Melanoma Breslow Thickness Classification Using Ensemble-Based Knowledge Distillation With Semi-Supervised Convolutional Neural Networks

Juan P. Dominguez-Morales, *Member, IEEE*, Juan-Carlos Hernández-Rodríguez, Lourdes Duran-Lopez, Julián Conejo-Mir, and Jose-Juan Pereyra-Rodriguez

***Abstract*—Melanoma is considered a global public health challenge and is responsible for more than 90% deaths related to skin cancer. Although the diagnosis of early melanoma is the main goal of dermoscopy, the discrimination between dermoscopic images of in situ and invasive melanomas can be a difficult task even for experienced dermatologists. Recent advances in artificial intelligence in the field of medical image analysis show that its application to dermoscopy with the aim of supporting and providing a second opinion to the medical expert could be of great interest. In this work, four datasets from different sources were used to train and evaluate deep learning models on in situ versus invasive melanoma classification and on Breslow thickness prediction. Supervised learning and semi-supervised learning using a multi-teacher ensemble knowledge distillation approach were considered and evaluated using a stratified 5-fold cross-validation scheme. The best models achieved AUCs of 0.8085±0.0242 and of 0.8232±0.0666 on the former and latter classification tasks, respectively. The best results were obtained using semi-supervised learning, with the best model achieving 0.8547 and 0.8768 AUC, respectively. An external test set was also evaluated, where semi-supervision achieved higher performance in all the classification tasks. The results obtained show that semi-supervised learning could improve the performance of trained models in different melanoma classification tasks compared to supervised learning. Automatic deep learning-based diagnosis systems could support medical professionals in their decision, serving as a second opinion or as a triage tool for medical centers.**

***Index Terms*—Breslow thickness, deep learning, knowledge distillation, melanoma, semi-supervision.**

I. Introduction

MELANOMA is a malignant skin neoplasm that emerges from melanocytes; these cells are located in the basal layer of the epidermis and produce the pigment melanin. Melanoma is responsible for more than 90% of deaths related to skin cancer [1]. Thus, it continues to be nowadays a global public health challenge, reaching in 2020 a total of 325,000 new melanoma cases and 57,000 deaths [2]. Unlike other solid malignant tumors, the incidence of malignant melanoma is still growing [3], with global incidence rates of 11.5 and 11.3 cases per 100,000 inhabitants in men and women, respectively.[1] Despite the fact that melanoma mortality appears to stabilize, it should be noted that incidence and prevalence in the United States have increased from the beginning of the 1990s to 2019. In the latter, melanoma was reported to be the skin cancer with the highest comorbidity, showing a disability-adjusted life-years rate of 64.4 [4]. The delay in diagnosis is directly correlated with a poor prognosis, making early detection the cornerstone of successful treatment of melanoma [5]. Dermoscopy is a non-invasive and cost-effective tool used in the daily clinical practice of dermatologists for the diagnosis of melanoma. It has shown its reliability and sensitivity in detecting early stage skin cancer, effectively minimizing unnecessary excisions or biopsies [6]. Dermoscopy allows the identification of asymmetry in the cutaneous structures, even before it could be clinically appreciated. This helps to detect melanomas even in the epidermal layer (in situ melanomas) [7]. From top to bottom, the epidermis is one of the layers of human skin, followed by the dermis

Received 11 February 2024; revised 15 July 2024 and 29 August 2024; accepted 16 September 2024. Date of publication 20 September 2024; date of current version 8 January 2025. This work was supported in part by Andalusian Regional Project (with FEDER support) DAFNE under Grant US-1381619 and in part by the EUROMELANOMA 2023 project from the Fundación Piel Sana de la Academia Española de Dermatología y Venereología. *(Juan P. Dominguez-Morales, Juan-Carlos Hernández-Rodríguez and Lourdes Duran-Lopez contributed equally to this work.) (Corresponding author: Juan-Carlos Hernández-Rodríguez.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Andalusian Review Board and Ethics Committee Virgen Macarena-Virgen del Rocio Hospitals under Application No. 0096-N-20.

Juan P. Dominguez-Morales and Lourdes Duran-Lopez are with the Robotics and Tech. of Computers Lab., ETSII-EPS, I3US, Universidad de Sevilla, 41012 Sevilla, Spain (e-mail: jpdominguez@us.es; lduran2@us.es).

Juan-Carlos Hernández-Rodríguez, Julián Conejo-Mir, and Jose-Juan Pereyra-Rodriguez are with the Departamento de Medicina, Facultad de Medicina, Universidad de Sevilla, 41009 Sevilla, Spain, and also with the Unidad de Gestión Clínica de Dermatología, Hospital Universitario Virgen del Rocío, 41013 Sevilla, Spain (e-mail: jhernandez7@us.es; jsconejomir@us.es; jpereyra@us.es).

Digital Object Identifier 10.1109/JBHI.2024.3465929

[1] https://gco.iarc.fr. Retrieved September 23, 2024

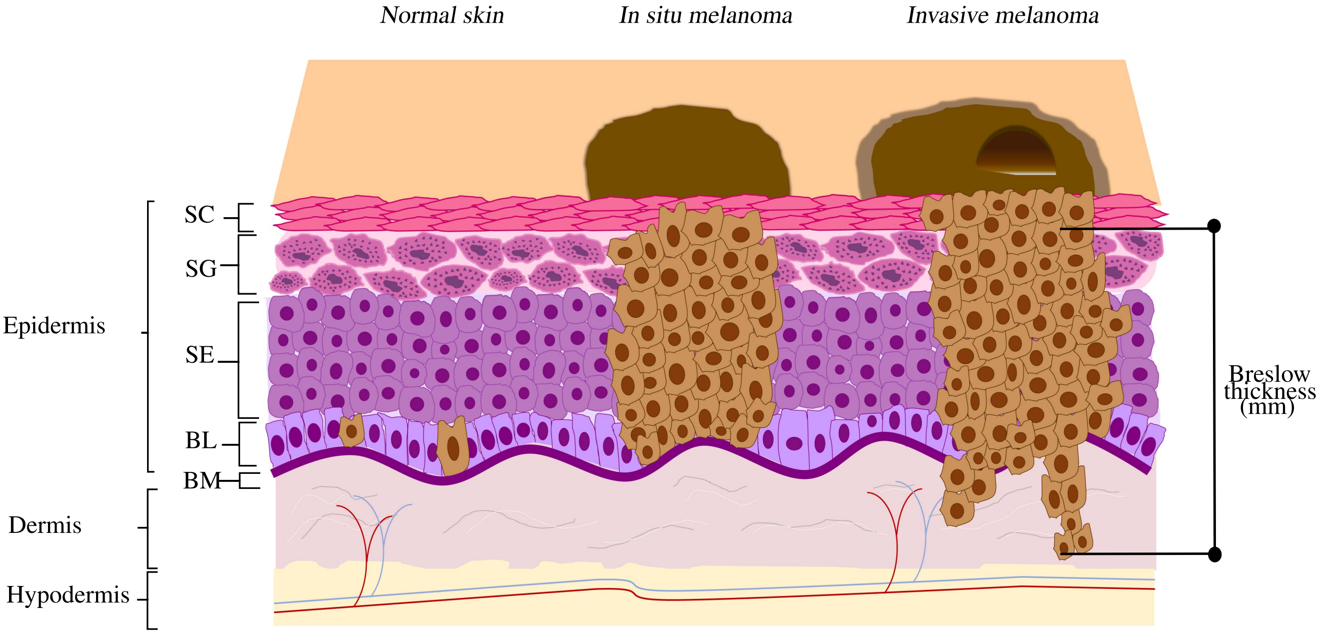


Fig. 1. Scheme of normal skin, in situ melanoma, invasive melanoma, and measurement of Breslow thickness. From top to bottom we represent all the layers of the skin (epidermis, dermis, and hypodermis) and all the sublayers of the epidermis. On the left, a scheme of the histological aspect of normal skin is represented. In the middle part of the image, an in situ melanoma can be seen, with all melanocytes confined to the epidermis without exceeding the basement membrane. The right side of the image shows an invasive melanoma with spreading tumor melanocytes that reach the basement membrane and invade the dermis. Breslow thickness is represented as the measurement from the stratum granulosum to the deepest invasive melanocyte of an invasive melanoma. Brown cells represent melanocytes. The stratum lucidum (between the stratum corneum and the stratum granulosum) was not represented, as it can only be found in the palms and soles. SC: Stratum corneum; SG: Stratum granulosum; SS: Stratum spinosum; BL: Basal layer; BM: Basement membrane; mm: Millimeters.

and hypodermis. Likewise, the epidermis is divided into five sublayers: the *stratum corneum*, the *stratum lucidum*, the *stratum granulosum*, the *stratum spinosum*, and the basal layer. The basement membrane is an essential structure that connects the epidermis to the dermis. When tumoral melanocytes are confined to the epidermis without exceeding the basement membrane, it is known as in situ melanoma (Mis). In contrast, if a tumor cell extends to the dermis, beyond the basement membrane, it is known as invasive melanoma (Miv) (Fig. 1).

Considering the significant correlation between melanoma thickness and prognosis, the detection of thinner melanomas could be related to a lower expected death rate and economic burden [8]. When a melanoma is suspected, an excisional biopsy removing the entire tumor in a first surgical step followed by histopathology is the gold standard for diagnosis and appropriate surgical treatment [9]. The thickness of the melanoma, the so-called Breslow thickness (BT), is objectively measured by a dermatopathologist using an optical micrometric scale. BT is considered the main prognostic factor for primary cutaneous melanoma, as it stands for microinvasion of the tumor in millimeters from the granular layer of the skin to the deepest layer of tumor invasion. In this sense, it is possible to differentiate a Mis from a Miv and quantify the microinvasion of Miv. Although an accurate diagnosis of early melanoma is the main goal of dermoscopy, the discrimination between dermoscopic images of Mis and Miv can be a difficult task even for experienced dermatologists. Dermoscopic structures such as irregular hyperpigmented areas and prominent skin margins could be predictors of Mis in a preoperative setting compared to atypical nevi, which is a benign lesion [10]. In the case of Miv, indicators such as a blue-white veil, multicomponent, or rainbow pattern could be found [11]. However, the dermoscopic features mentioned above are not sufficient to indirectly predict the microinvasion of melanoma. Decision making in the perioperative setting is one of the crucial points in the day-to-day of dermatology surgeons, since, depending on whether we are dealing with a Mis or a Miv, there could be variation in the surgical margins in the first surgical excision, triaging, and the patient's prognosis. In the case of Mis, after excisional biopsy of primary melanoma, current guidelines recommend performing a two-step surgical procedure to widen the surgical margins by 5 mm [1], [12]. Thus, one-step surgery could be a curative option, saving time and reducing costs and discomfort in patients [13]. On the other hand, in Miv, wider surgical margins are required, including sentinel node biopsy when the BT threshold is ≥ 0.8 mm or associated risk factors such as ulceration are present. Due to the difficulty in differentiating between Mis and Miv, artificial intelligence has emerged as a useful ancillary support tool to select optimal surgical margins and tumor staging in cutaneous melanoma based on dermoscopic images. However, histopathological confirmation of BT would be necessary once excision has been performed in all patients, as it is the gold standard. The field of deep learning research as a diagnostic decision support technique has increased in recent years in the context of dermatology [14]. Convolutional neural networks (CNNs) are the basis by which deep learning performs a computational analysis of images. CNNs take images as input that go through various hidden layers of artificial neurons to render a final output, like a prediction or a diagnosis. In the state-of-the-art, the performance shown by CNNs is similar to that of expert dermatologists in the diagnosis of skin cancer using clinical and dermoscopic images [15]. However, deep learning could involve a higher computational cost compared to other methods, such as deep transfer learning (DTL), which has become a popular option within the field of health. This significantly reduces the amount of required data and training time using pre-trained CNNs. The knowledge distillation-based approach (KD) is a novel deep learning method to reduce computational costs, making it easier to build decision support system models in a clinical setting. Khan et al. [16] successfully performed a KD approach to detect melanoma using dermoscopic images. However, to our knowledge, KD has not been developed or tested to predict microinvasion, measured as BT, using dermoscopic images of melanoma.

The main contributions of this study are as follows:

- The use of an offline response-based Multi-Teacher KD algorithm together with semi-supervised deep CNNs to predict the BT of melanoma.
- The models were trained and evaluated on heterogeneous data obtained from 4 different datasets using stratified 5-fold cross validation.
- Semi-supervision with the teacher-student paradigm outperforms supervised learning in both Mis versus Miv classification and BT classification tasks.
- All the code and data have been made available in a public repository, including a novel dataset.

The rest of the work is structured as follows: first, in Section II, the state-of-the-art in the classification of BT and the analysis of image-based melanoma is presented. Then, in Section III, the materials and methods used in this work are introduced, focusing on the datasets (Section III-A) and the pre-processing applied to the samples and the way they were partitioned (Section III-B), together with the different deep learning approaches considered for training and evaluating (Section III-C), as well as the metrics used to measure the performance of the models (Section III-D) and the framework used (Section III-F). Subsequently, the results of the two different classification tasks considered are presented in Section IV. Finally, in Section V, the results obtained are discussed, and in Section VI, the conclusions of this work are presented.

II. Related Works

A. Melanoma Classification

In recent years, several studies have developed de novo CNNs to classify between Mis or Miv, but, these methods require a significant amount of data to train and need a notable computational cost. Therefore, to overcome the mentioned disadvantages, DTL has been proposed as a machine learning technique that uses CNN models that have already been pre-trained on large image datasets such as ImageNet. This technique shares this knowledge for a related classification task, although at a lower computational cost. DTL is especially useful when there is a lack of training data, such as in the field of health, where obtaining large, annotated datasets is complex and time consuming. Regarding de novo CNNs, Polesie et al. [17] developed a CNN with seven convolutional layers and a single dense layer to differentiate Mis and Miv using 1551 close-up images of melanoma. Although this work took the first step towards automatic classification between Mis and Miv, the 0.72 area under the ROC curve (AUC) achieved by the model was outperformed by seven dermatologists with an AUC of 0.81. Gillstedt et al. [18] trained a de novo CNN using 1137 melanoma dermoscopic images, again outperformed by seven dermatologists with an AUC of 0.76 and 0.81, respectively. According to previous work, Gillstedt et al. [19] developed a de novo CNN combining close-up (6 convolutional layers) and dermoscopic images (7 convolutional layers) obtaining lower results than the six independent dermatologists, with an AUC of 0.73 compared to the 0.80 registered by human readers. Polesie et al. [20] conducted a direct comparison between a de novo CNN, the pre-trained CNN ResNet50, and 438 international readers, including dermatologists and non-dermatologists for the classification task between Mis and Miv and the prediction of BT using 1456 dermoscopic images. Despite the fact that de novo CNNs are outperformed by human readers, in this case, the performance of the fine-tuned pre-trained CNN ResNet50 was similar to that of human readers without a statistically significant difference. Chu et al. [21] used ResNet50 for Mis versus Miv in addition to the depth of microinvasion, although in acral melanomas, that is, a specific class of melanoma that occurs in palms, soles, and nail beds. Despite being in a different clinical setting, CNN effectively distinguished between < 0.8 mm and ≥ 0.8 mm BT, with an AUC of 0.90, in 57 dermoscopic images. Hernández-Rodríguez et al. [22] compared three DTL pretrained CNNs (ResNetV2, InceptionV3 and EfficientNetB6) for the classification between Mis and Miv and between melanomas with BT < 0.8 mm and ≥ 0.8 mm using 1315 dermoscopic images from a heterogeneous dataset. In this case, ten dermatologists outperformed the three models for the classification between Mis and Miv, although the pretrained CNNs ResNetV2 and InceptionV3 outperformed the ten dermatologists combined for the classification task between melanomas with BT < 0.8 mm and ≥ 0.8 mm. The authors established the BT threshold at 0.8 mm, since it is the cut-off point for performing a sentinel lymph node biopsy when other histological risk factors are present, such as ulceration, following the current 2022 European guidelines for melanoma [1], [9].

B. Knowledge Distillation and Semi-Supervision

In 2015, Hinton et al. [23] paved the way for the introduction of KD as a proper model with a high level of performance and with a lower required memory capacity. KD is based on the 'teacher-student' paradigm, in which the performance of a small and simple network known as the student model improves using the knowledge transferred from a large and complex network called the teacher model [24]. Only two authors have used KD approaches to differentiate between melanoma and non-melanoma. Khan et al. [16] carried out a KD approach with the aim of detecting melanoma from dermoscopic images. For this task, they used a pretrained ResNet50 (23M of trainable parameters) as the teacher model and developed a student model called the Distill Student Network (0.26 million parameters). They showed a reduction in the runtime compared to EfficientNet-B0 (4 million parameters) for the classification task between melanoma and non-melanoma lesions with acceptable accuracy compared to the teacher model, with 99.60% and 91.7%, respectively. Adepu et al. [25] performed a knowledge distilled light-weight CNN-based approach to deal with the high inter-class and low intra-class similarities in dermoscopic images for the classification of melanoma and non-melanoma in the ISIC-2020 dataset [26]. Focal Loss was used as a Cost-Sensitive Learning technique to tackle the high class-imbalance problem to improve sensitivity. EfficientNet-B5 was used as teacher and EfficientNet-B2 as the student model with an AUC of 0.92 for the classification task of differentiating melanoma and non-melanoma.

III. Materials and Methods

In this section, all the materials and methods used to perform the experiments are presented, together with the framework considered for this purpose. In particular, we first focus on the datasets used (Section III-A) and the way the images from the different datasets are processed and partitioned (Section III-B). Then, the different deep learning-based training methods considered are explained (Section III-C), together with the metrics used to evaluate the performance of the trained models (Section III-D). Finally, the deep learning framework, hyperparameters, CNN architecture, and libraries used in this work are presented (Section III-F).

A. Datasets

With the aim of expanding the generalizability of our work and achieving a context closer to daily clinical practice, our general dataset consisted of four independent subsets collected from different sources, with a total of 1449 images distributed as follows. (i) A total of 868 images labeled from 316 cases of melanoma from the dermoscopic image repository of Virgen del Rocio University Hospital (a tertiary hospital in Seville, Spain), prospectively collected between 2016 and 2023. The study protocol for the collection of this subset of images (ID 0096-N-20) was approved by the Andalusian Review Board and Ethics Committee of Virgen Macarena-Virgen del Rocio Hospitals. All patients gave written informed consent for participation of their case details and images. This dataset has been made publicly available for this work and can be requested on its website.[2] To increase external validity, we did not restrict the melanoma subtype or participants' phototype. (ii) A total of 193 labeled images of 184 cases of melanoma publicly available from Polesie et al. [27]. (iii) A total of 141 labeled images of 141 melanoma and 2,508 unlabeled images of 2,508 cases of melanoma from the International Skin Imaging Collaboration (ISIC) archive [26], which is an extensive open-source public-access archive of skin images. These images comprise a subset of images from the ISIC archive. To obtain labeled images, we applied the filters 'Lesion diagnosis-melanoma', 'Type of diagnosis-histopathology', 'Melanoma thickness (mm)' and 'Image type-dermoscopic', retrieving all the dermoscopic images of histopathologically confirmed melanoma in which BT metadata were available. For unlabeled images, we filtered by 'Lesions diagnosis-melanoma', 'Histopathological type of diagnosis-histopathologically' and "Image-type-dermoscopic", obtaining all the histopathologically confirmed cases of melanoma with available dermoscopic images. In the last subset, we removed the images in which BT was available. (iv) A total of 247 images of 247 cases of melanoma from the public-access dataset of Kawahara et al. [28]. In the (i) and (ii) subsets, there was more than one image for some of the cases of melanoma. Table I shows the characteristics of the melanoma cases from which the labeled images of subsets (i), (ii), (iii) and (iv) were extracted. Table II presents the distribution of the labels of the annotated samples in each of the datasets.

To evaluate the performance of the trained models with completely unseen data, an external test set was evaluated. This dataset consisted of a total of 153 labeled cases, of which 39 correspond to Mis and 114 to Miv. This leads to a total of 462 labeled images, of which 144 correspond to Mis, 199 to Miv with BT$<$0.8 mm, and 149 to Miv with BT$\geq$0.8 mm. Each of the three classification tasks (Miv vs Mis, BT classification, and multiclass classification) were evaluated on this external test set with both supervised and semi-supervised CNN models.

B. Image Pre-Processing and Data Partitioning

The images used in this work vary in size since they were sourced from different medical centers and different sensors were used to obtain them. This is a handicap for being used as input to deep CNNs, since they require a fixed image size as input. To address this problem, all the images were downsampled to 224×224pixels, since it is the input size required for the CNN models used in this work.

1) Data Augmentation: Data augmentation commonly refers to a widely used technique in deep learning that consists in artificially increasing the amount of samples in the training set by creating modified copies of the samples with minor changes with respect to the original ones in order to increase the heterogeneity of the dataset and reduce overfitting. In this work, two different operations were applied to augment the training set: 90° rotations (the augmented sample could be rotated by 90°, 180° or 270° with respect to the original one), horizontal flips and vertical flips. A probability of 0.5 was established for each of these transformations, meaning that, for each epoch, each of the images had a 50% chance of being transformed with each of the augmentations used (i.e., a 50% chance of being rotated, a 50% chance of being horizontally flipped, and a 50% chance of being vertically flipped). These transformations were applied per image and epoch, meaning that, during the same epoch, different images could be transformed in a different way based on the 50% chance used. This was done thanks to the Albumentations

[2] https://institucional.us.es/breslowdataset. Retrieved September 23, 2024.

TABLE I
CHARACTERISTICS (AGE, SEX DISTRIBUTION AND ANATOMIC LOCATION) OF THE MELANOMA CASES PRESENT IN EACH OF THE DATASET SUBSETS USED

Dataset	VRUH subset	ISIC archive subset	Polesie et al. subset	Kawahara et al. subset	External test
Age (average ± SD)	60.23 (± 15.66)	63.42 ± 14.51	67 (57-77 IQR)	n/a	57.76 (± 17.06)
Sex distribution	Male (42%) Female (58%)	Male (55%) Female (36%) Unspecified (9%)	Male (53%) Female (47%)	Male (46%) Female (54%)	Male (45%) Female (55%)
Anatomic location	Head and neck (9%) Anterior torso (10%) Posterior torso (46%) Upper extremities (18%) Lower extremities (17%)	Head and neck (6%) Anterior torso (11%) Posterior torso (11%) Upper extremities (16%) Lower extremities (9%) Unspecified (46%)	Trunk and upper extremities (88%)	Head and neck (9%) Anterior torso (19%) Posterior torso (29%) Upper extremities (13%) Lower extremities (30%)	Head and neck (13%) Anterior torso (13%) Posterior torso (41%) Upper extremities (10%) Lower extremities (23%)

IQR, interquartile range; ISIC, International Skin Imaging Collaboration; n/a, not available; VRUH, Virgen del Rocio University Hospital

TABLE II
DISTRIBUTION OF THE LABELED IMAGES FROM VIRGEN DEL ROCÍO, POLESIE ET AL. 2021, KAWAHARA ET AL. 2018 AND ISIC DATASETS

Dataset	In situ	Invasive	Breslow $<$ 0.8 mm	Breslow $\geq$ 0.8 mm	*Total*
VRUH	201	667	552	316	868
Polesie et al.	117	76	140	53	193
Kawahara et al.	64	183	166	81	247
ISIC archive	19	122	106	35	141
External test	114	348	312	150	462
Total	515	1396	1276	635	1911

The column with the total amount of samples is not the sum of the images shown in each column of a specific row, since most of the images are labeled both as either in situ or invasive and with their corresponding Breslow thickness.

library [29], which automatically performed these augmentation operations every time an image was loaded into memory.

2) Stratified K-Fold Cross-Validation: Cross-validation is one of the most common techniques to measure the generalizability of a trained model. There are different types of cross-validation in machine learning. In this work, we used the stratified k-fold cross-validation (where $k = 5$), which is an enhanced version of k-fold cross-validation. Although the dataset is also divided into k equal folds, the ratio of samples per class is the same in each of the folds, which is optimal for working with imbalanced datasets.

The entire dataset was split into five different folds, 4 of which were used to train the deep learning model, whereas the remaining one was used to evaluate its performance. This was done 5 times, rotating the folds used to train and evaluate to explore all the different combinations. As explained in Section IV, the results were evaluated as the average and standard deviation of the 5-fold cross-validation results for each of the metrics considered (see Section III-D). It is important to mention that, apart from taking the imbalance of the dataset into account, the samples were split between folds, also considering a patient-level distribution where samples from the same patient were not present in different folds.

C. Deep Learning Approaches

In this section, the different deep learning-based training approaches considered in this work are presented: full supervision and semi-supervision. These methods were used for two different classification tasks, including the classification of test samples as having a BT lesser than 0.8 mm or greater than or equal to 0.8 mm, and the classification of the test samples as Mis or Miv. The histological diagnosis of the melanoma cases sets the basis for the ground truth of the labeled images. For the BT classification task, we considered the specific cut-off of 0.8 mm of BT, as it is the threshold specified in the current 2022 European guidelines for melanoma to perform a sentinel lymph node biopsy when other risk factors (e.g., ulceration) are present.

1) Supervised Learning: Supervised learning includes methods that are developed to train machine learning algorithms with strongly annotated data. In the field of medical image analysis, this means samples that medical experts have manually annotated and given a specific label or class based on their experience and the identified finding.

Supervised learning is the most common training method in deep and machine learning. This is mainly due to two reasons. First, since samples are annotated by experts, the chances of them being incorrectly labeled are reduced. This allows for a faster training process with a smaller amount of images needed to obtain results that are similar to or better than other training approaches. Second, supervised learning is the easiest method to implement and the most common approach to fast and easy training of deep learning models to test their ability in a classification task for a specific application.

The main drawback of supervised learning in medical image analysis comes from the fact that obtaining annotations from medical experts is not easy, since it is a time-consuming process that requires additional effort from them apart from their daily work. This is also the reason why only a few datasets with these types of label are publicly available, each of them containing a reduced number of images.

2) Ensemble-Based Knowledge Distillation Annotator: KD is referred to as the process of transferring the knowledge acquired by a model in the training phase to a different model [23]. Since it was first formulated, this approach has gained popularity and is now being used in a wide range of applications, including medical imaging [30].

There are currently many different KD algorithms, including Graph-Based [31], [32], Cross-Modal [33], [34], Attention-Based [35], [36], Cross-Layer [37] and Multi-Teacher knowledge distillation [38], [39], [40], among many others.

In this work, we used an offline response-based Multi-Teacher KD algorithm, in which 5 teacher models (the 5 different supervised models trained using cross-validation presented in Section III-C-1) were used to pseudo-annotate a set of unlabeled images (images from the ISIC dataset that had no label, as described in Section III-A). Instead of pseudo-annotating the samples with the average response of the five teacher models as presented in [23], a majority voting algorithm was used. This means that the label assigned to an image was set based on the winning class predicted by the 5 teacher models (if at least 3 of the models agree on the prediction of an image, the class that was predicted as label is assigned to that image).

3) Semi-Supervision Using Teacher-Student Paradigm: Semi-supervised learning is a training method that combines both labeled and unlabeled data during the training process of the machine learning model [41]. Typically, the number of unlabeled data used in this approach is much higher than that of labeled data due to the greater complexity in obtaining the latter, as mentioned in Section III-C-1. This is also one of the main advantages of semi-supervision, since it allows exploiting unlabeled data, alleviating the fact of having a limited number of labeled samples available.

Semi-supervised learning mainly relies on the development of an algorithm that automatically annotates unlabeled data, reducing the need of large labeled datasets and, thus, the effort needed by medical experts to manually annotate a large amount of samples. Among the different training variants within semi-supervision, one of the best known is the teacher-student paradigm, in which two or more models are involved. In this case, the annotator role (also called Teacher) was performed by a set of five models (see Section III-C-2), which are in charge of annotating unlabeled data. After that, these automatically-labeled data are exploited by the Student model in the training phase, together with the labeled data. Although Teacher models tend to be implemented with models that are deeper than the Student model [42], they can also be implemented using the same architecture, as in [43], which is the particular case of this work.

The teacher-student learning approach has been used in many different deep learning tasks, including medical image analysis [44], [45], [46], [47], [48], where relevant results were obtained compared to other training approaches while improving the generalization of models, because they were able to exploit unlabeled data.

Fig. 2 shows a diagram representing the whole teacher-student pipeline where teacher training, the ensemble-based KD annotator and the student training are depicted.

In a 5-fold cross-validation multi-teacher ensemble KD approach, the dataset is split into 5 parts (D_1-D_5). Each fold is trained using 4 of them and evaluated on the remaining one. After this, each of the folds (i.e., Teacher models) is used to predict a completely different set of data (not part of D_1-D_5) in which the images do not have associated labels. Let us call this unlabeled set U. Each of the Teacher models (T_1-T_5) generates a prediction for each of the images in U. A combined pseudo-label report is obtained, where the label of a specific image in U (U_i) is assigned based on the majority voting among the prediction performed by the 5 Teachers. This means that, if T_1-T_3 predicted U_i as class 0 and T_4-T_5 predicted U_i as class 1, U_i will be pseudo-labeled as class 0. After the majority voting knowledge distillation approach is performed and all the U_i in U are pseudo-labeled, the student models are trained. Focusing on a single Teacher model from the cross validation (T_1), let us assume T_1 was trained using D_1-D_4 and validated using D_1. Student Model 1 (or S_1), uses D_1-D_4 as training data, together with all the external images (U) that were pseudo-labeled, and then the model is evaluated on D_5. This means that Student models use the same training and validation data as the Teacher models, but the training data is increased with the whole unlabeled dataset that was pseudo-labeled by the Teacher models.

D. Evaluation Metrics

In order to evaluate the performance of the models, different well-known metrics were used. These are: precision (1), recall or sensitivity (2), specificity (3), F1-score (4), and AUC of the ROC curve.

$$\text{Precision} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = 100 \times \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specificity} = 100 \times \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives. The ROC curve shows the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The AUC measures the area that is under the ROC curve, where an area of 1 means perfect agreement between the ground truth and the predicted value.

E. Gradient Maps

Grad-CAM++ images and their corresponding integrated gradients were rendered for visual comprehension and readability of the models' outputs. Grad-CAM++ is a computer vision technique that provides visual explanations of CNN model predictions, allowing better localization of multiple object instances in

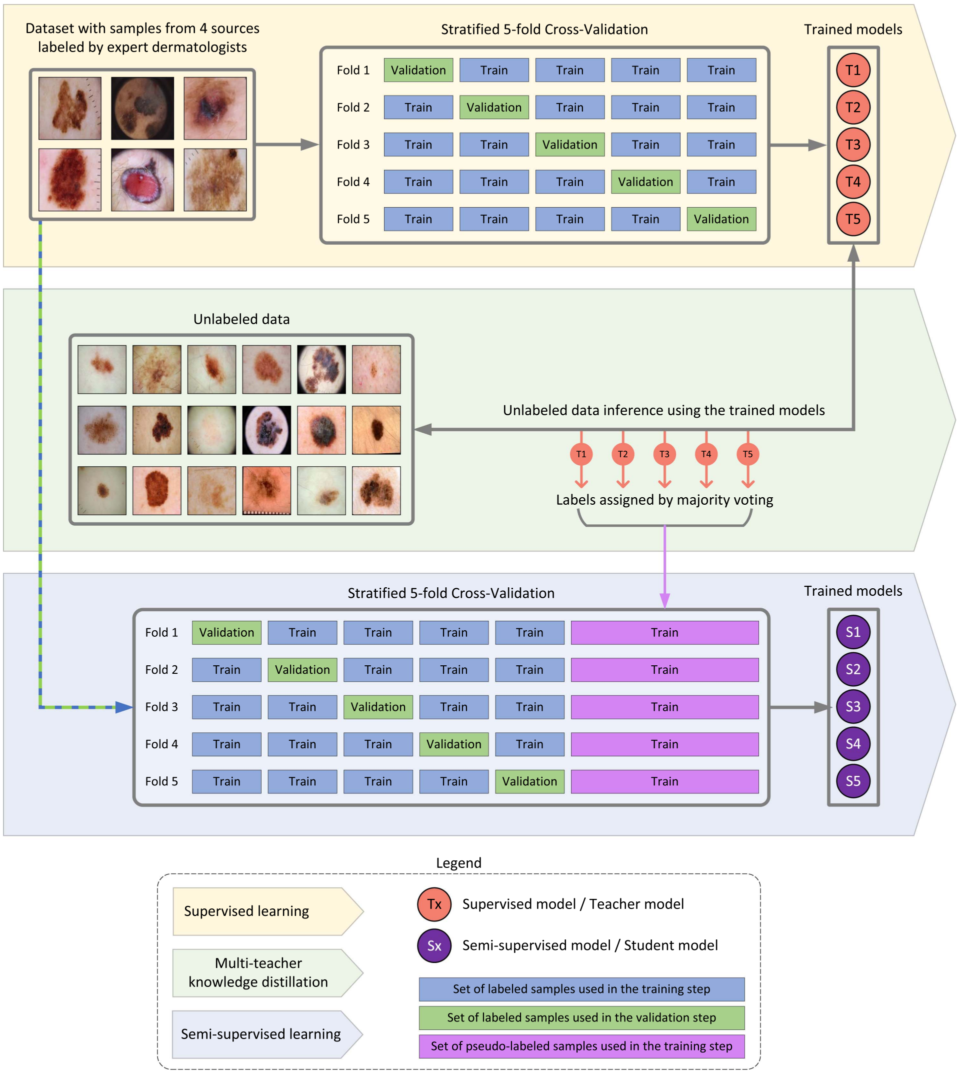

Dataset with samples from 4 sources labeled by expert dermatologists
Stratified 5-fold Cross-Validation
Trained models
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Validation
Train
T1
T2
T3
T4
T5
Unlabeled data
Unlabeled data inference using the trained models
Labels assigned by majority voting
S1
S2
S3
S4
S5
Legend
Supervised learning
Multi-teacher knowledge distillation
Semi-supervised learning
Tx Supervised model / Teacher model
Sx Semi-supervised model / Student model
Set of labeled samples used in the training step
Set of labeled samples used in the validation step
Set of pseudo-labeled samples used in the training step

Fig. 2. Block diagram of the whole deep learning approach considered. Firstly, the labeled images from the dataset are used to train five CNN models in a supervised manner using a stratified 5-fold cross-validation scheme. The results from these models are obtained and are then used as teachers for predicting unlabeled samples and assigning them pseudo-annotations after performing a majority voting algorithm. These pseudo-annotations are later used as part of the training set of a semi-supervised stratified 5-fold cross validation scheme where the same partitions used in the supervised learning step are considered. Finally, the trained student models are evaluated.

dermoscopic images [49]. In this case, Grad-CAM++ highlights influential areas of dermoscopic images, where the red color indicates high attribution areas for specific predictions.

F. Deep Learning Framework

PyTorch [50] was used for designing, training, and evaluating all the models proposed in this work, as well as for performing the different experiments that were carried out. PyTorch is a Python-based open-source AI framework developed by Meta, which, together with TensorFlow [51], has become one of the most used machine learning packages. Its high versatility allows defining and training specific models focused on particular applications with a user-friendly high-level code. Python 3.9.7 was used in this work, together with PyTorch 1.8.2 with GPU support. An NVIDIA A100 GPU was used as the hardware platform to accelerate both model training and inference.

DenseNet121 [52], ResNet50 [53] and VGG16 [54] models were used in all experiments, which were initialized with ImageNet weights [55], was obtained from PyTorch vision.[3] The model architecture was modified, removing the last layer (which consisted of 1000 output neurons, one per class) and replacing it with a Dense layer with 128 neurons and the output layer with two output neurons. These two neurons in the modified classifier correspond to the output classes of the network, which represent either BT < 0.8 mm and BT ≥ 0.8 mm or Mis and Miv, depending on the task to be performed.

[3] https://pytorch.org/hub/pytorch_vision_densenet. Retrieved September 23, 2024.

Regarding the hyperparameters used, Adam[4] optimizer was selected with learning rates ranging from 10^{-2} to $^{-4}$, and CrossEntropyLoss[5] was used as loss function. A batch size of 64 was used in the training phase. A maximum of 15 epochs were set to train the models by visual inspection and prior experimentation. To select and fine-tune the different hyperparameters of the model and the learning process, including batch size, learning rate and optimizer, a Grid Search algorithm [56] was used, where the configuration that achieved the best result was selected.

Scikit-learn's compute_class_weight function was used to weigh the loss function during the training process based on the class representation, since the number of samples per class was not balanced. This helped avoid overfitting in the most represented class (BT $<$ 0.8 mm and Mis). After each validation step (the model was evaluated at the end of each training epoch), the model was saved only if the validation loss improved from that achieved in previous validation steps. An early stop of 5 epochs was set, meaning that the training phase was stopped when the error on the validation set was higher than the previous best one for 5 consecutive epochs.

IV. Results

Two different classification tasks were evaluated in this work: firstly, a melanoma BT classification task where samples are labeled as either $<$ 0.8 mm or $\geq$ 0.8 mm thickness (Section IV-A), and, secondly, a classification between Mis and Miv samples (Section IV-B). For each of these tasks, both supervised learning and semi-supervised learning were used to evaluate their performance and compare them.

Supervised and semi-supervised models were trained following a stratified 5-fold cross-validation scheme, as explained in Section III-B-2, where each fold was trained using 80% of the samples and evaluated using the remaining 20%. The number of samples in the training and test subset for each fold was not the same, since the partition was made based on patients, not samples. Semi-supervised models were trained using the same partitions used in the corresponding supervised models, adding the pseudo-labeled data obtained from the annotation process (e.g., the semi-supervised CNN model from Fold 1 was trained and evaluated with the same partitions used in the supervised Fold 1 with the difference that the pseudo-labeled images were added to the training set).

The performance of the models was evaluated using the evaluation metrics presented in Section III-D, reporting the average and standard deviation of each metric among the five different folds to provide more realistic and fair results.

A. Breslow Thickness Classification

After training the supervised CNN models using the datasets presented in Section III-A following the 5-fold cross-validation scheme (Section III-B-2), the models were evaluated and then used to annotate the unlabeled data from the ISIC dataset. Semi-supervised models were trained with the pseudo-annotations and the same labeled data as the supervised CNNs, and the same test partitions were used in both training approaches. Table III presents the results achieved with the supervised and semi-supervised models. The results are shown per fold, and the last row shows the average and standard deviation of the folds. The evaluation metrics reported are explained in Section III-D. Semi-supervised models achieve higher performance than supervised models in the BT classification task, improving them by 5 points on the AUC on average. The confusion matrices and ROC curves of the models that achieved the best performance are shown in Fig. 3, where the best supervised model achieves an AUC of 0.8329 and the best semi-supervised model achieves an AUC of 0.8768. Table IV presents the results specified for each dataset.

B. In Situ Versus Invasive Melanoma Classification

For the Miv versus Mis task, the same procedure followed in Section IV-A was followed. Table V presents the results achieved on the supervised and semi-supervised models for the aforementioned classification task. Semi-supervised models achieve higher performance than supervised models in the Mis versus Miv classification task, improving them by 6 points on the AUC on average. Table VI presents the results specified for each dataset. The confusion matrices and ROC curves of the models that achieved the best performance are shown in Fig. 4, where the best supervised model achieves an AUC of 0.7743 and the best semi-supervised model achieves an AUC of 0.8547.

C. Multiclass Approach

In order to compare the performance of the binary classification tasks evaluated in previous sections (Sections IV-A and IV-B) with a multiclass classification task (Mis, Miv with BT $<$ 0.8 mm, and Miv with BT $\geq$ 0.8 mm), full-supervision and semi-supervision were used to train and evaluate CNN models following the same approach. For this end, the last layer of the CNN model was modified with three output neurons. Table VII presents the results achieved on the supervised and semi-supervised models for the multiclass classification task. As in the two binary classification tasks evaluated, semi-supervised models achieve higher performance than supervised models, improving them by 4 points on the AUC on average. Table VIII presents the results specified for each dataset. The confusion matrices and ROC curves of the models that achieved the best performance are shown in Fig. 5, where the best supervised model achieves an AUC of 0.7493 and the best semi-supervised model achieves an AUC of 0.7787.

D. External Test Set

Table IX presents a summary of all the results obtained. The best performance across all the three tasks is achieved with semi-supervised models. In particular, an average AUC of 0.8024±0.0080 is obtained in the BT classification task, an average AUC of 0.6797±0.0168 is obtained in the Miv vs Mis

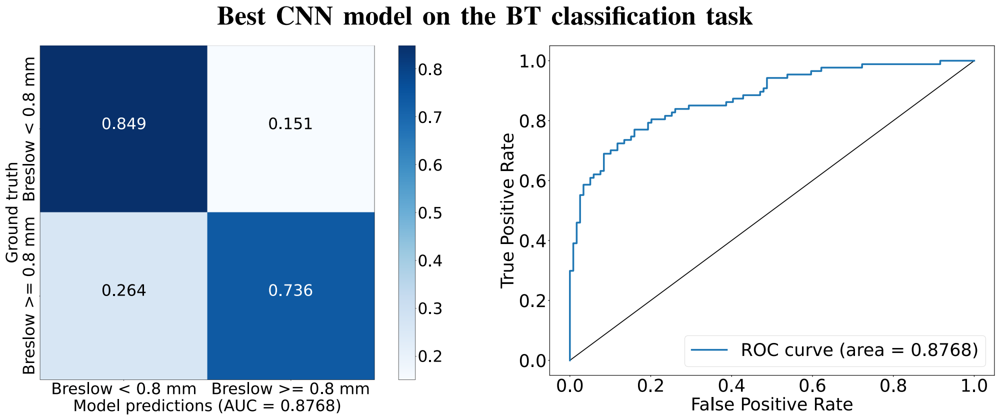
[4] https://pytorch.org/docs/stable/generated/torch.optim.Adam.html. Retrieved September 23, 2024

[5] https://pytorch.org/docs/stable/generated/torch.nn.CrossEntropyLoss.html. Retrieved September 23, 2024

TABLE III
RESULTS OBTAINED FOR THE SUPERVISED AND SEMI-SUPERVISED LEARNING APPROACHES ON THE BT CLASSIFICATION TASK

	Balanced acc	F1-score	Kappa score	Precision	AUC	Recall	Specificity
Supervised learning	0.7148 ± 0.0620	0.7244 ± 0.0656	0.4314 ± 0.1249	0.7302 ± 0.0590	0.7752 ± 0.0532	0.6233 ± 0.1029	0.8062 ± 0.0323
Semi-supervised learning	0.7350 ± 0.0719	0.7407 ± 0.0762	0.4724 ± 0.1377	0.7585 ± 0.0526	0.8232 ± 0.0666	0.6155 ± 0.1733	0.8544 ± 0.0371

The performance of the models is evaluated using different metrics, which are reported with the average and standard deviation calculated across the different cross-validation folds.



(a) Normalized confusion matrix.

(b) ROC curve.

Fig. 3. Confusion matrix (left) and ROC curve (right) obtained for the best CNN model on the classification task between BT $<$ 0.8 mm and BT $\geq$ 0.8 mm. The confusion matrix is normalized and represents all the labeled samples in the test set of that fold. The best results are obtained with semi-supervision, achieving an AUC of 0.8768.

TABLE IV
RESULTS OBTAINED PER DATASET ON THE BT CLASSIFICATION TASK

	Dataset	Balanced acc	F1-score	Kappa score	Precision	AUC	Recall	Specificity
Supervised learning	VRUH	0.7022 ± 0.0643	0.7031 ± 0.0691	0.3979 ± 0.1325	0.7151 ± 0.0584	0.7585 ± 0.0752	0.6466 ± 0.1171	0.7578 ± 0.0142
	Polesie et al.	0.7141 ± 0.0870	0.7074 ± 0.0846	0.4045 ± 0.1576	0.7359 ± 0.1001	0.8403 ± 0.0629	0.6339 ± 0.0699	0.7942 ± 0.1353
	Kawahara et al.	0.7154 ± 0.1446	0.8520 ± 0.0822	0.4995 ± 0.3284	0.8698 ± 0.0902	0.8978 ± 0.0824	0.4489 ± 0.2569	0.9818 ± 0.0364
	ISIC	0.5739 ± 0.0865	0.6636 ± 0.0788	0.1452 ± 0.1784	0.6593 ± 0.0877	0.5189 ± 0.1574	0.3561 ± 0.1674	0.7918 ± 0.0330
Semi-supervised learning	VRUH	0.7085 ± 0.0757	0.7054 ± 0.0841	0.4091 ± 0.1480	0.7233 ± 0.0657	0.7714 ± 0.0777	0.6334 ± 0.1826	0.7836 ± 0.0483
	Polesie et al.	0.8292 ± 0.0972	0.7980 ± 0.1151	0.6205 ± 0.2057	0.8680 ± 0.0706	0.8790 ± 0.0747	0.6806 ± 0.1915	0.9778 ± 0.0444
	Kawahara et al.	0.7512 ± 0.1649	0.8500 ± 0.1109	0.5228 ± 0.3326	0.8670 ± 0.0913	0.9415 ± 0.0527	0.5211 ± 0.3431	0.9814 ± 0.0228
	ISIC	0.7581 ± 0.1225	0.8289 ± 0.0738	0.5443 ± 0.2545	0.8351 ± 0.0813	0.8290 ± 0.0779	0.5833 ± 0.2021	0.9329 ± 0.0510

The performance of the models is evaluated using different metrics, which are reported with the average and standard deviation calculated across the different cross-validation folds.

TABLE V
RESULTS OBTAINED FOR THE SUPERVISED AND SEMI-SUPERVISED LEARNING APPROACHES ON THE MIS VERSUS MIV CLASSIFICATION TASK

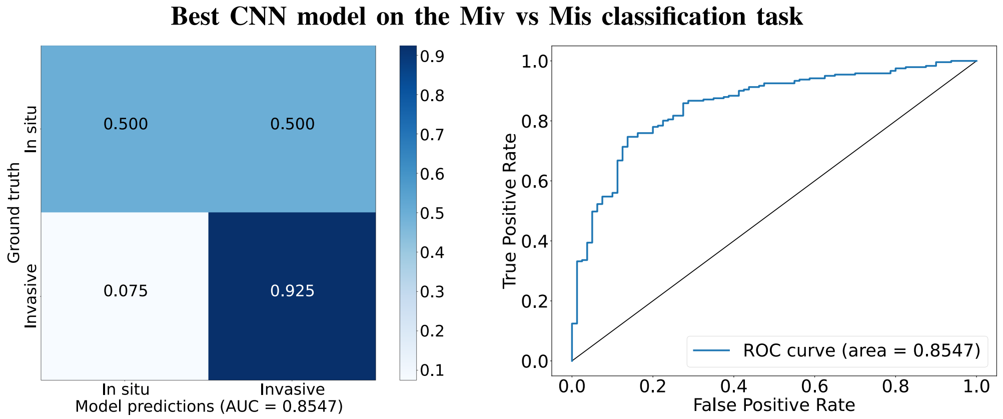
	Balanced acc	F1-score	Kappa score	Precision	AUC	Recall	Specificity
Supervised learning	0.6429 ± 0.0545	0.7315 ± 0.0335	0.2905 ± 0.0811	0.7442 ± 0.0427	0.7492 ± 0.0236	0.8527 ± 0.0705	0.4331 ± 0.1716
Semi-supervised learning	0.6540 ± 0.0399	0.7617 ± 0.0405	0.3499 ± 0.0800	0.7648 ± 0.0350	0.8085 ± 0.0242	0.9231 ± 0.0088	0.3849 ± 0.0826

The performance of the models is evaluated using different metrics, which are reported with the average and standard deviation calculated across the different cross-validation folds.

TABLE VI
RESULTS OBTAINED PER DATASET ON THE MIS VERSUS MIV CLASSIFICATION TASK

	Dataset	Balanced acc	F1-score	Kappa score	Precision	AUC	Recall	Specificity
Supervised learning	VRUH	0.6123 ± 0.0545	0.7443 ± 0.0493	0.2385 ± 0.0891	0.7617 ± 0.0716	0.7494 ± 0.0287	0.8880 ± 0.0583	0.3367 ± 0.1618
	Polesie et al.	0.7119 ± 0.0704	0.6874 ± 0.0942	0.4093 ± 0.1400	0.7597 ± 0.0508	0.7797 ± 0.0641	0.7408 ± 0.1535	0.6830 ± 0.2476
	Kawahara et al.	0.6838 ± 0.0718	0.7157 ± 0.0276	0.3571 ± 0.1044	0.7422 ± 0.0452	0.7346 ± 0.0515	0.7994 ± 0.1074	0.5682 ± 0.2481
	ISIC	0.5209 ± 0.0879	0.7016 ± 0.0735	0.0075 ± 0.1538	0.7691 ± 0.0791	0.4962 ± 0.1089	0.7419 ± 0.1531	0.3000 ± 0.2449
Semi-supervised learning	VRUH	0.6209 ± 0.0538	0.7656 ± 0.0532	0.2847 ± 0.1078	0.7774 ± 0.0406	0.7924 ± 0.0355	0.9455 ± 0.0229	0.2963 ± 0.1200
	Polesie et al.	0.6786 ± 0.0828	0.6860 ± 0.0738	0.3629 ± 0.1595	0.7156 ± 0.0882	0.8304 ± 0.0693	0.8762 ± 0.0680	0.4810 ± 0.1323
	Kawahara et al.	0.6266 ± 0.0534	0.6794 ± 0.0577	0.2613 ± 0.0984	0.6806 ± 0.0598	0.7105 ± 0.0476	0.8119 ± 0.0685	0.4414 ± 0.1225
	ISIC	0.7622 ± 0.1622	0.8744 ± 0.0731	0.4739 ± 0.2583	0.8748 ± 0.0837	0.8926 ± 0.0637	0.9345 ± 0.0199	0.5900 ± 0.3121

The performance of the models is evaluated using different metrics, which are reported with the average and standard deviation calculated across the different cross-validation folds.



(a) Normalized confusion matrix.

(b) ROC curve.

Fig. 4. Confusion matrix (left) and ROC curve (right) obtained for the best CNN model on the classification task between Mis and Miv. The confusion matrix is normalized and represents all the labeled samples in the test set of that specific fold. The best results are obtained with semi-supervision, achieving an AUC of 0.8547.

TABLE VII

Results Obtained for the Supervised and Semi-Supervised Learning Approaches on the Multiclass Classification Task (Mis Versus Miv With BT < 0.8 mm Versus Miv With BT ≥ 0.8 mm

	Balanced acc	F1-score	Kappa score	Precision	AUC
Supervised learning	0.5336 ± 0.0298	0.5494 ± 0.0412	0.4419 ± 0.0289	0.5648 ± 0.0405	0.7202 ± 0.0214
Semi-supervised learning	0.5697 ± 0.0335	0.5852 ± 0.0491	0.4854 ± 0.0733	0.6016 ± 0.0432	0.7571 ± 0.0271

The performance of the models is evaluated using different metrics, which are reported with the average and standard deviation calculated across the cross-validation folds.

TABLE VIII

Results Obtained per Dataset on the Multiclass Classification Task

	Dataset	Balanced acc	F1-score	Kappa score	Precision	AUC
Supervised learning	VRUH	0.5019 ± 0.0439	0.5301 ± 0.0695	0.4056 ± 0.0274	0.5534 ± 0.0709	0.7339 ± 0.0206
	Polesie et al.	0.6307 ± 0.0458	0.6557 ± 0.0594	0.6555 ± 0.1054	0.6755 ± 0.0727	0.7960 ± 0.0604
	Kawahara et al.	0.5972 ± 0.0597	0.5841 ± 0.0521	0.4865 ± 0.0453	0.6337 ± 0.0865	0.6901 ± 0.0288
	ISIC	0.2933 ± 0.0675	0.4147 ± 0.0690	-0.0573 ± 0.1132	0.4603 ± 0.0802	0.5114 ± 0.0847
Semi-supervised learning	VRUH	0.5459 ± 0.0118	0.5784 ± 0.0347	0.4391 ± 0.0578	0.6191 ± 0.0317	0.7546 ± 0.0062
	Polesie et al.	0.6093 ± 0.0285	0.5937 ± 0.0495	0.6306 ± 0.0742	0.6358 ± 0.0820	0.7622 ± 0.0566
	Kawahara et al.	0.5861 ± 0.0935	0.5930 ± 0.1040	0.4691 ± 0.1040	0.6427 ± 0.0918	0.7297 ± 0.0305
	ISIC	0.5629 ± 0.1871	0.6198 ± 0.2008	0.3682 ± 0.2813	0.6541 ± 0.1893	0.6973 ± 0.1369

The performance of the models is evaluated using different metrics, which are reported with the average and standard deviation calculated across the different crossvalidation folds.

TABLE IX

Summary of the Results Obtained on the External Test Set for Each of the Training Methods and Tasks

	Task	Balanced acc	F1-score	Kappa score	Precision	AUC	Recall	Specificity
Supervised learning	BT<0.8 vs ≥0.8 mm	0.7188 ± 0.0109	0.7269 ± 0.0073	0.4417 ± 0.0173	0.7302 ± 0.0092	0.7721 ± 0.0139	0.6456 ± 0.0619	0.7920 ± 0.0451
	Miv vs Mis	0.5679 ± 0.0235	0.6734 ± 0.0387	0.1406 ± 0.0535	0.6961 ± 0.0377	0.6162 ± 0.0305	0.7902 ± 0.1235	0.3456 ± 0.1278
	Multiclass	0.5270 ± 0.0162	0.5471 ± 0.0125	0.3944 ± 0.0290	0.5484 ± 0.0102	0.6978 ± 0.0114	-	-
Semi-supervised learning	BT<0.8 vs ≥0.8 mm	0.7498 ± 0.0111	0.7550 ± 0.0127	0.5011 ± 0.0227	0.7597 ± 0.0084	0.8024 ± 0.0080	0.7047 ± 0.0644	0.7950 ± 0.0612
	Miv vs Mis	0.5822 ± 0.0343	0.7193 ± 0.0241	0.1963 ± 0.0746	0.7177 ± 0.0248	0.6787 ± 0.0168	0.9241 ± 0.0232	0.2404 ± 0.0850
	Multiclass	0.5444 ± 0.0123	0.5651 ± 0.0160	0.4134 ± 0.0179	0.5816 ± 0.0166	0.7215 ± 0.0126	-	-

Each cell represents the average and standard deviation of the results obtained when evaluating the external test set with the 5 models in the cross-validation for a specific metric and task.

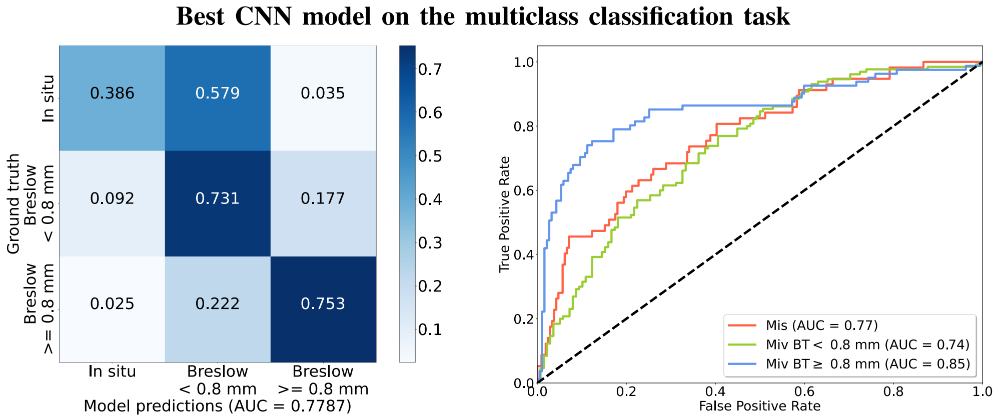
task, and an average AUC of 0.7215±0.0126 is obtained in the multiclass classification task.

E. Gradient Maps

Fig. 2 of the supplementary material shows a matrix of illustrative images in which semi-supervised learning improved the performance of fully supervised learning for the classification task between Mis and Miv.

V. Discussion

AI and deep learning have become very popular in many different fields, including medical image analysis, since they



(a) Normalized confusion matrix.

(b) ROC curve.

Fig. 5. Confusion matrix (left) and ROC curve (right) obtained for the best CNN model on the multiclass classification task. The confusion matrix is normalized and represents all the labeled samples in the test set of that specific fold. The best results are obtained with semi-supervision, achieving an AUC of 0.7787.

could help experts when diagnosing complex cases where the expert could be aided by a second opinion on the analysis, or even in routine cases, reducing the time dedicated by medical experts and allowing them to focus on more problematic cases.

In this regard, two different classification problems were studied in this work: the classification of melanoma samples as Mis or Miv, and as having a BT below 0.8 mm or greater than or equal to 0.8 mm. As mentioned previously, the fact of considering 0.8 mm as the cut-off is due to the fact that it is the threshold used in Europe to perform a sentinel lymph node biopsy when associated with additional histological risk factors such as ulceration. Moreover, we also considered a multiclass classification task combining Mis and Miv, together with the BT information, in order to analyze its performance compared to the binary tasks.

Publicly available datasets with ground-truth labels assigned by histopathological diagnosis are hard to find. We collected four different datasets from different sources in order to have a highly-heterogeneous dataset, facilitating the generalization of the CNN models during the training step. Furthermore, we provide the VRUH dataset as one of the largest labeled images repositories open available in the literature.

We tested both supervised learning and semi-supervised learning with an ensemble-based multi-teacher KD approach on both classification tasks. Supervised learning is the most common deep learning paradigm when it comes to training AI models. It enables fast training and accurate predictions with few data, as it requires samples to be annotated in order to train the models. This is also a counterpart of this paradigm, since annotated samples are not always available (or at least not as many as unlabeled data), and this approach does not allow benefiting from unlabeled samples in the training process.

Semi-supervised learning requires a more complex setup with multiple steps prior to training the models, which is much more time consuming than supervised learning. However, it allows for the use of unlabeled samples, which are easier to find, and thus improve the generalization capability of the model compared to supervised learning. In this case, we followed an ensemble-based multi-teacher KD approach, in which the five supervised models trained were used to pseudo-annotate unlabeled data with a majority voting algorithm in order to obtain more accurate labels. This approach led semi-supervised learning to achieve the highest results, improving those obtained by supervised learning in all the tasks performed, including the evaluation with an external dataset. Likewise, it also outperforms the results obtained in previous studies that used supervised learning models for the same classification tasks [17], [18], [20], [22]. This is probably due to the amount of unlabeled samples that could be introduced in the training step, which were able to increase the ability of the models to predict new unseen images. However, it should be taken into account that a 3-step process was conducted to this end: firstly, the teacher models were trained with supervised learning; then, the teachers were used to predict the unlabeled data; and finally, the student models were trained using the annotated and pseudo-annotated data. This 3-step process is definitely more time-consuming than the single-step supervised learning, and it should be taken into account when using this approach if the result obtained is not worth the effort. In the field of dermatology, only three groups of authors have used KD approaches as a decision-making support. Wang et al. [57] employed a KD framework called SSD-KD, focusing on multiclassification of dermoscopic images representing various skin conditions within the ISIC 2019 dataset, using lightweight deep learning models. Addressing melanoma detection, Khan et al. [16] utilized the pretrained CNN ResNet50 as the teacher model and developed what they named the 'Distill Student Network' as the student model, thus reducing computational complexity compared to other pre-trained models like EfficientNetB0, achieving acceptable accuracy in detecting melanoma. Adepu et al. [25] applied a KD approach to manage the high level of inter-class similarity and low intra-class similarity in dermoscopic images for binary classification between melanoma and non-melanoma within the ISIC archive 2020 dataset. In our study, the ensembled-based KD with semisupervised CNN was focused on the two classification tasks for the differentiation of Mis or Miv and the classification by Breslow thickness, together with the multiclass classification task. In the three cases semi-supervised models overcome those metrics of the supervised.

In order to analyze and compare the performance of both binary classification tasks performaed with a combined approach, a multiclass classification evaluation was carried out, where the models were trained to classify between Mis, Miv with BT $<$ 0.8 mm and Miv with BT ≥ 0.8 mm. The results obtained show lower performance than those of binary tasks, however, the main aspect to take into account is its low precision in the Mis class. The results obtained show lower performance than those of binary tasks; however, the main aspect to take into account is its low precision in the Mis class. It could be caused by the similarity between Mis and thin melanomas (which are the majority of invasive melanomas) in most of their clinical and dermoscopic features, that could limit the prediction both for dermatologist and CNN models [10]. This observation prompted us to consider the possibility of comparing the multiclass approach with a dual-stage execution pipeline. In this proposed pipeline, the image would first be classified between Mis and Miv, and those classified as Miv would then be used as input to a secondary model that classifies BT. This approach will be considered in future work.

Different KD approaches could be followed instead of the one we used. We used the multi-teacher with majority voting, since stratified 5-fold cross-validation was already used on the supervised learning side, although it is not the only one we could have used. Better results could have been obtained with other solutions; however, it was not the focus of this study to achieve the highest results possible, but to perform a comparison between supervised learning and semi-supervision in the two classification tasks considered, which we believe is a novelty in the field of melanoma image analysis.

As a future work, we will also explore training regression models rather than classification ones to predict BT, since they would bring much potential for expert dermatologists. However, for that purpose, a larger, more heterogeneous and better represented dataset would be needed in order to perform a regression task. Even though we have used what we believe is the current largest melanoma dataset with Breslow thickness annotations, we consider it still being far from optimal for training models on a regression task. Moreover, most of the datasets used do not report a specific BT value, but a wide range between two thicknesses, which, combined with the previous aforementioned aspect, makes the amount of samples with specific BT labels available even lower. Still, we find the idea of performing regression over the BT value very interesting and we would focus on increasing the number of labeled samples that are publicly available for future research on it.

VI. Conclusion

In this work, an offline response-based Multi-Teacher KD algorithm was used together with semi-supervised deep CNNs for two different classification tasks related to melanoma classification: a BT classification task and a Mis vs Miv classification task. This approach was compared to the standard supervised learning approach using stratified 5-fold cross-validation, which were the same models used as teachers for the semi-supervised learning. A total of 4 datasets from different sources were used to increase the heterogeneity of the data and the generalization of the trained models. The dataset sourced from Virgen del Rocio University Hospital has been made publicly available for this work and can be accessed through its website, allowing other researchers to use it.

The performance of the different methods considered was analyzed using different evaluation metrics, where semi-supervision showed the best results in both classification tasks. Moreover, semi-supervision allows exploiting unlabeled data, which is not possible with supervised learning, where only strongly-annotated data can be used. As a counterpart to the use of semi-supervision, it is more time-consuming than supervised learning, since it first needs a supervised model to infer on the unlabeled data. This aspect is even worse in our case, since a Multi-Teacher KD approach was used.

The results presented in this work show the potential of semi-supervised learning in BT classification in melanoma and could be of great use for expert dermatologists to speed up the analysis of routine cases and serve as a second opinion.

The source code used in this work is available on an open source GitHub repository.[6] Images from the Virgen del Rocío University Hospital dataset are available upon request.[7]

References

[1] C. Garbe et al., "European consensus-based interdisciplinary guideline for melanoma. Part 1: Diagnostics: Update 2022," *Eur. J. Cancer*, vol. 170, pp. 236–255, 2022.

[2] M. Arnold et al., "Global burden of cutaneous melanoma in 2020 and projections to 2040," *JAMA Dermatol.*, vol. 158, no. 5, pp. 495–503, 2022.

[3] A. M. Eggermont, A. Spatz, and C. Robert, "Cutaneous melanoma," *Lancet*, vol. 383, no. 9919, pp. 816–827, 2014.

[4] P. Aggarwal, P. Knabel, and A. B. Fleischer Jr, "United States burden of melanoma and non-melanoma skin cancer from 1990 to 2019," *J. Amer. Acad. Dermatol.*, vol. 85, no. 2, pp. 388–395, 2021.

[5] A. C. Geller, S. M. Swetter, and M. A. Weinstock, "Focus on early detection to reduce melanoma deaths," *J. Invest. Dermatol.*, vol. 135, no. 4, pp. 947–949, 2015.

[6] C. Ring, N. Cox, and J. B. Lee, "Dermatoscopy," *Clin. Dermatol.*, vol. 39, no. 4, pp. 635–642, 2021.

[7] L. Thomas and S. Puig, "Dermoscopy, digital dermoscopy and other diagnostic tools in the early detection of melanoma and follow-up of high-risk skin cancer patients," *Acta dermato-venereologica*, vol. 97, pp. 14–21, 2017.

[8] L. K. Ferris, "Early detection of melanoma: Rethinking the outcomes that matter," *JAMA Dermatol.*, vol. 157, no. 5, pp. 511–513, 2021.

[9] C. Garbe et al., "European consensus-based interdisciplinary guideline for melanoma. Part 2: Treatment-update 2022," *Eur. J. Cancer*, vol. 170, pp. 256–284, 2022.

[10] A. Lallas et al., "Accuracy of dermoscopic criteria for the diagnosis of melanoma in situ," *JAMA Dermatol.*, vol. 154, no. 4, pp. 414–419, 2018.

[11] E. Rodríguez-Lomba et al., "'Rainbow pattern': A dermoscopic sign of invasive melanoma," *Clin. Exp. Dermatol.*, vol. 47, no. 3, pp. 529–533, 2022.

[12] S. M. Swetter et al., "Guidelines of care for the management of primary cutaneous melanoma," *J. Amer. Acad. Dermatol.*, vol. 80, no. 1, pp. 208–250, 2019.

[13] F. R.-d. l. Torre, "One-step surgical removal of a cutaneous melanoma: Current evidence," *Actas Dermo-Sifiliograficas*, vol. 111, no. 7, pp. 541–544, 2020.

[6] https://github.com/jpdominguez/Breslow_Melanoma_DeepLearning. Retrieved September 23, 2024.

[7] https://institucional.us.es/breslowdataset. Retrieved September 23, 2024.

[14] R. C. Maron et al., “Artificial intelligence and its effect on dermatologists’ accuracy in dermoscopic melanoma image classification: Web-based survey study,” *J. Med. Internet Res.*, vol. 22, no. 9, 2020, Art. no. e18091.

[15] S. Haggenmüller et al., “Skin cancer classification via convolutional neural networks: Systematic review of studies involving human experts,” *Eur. J. Cancer*, vol. 156, pp. 202–216, 2021.

[16] M. S. Khan, K. N. Alam, A. R. Dhruba, H. Zunair, and N. Mohammed, “Knowledge distillation approach towards melanoma detection,” *Comput. Biol. Med.*, vol. 146, 2022, Art. no. 105581.

[17] S. Polesie et al., “Discrimination between invasive and in situ melanomas using clinical close-up images and a de novo convolutional neural network,” *Front. Med.*, vol. 8, 2021, Art. no. 723914.

[18] M. Gillstedt, E. Hedlund, J. Paoli, and S. Polesie, “Discrimination between invasive and in situ melanomas using a convolutional neural network,” *J. Amer. Acad. Dermatol.*, vol. 86, no. 3, pp. 647–649, 2022.

[19] M. Gillstedt et al., “Evaluation of melanoma thickness with clinical close-up and dermoscopic images using a convolutional neural network,” *Acta Dermato-Venereologica*, vol. 102, 2022, Art. no. 102:adv00790.

[20] S. Polesie, M. Gillstedt, H. Kittler, C. Rinner, P. Tschandl, and J. Paoli, “Assessment of melanoma thickness based on dermoscopy images: An open, web-based, international, diagnostic study,” *J. Eur. Acad. Dermatol. Venereol.*, vol. 36, no. 11, pp. 2002–2007, 2022.

[21] Y. S. Chu et al., “Deep learning algorithms for predicting breslow thickness from dermoscopic images of acral lentiginous melanomas,” *J. Invest. Dermatol.*, vol. 142, pp. 2268–2271, 2022.

[22] J.-C. Hernández-Rodríguez, L. Durán-López, J. P. Domínguez-Morales, J. Ortiz-Álvarez, J. Conejo-Mir, and J.-J. Pereyra-Rodriguez, “Prediction of melanoma breslow thickness using deep transfer learning algorithms,” *Clin. Exp. Dermatol.*, vol. 48, 2023, Art. no. llad107.

[23] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, *arXiv:1503.02531*.

[24] H. Choi, Y. Lee, K. C. Yow, and M. Jeon, “Block change learning for knowledge distillation,” *Inf. Sci.*, vol. 513, pp. 360–371, 2020.

[25] A. K. Adepu, S. Sahayam, U. Jayaraman, and R. Arramraju, “Melanoma classification from dermatoscopy images using knowledge distillation for highly imbalanced data,” *Comput. Biol. Med.*, vol. 154, 2023, Art. no. 106571.

[26] V. Rotemberg et al., “A patient-centric dataset of images and metadata for identifying melanomas using clinical context,” *Sci. Data*, vol. 8, no. 1, 2021, Art. no. 34.

[27] S. Polesie et al., “Can dermoscopy be used to predict if a melanoma is in situ or invasive?” *Dermatol. Practical Conceptual*, vol. 11, no. 3, 2021, Art. no. e2021079.

[28] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, “Seven-point checklist and skin lesion classification using multitask multimodal neural nets,” *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 538–546, Mar. 2019.

[29] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, 2020, Art. no. 125.

[30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge distillation: A survey,” *Int. J. Comput. Vis.*, vol. 129, pp. 1789–1819, 2021.

[31] J. Liu, T. Zheng, G. Zhang, and Q. Hao, “Graph-based knowledge distillation: A survey and experimental evaluation,” 2023, *arXiv:2302.14643*.

[32] S. Lee and B. C. Song, “Graph-based knowledge distillation by multi-head attention network,” 2019, *arXiv:1907.02226*.

[33] S. Gupta, J. Hoffman, and J. Malik, “Cross modal distillation for supervision transfer,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2827–2836.

[34] S. Roheda, B. S. Riggan, H. Krim, and L. Dai, “Cross-modality distillation: A case for conditional generative adversarial networks,” in *2018 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2926–2930.

[35] M. Ji, B. Heo, and S. Park, “Show, attend and distill: Knowledge distillation via attention-based feature matching,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 7945–7952.

[36] P. Passban, Y. Wu, M. Rezagholizadeh, and Q. Liu, “ALP-KD: Attention-based layer projection for knowledge distillation,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 13657–13665.

[37] D. Chen et al., “Cross-layer distillation with semantic calibration,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 7028–7036.

[38] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, J. Cui, and B. Ramabhadran, “Efficient knowledge distillation from an ensemble of teachers,” in *Proc. Interspeech*, 2017, pp. 3697–3701.

[39] J. M. Noothout et al., “Knowledge distillation with ensembles of convolutional neural networks for medical image segmentation,” *J. Med. Imag.*, vol. 9, no. 5, 2022, Art. no. 052407.

[40] F. Yuan et al., “Reinforced multi-teacher selection for knowledge distillation,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 14284–14291.

[41] J. E. V. Engelen and H. H. Hoos, “A survey on semi-supervised learning,” *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[42] H. Chen, Y. Wang, C. Xu, C. Xu, and D. Tao, “Learning student networks via feature embedding,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 25–35, Jan. 2021.

[43] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, “Dual student: Breaking the limits of the teacher in semi-supervised learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6727–6735.

[44] D. Zhang, B. Chen, J. Chong, and S. Li, “Weakly-supervised teacher-student network for liver tumor segmentation from non-enhanced images,” *Med. Image Anal.*, vol. 70, 2021, Art. no. 102005.

[45] Z. Xiao, Y. Su, Z. Deng, and W. Zhang, “Efficient combination of CNN and transformer for dual-teacher uncertainty-guided semi-supervised medical image segmentation,” *Comput. Methods Programs Biomed.*, vol. 226, 2022, Art. no. 107099.

[46] N. Marini, S. Otálora, H. Müller, and M. Atzori, “Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification,” *Med. Image Anal.*, vol. 73, 2021, Art. no. 102165.

[47] K. Wang et al., “Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation,” in *Proc. Med. Image Comput. Comput. Assist. Intervention– 2021: 24th Int. Conf., Proc., Part II 24*, Strasbourg, France, Sep. 27–Oct. 1, 2021, pp. 450–460.

[48] J. P. Dominguez-Morales et al., “A systematic comparison of deep learning methods for gleason grading and scoring,” *Med. Image Anal.*, vol. 95, 2024, Art. no. 103191.

[49] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++ : Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE winter Conf. Appl. Comput. Vis.*, 2018, pp. 839–847.

[50] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, vol. 32, 2019.

[51] M. Abadi et al., “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” 2016, *arXiv:1603.04467*.

[52] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[54] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conf. Comput. Vis. pattern Recognit.*, 2009, pp. 248–255.

[56] F. J. Pontes, G. Amorim, P. P. Balestrassi, A. Paiva, and J. R. Ferreira, “Design of experiments and focused grid search for neural network parameter optimization,” *Neurocomputing*, vol. 186, pp. 22–34, 2016.

[57] Y. Wang, Y. Wang, J. Cai, T. K. Lee, C. Miao, and Z. J. Wang, “SSD-KD: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images,” *Med. Image Anal.*, vol. 84, 2023, Art. no. 102693.