An ensemble-based convolutional neural network model powered by a genetic algorithm for melanoma diagnosis

Eduardo Pérez1,2 · Sebastián Ventura1,2,3

Received: 31 March 2021 / Accepted: 27 October 2021 / Published online: 19 November 2021 © The Author(s) 2021

Abstract
Melanoma is one of the main causes of cancer-related deaths. The development of new computational methods as an important tool for assisting doctors can lead to early diagnosis and effectively reduce mortality. In this work, we propose a convolutional neural network architecture for melanoma diagnosis inspired by ensemble learning and genetic algorithms. The architecture is designed by a genetic algorithm that finds optimal members of the ensemble. Additionally, the abstract features of all models are merged and, as a result, additional prediction capabilities are obtained. The diagnosis is achieved by combining all individual predictions. In this manner, the training process is implicitly regularized, showing better convergence, mitigating the overfitting of the model, and improving the generalization performance. The aim is to find the models that best contribute to the ensemble. The proposed approach also leverages data augmentation, transfer learning, and a segmentation algorithm. The segmentation can be performed without training and with a central processing unit, thus avoiding a significant amount of computational power, while maintaining its competitive performance. To evaluate the proposal, an extensive experimental study was conducted on sixteen skin image datasets, where state-of-the-art models were significantly outperformed. This study corroborated that genetic algorithms can be employed to effectively find suitable architectures for the diagnosis of melanoma, achieving in overall 11% and 13% better prediction performances compared to the closest model in dermoscopic and non-dermoscopic images, respectively. Finally, the proposal was implemented in a web application in order to assist dermatologists and it can be consulted at http://skinensemble.com.

Keywords Convolutional neural networks · Melanoma diagnosis · Ensemble learning · Genetic algorithm · Lesion segmentation

1 Introduction
Melanoma is the most serious form of skin cancer that begins in cells known as melanocytes. Melanoma has an increasing incidence, where just in Europe were estimated 144,200 cases and 20,000 deaths in 2018 [1], whereas in USA, 106,110 new cases of invasive melanoma will be diagnosed (62,260 in men and 43,850 in women) and 7180 deaths are expected in 2021 (4600 men and 2580 women) [2]. The lesion is first diagnosed through an initial clinical screening, and then potentially through a dermoscopic analysis, biopsy and histopathological examination [3]. Despite the expertise of dermatologists, early diagnosis of melanoma remains a challenging task since it is presented in many different shapes, sizes and colors even between samples in the same category [4]. Providing a comprehensive set of tools is necessary for simplifying diagnosis and assisting dermatologists in their decision-making processes [3].

Several automated computer image analysis strategies have been used as tools for medical practitioners to provide accurate lesion diagnostics, including descriptor-based methods [5, 6] and convolutional neural networks (CNNs) [3, 7, 8]. Descriptor-based methods require the previous
CNN models have shown to be effective in solving several complex problems [12, 13]. However, they still present several issues which hamper their accuracy in diagnosing skin conditions. CNN models can learn from a wide variety of nonlinear data points. As such, they are prone to overfitting on datasets with small numbers of samples per category, thus attaining a poor generalization capacity. It is noteworthy that so far, most of the existing public skin datasets only encompass a few hundreds or thousands of images. This can limit the learning capacity of CNN models. On the other hand, CNN models are sensitive to some characteristics in data, such as large inter-class similarities and intra-class variances, variations in viewpoints, changes in lighting conditions, occlusions, and background clutter [14]. These days, the majority of skin datasets are made up of dermoscopic images reviewed by expert dermatologists. However, bear in mind that there is an increased tendency to collect images taken by common digital cameras [15]. The above can reduce invasive treatments and their associated expenses in addition to augmenting the development of modern tools for cheaper and better melanoma diagnoses. Finally, CNN models are approximately invariant with regard to small translations to the input, but they are not rotation, color or lighting-invariant [16, 17]. Invariance is an important concept in the area of image recognition. It means that if you take the input and transform it, the representation you get is the same as the representation of the original. Due to the vast variability in the morphology of moles, this is an important issue to resolve in order to attain a more effective melanoma diagnosis, thus allowing a model to detect rotations or changes to proportion and adapt itself in a way that the learned representation is the same.

Several techniques can be applied to overcome some of these issues, but the most proven include data augmentation [18], transfer learning [3], ensemble learning [19] and, more recently, generative adversarial networks [20] and multi-task learning [21]. However, in most cases, researchers rely on their expertise to select which techniques to apply, and there is no specific pattern to follow that will definitively produce a model with a high level of reliability. Furthermore, most research does not follow a standard experimental study and only includes a limited number of datasets.

As a consequence of the above, in this work, a novel approach for diagnosing melanoma via the use of images is proposed. First, the proposal is inspired by ensemble learning and is built via a genetic algorithm. The genetic algorithm is designed to find the optimal members of such ensembles while considering the entire training phase of the possible members. In addition, the abstract features of all models in the ensemble are merged, resulting in an additional prediction. Next, these individual predictions are combined and a final diagnosis is made. This approach can be seen as a double ensemble, in which all predictive components are double related. The aim is to find the models that best contribute to the ensemble, rather than the individual level. As a result, the state of each CNN model that best trains, generalizes and suits with the other CNN models is selected [22]. In this manner, the training process is implicitly regularized, which has shown better convergence, mitigates the overfitting of the model and improves the generalization performance [19, 23]. To the best of our knowledge, this is the first attempt at intelligently constructing ensembles of CNN models by following a genetic algorithm to better solve the challenge of diagnosing melanoma through the use of images. Second, a novel lesion segmentation method, which is capable of efficiently obtaining reliable segmentation masks without prior information through the use of just a CPU, is applied. Bear in mind that state-of-the-art biomedical segmentation methods commonly require a prior training stage and the use of GPUs. In addition, common techniques such as transfer learning and data augmentation have been applied to further improve performance. To evaluate the suitability of this proposal, an extensive experimental study was conducted on sixteen melanoma-image datasets, enabling a better analysis of the effectiveness of the model. The results showed that the proposed approach achieved promising results and was competitive compared to six state-of-the-art CNN models which have previously been used for diagnosing melanoma [3, 10, 24–26]. These works were summarized in Pérez et al. [8], and the most relevant are explained in the next section. To perform a fair comparison, the above architectures were assessed in Sect. 5.5 by using the same large number of datasets and a quantification of the results was included. Finally, a web application was developed in order to support dermatologists in decision making.

The rest of this work is organized as follows: Sect. 2 briefly presents the state of the art in solving melanoma diagnosis problems mainly by using CNN models; Sect. 3 describes the proposed ensemble convolutional architecture; Sect. 4 describes the genetic algorithm application;
the analysis and discussion of the results are portrayed in Sect. 5; and finally, Sect. 6 outlines conclusions and future works.

2 Related works

The popularity of CNN models increased when AlexNet [27] won the well-known ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012, reducing the top-5 labels error rate from 26.1% to 15.3%. Since then, CNN models are widely applied in image classification tasks [28–30]. However, most authors apply other sophisticated techniques over CNN models to achieve a better performance in melanoma diagnosis, such as segmentation [31], data augmentation [18], transfer learning techniques [3], and CNN-based ensembles [19].

Data augmentation is employed to add new data to the input space, which helps reducing overfitting [32] and obtaining transformation-invariant models [17]. This technique is usually performed by means of random transformations [33]. Also, most of the datasets available for melanoma diagnosis lack of balance between the categories, so data augmentation can help to tackle the imbalance issue [34]. For example, Esteva et al. [3] showed the suitability of CNN models as a powerful tool for melanoma diagnosis. The authors augmented the images by a factor of 720× using basic transformations, such as rotation, flips and crops. Also, they compared the performance of one CNN model to 21 board-certified dermatologists on biopsy-proven clinical images. The results showed that the CNN model achieved a performance on par with experts. Pérez et al. [8] showed that 12 CNN models achieve better performance when applying data augmentation. For example, Xception [35] increased its average performance by 76%.

Transfer learning has been successfully applied in image classification tasks [3, 30, 36, 37]. For example, Shin et al. [30] applied it specifically in thoraco abdominal lymph node detection and interstitial lung disease classification, in both cases the weights obtained after training with ImageNet were beneficial. Esteva et al. [3] used Google’s InceptionV3 [38] architecture pretrained on ImageNet. The authors removed the final classification layer and then they re-trained with 129,450 skin lesions images. Pérez et al. [8] studied the impact of applying transfer learning in skin images, where MobileNet [39] increased its average performance by 42%.

On the other hand, ensemble learning [40] has also shown to be effective in solving complex problems [41, 42]. The combination of several classifiers built from different hypothesis spaces can reach better results than a single classifier [43]. For example, Mahbod et al. [19] proposed an ensemble-based approach where two different CNN architectures were trained with skin lesion images. The results showed to be competitive compared to the state-of-the-art methods for melanoma diagnosis. Harangi et al. [23] proposed an ensemble composed by the well-known CNN architectures AlexNet [27], VGG and InceptionV1 [44]. The ensemble was assessed on the International Symposium on Biomedical Imaging (ISBI) 2017 challenge and obtained very competitive results. However, the ensemble models are usually built trusting in prior knowledge.

In order to achieve better performance, not only new strategies to train the models have been developed, but also efforts have been made to improve the input data. Skin lesion segmentation plays an important role in melanoma diagnosis. It isolates the region of interest and significantly improves the performance of the model. This is a highly complex task, and it is important because some areas not related to the lesion can lead CNN models to misclassify samples. However, some authors decided not to use some of these preprocessing techniques. Mahbod et al. [19] ignored complex preprocessing steps, as well segmentation methods, but applied basic data augmentation techniques to prevent overfitting. Nevertheless, if irrelevant information is removed from the image, the models could be able to achieve better performance. Since 2016, The International Skin Imaging Collaboration1 (ISIC) project annually organizes a challenge in which more than 180 teams have already participated. From 2016 to 2018, there was a special task about lesion segmentation. In ISIC-2016, considering the top performances, the use of segmentation obtained 8% better sensitivity compared to its non-use. Several segmentation methods can be found. For example, Ronneberger et al. [45] designed a CNN model (U-Net) for biomedical image segmentation. The model relies on data augmentation to use the available labeled images more efficiently. The U-Net architecture achieved high performance on different biomedical segmentation applications, such as neuronal structures in electron microscopic recordings and skin lesion segmentation [46, 47]. However, U-Net requires a costly training process using GPU and more important, a prior knowledge from data already segmented by expert dermatologists. Furthermore, Alom et al. [48] proposed a recurrent residual U-Net model (R2U-Net). The model was tested on blood vessel segmentation in retinal images, skin cancer segmentation, and lung lesion segmentation. The results showed better performance compared to the U-Net model. Huang et al. [31] proposed a new segmentation method based on end-to-end object scale-oriented fully convolutional networks. The authors achieved 92.5% of sensitivity and outperformed all

---

1 https://www.isic-archive.com.
CNN models in their study, which was 1.4% better compared to the winner in ISIC-2016.

Considering the above, it would be interesting to design a deep learning model that combines features from different approaches such as segmentation, data augmentation, transfer learning and ensemble learning. We hypothesized that evolutionary optimization methods could be an effective approach to find the optimal combination of CNN models [49, 50], considering all states from all the models. Evolutionary methods have proven to be useful in solving many complex problems, finding and optimizing architectures of neural networks [51], and mining imbalanced data [52]. Regarding segmentation methods, in this work it is applied an extension of the Chan-Vese segmentation algorithm [53] and it is evaluated using specialized datasets. To augment data, it is important to perform a data augmentation both on training and test phases [18], which can increase the performance significantly.

3 Ensemble-based convolutional architecture

First of all, the source datasets were preprocessed following an extension of the Chan-Vese segmentation algorithm, as shown in Fig. 1. This algorithm is designed to segment objects without clearly defined boundaries and is based on techniques of curve evolution, Mumford-Shah [54] functional for segmentation and level sets. Chan-Vese has been previously applied in skin image segmentation [55], being an effective starting point to accurately segment skin images. First, Chan-Vese is applied on each input image and as a result, a mask with positive and negative values is obtained. After that, the positive pixels within 40% of the center are selected (cluster of pixels $P$). After using several recognized skin lesion detection applications such as SkinVision2, we realized that most of them demand that images must be centered in the lesion in order to perform an accurate diagnosis [56]. In addition, after reviewing a large number of skin image datasets, it was noticed that most images are centered in the lesion, which is an advantage when applying segmentation. Third, all positive clusters ($Q$) that intersect $P$ are merged, obtaining a new segmentation mask $M$, $M = Q_1 \cup Q_2 \cup \ldots \cup Q_n, P \cap Q_i \neq \emptyset$. Finally, the segmented image is obtained after applying the mask $M$ on the original image. In Sect. 5, the proposed segmentation method is compared to other state-of-the-art biomedical segmentation methods to corroborate its effectiveness.

Next, the segmented images are used as input to the architecture described below.

Let us say $\Phi$ is a model with $m$ independent CNN models, which learn the representations from the same feature space, and an extra prediction block (stacking of dense layers) that yields an extra prediction. The prediction block is obtained by first concatenating all the representations learned by the individual CNN models. Figure 2 shows a specific example of the proposal. An image $i$ is passed as input to the $j$th CNN model of $\Phi$, and a chromosome indicates which tuples of CNN models and their weights determine the ensemble’s architecture. Each CNN outputs the learned representation ($r_i^j$) and a partial prediction ($g_i^j$) for the label of this sample by considering the weights required in the chromosome, e.g., the weights from DenseNet trained until epoch 33 ($g_1 = 33$). Thereafter, the representations $r_i^j$ learned by each CNN model are flattened, then concatenated and then passed to the prediction block of the model. A late fusion approach is used to concatenate all the representations learned by the CNN models (Fig. 2c), which has proven to obtain better performance compared to merely combining the individual predictions [57]. Nevertheless, in Sect. 5.5 the prediction block is assessed versus its non-use. Thereafter, we freeze the weights from the individual CNN models and only the prediction block is trained during 20 epochs, providing an additional prediction ($g_m^{m+1}$) for the label of the sample. In this work, the prediction block is composed of Dense(512 RelUs) $\rightarrow$ Dense(256 RelUs) $\rightarrow$ Dense(128 RelUs) $\rightarrow$ Dense(1 unit - sigmoid). Then, $\Phi$ predicts the label of an image by using a soft-voting procedure based on the individual predictions. Regarding efficiency, each CNN model was trained just once and its weights along every epoch of interest were stored in a hard drive.

Once the prediction for a given training sample is computed, the losses produced by each CNN model and the prediction block of $\Phi$ are calculated. The loss produced by the $j$th CNN model of $\Phi$ on the $i$th training image (denoted as $L^j(i)$) is computed by means of a binary cross-entropy. The goal is to iteratively minimize the prediction errors on the training samples. Mini-batch Gradient Descent (MGD) [58] can be applied to solve this optimization problem, since the first derivative of the loss function is well-defined. This algorithm consists in randomly splitting the training set in small batches in order to calculate the model error. The above method has several advantages, such as the model update frequency is higher than batch gradient descent which allows for a more robust convergence, avoiding local minima; batch-based updates provide a computationally more efficient process than stochastic gradient descent and the split in small batches allows the efficiency of not having all training data in memory.
4 Ensemble model via genetic algorithm

In this work, a genetic algorithm (GA) was designed to find the optimal members of an ensemble model. Next, the different components and steps of the proposed method are explained in detail.

4.1 Individuals and chromosome codification

Let say that the population of the GA has \( q \) individuals \( \{I_1, I_2, \ldots, I_q\} \), where the \( j \)-th individual \( (I_j) \) has a chromosome encoded as a list of integer values (from 0 to 150), as shown in Fig. 2a. A chromosome is composed of \( K \) genes, where \( g^j_k \) represents the \( k \)-th gene of the individual \( I_j \). Each gene index \( k \) is independent of the others and is related to a specific CNN model. Its value means the epoch of the CNN model to be considered as part of the ensemble; if \( g^j_k > 0 \), the CNN model is selected; otherwise, it is not selected. Consequently, using this encoding, each individual of the population represents a full solution to the problem.

4.2 Fitness function

The fitness function used to evaluate the individual \( I_j \) can be calculated as

\[
f_j = \frac{1}{n \times p} \sum_{i=1}^{n} \left[ L^i_{m+1}(i) + \sum_{k=1}^{m} L^i_k(i) \right] \text{ if } g^j_k > 0,
\]

where \( L^i_k(i) \) calculates the loss values for the \( i \)-th image in the CNN model with index \( k \), and \( L^i_{m+1}(i) \) means the loss value of the prediction block; \( p \) means the number of times that \( g^j_k > 0 \) is fulfilled; \( m \) represents the total number of CNN models encoded in the chromosome; and \( n \) is the total number of images. In summary, for each individual, the fitness function calculates the average loss value between the chosen CNN models and the prediction block; lower average means a more desirable individual.

4.3 Creation of the initial population

Maintaining a diverse population, especially in the early iterations of the algorithm, is crucial to ensure a good exploration of the search space. The training epochs \( e \) and the CNN models \( m \) determine the total number of possible combinations in the search space. To guarantee the diversity, the chromosome of each individual \( I_j \) of the population is randomly created, but repeated individuals are not allowed. In this manner, it is possible to avoid the early convergence of the method to local minima. In this work, we have considered an individual as repeated when all genes \( g^A_k \) in the chromosome are identical. For example, let us say \( g^A_k \) and \( g^B_k \) are genes from chromosomes \( A \) and \( B \),
respectively; \( k \) means the index related to a CNN model (e.g., DenseNet, InceptionV3 and Xception) and \( m \) means the maximum number of CNN models in each ensemble. \( A \) and \( B \) are identical if for every \( k = 1, 2, \ldots, m \), \( g^A_1 = g^B_1 \), \( g^A_2 = g^B_2 \), \( g^A_k = g^B_k \), \ldots, \( g^A_m = g^B_m \). Having said the above, the individuals are repeated if they use a different CNN in any of their genes, i.e., two models are considered different if they have different architectures or the same architecture but with different weights.

### 4.4 Parent selection

The parents are selected by a tournament selection procedure to create the intermediate population [59]. A tournament size equal 2 was used in this work; the smaller the tournament size, the lower the selection pressure and the search space can be widely considered. To achieve this, two individuals are randomly selected. Then, the individuals are compared and the best individual is selected with replacement, i.e., this individual could be selected in further rounds. This process is repeated until the number of individuals is completed. As the generations increase, the algorithm can focus more on promising regions of the search space.

### 4.5 Genetic operators

Figure 3 shows the genetic operators applied. First, a custom Flat crossover [60, 61] was performed with a crossover rate \( p_c \), where its general mechanism is adapted for an integer representation scheme. Flat crossover is applied in each pair of genes that are located in the same locus and represents the same CNN model. As a result, a random integer value is chosen from the interval \([g^A_1; g^B_2]\). Once the new offspring is generated, an one-point mutator operator is applied with a probability \( p_m \). The mutation operator randomly selects a gene, then, the value of epoch is changed by a new value selected from a range of possible epochs, e.g., [0; 150]. Finally, the offspring was generated.

It is noteworthy that valid individuals are always obtained after performing these operators.

### 4.6 Population update

In this work, a generational elitist algorithm [62] was used to update the population passed from one generation to the next one. As a result, the best individual in the last generation is the best individual of the evolution. To achieve this, the population in each generation keeps all new individuals, as long as the best parent is not better than all the children. In such cases, the best parent replaces the worst child. The worst child is determined by sorting the individuals of the new population according to the fitness value. After replacing the worst child, the new population replaces the previous one. At the end of the algorithm, the best individual in the last generation will be the best ensemble. In this work, the number of individuals generated in each generation is 100 and the population size is kept constant in order to alleviate the computational cost of CNN. However, these values could be tuned depending on the context, including the possibility of increasing/decreasing them. Regarding efficiency, a map with each explored individual and its corresponding fitness value is cached.
4.7 Integration of the genetic algorithm in the training phase

Figure 4 shows how the training phase was conducted. First, the source images are preprocessed by using the method explained above. Second, given a set of CNN models the initial population of the GA is created. For example, if we have an individual with \( g_1^i = 50 \) and another one is \( g_1^{i+1} = 51 \) \((z > 0)\), the last one is obtained after training \( g_1^j \) for one more epoch. In general, each CNN model was trained just one time when needed and its weights along every demanded epoch were stored in a hard drive. In the worst case, we train each CNN model 150 epochs once. Third, each ensemble represented by a chromosome is trained for \( n \) epochs, bearing in mind that only the prediction block is updated. Then, the fitness of each individual of the current population is calculated by considering the loss obtained in the training images by the members of the ensemble. Four, the parents are selected, then, a crossover and a mutation are performed, and after that a new population is created. Five, the population of the GA is updated. Steps from three to five are repeated until one of the stopping criteria is satisfied. In this work, we have applied the most frequently used stopping criterion which is a specified maximum number of generations. In addition, we stopped the search when an individual had achieved the top performance. Finally, the ensemble model \( \Phi \) is obtained.

5 Experimental study

This section describes the experimental study carried out in this work. First, the datasets and the experimental protocol are portrayed, and then, the experimental results and a discussion of them are presented.

5.1 Datasets

Table 1 shows a summary of the benchmark datasets. UDA [66], MSK [67], HAM10000 [65] and BCN20000 [63] datasets are included in the ISIC repository. The images are composed strictly of melanocytic lesions that are biopsy-proven and annotated as malignant or benign. Also, images from BCN20000 would be considered hard-to-diagnose and had to be excised and histopathologically diagnosed. PH2\(^4\) [69] comprises dermoscopic images, clinical diagnosis and the identification of several dermoscopic structures. The Dermofit Image Library\(^5\) [64] gathers 1,300 focal high-quality skin lesion images under standardised conditions. Each image has a diagnosis based on expert opinion and like PH2, it includes a binary segmentation mask that denotes the lesion area. MED-NODE\(^6\) [68] collects 170 non-dermoscopic images from common digital cameras; this type of image is very important to prove the models with data from affordable devices. SDC-198\(^7\) [70] contains 6,584 real-world images from 198 categories to encourage further research and its application in real-life scenarios. Finally, DERM7PT8 [21] is a benchmark dataset composed of clinical and dermoscopic images, allowing to assess how different it is to use dermoscopic images versus images taken with digital cameras.

Only the images labeled as melanoma and nevus were considered, being in total 36,703 images. Most datasets present a high imbalance ratio (ImbR), up to ten times in the case of MSK-3, commonly hampering the learning process. The intra-class (IntraC) and inter-class (InterC) metrics show the average distances between images belonging to different classes, as well as between images belonging to the same class. Both metrics were computed using the Euclidean distance; each image \( i \) was represented as a vector. Then, the ratio (DistR) between these metrics showed that both distances are similar, which commonly indicates a high degree of overlapping between classes. Finally, the silhouette score (Silho) [71] was calculated, representing how similar an image is to its own cluster compared to other clusters. The results indicated that images were not well matched to their own cluster, and even samples belonging to different clusters are close in the feature space.

5.2 Experimental settings

Firstly, the proposed segmentation algorithm was applied over PH2 and DERM-LIB, where both datasets were manually segmented by expert dermatologists and they are commonly used as benchmarks due its quality [64, 69]. The aim was to find how much the method is close to expert segmentation. Also, we compared the segmentation algorithm to U-Net and R2U-Net, which are CNN architectures designed for biomedical image segmentation. Bear in mind that the above CNN architectures need prior training. In order to obtain the segmentation mask of an image \( X \), the CNN architectures were trained during 150 epochs with 10% of images as validation set and the rest as training set. The best model obtained during validation was applied on the image \( X \) and a segmentation mask was obtained.

---

\(^4\) https://www.fc.up.pt/addi/ph2%20database.html.
\(^5\) https://bit.ly/3gi6rKR.
\(^6\) http://www.cs.rug.nl/imaging/databases/melanoma_naevis/.
\(^7\) https://bit.ly/2TdzWQ6.
\(^8\) http://derm.cs.sfu.ca.
above procedure was applied in both PH2 and DERM-LIB datasets.

The second phase aimed at analyzing the GA’s hyperparameters on the model’s performance. The model was dubbed as Genetic Algorithm Programming-based Ensemble CNN model (GAPE-CNN). It should be noted that the main aim of this work was to perform a preliminary study to assess the effectiveness that can be attained by using the proposed architecture in melanoma diagnosis. Consequently, only the hyperparameter values listed in Table 2 values were considered, because including more settings requires high computational resources during training CNN models. Three different number of generations were evaluated, where larger values can lead to a larger number of possible ensembles, but increasing the training cost. Three crossover probabilities were tested, where higher values increase the probability that the genetic information of parents can be combined to generate new individuals. Three mutation probabilities were also evaluated, where higher values increase the probability that the genetic information of parents can be combined to generate new individuals. Three mutation probabilities were also evaluated, where higher values allow to escape from local minima and add more exploration of the search space.

In the third phase, the proposal was compared to the following state-of-the-art CNN models that have previously been used in melanoma diagnosis: InceptionV3 [3], DenseNet [25], VGG16 [10], MobileNet [24], Xception [26] and NASNetMobile [72, 73]. Table 2 shows the configuration used to train all the models: the learning rate (α) was equal to 0.01 and it was reduced by a factor of 0.2 if an improvement in predictive performance was not observed during 10 epochs; the weights of the networks were initialized using Xavier method [74] in those cases where transfer learning was not present, e.g., the prediction block and baseline CNN models; a batch of size 8 was used due the medium size of the used datasets and the models were trained along 150 epochs. Mini-batch gradient descent was used for training the models, which is one of the most used optimizers for training CNNs. Despite its simplicity, it performs well across a variety of applications [75] and has been successfully applied for training networks in melanoma diagnosis [7, 18, 76]. In this work, a tuning process was not carried out and so the results could not be conferred to an over-adjustment. The datasets utilized in this work correspond to binary classification problems, so the cost function used for training the models was defined as the average of the binary cross-entropy along all training samples.

Data augmentation technique was mainly applied to tackle the imbalance problem in melanoma diagnosis by applying and combining random rotation-based, flip-based and crop-based transformations over the original images. Bear in mind that color-based transformations were not considered in order not to alter the color space, which is important for the diagnosis of melanoma [77]. As a result, the changes performed during data augmentation do not change the labels of the samples. The data augmentation process was assessed in both training and test data, which can increase the performance significantly [18]. After splitting a dataset into training and test sets, training data

| Dataset | Source | Img | ImbR | IntraC | InterC | DistR | Silho |
|---------|--------|-----|------|--------|--------|-------|------|
| BCN20000 | [63] | 17,393 | 2.848 | 9014 | 10,107 | 0.892 | 0.153 |
| DERM-LIB | [64] | 407 | 4.355 | 7171 | 9163 | 0.783 | 0.270 |
| DERM7PT-C | [21] | 827 | 2.282 | 15,442 | 16,318 | 0.946 | 0.086 |
| DERM7PT-D | [21] | 827 | 2.282 | 15,971 | 16,866 | 0.947 | 0.087 |
| HAM10000 | [65] | 7818 | 6.024 | 8705 | 9770 | 0.891 | 0.213 |
| ISBI2016 | [66] | 1273 | 4.092 | 10,553 | 10,992 | 0.960 | 0.101 |
| ISBI2017 | [67] | 2745 | 4.259 | 9280 | 9674 | 0.959 | 0.089 |
| MED-NODE | [68] | 170 | 1.429 | 9029 | 9513 | 0.949 | 0.068 |
| MSK-1 | [67] | 1088 | 2.615 | 11,753 | 14,068 | 0.835 | 0.173 |
| MSK-2 | [67] | 1522 | 3.299 | 9288 | 9418 | 0.986 | 0.062 |
| MSK-3 | [67] | 225 | 10.842 | 8075 | 8074 | 1.000 | 0.112 |
| MSK-4 | [67] | 943 | 3.366 | 6930 | 7162 | 0.968 | 0.065 |
| PH2 | [69] | 200 | 4.000 | 12,688 | 14,928 | 0.850 | 0.210 |
| SDC-198 | [70] | 648 | 4.735 | 14,054 | 14,840 | 0.947 | 0.116 |
| UDA-1 | [66] | 557 | 2.503 | 11,730 | 12,243 | 0.958 | 0.083 |
| UDA-2 | [66] | 60 | 1.609 | 11,297 | 11,601 | 0.974 | 0.020 |
were balanced by creating new images until the number of melanoma images was equal to the normal ones, and the generated training images were considered as independent from the original ones. On the other hand, test data were expanded by randomly augmenting each test image at least ten times, but the generated images remained related to the original ones. Consequently, given an original test image X, the classes’ probabilities for X and its related set of images SX were averaged to yield the final prediction; so any CNN model performs like an ensemble one, where the final probabilities for a test image was computed using a soft-voting strategy.

### 5.3 Evaluation process

In order to evaluate quantitatively the segmentation method, several performance metrics are considered, including accuracy (ACC), F1-score, Dice coefficient (DC), Jaccard similarity (JS), and the U-Net score (UNS), which is the average between DC and JS. Accuracy and F1-score are calculated using Eqs. 2, 3 and Dice coefficient and Jaccard similarity are calculated using Eqs. 4 and 5.

\[
ACC = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)},
\]

\[
DC = \frac{2|GT \cap SR|}{|GT| + |SR|},
\]

\[
JS = \frac{|GT \cap SR|}{|GT \cup SR|},
\]

\[
UN = \frac{DC + JS}{2},
\]

where TP, FP, TN and FN represent well-selected pixels, mis-selected pixels, well-discarded pixels and mis-discarded pixels, respectively; GT and SR mean the ground truth pixels and the segmentation result, respectively. The comparison between the segmentation methods is performed using UNS, summarizing DC and JS. The above metrics have been used before in the segmentation task of the ISIC-2018 contest.

Regarding the evaluation process of the CNN models, a 3-times 10-fold cross validation process was performed on the datasets, and the results were averaged across all fold executions. In each fold, Matthews Correlation Coefficient (MCC) was used to measure the predictive performance of the models. MCC is widely used in Bioinformatics as a performance metric [78], and it is specially designed to analyze the predictive performance on unbalanced data. MCC is computed as:

\[
MCC = \frac{t_p \times t_n - f_p \times f_n}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_n)(t_n + f_p)}}
\]

where \(t_p\), \(t_n\), \(f_p\), and \(f_n\) are the number of true positives, true negative, false positives, and false negatives, respectively. MCC value is always in the range \([-1, 1]\], where 1 represents a perfect prediction, 0 indicates a performance similar to a random prediction, and -1 an inverse prediction.

Finally, non-parametric statistical tests were used to detect whether there was any significant difference in predictive performance. Friedman’s test [79] was conducted in cases where a multiple comparison was carried out, Hommel’s post hoc test [80] was employed to perform a multiple comparison with a control method, Shaffer post hoc test [81] was employed to perform pairwise comparisons, and finally Wilcoxon Signed-Rank test [82] was performed in those cases where only two individual methods were compared. All hypothesis testing was conducted at 95% confidence.

### 5.4 Software and hardware

The experimental study (training and test) was executed with Ubuntu 18.04, four GPUs NVIDIA GeForce RTX 2080-Ti with 11 GB DDR6 each one and four GPUs NVIDIA GeForce RTX 1080-Ti with 11 GB DDR5X each one. All the experiments were implemented in Python v3.6, and the CNN models were developed by using Keras.
5.5 Results and discussion

In this section, the main results are presented. Additional material for supporting this work can be found at the available web page.\footnote{http://www.uco.es/kdis/ensemble-based-cnn-melanoma/}

5.5.1 Segmentation methods

Table 4 shows the segmentation performance obtained from a small sample (594 images in total). Both datasets contain high-quality images and manual segmentation performed by expert dermatologists. Each image was evaluated applying UNS and other metrics. The proposed segmentation method obtained the best average performance in both datasets with 81% and 80% UNS in PH2 and DERM-LIB, respectively. In addition, Fig. 5 shows Friedman’s test ranking, where the proposed method outperformed both U-Net and R2U-Net. Then, Friedman’s test rejected the null hypothesis with a p-value < 2.2E-16 in DERM-LIB dataset. The proposal was ranked first, and afterward, the Shaffer’s post hoc test was conducted, where the proposal achieved significantly better performance compared to U-Net and R2U-Net. In addition, U-Net significantly outperformed R2U-Net. On the other hand, in PH2 no significant differences were encountered between the three methods, the Friedman’s test did not reject the null hypothesis with a p-value equal to 6.476E-1. (The test was conducted with two degrees of freedom, and the Friedman’s statistic was equal to 86.898E-2.) However, the proposal was ranked first and it attained 111% and 74% less variance compared to U-Net and R2U-Net, respectively. In this way, the CNN models can focus on relevant pixels, so easing the learning of better abstract and discriminative features for melanoma diagnosis. Also, the proposal did not require prior training, which is a clear advantage compared to those based on CNN models. In addition, the proposed segmentation method can be used with only a CPU, avoiding a significant amount of computational power for training like those using GPU. Finally, R2U-Net obtained a better performance compared to U-Net in PH2, but the opposite occurred in DERM-LIB. A larger number of datasets are needed to validate the differences between both methods regarding skin lesion diagnosis.

In addition, we corroborated that the CNN models trained with segmented data were able to achieve better performance. All models achieved the best performance using segmented data, and on average the best ones were MobileNet, DenseNet, and InceptionV3, in that order. Despite MobileNet was designed with efficiency in mind, it managed to overcome more complex CNN models. Nevertheless, DenseNet and InceptionV3 surpassed MobileNet in BCN20000, which is the largest and one of the more complex datasets. Skin image datasets are commonly small, which is where MobileNet usually achieved its best performance\footnote{https://aws.amazon.com/}. Although these results could be also caused by the use of default hyperparameters, all CNN models shared the same condition in order to avoid advantages between them. On the other hand, the number of epochs could be another cause. This is traditionally large.

### Table 3 Summary of the computational resources in deployment time; “D” and “L” mean using default tensorflow-cpu and tensorflow-lite; HDD includes both architecture and weights.

| CNN            | HDD (MB) D | RAM (MB) D | Time (ms) D | HDD (MB) L | RAM (MB) L | Time (ms) L |
|----------------|------------|------------|-------------|------------|------------|-------------|
| DenseNet201     | 141        | 69         | 1434        | 99         | 255        | 154         |
| InceptionV3     | 167        | 83         | 974         | 109        | 92         | 111         |
| MobileNet       | 25         | 12         | 337         | 40         | 70         | 47          |
| NASNetMobile    | 36         | 16         | 1126        | 40         | 156        | 85          |
| Xception        | 160        | 79         | 778         | 118        | 146        | 200         |
| VGG16           | 538        | 537        | 1229        | 104        | 161        | 398         |
| Total           | 1,067      | 796        | 5878        | 510        | 880        | 995         |

The best value for each column is highlighted in bold typeface, e.g., MobileNet is the one that consumes less RAM, less hard drive space, and needs less inference time (the aim is to minimize all metrics).
when training CNN models, often hundreds or thousands, allowing the learning algorithm to run until the error from the model has been sufficiently minimized. It is common to find examples in the literature and in tutorials where the number of epochs is set to 500, 1000, and larger. However, in this work we used 150 epochs, mainly because of the number of available images, and even then all models and the proposal achieved competitive performances.

UDA-2 was the most challenging dataset. On average, the models achieved only 44% MCC; UDA-2 has the lowest Silhouette value (0.020), indicating a high overlapping level between classes and increasing the difficulty. The overall best performance was achieved in DERM-LIB, PH2 and HAM10000 datasets with a 90%, 85% and 79% MCC, respectively. The above datasets have the three highest Silhouette values, meaning that they have a low overlapping level between images, which makes easier the task. Table 5 and Fig. 6 summarize the results obtained after training the six CNN models with the segmentation method. The models in Fig. 6 are ordered from left to right according to the ranking computed by Friedman’s test, where the proposed segmentation method achieved the best performance in all datasets, followed by TDA. Friedman’s test rejected the null hypothesis on all CNN models. It was observed that overall all CNN models using the segmented data presented better results compared to its non-use, proving to be suitable for melanoma diagnosis. Shaffer’s post hoc test found that InceptionV3 and Xception significantly improved their performance when segmented data was used compared to the baseline and transfer-learning combined with data augmentation. All baseline CNN models were significantly surpassed by the other techniques. Results showed the proposed segmentation method was able to improve the performance in all CNN models considered in this work. In the following sections, all CNN models are compared using segmented data, which already proved to achieve the best performance.

5.5.2 Analyzing the impact of three main hyperparameters

First of all, the advantages of using the extra prediction block were analyzed. The best epochs from each CNN model were combined and all possible ensembles were evaluated. In the end, 57 ensembles were obtained after discarding the empty set and the one-element sets. Then, the best one was selected as baseline (BL) and included in the comparison. In addition, it was obtained another baseline model applying the same above procedure by simply combining the predictions of the CNN models without using the prediction block (BLs). Table 6 shows the average MCC values on test sets comparing BL versus BLs. Results show that models using the extra prediction block obtained the best performance in all datasets compared to BLs. Finally, the Wilcoxon’s test rejected the null hypothesis with a p-value equal to 2.189E-4, confirming the benefit and effectiveness of using the prediction block for combining the features from the individual CNN models. Henceforth, the rest of the experimental study was executed using the proposal architecture.

Table 7 shows the average MCC results attained by the proposed genetic algorithm using different values of g, p_c and p_m. For each dataset, the best MCC value is highlighted in bold typeface. Overall, all models obtained a high performance, being the lowest and the highest average performance 96.9% and 97.8%, respectively. The models attained their best performance in MSK-3, PH2, DERM-LIB and SDC-198 datasets. Regarding the parameters that

| Image id | U-Net | R2U-Net | Proposed segmentation method |
|----------|-------|---------|-----------------------------|
|          | ACC   | F1      | JS   | DC   | UN   | ACC   | F1      | JS   | DC   | UN   | ACC   | F1      | JS   | DC   | UN   |
| P54      | 0.913 | 0.892   | 0.805 | 0.892 | 0.849 | 0.926 | 0.909   | 0.833 | 0.909 | 0.871 | 0.968 | 0.962   | 0.927 | 0.962 | 0.945 |
| P55      | 0.881 | 0.725   | 0.569 | 0.725 | 0.647 | 0.852 | 0.686   | 0.522 | 0.686 | 0.604 | 0.895 | 0.810   | 0.680 | 0.810 | 0.745 |
| P57      | 0.909 | 0.856   | 0.748 | 0.856 | 0.802 | 0.873 | 0.788   | 0.650 | 0.788 | 0.719 | 0.964 | 0.948   | 0.901 | 0.948 | 0.925 |
| P63      | 0.906 | 0.857   | 0.749 | 0.857 | 0.803 | 0.895 | 0.836   | 0.718 | 0.836 | 0.777 | 0.950 | 0.928   | 0.865 | 0.928 | 0.896 |
| P75a     | 0.791 | 0.610   | 0.439 | 0.610 | 0.525 | 0.781 | 0.680   | 0.516 | 0.680 | 0.598 | 0.918 | 0.879   | 0.784 | 0.879 | 0.831 |
| Average  | 0.909 | 0.821   | 0.726 | 0.821 | 0.773 | 0.885 | 0.770   | 0.659 | 0.770 | 0.715 | 0.915 | 0.844   | 0.761 | 0.844 | 0.803 |
Table 5  Average MCC values obtained by using six state-of-the-art CNN models.

| Dataset         | DenseNet | InceptionV3 | MobileNet | NASNet | VGG16 | Xception |
|-----------------|----------|-------------|-----------|--------|-------|----------|
|                 | B        | TDA         | STDAM     | B      | TDA   | STDAM    | B        | TDA   | STDAM |
| BCN20000        | 0.642    | 0.929       | 0.933     | 0.633  | 0.948 | 0.938    | 0.588    | 0.919 | 0.928 |
| DERM-LIB        | 0.888    | 0.934       | 0.993     | 0.807  | 0.936 | 0.965    | 0.859    | 0.954 | 0.992 |
| DERM7PT-C       | 0.389    | 0.803       | 0.813     | 0.317  | 0.728 | 0.771    | 0.315    | 0.806 | 0.812 |
| DERM7PT-D       | 0.502    | 0.890       | 0.848     | 0.455  | 0.838 | 0.843    | 0.458    | 0.893 | 0.903 |
| HAM10000        | 0.628    | 0.954       | 0.960     | 0.591  | 0.940 | 0.942    | 0.463    | 0.945 | 0.947 |
| ISBI2016        | 0.367    | 0.850       | 0.878     | 0.307  | 0.802 | 0.825    | 0.309    | 0.850 | 0.854 |
| ISBI2017        | 0.217    | 0.854       | 0.864     | 0.225  | 0.829 | 0.839    | 0.095    | 0.875 | 0.875 |
| MED-NODE        | 0.508    | 0.698       | 0.699     | 0.567  | 0.732 | 0.766    | 0.533    | 0.741 | 0.768 |
| MSK-1           | 0.555    | 0.880       | 0.940     | 0.574  | 0.868 | 0.873    | 0.494    | 0.886 | 0.899 |
| MSK-2           | 0.325    | 0.830       | 0.889     | 0.282  | 0.805 | 0.817    | 0.263    | 0.860 | 0.870 |
| MSK-3           | 0.165    | 1.000       | 0.959     | 0.080  | 0.959 | 0.969    | 0.124    | 1.000 | 0.966 |
| MSK-4           | 0.393    | 0.864       | 0.868     | 0.365  | 0.844 | 0.847    | 0.269    | 0.890 | 0.850 |
| PH2             | 0.771    | 0.960       | 0.967     | 0.647  | 0.963 | 0.964    | 0.700    | 0.963 | 0.987 |
| SDC-198         | 0.610    | 0.951       | 0.956     | 0.444  | 0.908 | 0.929    | 0.496    | 0.971 | 0.979 |
| UDA-1           | 0.394    | 0.764       | 0.808     | 0.336  | 0.720 | 0.762    | 0.360    | 0.781 | 0.813 |
| UDA-2           | 0.596    | 0.522       | 0.513     | 0.540  | 0.413 | 0.410    | 0.375    | 0.577 | 0.580 |

The best MCC values by dataset were highlighted in bold typeface. "B" means a baseline performance (i.e., none technique was applied). "TDA" and "STDAM" represent the model applying transfer learning and data augmentation, and combining segmented images with transfer learning and data augmentation, respectively; data augmentation was applied both in training and test sets.
control the GA, the results showed the higher the number of generations, the better was the predictive performance. A higher number of generations could imply more exploration and higher possibilities to create offspring with a more accurate set of CNN models. Furthermore, the best results peaked when using a high mutation probability, denoting that a higher mutation probability could imply a better exploration of good sets of epochs. Finally, it was observed that 250 generations are good enough to obtain a performance which is significantly similar to the top performance obtained by using 400 generations.

The Rank column shows the average ranking computed by Friedman’s test, and this ranking found that $g = 400$, $p_c = 90\%$ and $p_m = 30\%$ were the best settings for the GA, obtaining on average the best predictive performance. The Friedman’s test rejected the null hypothesis with a $p$-value equal to 8.771E-15; Friedman’s statistic was equal to 124.28 with 26 degrees of freedom. Afterward, the Hommel’s post hoc test was conducted by considering the GA as the control method, which significantly outperformed 50% of all other considered configurations. Next, the best GA is compared to several state-of-the-art CNN models that have previously been used in melanoma diagnosis and at the same time are the baseline to build the ensemble.

5.5.3 Comparing with state-of-the-art CNN models

Table 8 shows the fold changes between the proposed model and each of the state-of-the-art CNN models. The BL ensemble overcame all individual state-of-the-art models, except in DERM-LIB, where DenseNet201 and MobileNet surpassed it by a small margin. It should be noticed that these models are the same considered by the GA to build the ensemble, so this comparison plays another role, which is to corroborate that ensemble learning is more suitable for melanoma diagnosis tasks compared to individual models. The results were very promising since the proposal GAPE-CNN achieved the highest MCC values in all datasets. It is noteworthy that the proposal achieved a predictive performance 165% and 130% higher than Xception and InceptionV3 on UDA-2 dataset, respectively. Also, it achieved high performance in the two largest
datasets. The best average predictive performance was observed on DERM-LIB and PH2 datasets, where it was obtained 98% and 96% MCC values, respectively. The lowest overall predictive performance was observed on the dataset UDA-2 with 56% of MCC. However, the proposal was at least 62% and 21% better than all the individual CNN models and the BL, respectively. Overall, the worst performance was attained by VGG16.

Table 9 summarizes the average MCC values on test sets comparing the baseline versus GAPE-CNN. The proposal surpassed BL in all datasets, and finally, the Wilcoxon’s test rejected the null hypothesis with a $p$-value equal to 2.189E-4.

The best MCC values are highlighted in bold typeface. Wilcoxon’s test rejected the null hypothesis with a $p$-value equal to 2.189E-4.

| Dataset       | BLs | BL  |
|---------------|-----|-----|
| BCN20000      | 0.934 | 0.949 |
| DERM-LIB      | 0.982 | 0.990 |
| DERM7PT-C     | 0.852 | 0.868 |
| DERM7PT-D     | 0.916 | 0.929 |
| HAM10000      | 0.963 | 0.970 |
| ISBI2016      | 0.911 | 0.933 |
| ISBI2017      | 0.919 | 0.963 |
| MED-NODE      | 0.808 | 0.890 |
| MSK-1         | 0.943 | 0.953 |
| MSK-2         | 0.902 | 0.925 |
| MSK-3         | 0.960 | 0.970 |
| MSK-4         | 0.938 | 0.961 |
| PH2           | 0.960 | 0.980 |
| SDC-198       | 0.955 | 0.995 |
| UDA-1         | 0.822 | 0.890 |
| UDA-2         | 0.657 | 0.776 |

5.5.4 Dermoscopic versus non-dermoscopic images

Figure 7 shows the average performance attained by the models by grouping the datasets in dermoscopic and non-dermoscopic ones. The results showed that the proposed architecture attained the best performance whatever the type of image, denoting the effectiveness of the approach. All CNN models attained their best performance using dermoscopic images by a slight margin. NASNetMobile was the architecture most benefited from using dermoscopic images, with an improvement of about 8%. The biggest improvement was found comparing with VGG; the proposal outperformed VGG considering dermoscopic and non-dermoscopic images in 31% and 30%, respectively. The second best performance behind the proposal was achieved by the BL in both types of images; the proposal surpassed it by 5% in both types of images. To sum up, the results obtained through the experimental study revealed that GAPE-CNN was effective for diagnosing melanoma, attaining better predictive performance with respect to the state-of-the-art models.

6 Conclusions

It is clear that automatic melanoma diagnosis via deep learning models is a challenging task, mainly due to a lack of data and differences even between samples from the same category. We addressed these problems via a series of contributions. First, to preprocess data, we validated and applied an extension of the Chan-Vese algorithm. The segmentation masks indicated that the proposal achieved a better performance compared to state-of-the-art segmentation methods. Also, the results showed that all CNN models improved their performance by using segmented data. Second, the training and testing data were enriched using data augmentation, reducing overfitting and obtaining transformation-invariant models. Third, we increased the performance by applying transfer learning from the pre-trained ImageNet. Also, transfer learning alleviated the requirement for a large number of training data. Finally, to further improve the discriminative power of CNN models, we proposed a novel ensemble method based on a genetic algorithm, which finds an optimal set of CNN models. An extensive experimental study was conducted on sixteen image datasets, demonstrating the utility and effectiveness of the proposed approach, attaining a high predictive performance even in datasets with complex properties. Results also showed the proposed ensemble model is competitive with regard to state-of-the-art computational methods.
Table 7 Average MCC values on test sets; $g$, $p_c$ and $p_m$ represent generations, crossover rate and mutation rate, respectively

| $g$ | $p_c$ | $p_m$ | BCN | DER | DER-C | DER-D | HAM | IS-16 | IS-17 | MED | MS-1 | MS-2 | MS-3 | MS-4 | PH2 | SDC | UD-1 | UD-2 | Rank | $p$-value |
|-----|------|------|-----|-----|-------|-------|-----|------|------|-----|------|------|------|------|-----|-----|------|------|------|-------|
| 0.1 | 0.977 | 1.000 | 0.926 | 0.944 | 0.997 | 0.990 | 0.992 | 0.931 | 0.977 | 0.986 | 1.000 | 0.979 | 1.000 | 1.000 | 0.939 | 0.915 | 1.600E+01 | 1.877E-02 |
| 0.1 | 0.978 | 1.000 | 0.915 | 0.945 | 0.997 | 0.991 | 0.991 | 0.942 | 0.979 | 0.986 | 1.000 | 0.976 | 1.000 | 1.000 | 0.935 | 0.915 | 1.703E+01 | 6.271E-03 |
| 0.1 | 0.978 | 1.000 | 0.918 | 0.950 | 0.998 | 0.991 | 0.993 | 0.931 | 0.977 | 0.984 | 1.000 | 0.976 | 1.000 | 1.000 | 0.935 | 0.915 | 1.575E+01 | 2.549E-02 |
| 0.1 | 0.978 | 1.000 | 0.920 | 0.942 | 0.997 | 0.989 | 0.992 | 0.920 | 0.977 | 0.985 | 1.000 | 0.977 | 1.000 | 0.995 | 0.934 | 0.912 | 2.000E+01 | 8.308E-05 |
| 0.1 | 0.976 | 1.000 | 0.918 | 0.945 | 0.998 | 0.989 | 0.993 | 0.931 | 0.977 | 0.985 | 1.000 | 0.976 | 1.000 | 1.000 | 0.939 | 0.941 | 1.606E+01 | 1.737E-02 |
| 0.1 | 0.977 | 0.992 | 0.912 | 0.945 | 0.997 | 0.988 | 0.993 | 0.919 | 0.975 | 0.984 | 1.000 | 0.973 | 1.000 | 0.995 | 0.931 | 0.915 | 2.231E+01 | 1.262E-06 |

Rank means the average ranking computed by Friedman’s test; Hommel’s $p$-values are shown in the last column where strike values represent the ones do not have significant differences compared to the best ranked (last row). The best MCC values for each dataset are highlighted in bold typeface; dataset names were shortened and placed horizontally to save space.
| Dataset    | DenseNet201 | % | InceptionV3 | % | MobileNet | % | NASNetMobile | % | VGG16 | % | Xception | % | BL | % | GAPE-CNN |
|------------|-------------|---|-------------|---|-----------|---|--------------|---|-------|---|----------|---|----|---|---------|
| BCN20000   | 0.933       | 5 | 0.938       | 5 | 0.928     | 6 | 0.916        | 7 | 0.809 | 21| 0.919   | 7 | 0.949| 3 | 0.982   |
| DERM-LIB   | 0.993       | 1 | 0.965       | 4 | 0.992     | 1 | 0.978        | 2 | 0.941 | 6 | 0.977   | 2 | 0.990| 1 | 1.000   |
| DERM7PT-C  | 0.813       | 16| 0.771       | 22| 0.812     | 16| 0.728        | 29| 0.722 | 30| 0.741   | 27| 0.868 | 8 | 0.940   |
| DERM7PT-D  | 0.848       | 13| 0.843       | 13| 0.903     | 6 | 0.854        | 12| 0.738 | 29| 0.849   | 12| 0.929 | 3 | 0.955   |
| HAM10000   | 0.960       | 4 | 0.942       | 6 | 0.947     | 6 | 0.935        | 7 | 0.835 | 20| 0.894   | 12| 0.970 | 3 | 1.000   |
| ISBI2016   | 0.878       | 13| 0.825       | 21| 0.854     | 17| 0.829        | 20| 0.667 | 49| 0.843   | 18| 0.933 | 7 | 0.995   |
| ISBI2017   | 0.864       | 15| 0.839       | 18| 0.875     | 13| 0.849        | 17| 0.805 | 23| 0.882   | 13| 0.963 | 3 | 0.993   |
| MED-NODE   | 0.699       | 36| 0.766       | 25| 0.768     | 24| 0.665        | 43| 0.675 | 41| 0.759   | 26| 0.890 | 7 | 0.954   |
| MSK-1      | 0.940       | 4 | 0.873       | 12| 0.899     | 9 | 0.852        | 15| 0.683 | 44| 0.860   | 14| 0.953 | 3 | 0.982   |
| MSK-2      | 0.889       | 11| 0.817       | 20| 0.870     | 13| 0.867        | 13| 0.581 | 69| 0.832   | 18| 0.925 | 6 | 0.984   |
| MSK-3      | 0.959       | 4 | 0.969       | 3 | 0.966     | 4 | 0.808        | 24| 0.814 | 23| 0.938   | 7 | 0.970 | 3 | 1.000   |
| MSK-4      | 0.868       | 14| 0.847       | 17| 0.850     | 16| 0.823        | 20| 0.799 | 24| 0.829   | 19| 0.961 | 3 | 0.988   |
| PH2        | 0.967       | 3 | 0.964       | 4 | 0.987     | 1 | 0.937        | 7 | 0.911 | 10| 0.971   | 3 | 0.980 | 2 | 1.000   |
| SDC-198    | 0.956       | 5 | 0.929       | 8 | 0.979     | 2 | 0.922        | 8 | 0.836 | 20| 0.941   | 6 | 0.995 | 1 | 1.000   |
| UDA-1      | 0.808       | 16| 0.762       | 23| 0.813     | 15| 0.716        | 31| 0.690 | 36| 0.734   | 28| 0.890 | 6 | 0.939   |
| UDA-2      | 0.513       | 83| 0.410       | 130| 0.580    | 62| 0.441       | 113| 0.437 | 115| 0.355   | 165| 0.776 | 21| 0.941   |

The % columns represent the fold changes after comparing the proposed model with the others, e.g., GAPE-CNN attained a 21% of improvement over VGG16 on BCN20000 dataset. The best MCC values for each dataset are highlighted in bold typeface.
Future works will conduct more extensive experiments to validate the full potential of the proposed architecture, for example by considering a wide set of hyperparameters to be tuned as well as a larger number of datasets. Finally, it is noteworthy that our approach benefits from combining different CNN models—abstract features are combined in the extra prediction block, and individual predictions from the ensemble and the mentioned block are combined to obtain a diagnosis. This approach is not restricted to melanoma diagnosis problems and could be applied on other real-world problems in the future.

### Acknowledgements

This research was supported by the i-PFIS contract no. IFI17/00015 granted by Health Institute Carlos III of Spain. It was also supported by the University of Cordoba and the European Regional Development Fund, project UCO-FEDER 18 REF.1263116 MOD.A.

### Funding

Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Consejería de Transformación Económica, Industria, Conocimiento y Universidades. Junta de Andalucía (ES) (Grant No. UCO-126278) and ministerio de ciencia e innovación (Grant No. PID2020-115832GB-I00).

### Declarations

#### Conflict of interest

The authors declare that they have no conflict of interest.

#### Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

### References

1. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, Gavin A, Visser O, Bray F (2018) Cancer incidence and mortality patterns in Europe: estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 103:356–387
2. American Cancer Society: Cancer Facts and Figures (2021). https://bit.ly/3gNDBV. Consulted on June 22, 2021
3. Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118
4. Geller AC, Swetter SM, Brooks K, Demierre MF, Yaroch AL (2007) Screening, early detection, and trends for melanoma: current status (2000–2006) and future directions. J Am Acad Dermatol 57(4):555–572
5. Rastgoo M, Lemaitre G, Morel O, Massich J, Garcia R, Mériaudeau F, Marzani F, Sidibé D (2016) Classification of melanoma lesions using sparse coded features and random forests. In: progress in biomedical optics and imaging - proceedings of SPIE, vol. 9785. San Diego, California, USA
6. Sánchez-Monedero J, Pérez-Ortiz M, Sáez A, Gutiérrez PA, Hervás-Martínez C (2018) Partial order label decomposition approaches for melanoma diagnosis. Appl Soft Comput 64:341–355
7. Li X, Yu L, Fu C.W., Heng P.A. (2018) Deeply supervised rotation equivariant network for lesion segmentation in dermoscopy images. Lect Notes Comput Sc 11041 LNCS, 235–243
8. Pérez E, Reyes O, Ventura S (2021) Convolutional neural networks for the automatic diagnosis of melanoma: an extensive experimental study. Med Image Anal 67:101858
9. Jin L, Gao S, Li Z, Tang J (2015) Hand-crafted features or machine learnt features? together they improve RGB-D object

### Table 9

Average MCC values on test sets; BL represents the best ensemble obtained by combining the individual best model from each architecture and GAPE-CNN represents our proposal

| Dataset     | BL  | GAPE-CNN |
|-------------|-----|----------|
| BCN20000    | 0.949 | 0.982 |
| DERM-LIB    | 0.990 | 1.000 |
| DERM7PT-C   | 0.868 | 0.940 |
| DERM7PT-D   | 0.929 | 0.955 |
| HAM10000    | 0.970 | 1.000 |
| ISBI2016    | 0.933 | 0.995 |
| ISBI2017    | 0.963 | 0.993 |
| MED-NODE    | 0.890 | 0.954 |
| MSK-1       | 0.953 | 0.982 |
| MSK-2       | 0.925 | 0.984 |
| MSK-3       | 0.970 | 1.000 |
| MSK-4       | 0.961 | 0.988 |
| PH2         | 0.980 | 1.000 |
| SDC-198     | 0.995 | 1.000 |
| UDA-1       | 0.890 | 0.939 |
| UDA-2       | 0.776 | 0.941 |

The best MCC values are highlighted in bold typeface. Wilcoxon’s test rejected the null hypothesis with a p-value equal to 2.189E-4.

### Fig. 7

Average MCC values on test sets by grouping the datasets in dermoscopic (green bars) and non-dermoscopic (orange bars); “D”: DenseNet201; “I”: InceptionV3; “M”: MobileNet; “N”: NASNetMobile; “V”: VGG16; “X”: Xception; “B”: base line; “G”: our proposal

Future works will conduct more extensive experiments to validate the full potential of the proposed architecture, for example by considering a wide set of hyperparameters to be tuned as well as a larger number of datasets. Finally, it is noteworthy that our approach benefits from combining different CNN models—abstract features are combined in the extra prediction block, and individual predictions from the ensemble and the mentioned block are combined to obtain a diagnosis. This approach is not restricted to melanoma diagnosis problems and could be applied on other real-world problems in the future.
10. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr S, Jafari M, Ward K, Najarian K (2016) Melanoma detection by analysis of clinical images using convolutional neural network. In: proceedings of the IEEE EMBS, pp. 1373–1376. Florida, USA

11. Brinker TJ, Hekler A, Enk AH, Klode J, Haushchild A (2019) A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 111:148–154

12. Liu Y, Chen X, Peng H, Wang Z (2017) Multi-focus image fusion with a deep convolutional neural network. Inform Fusion 36:191–207

13. Hossain MS, Muhammad G (2019) Emotion recognition using deep learning approach from audio-visual emotional big data. Inform Fusion 49:69–78

14. Asif U, Bennamoun M, Sohel F (2018) A multi-modal, discriminative and spatially invariant CNN for RGB-D object labeling. IEEE T Pattern Anal 40(9):2051–2065

15. Ericsson: on the pulse of the networked society. Tech. rep.

16. Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between symmetric hopfield networks. arXiv:1710.05285

17. Lenc K, Vedaldi A (2019) Understanding image representations by measuring their equivariance and equivalence. Int J Comput Vision 127(5):456–476

18. Perez F, Vasconcelos C, Avila S, Valle E (2018) Data augmentation for skin lesion analysis. In: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis, pp. 303–311. Springer, Granada, Spain

19. Mahbod A, Schaeffer G, Ellinger I, Ecker R, Pittot A, Wang C (2019) Fusing fine-tuned deep features for skin lesion classification. Comput Med Imag Graph 71:19–29

20. Baur C, Albarqouni S, Navab N (2018) MelanoGANs: high resolution skin lesion synthesis with GANs. arXiv preprint: arXiv:1804.04338

21. Kawahara J, Daneshvar S, Argenzano G, Hamarneh G (2019) Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE J Biomed Health Inform 23(2):538–546

22. Zeng H, Haleem H, Plantaz X, Cao N, Qu H (2017) CNNComparator: Comparative analytics of Convolutional Neural Networks. arXiv preprint: arXiv:1710.05285

23. Harangi B, Baran A, Hajdu A (2018) Classification of skin lesions using an ensemble of deep neural networks. In: proceedings of the annual international conference of the IEEE EMBS, vol. 2018-July, pp. 2575–2578. Honolulu, HI, USA

24. Sahu P, Yu D, Qin H (2018) Apply lightweight deep learning on internet of things for low-cost and easy-To-Access skin cancer detection. In: progress in biomedical optics and imaging - proceedings of SPIE, vol. 10579. Houston, Texas, USA

25. Zeng G, Zheng G (2018) Multi-scale fully convolutional denseNets for automated skin lesion segmentation in dermoscopy images. Lect Notes Comput Sci 10882 LNCS, 513–521

26. Zhao XY, Wu X, Li FF, Li Y, Huang WH, Huang K, He XY, Fan W, Wu Z, Chen ML, Li J, Luo ZL, Su J, Xie B, Zhao S (2019) The application of deep learning in the risk grading of skin tumors for patients using clinical images. J Med Syst 43(8):1–7

27. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst, vol 2. Harrahs and Harveys, Lake Tahoe, NV, USA, pp 1097–1105

28. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F.F. (2014) Large-scale video classification with convolutional neural networks. In: proceedings of the IEEE computer society CVPR, pp. 1725–1732. Washington, USA

29. He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE T Pattern Anal 37(9):1904–1916

30. Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE T Med Imaging 35(5):1285–1298

31. Huang L, Zhao YG, Yang TJ (2019) Skin lesion segmentation using object scale-oriented fully convolutional neural networks. Signal Image Video Process 13(3):431–438

32. Ciresan DC, Meier U, Gambardella LM, Schmidhuber J (2010) Deep, big, simple neural nets for handwritten digit recognition. Neural Comput 22(12):3207–3220

33. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv preprint. arXiv:1712.04621

34. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, London

35. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: proceedings of the 30th IEEE CVPR-2017, pp. 1800–1807. Honolulu, HI, USA

36. Perez F, Vasconcelos C, Avila S, Valle E (2018) Data augmentation for skin lesion analysis. In: OR 2.0 context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis, pp. 303–311. Springer, Granada, Spain

37. Sa I, Ge Z, Dayoub F, Uprcft B, Perez T, McCool C (2016) Deepfruits: a fruit detection system using deep neural networks. Sensors (Switzerland) 16(8):1222

38. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: 31st AAAI-2017. California, USA, San Francisco, pp 4278–4284

39. Howard A.G., Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint. arXiv:1704.04861

40. Dietterich T (2017) Ensemble methods in machine learning, vol. 1857 LNCS (2000)

41. Singh B, Davis L.S. (2018) An analysis of scale invariance in object detection - SNIP. In: proceedings of the IEEE computer society CVPR, pp. 3578–3587. Utah, USA

42. Zhang C, Pan X, Li H, Gardiner A, Sargent I, Hare J, Atkinson PM (2018) A hybrid MLP-CNN classifier for very fine resolution remotely sensed scene image classification. J Photogramm Remote Sens 140:133–144

43. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: proceedings of the IEEE CVPR, pp. 770–778. Las Vegas, NV, USA

44. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: proceedings of the IEEE CVPR, vol. 07-12-June-2015, pp. 1–9. Boston, Massachusetts, USA

45. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: international conference on medical image computing and computer-assisted intervention. Springer, Munich, Germany, pp 234–241
85. Gavai N.R., Jakhade Y.A., Tribhuvan S.A., Bhattad R (2017) Mobilenets for flower classification using tensorflow. In: 2017 BID, pp. 154–158
86. Liu X, Jia Z, Hou X, Fu M, Ma L, Sun Q (2019) Real-time marine animal images classification by embedded system based on mobilenet and transfer learning. In: OCEANS 2019 - Marseille, pp. 1–5. https://doi.org/10.1109/OCEANSE.2019.8867190
87. Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1–2):1–39
88. Menegola A, Tavares J, Fornaciali M, Li L.T., Avila S, Valle E (2017) RECOD Titans at ISIC Challenge 2017. arXiv:1703.04819

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