Ensembling CNNs for dermoscopic analysis of suspicious skin lesions

Yali Nie
dept. of Electronics Design
Mid Sweden University
Sundsvall, Sweden
yali.nie@miun.se

Matteo Ferro
dept. of Industrial Engineer
University of Salerno
Salerno, Italy
mferro@unisa.it

Paolo Sommella
dept. of Industrial Engineer
University of Salerno
Salerno, Italy
psommella@unisa.it

Marco Carratù
dept. of Industrial Engineer
University of Salerno
Salerno, Italy
mcarratu@unisa.it

Sara Cacciapuoti
dept. of Medicine and Clinical Surgery
University Federico II of Naples
Naples, Italy
sara.cacciapuoti@libero.it

Giuseppe Di Leo
dept. of Industrial Engineer
University of Salerno
Salerno, Italy
gdileo@unisa.it

Jan Lundgren
dept. of Electronics Design
Mid Sweden University
Sundsvall, Sweden
jan.lundgren@miun.se

Gabriella Fabbrocini
dept. of Medicine and Clinical Surgery
University Federico II of Naples
Naples, Italy
gafabbro@unina.it

***Abstract*—Deep Convolution Neural Networks (CNN) enable advanced methods to predict the skin cancer classes through the automatic analysis of digital dermoscopic images. However, small datasets' availability often allows the models to be characterized by low prediction accuracy and poor generalization ability, which significantly influences clinical decisions. This paper proposes to use an original ensembling of multiple CNNs as feature extractors able to detect and measure skin lesions atypical criteria according to the well-known diagnostic method 7-Point Check List. The experimental results show that the Artificial Intelligence-based model can suitably manage the classification uncertainty of the single CNNs and finally distinguish melanomas from benignant nevi. Diagnostic performance is promising in terms of sensitivity and specificity towards a decision-supporting system used by a dermatologist with low experience during clinical practice.**

Keywords—DCNNs, melanoma, K-Fold, classification

I. INTRODUCTION

The American Cancer Organization estimates that 207,390 cases of melanoma will be diagnosed in the U.S., and an estimated 7,180 people will die of melanoma in 2021 [1]. Melanoma of the skin is the 19th most commonly occurring cancer in men and women. There were nearly 300,000 new cases in 2018 [2]. Considering the national estimates melanoma heatmap in 185 countries by age-standardized incidence rates in 2018, the most painful areas are concentrated in Europe, North America, and Oceania.

For skin cancer diagnoses, three different types of images are adopted (an example is reported in Fig. 1): dermoscopic, clinical, and histopathological images to present significant differences related to the level of details available in each image. In this paper, we focus on dermoscopic image classification.

(a) Dermoscopic (b) Clinical (c) Histopathological

Fig. 1 The difference between dermoscopic [4] clinical [5], and histopathological [6] images of a skin cancer

Dermoscopy (also referred to as epiluminescence microscopy) images have played a significant role in skin cancer by increasing patients' survival rate [3]. Indeed, it has been possible to assist clinicians efficiently since dermatoscopy captures the dermal features and eliminates the surface glare. Dermatoscope is a non-invasive technique where a gel is applied on the skin lesion's surface, and an enhanced image is acquired using the digital imaging technique dermatoscopy. The device magnifies otherwise invisible structures to naked eyes, thus helping to detect melanoma from other types of skin cancer.

Over the past decades, a significant research effort was also addressed to the dermoscopic analysis automation: most image classification techniques utilize classical machine learning methods to extract features relying on hand-crafted features. Aitken et al. classified pigmented skin lesions, including their color, size, shape, and distinctness of boundary [4]. Based on the criteria asymmetry (A), border (B), color (C), and differential structure (D), Nachbar et al. developed the "ABCD" rule [5] to make the classification of pigmented skin lesions. Yuan et al. employed a support vector machine (SVM) based on texture information [6] to achieve early melanoma detection. The 7-point checklist [7]-[8] is a typical method for diagnosing melanoma skin cancer. Here, by the simple addition of the individual scores, a minimum total score of 3 is required for malignant diagnosis, whereas a total score of less than 3 indicates a benign.

Recently, Convolution Neural Networks (CNNs) have become a popular approach to handle the different medical image analysis tasks [9]-[12]. They can learn more complex feature hierarchies from raw data. In 2012, image classification makes a big progress as the publication of [13]. This paper was used to win the 2012 ILSVRC (ImageNet Large-Scale Visual Recognition Challenge). 2012 marked the first year where a CNN was used to achieve a top 5 test error rate of 15.4%. However, the next best entry an error of 26.2%, a remarkable improvement that pretty much shocked the computer vision community.

Starting from previous experience and works concerning both the application of image processing techniques to the measurement of dermoscopic parameters [14]-[16] and the optimization of Deep Learning approaches [17],[18], the present study represents the first step towards the development of a methodology that allows the potentials of CNNs to be

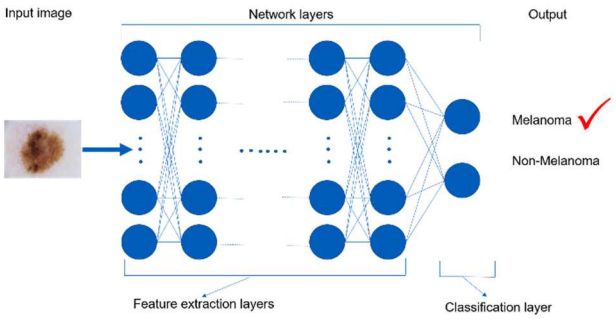

Fig. 2 The melanoma classification flow by Deep Learning

fully exploited for an efficient skin cancer classification when a small dermoscopy dataset is available. In detail, an original approach is introduced to classify malignant and benign skin lesions using an ensemble of three well-known CNN models. The idea being investigated is that the ensemble may obtain better performance than the individual predictions from the CNNs trained to recognize the presence or absence of 7 diagnostic features commonly used by dermatologists. The predictions from the CNNs are combined using logic rules and a final decision is made by a Logistic Model Tree (LMT).

The layout of the paper continues with the the state of the art about the development and application of CNNs for image processing in Section II. In Section III, the landmark literature about the CNN application to dermoscopic issues is critically reviewed to disclose the novel proposal. The experiment and result pertaining to melanoma skin cancer data, training process, and the results' analysis are detailed in Section IV. The last section presents the conclusions and plans of this research.

II. State of the Art

Deep Learning (DL) is fast becoming a key instrument in Artificial Intelligence applications such as feature extraction, pattern recognition, and classification. This innovative Machine Learning (ML) technique utilizes many layers of nonlinear information processing. When DL takes the raw image as input, do the heaviest lifting in image processing, and automatically extract useful features from the dataset. However, traditional ML usually uses several complex procedures, such as pre-processing, segmentation, and hand-crafted features.

DL has overcome the difficulty of hand-crafted computing features and learned from its errors. Instead of human intervention and a small dataset, DL discovers the informative feature representation automatically with large data sets. DL retains image features through the stack layer of the neural network. The deeper, the more it learns. Hence, the name DL whose power is enormous. DL analyzes data with a neural network structure similar to how a human draws a conclusion. In detail, CNNs is one deep learning algorithm that is very efficient for solving general and highly variable tasks with large datasets. It can extract the regional characteristics of the original image based on the local feature extraction. Generally, by adding more layers, the learning model can get high precision through Supervised Learning. To understand it well, we take Fig. 2 as an example. The input melanoma image passes through the layer-by-layer. Early layers detect patterns in the image, and then later layers detect patterns with patterns until it creates the features vector. The last layer usually acts as the classifier that outputs the final predicted class label.

A considerable variety of approaches and Deep Neural Network models have been introduced in the last few years [19],[20]. Training a model on a large image-set gives it an ability to match the human-level vision, given the diversity of data. The plot in Fig. 3 illustrates how the models have improved on ImageNet (consisting of more than 14 million images and comprising classes such as animals, flowers, everyday objects, people, and many more) over the years from 2011 to 2020. Here, it shows some top-performing methods according to papers with code on the widely popular datasets used for benchmarking the image classification models. These models can be used as the basis for transfer learning in computer vision applications. Transfer learning generally refers to a process where a model trained on one problem is used in some way on a second related problem.

AlexNet is the first CNN network to try to solve large-scale and complex tasks. It has introduced many subsequent network modules that have been widely adopted: Rectified Linear Units (ReLU), Local Response Normalization (LRN), and Overlapping Pooling. On the other hand, AlexNet provides practical measures to prevent overfitting: Data Augmentation, Dropout, and network training methods. VGG is an enhanced version of AlexNet. Compared with AlexNet, on the one hand, VGG decomposes the 11×11 and 5×5 convolution kernels in AlexNet into a series of 3×3 small convolution kernels and increases the depth of the network. GoogLeNet introduces the Inception module, which increases the ability of multi-scale feature learning while considering the impact of network depth and width on network performance. It introduces a 1×1 convolution kernel to reduce the number of channels input to the large convolution kernel's feature map. ResNet does not directly fit the objective function but instead fits the residual of the objective function, which solves the problem that simply increasing the network depth will degrade the network's performance. The rest of the networks are based on the enhanced version or deformation of the above networks. Some models are lightweight, such as SqueezeNet, MobileNet, ShuffleNet, Xception, and MnasNet.

III. Related Works and New Proposal

About the classification of skin cancer through the automatic analysis of dermoscopic images, the different CNN based approaches proposed in literature should be revised and compared not only regarding the corresponding accuracy but also considering the actual possibilities to create algorithms representing diverse patient populations; ensure algorithm

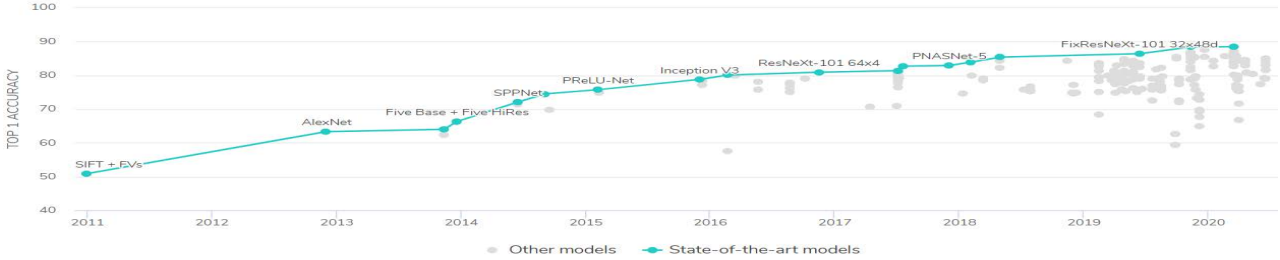

Fig. 3 Comparison of the proposed Deep Learning Models

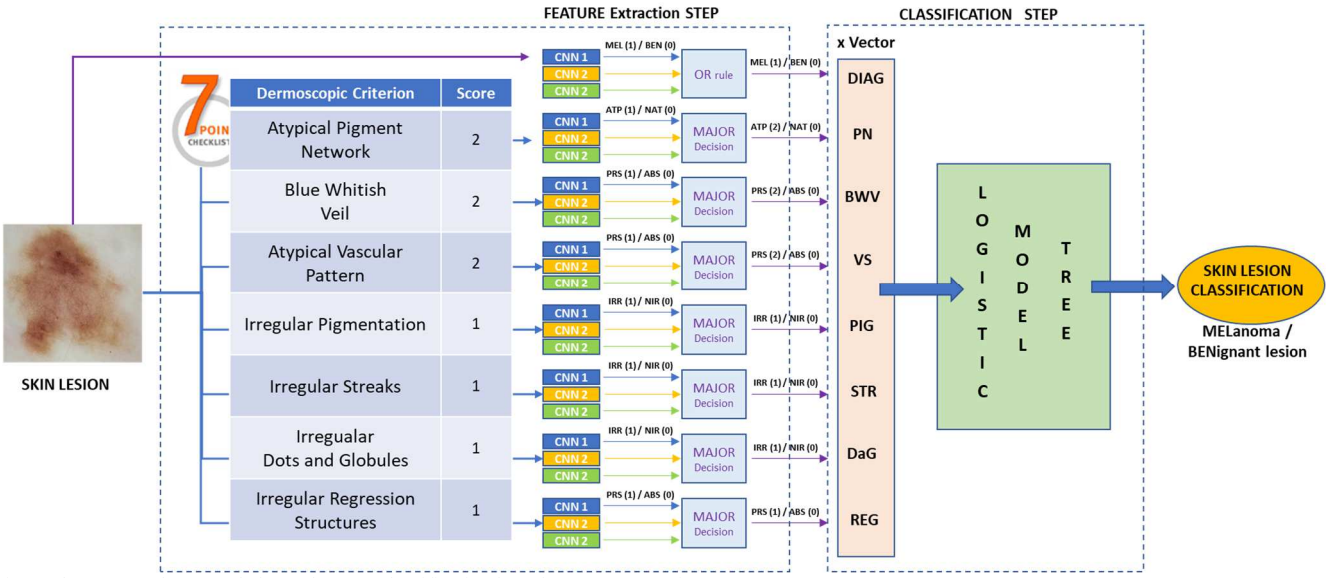

Fig. 4 The proposed approach for melanoma classification based on Deep Learning

output is ultimately interpretable; validate algorithm performance prospectively; preserve human-patient interaction when necessary, and demonstrate validity in the eyes of regulatory bodies.

From this point of view, an exciting overview and categorization of the presented CNNs methods are reported in [21]: CNNs can be used to classify skin lesions in two fundamentally different ways.

On the one hand, a CNN pre-trained on another large dataset, such as ImageNet, can be applied as a feature extractor. In this case, classification is performed by another classifier, such as k-Nearest Neighbors (kNN), Support Vector Machines (SVM), or Artificial Neural Networks (ANN).

On the other hand, a CNN can directly learn the relationship between the raw pixel data and the class labels through end-to-end learning. In contrast with the classical workflow typically applied in machine learning, feature extraction becomes an integral part of classification and is no longer considered a separate, independent processing step. If the CNN is trained by end-to-end learning, the research can be divided into two different approaches: learning the model from scratch or transfer learning.

To the latter approach belongs the landmark publication by Esteva et al. [22]. The proposed CNN model adopts the GoogLeNet Inception v3 model pre-trained with the large image database ImageNet and then fine-tuned to classify skin lesions using transfer learning concerning more than 120,000 clinical images. The model achieved a value for the Area Under Curve (AUC) of the corresponding Receiver Operating Characteristics (ROC) AUC equal to .94 about skin lesions classified exclusively with dermatoscopic images. The very similar approach presented in [23] (where the modified version of the GoogleNet Inception CNN architecture was additionally trained with more than 100,000 digital images) showed a significant worseness of the diagnostic accuracy (0.86 achieved as AUC for the classification task of melanomas versus benign nevi) when a comparison of the diagnostic performance between the CNN model and a group of dermatologist was carried out about a collection of 100 dermoscopic images representing the spectrum of melanocytic lesions as typically encountered in daily clinical routine.

About the former approach, the most recent works and meta-analysis carried out by experts in both the domains of computer science and dermatology highlight the exploitation of the CNN as feature extraction may lead to satisfying diagnostic performance (similar to the behavior shown by physicians with short clinical experience) also when the transfer learning is applied to the small proprietary datasets (typically including less than 2000 dermoscopic images and the corresponding expert annotations and biopsy results) that are often available from the involved clinical institution.

Indeed, most results show an excellent sensitivity rather than a satisfying specificity (like the performance exhibited by the automatic classification systems developed based on traditional image processing and Machine Learning techniques), probably as a consequence of the typical approach of focusing on the atypical features that characterize the malignant skin lesions (that leads to unbalanced datasets). However, the exploitation of the CNNs as feature extraction and basis for the following classification in the author's opinion is still the most interesting approach for the future improvements both in diagnostic performance (that may be a guarantee, for example, by further analysis applied to pre-processing step, such the suitable selection of different color spaces and/or correction algorithms and the data augmentation provided by the image splitting into smaller sub-areas for better exploiting the original resolution of the image-set) and in supporting the dermatologist with an automatic second opinion (by using the extracted features according to well-known diagnostic methods thus resulting in interpretable classification).

In this framework, the works reported in [24]-[25] are auspicious. They adopt suitable CNNs for automatic detection of the dermoscopic criteria included in the 7-Point Check List score to be adopted as the basis of the ensembling approach for skin cancer classification.

Following this research trend, the authors introduce the second level of processing that intends to ensemble the classification results of multiple CNNs for both the feature extraction and the dichotomous skin lesion classification (melanoma vs. benignant nevus) to exploit the inner uncertainty of the adopted DL and the unbalanced imageset. The proposed methodology is depicted in the scheme of Fig. 4, where three different CNN models are adopted for each one

TABLE I DETAILS OF THE DATASET ADOPTED FOR EXPERIMENTAL ANALYSIS.

abbrev.	Original clinical Diagnosis	# images of sub-case	# images
MEL	Melanoma / metastatis	5	252
	Melanoma in situ	64	
	Melanoma (less than 0.76 mm)	102	
	Melanoma (0.76 to 1.5 mm)	53	
	Melanoma (more than 1.5 mm)	28	
BEN	Basal Cell Carcinoma	42	759
	Blue / Clark / Congenital / Dermal / Recurrent / Reed/Spitz nevus	575	
	Dermatofibroma	20	
	Lentigo / Melanosis / Miscellaneous	48	
	Vascular lesion	29	
	Seborrheic keratosis	45	
		TOTAL	1011

of the seven dermoscopic criteria (three major and four minor criteria are taken into account with a score equal to 2 and 1, respectively) and the skin lesion classification. The results from the corresponding CNNs are fused according to a decision policy (including both the or-rule and the majority decision) and then processed (as vector x including eight components) according to a suitable Logistic Model Tree [26] that can classify the dermoscopic image as melanoma or benignant lesion.

IV. Proposal Validation

The methodology proposed for computerized image-based prediction of the 7-point skin lesion malignancy checklist has been developed and tested with reference to a publicly available dataset [24] for training and evaluating computer-aided diagnosis systems. The dataset including over 1000 dermoscopy colour images (sized to $512 \times 512 \times 3$) corresponding to different cases has been noted to have "excellent interobserver agreements" and was used to teach dermatologists [27], [28], suggesting that it is a suitable source for training machine learning algorithms. In detail, dermoscopic images are taken with a dermatoscope and offer a standardized field of view; moreover, they have been captured under controlled conditions like standard illumination, lighting, and contrast.

The original dataset contains labels at a very granular level. As some labels occur infrequently and many labels have a similar clinical interpretation (e.g. types of benign nevi) and the paper is focused on the dichotomous classification problem, we have grouped the clinical cases as reported in Table I. The same approach has been applied for the 7-point criteria where the labels with similar clinical meaning and melanoma score contributions are grouped (see Table II). The final label grouping adopted in the further experimental analysis is shown in the *abbrev* column for both interest tables.

According to the proposed approach, three different CNN models have been adopted as three independent observers, namely the pre-trained versions of GoogleNet, Inception-v3, and ResNet-101 networks, all trained on more than a million images from the ImageNet database. These models have been

TABLE II DETAILS OF THE DATASET ADOPTED FOR EXPERIMENTAL ANALYSIS.

Dermoscopic Criterion	abbrev	Original interpretation	# images
(PN) Pigment Network	ATP	Atypical	230
	NAT	Absent / Atypical	781
(BWV) Blue Whitish Veil	PRS	Present	195
	ABS	Absent	816
(VS) Vascular Structures	PRS	Arborizing / Comma/ Hairpin / within regression / Dotted / Linear Irregular	188
	ABS		823
(PIG) Pigmentation	IRR	Diffuse / Localized Irregular	305
	NIR	Absent / Diffuse / Localized Regular	706
(STR) Streaks	IRR	Irregular	251
	NIR	Absent / Regular	760
(DaG) Dots and Globules	IRR	Irregular	448
	NIR	Absent / Regular	663
(REG) Regression Structures	PRS	Blue / White Areas / Combinations	253
	ABS	Absent	758

selected based on the previous literature results [29]-[30]: they represent a good compromise between the complexity (reduced parameters if compared with other extensively adopted networks such as VGG-16 and VGG-19 networks) and prediction performance when the transfer learning is limited to small dermoscopic imagesets.

About the prediction indexes, the development and testing of the models for each feature to be extracted and for the global classification have considered the *sensitivity* (*SE*) and *specificity* (*SP*) defined as the ratio of the correct decisions to the number of the cases where the corresponding dermoscopic criterion is present or absent, respectively. Moreover, the prediction accuracy (*ACC*) has been considered, which represents the weighted mean of sensitivity and specificity according to the distribution of the dichotomous cases within the dataset. To highlight the reliability of the prediction, the Cohen's kappa coefficient (k) has been considered according to the interpretation reported in Table III.

Since the distribution of the cases for most 7-point criteria is unbalanced, suitable Training and Test sets have been manually selected to consider the occurrence ratio of the present issues to the absent ones not lower than 1:3, as reported in Table IV. The adopted strategy was expected to prevent the training of feature extraction CNN models not able to detect the single dermoscopic criterion. An analogous approach has been adopted to split the starting imageset in the Training and Testing datasets for melanoma classification that share the same case distribution (occurrence ratio of melanomas to benign nevi approximately equal to 0.33).

TABLE III INTERPRETATION OF COHEN'S KAPPA.

Value of k	*0 -.39*	*.40 -- .59*	*.60 -- .79*	*.80 -- 1.0*
Level of Agreement	minimal	weak	moderate	strong

TABLE IV PREDICTION PERFORMANCE OF THE TRAINED CNNS.

Feature	Label	Train	Test	GoogleNet			Inception-v3			ResNet-101			OR-Rule			SPLIT-Decision		
		# img	# img	SE	SP	k	SE	SP	k	SE	SP	k	SE	SP	k	SE	SP	k
PN	ATP	138	92	29	85	.25	54	89	.52	44	79	.29	73	70	.38	34	87	.32
	NAT	400	270															
BWV	PRS	117	78	25	97	.52	25	93	.45	28	96	.44	39	91	.34	22	98	.49
	ABS	350	230															
VS	PRS	112	76	15	94	.21	44	85	.16	67	75	.25	74	68	.12	41	89	.20
	ABS	325	225															
PIG	IRR	183	122	43	80	.32	47	79	.34	50	76	.31	71	60	.32	45	80	.35
	NIR	370	250															
STR	IRR	150	101	51	83	.27	59	79	.32	72	63	.25	81	48	.18	61	81	.32
	NIR	450	300															
DaG	IRR	268	180	63	74	.44	73	68	.47	69	68	.35	87	49	.33	69	72	.48
	NIR	350	210															
REG	PRS	150	103	54	77	.32	67	58	.25	67	58	.28	85	41	.22	66	65	.32
	ABS	450	308															
DIAG	MEL	151	101	63	67	.30	79	73	.46	74	72	.44	89	51	.32	74	74	.46
	BEN	455	304															

About the main training parameters, the number of epochs (25), the batch size (32), learning rate (0.001), the optimizer (Adam) that have experimentally guarantee the best validation results in terms of the prediction accuracy and the function loss (cross-entropy) by the models trained for detecting the Blue-Whitish Veil (the least frequent criterion together with Vascular Structures) have been adopted for all the CNNs.

The trained models' prediction capabilities for the feature extraction and melanoma classification with respect to the corresponding selected Test sets are summarized in Table IV. As expected, about the main classification problem (melanoma vs benign nevus), the achieved CNNs exhibit comparable performance when the synthetic indexes are evaluated: a satisfying specificity for all the models and an increasing sensitivity guaranteed by the ResNet-101 and Inception-v3 (that seems to provide the best and more reliable results as reported by most literature). However, the analysis of disaggregated data (i.e., the single dermoscopic images classification) shows the three CNNs are able to highlight different patterns of each other and may be considered as "different observers". Thus, a kind of information gain may be achieved through a suitable combination of the models' prediction results as intended by the proposed approach. As reported in the last columns of Table IV, when a suitable decision rule is applied (i.e. the skin lesion is classified as melanoma if it is predicted by one or two CNNs according to the OR and Split rule, respectively) a slight increase of both sensitivity and specificity is achieved leading to greater accuracy (typically equal to the diagnostic professional of dermatologist with a medium level of experience in dermoscopy).

About the prediction capability of the CNNs trained to detect the presence of 7-Point Check List criteria, poor performance is achieved in terms of Cohen's kappa and sensitivity, especially for the dermoscopic atypical features that are less frequent (BWV and VS) whereas, all the types of CNN models generally obtain a reasonable specificity. This behavior leads to apply a Split rule for the feature classification of the single lesion based on the CNN results. The decision strategy aims to improve the sensitivity related to the major dermoscopic criteria and maintain both the specificity and reliability as high as possible for all features.

Thus, as an intermediate result of the proposed feature extraction, an x vector is achieved, including 8 numeric variables, and considered for a suitable Logistic Model Tree training by adopting the whole database and 10-fold validation [32]. In detail, the Decision Tree takes into account both the classification resulting from the fusion of CNNs' prediction and the 7-Point Check List score as computed by the CNNs during the feature extraction task. Moreover, two Decision Trees have been trained and tested corresponding to the MEL variable's different values resulting from the rule (OR versus SPLIT) applied at the previous step. The results are summarized in Table V. They highlight the dermoscopic analysis based on 7-Point Check List as performed by CNNs can improve both the prediction reliability and the diagnostic accuracy provided by the simple analysis of the digital image through the Deep Learning. In detail, the agreement degree between each model and the reference (expert dermatologists) becomes moderate (from weak) and specificity exceeds 90% whereas, the sensitivity still remains at satysfyng level (greater than the performance typically exhibited by unexeperienced physicians). Finally, the adoption of the most common CNN for the feature extraction through the split rule and LMT as classifier is suggested as prediction tool able to support the clinical practice.

V. Conclusion and Future Work

Artificial Intelligence in dermatology is not replacing specialists or placing decision-making in the hands of a nonexpert. The near future will follow what is already happening in radiology, with AI useful for triaging and improving workflow efficiency by prioritizing tasks, which is the current direction for the most significant research efforts. At this point, in very few cases, CNNs have proven they make physicians better at diagnosing skin cancer concerning real-world clinical data. The amount of training data strongly affects the level of accuracy in classifying benign and malignant images. To produce the best accuracy, an original approach is proposed that ensembles multiple CNNs models trained to implement a well-known diagnostic method typically adopted by a dermatologist during clinical practice. The original contribution is represented by the ensembling idea, that, at our knowledge, has not yet enough investigated by the past literature. Previous works have been focused on the proposal of the best CNN architecture to the clinical

TABLE V DIAGNOSTIC PERFORMANCE OF THE CNN-BASED CLASSIFICATION.

Classification Type	SE	SP	k	ACC
CNN Fusion (OR rule) + LMT	76.2	91.7	.67	87.8
CNN Fusion (SPLIT rule) + LMT	83.7	93.0	.76	90.7

domain (classification of melanoma and benignant nevi) by varying the different layers and training parameters as well as modifying large dataset (for example through data augmentation). No works has addressed the idea to achieve better classification performance by comparing the results of different CNNs when trained on the same database. The proposal instead tries to enrich the knowledge from the data following the typical metrological approach focusing on the concept of measurement uncertainty (the quantitative information fundamental for understanding the reliability of the measurement data). At the same way, ensembling the "uncertain" classification of different CNNs (when applied to the same measurand, i.e. the suspicious lesion) leads to more reliable diagnostic results. The aim is to properly improve the dermatological workflow at clinical institutions (hospitals and/or excellence centers) by providing a computer-aided triage (that scans which pigmented lesion might need prompt evaluation by a dermatologist) and supporting young professionals in classification tasks. Starting from the auspicious results of the present work, the future research will be addressed to the improvement of the feature extraction step. In detail, the image analysis through the CNNs may be limited to the boundingbox surrounding the lesion (i.e. resulting in the better exploiting of the digital image resolution) and suitable fusion of both traditional and innovative (Deep Semantic Segmentation) image processing techniques may be applied in order to complete the intended measurement and classification system for dermoscopic analysis of suspicious skin lesions.

REFERENCES

[1] Skin Cancer Facts & Statistics- What You Need to Know (World Health Organization (WHO)). Last accessed on 1 January 2021.

[2] https://melanoma.canceraustralia.gov.au/statistics. Last accessed on 17 January 2020.

[3] M. E. Vestergaard, P. Macaskill, P. E. Holt, and S. W. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," *British Journal of Dermatology*, 159(3), pp. 669 – 676, Sep. 2008.

[4] J. F. Aitken, J. Pfitzner, D. Battistutta, P. K. O'Rourke, A. C. Green and N. G.Martin, "Reliability of computer image analysis of pigmented skin lesions of Australian adolescents," *Cancer: Interdisciplinary International Journal of the American Cancer Society* 78.2 (1996): 252-257.

[5] F. Nachbar, W. Stolz et alii, "The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions", *J. Am. Acad. Dermatol.* Vol. 30 N. 4, pp. 551-559, 1994.

[6] X. Yuan, Z. Yang, G. Zouridakis, and N. Mullani, "SVM-based texture classification and application to early melanoma detection," in Proc. *IEEE 28th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBS)*, Aug. 2006, pp. 4775–4778.

[7] http://www.dermoscopy.org/consensus/2d.asp. Last accessed in 19 January 2020.

[8] G. Fabbrocini, G. Betta, G. Di Leo et alii, "Epiluminescence Image Processing for Melanocytic Skin Lesion Diagnosis Based on 7-Point Check-List: A Preliminary Discussion on Three Parameters", *Open Dermatology Journal*, pp. 110-115, 2010.

[9] A. GC. Pacheco, A. R. Ali and T. Trappenberg, "Skin cancer detection based on deep learning and entropy to detect outlier samples". arXiv preprint arXiv: 1909.04525, Sep 10. 2019.

[10] R. Refianti, A. B. Mutiara and R. Poetri Priyandini, "Classification of Melanoma Skin Cancer using Convolutional Neural Network," *Int. J. of Advanced Computer Science and Applications* 10.3 (2019).

[11] R. D. Seeja and A. Suresh, "Deep Learning Based Skin Lesion Segmentation and Classification of Melanoma Using Support Vector Machine (SVM)," *Asian Pacific Journal of Cancer Prevention: APJCP 20*.5 (2019): 1555.

[12] N.C. Codella, Q.B. Nguyen, S. Pankanti, D.A. Gutman, B. Helba, A.C. Halpern and J.R. Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development* 61.4/5 (2017): 5-1.

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.

[14] G. Betta, G. Di Leo, G. Fabbrocini, A. Paolillo, P. Sommella, "Dermoscopic image-analysis system: Estimation of atypical pigment network and atypical vascular pattern", *IEEE International Workshop on Medical Measurement and Applications MeMeA 2006*, pp. 63-67.

[15] G. Di Leo, A. Paolillo, P. Sommella, G. Fabbrocini, O. Rescigno, "A software tool for the diagnosis of melanomas automatic implementation of the 7-point check list method", *2010 IEEE International Instrumentation and Measurement Technology Conference 12MTC 2010*, pp. 886-891, 3–6 May 2010.

[16] G. Di Leo, G. Fabbrocini, A. Paolillo, P. Sommella, " A web-based application for dermoscopic measurements and learning", *2015 IEEE International Symposium on Medical Measurements and Applications, MeMeA 2015;* Torino; Italy; 7 May 2015 through 9 May 2015;

[17] Y. Nie Y., P. Sommella, M. O'Nils, C. Liguori, J. Lundgren, "Automatic detection of melanoma with yolo deep convolutional neural networks", *2019 7th E-Health and Bioengineering Conference, EHB 2019*. p. 1-4, 2019, Institute of Electrical and Electronics Engineers Inc., ISBN: 978-1-7281-2603-6.

[18] Y. Nie, L. De Santis, M. Carratù, M. O'Nils, P. Sommella, J. Lundgren, "Deep Melanoma classification with K-Fold Cross-Validation for Process optimization*", IEEE Medical Measurements and Applications, MeMeA 2020*, pp. 1-6, 2020.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, Sep 4. 2014.

[20] Y. LeCun, Y., Y. Bengio and G. Hinton, 2015, "Deep Learning", *Nature* 521, no. 7553 (2015): 436-444.

[21] T.J. Brinker, A. Hekler, J.S. Utikal et alii, "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review", *J. Medical Internet Res 2018*, Vol. 20 N. 10, pp. 1-8.

[22] A. Esteva, B. Kuprel, R.A. Novoa et al.," Dermatologist-level classification of skin cancer with deep neural networks", *Nature* 2017 Dec 02;542(7639), pp-115-118.

[23] H. Haenssle, C. Fink, R. Schneiderbauer et al., " Reader Study Level-I and Level-II Groups. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists", *Ann Oncol 2018*, Vol. 29 N. 8), pp.1836-1842, 2018.

[24] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, 2019.

[25] S. Alzahrani, W. Al-Nuaimy, B. Al-Bander, "Seven-Point Checklist with Convolutional neural Networks for Melanoma Diagnosis", 2019 8th European Workshop on Visual Information Processing (EUVIP), pp. 211-216, 2019."

[26] N. Landwehr, M. Hall, E. Frank, "Logistic Model Trees", *Machine Learning*, N 59, pp. 161, 2005

[27] P. Carli et al., "Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology," Br. J. Dermatol., vol. 148, no. 5, pp. 981–984, 2003.

[28] P. A. Lio and P. Nghiem, "Interactive atlas of dermoscopy," J. Am. Acad. Dermatol., vol. 50, no. 5, pp. 807–808, 2004.

[29] S. Abhinav, J. Dheeba, "Convolutional Neural Networks for Classifying Melanoma Images", Vello Institute of Thechnology.

[30] S.H. Kassani, P.H. Kassani, "A comparative study of deep learning architectures on melanoma detction", Tissue and Cell, Vol 58, 2019.

[31] G. Fabbrocini, S. Cacciapuoti, G. De Fata Salvatores, et alii, "A new "hub and spoke" teledermoscopy system involving general practitioners and dermatologists for the early detection of cutaneous melanoma: a pilot study", AIMS Bioengineering, 2020, 7(4): 306-320. doi:10.3934/bioeng.2020025.

[32] https://machinelearningmastery.com/k-fold-cross-validation/. Last accessed in 19 January 2020.