma patients: prognostic and predictive markers. J Surg Oncol. 2019;119:175–86. https://doi.org/10.1002/jso.25319.
5. Quaresmini D, Guida M. Neoangiogenesis in Melanoma: an issue in Biology and systemic treatment. Front Immunol. 2020;11. https://doi.org/10.3389/fimmu.2020.584903.
6. Balch CM, Gershenwald JE, Soong S, et al. Final version of 2009 AJCC melanoma staging and classification. J Clin Oncol. 2009;27:6199–206. https://doi.org/10.1200/JCO.2009.23.4799.
7. Yushak M, Mehnert J, Luke J, Poklepovic A. Approaches to high-risk Resected Stage II and III Melanoma. Am Soc Clin Oncol Educ B. 2019;e207–11. https://doi.org/10.1200/EDBK_239283.
8. Long GV, Hauschild A, Santinami M, et al. Adjuvant dabrafenib plus Trametinib in Stage III BRAF -Mutated Melanoma. N Engl J Med. 2017;377:1813–23. https://doi.org/10.1056/NEJMoa1708539.
9. Eggermont AMM, Blank CU, Mandala M, et al. Adjuvant Pembrolizumab versus Placebo in Resected Stage III Melanoma. N Engl J Med. 2018;378:1789–801. https://doi.org/10.1056/NEJMoa1802357.
10. Weber J, Mandala M, Del Vecchio M, et al. Adjuvant Nivolumab versus Ipilimumab in Resected Stage III or IV Melanoma. N Engl J Med. 2017;377:1824–35. https://doi.org/10.1056/NEJMoa1709030.
11. Luke JJ, Rutkowski P, Queirolo P, et al. Pembrolizumab versus placebo as adjuvant therapy in completely resected stage IIB or IIC melanoma (KEYNOTE-716): a randomised, double-blind, phase 3 trial. Lancet. 2022;399:1718–29. https://doi.org/10.1016/S0140-6736(22)00562-1.
12. Kirkwood JM, Del Vecchio M, Weber J, et al. Adjuvant nivolumab in resected stage IIB/C melanoma: primary results from the randomized, phase 3 CheckMate 76K trial. Nat Med. 2023;29:2835–43. https://doi.org/10.1038/s41591-023-02583-2.
13. Bera K, Schalper KA, Rimm DL, et al. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16:703–15. https://doi.org/10.1038/s41571-019-0252-y.
14. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15:20170387. https://doi.org/10.1098/rsif.2017.0387.
15. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol. 2019;20:e253–61. https://doi.org/10.1016/S1470-2045(19)30154-8.
16. Adeuyan O, Gordon ER, Kenchappa D, et al. An update on methods for detection of prognostic and predictive biomarkers in melanoma. Front Cell Dev Biol. 2023;11. https://doi.org/10.3389/fcell.2023.1290696.

17. Comes MC, Fucci L, Mele F, et al. A deep learning model based on whole slide images to predict disease-free survival in cutaneous melanoma patients. Sci Rep. 2022;12:20366. https://doi.org/10.1038/s41598-022-24315-1.
18. Hu J, Cui C, Yang W, et al. Using deep learning to predict anti-PD-1 response in melanoma and lung cancer patients from histopathology images. Transl Oncol. 2021;14:100921. https://doi.org/10.1016/j.tranon.2020.100921.
19. Kulkarni PM, Robinson EJ, Pradhan JS, et al. Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. Clin Cancer Res. 2020;26:1126–34. https://doi.org/10.1158/1078-0432.CCR-19-1495.
20. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): toward Medical XAI. IEEE Trans Neural Networks Learn Syst. 2020;32:4793–813. https://doi.org/10.1109/tnnls.2020.3027314.
21. Ribeiro MT, Singh S, Guestrin C. (2016) Why Should I Trust You? Explaining the Predictions of Any Classifier. NAACL-HLT 2016–2016 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol Proc Demonstr Sess 97–101. https://doi.org/10.18653/v1/n16-3020
22. Mann HB, Whitney DR. On a test of whether one of two Random variables is stochastically larger larger than the other. Ann Math Stat. 1947;18:50–60.
23. Pandis N. The chi-square test. Am J Orthod Dentofac Orthop. 2016;150:898–9. https://doi.org/10.1016/j.ajodo.2016.08.009.
24. Pantanowitz L, Farahani N, Parwani A. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathol Lab Med Int 23. 2015. https://doi.org/10.2147/PLMI.S59826.
25. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017;7:1–7. https://doi.org/10.1038/s41598-017-17204-5.
26. Macenko M, Niethammer M, Marron JS et al. (2009) A Method For Normalizing Histology Slides For Quantitative Analysis 3 statistics and Operations Research, 4 Lineberger Comprehensive Cancer Center, 5 Renaissance Computing Institute, 6 Pathology and Laboratory Medicine, 7 Dermatology University of nor. IEEE Int Symp Biomed Imaging 1107–10.
27. Comes MC, Fanizzi A, Bove S, et al. Early prediction of neoadjuvant chemotherapy response by exploiting a transfer learning approach on breast DCE-MRIs. Sci Rep. 2021;11. https://doi.org/10.1038/s41598-021-93592-z.
28. Chollet F. Xception: deep learning with depthwise separable convolutions. Proc – 30th IEEE Conf Comput Vis Pattern Recognit CVPR 2017. 2017;2017–Janua:1800–7. https://doi.org/10.1109/CVPR.2017.195.
29. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-Decem. 2016;2818–26. https://doi.org/10.1109/CVPR.2016.308.
30. He K. (2015) Deep Residual Learning for Image Recognition ResNet @ ILSVRC & COCO 2015 Competitions. 1–9.
31. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. (2017) Densely connected convolutional networks. Proc – 30th IEEE conf Comput Vis Pattern Recognition, CVPR 2017 2017-Janua:2261–9. https://doi.org/10.1109/CVPR.2017.243
32. Dosovitskiy A, Beyer L, Kolesnikov A et al. (2021) An image is worth 16X16 words: transformers for Image Recognition at Scale. ICLR 2021–9th Int Conf Learn Represent.
33. Mencattini A, Spalloni A, Casti P, et al. NeuriTES. Monitoring neurite changes through transfer entropy and semantic segmentation in bright-field time-lapse microscopy. Patterns. 2021;2:100261. https://doi.org/10.1016/j.patter.2021.100261.
34. Lin TY, Goyal P, Girshick R, et al. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2020;42:318–27. https://doi.org/10.1109/TPAMI.2018.2858826.
35. Kingma DP, Ba JL. (2015) Adam: A method for stochastic optimization. 3rd int conf learn represent ICLR 2015 -. Conf Track Proc 1–15.
36. Amoroso N, Errico R, Bruno S, et al. Hippocampal unified multi-atlas network (HUMAN): protocol and scale validation of a novel segmentation tool. Phys Med Biol. 2015;60:8851–67. https://doi.org/10.1088/0031-9155/60/22/8851.
37. Garbe C, Keim U, Amaral T, et al. Prognosis of patients with primary Melanoma Stage I and II according to American Joint Committee on Cancer Version 8 validated in two independent cohorts: implications for adjuvant treatment. J Clin Oncol. 2022;40:3741–9. https://doi.org/10.1200/JCO.22.00202.
38. Bleicher J, Swords DS, Mali ME, et al. Recurrence patterns in patients with stage II melanoma: the evolving role of routine imaging for surveillance. J Surg Oncol. 2020;122:1770–7. https://doi.org/10.1002/jso.26214.
39. Lee R, Mandala M, Long GV, et al. Adjuvant therapy for stage II melanoma: the need for further studies. Eur J Cancer. 2023;189:112914. https://doi.org/10.1016/j.ejca.2023.05.003.
40. Simiczyjew A, Dratkiewicz E, Mazurkiewicz J, et al. The influence of Tumor Microenvironment on Immune escape of Melanoma. Int J Mol Sci. 2020;21:8359. https://doi.org/10.3390/ijms21218359.
41. Liu D, Yang X, Wu X. Tumor Immune Microenvironment characterization identifies prognosis and immunotherapy-related gene signatures in Melanoma. Front Immunol. 2021;12. https://doi.org/10.3389/fimmu.2021.663495.
42. Moldoveanu D, Ramsay L, Lajoie M, et al. Spatially mapping the immune landscape of melanoma using imaging mass cytometry. Sci Immunol. 2022;7. https://doi.org/10.1126/sciimmunol.abi5072.
43. Serratì S, Di Fonte R, Porcelli L, et al. Circulating extracellular vesicles are monitoring biomarkers of anti-PD1 response and enhancer of tumor progression and immunosuppression in metastatic melanoma. J Exp Clin Cancer Res. 2023;42:251. https://doi.org/10.1186/s13046-023-02808-9.

Publisher's note

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