

Article

A Hybrid Artificial Intelligence Framework for Melanoma Diagnosis Using Histopathological Images

Alberto Nogales ^{1,*}, María C. Garrido ², Alfredo Guitian ¹, Jose-Luis Rodriguez-Peralto ², Carlos Prados Villanueva ², Delia Díaz-Prieto ¹ and Álvaro J. García-Tejedor ¹

¹ CEIEC Research Institute, Universidad Francisco de Vitoria, Ctra. M-515 Pozuelo-Majadahonda km. 1, 200, Pozuelo de Alarcón, 28223 Madrid, Spain; alfredoguitian@gmail.com (A.G.); deliadiazprieto@gmail.com (D.D.-P.)

² Hospital Universitario 12 de Octubre, Universidad Complutense de Madrid, Avda de Cordoba s/n, 28041 Madrid, Spain; mariaconcepcion.garrido@salud.madrid.org (M.C.G.); jrperalto@salud.madrid.org (J.-L.R.-P.); carlos.prados@salud.madrid.org (C.P.V.)

* Correspondence: alberto.nogales@ceiec.es

Abstract

Cancer remains one of the most significant global health challenges due to its high mortality rates and the limited understanding of its progression. Early diagnosis is critical to improving patient outcomes, especially in skin cancer, where timely detection can significantly enhance recovery rates. Histopathological analysis is a widely used diagnostic method, but it is a time-consuming process that heavily depends on the expertise of highly trained specialists. Recent advances in Artificial Intelligence have shown promising results in image classification, highlighting its potential as a supportive tool for medical diagnosis. In this study, we explore the application of hybrid Artificial Intelligence models for melanoma diagnosis using histopathological images. The dataset used consisted of 506 histopathological images, from which 313 curated images were selected after quality control and preprocessing. We propose a two-step framework that employs an Autoencoder for dimensionality reduction and feature extraction of the images, followed by a classification algorithm to distinguish between melanoma and nevus, trained on the extracted feature vectors from the bottleneck of the Autoencoder. We evaluated Support Vector Machines, Random Forest, Multilayer Perceptron, and K-Nearest Neighbours as classifiers. Among these, the combinations of Autoencoder with K-Nearest Neighbours achieved the best performance and inference time, reaching an average accuracy of approximately 97.95% on the test set and requiring 3.44 min per diagnosis. The baseline comparison results were consistent, demonstrating strong generalisation and outperforming the other models by 2 to 13 percentage points.

Keywords: hybrid models; deep learning; machine learning; diagnostic models; skin cancer; melanoma



Academic Editors: George F. Fragulis, Masateru Kawakubo and Tamás Haidegger

Received: 2 June 2025

Revised: 17 July 2025

Accepted: 25 July 2025

Published: 1 August 2025

Citation: Nogales, A.; Garrido, M.C.; Guitian, A.; Rodriguez-Peralto, J.-L.; Villanueva, C.P.; Díaz-Prieto, D.; García-Tejedor, Á.J. A Hybrid Artificial Intelligence Framework for Melanoma Diagnosis Using Histopathological Images. *Technologies* **2025**, *13*, 330. <https://doi.org/10.3390/technologies13080330>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cancer is widely recognised as one of the most challenging diseases for physicians, due to its multifactorial complexity. Its genetic nature, rooted in DNA mutations and alterations, leads to high heterogeneity across cases. Additionally, environmental factors significantly influence tumour growth and progression, further complicating its diagnosis and treatment. Moreover, the variability in individual responses to cancer requires the

adoption of personalised therapeutic approaches. Moreover, cancer requires a multidisciplinary approach, involving the collaboration of oncologists, surgeons, pathologists, and other specialists to provide comprehensive care. The diversity of cancer types adds another layer of complexity, with over 200 distinct forms identified [1], each demanding specialised expertise for proper diagnosis and treatment. Among these, skin cancer stands out as one of the most common forms in the United States, with its global prevalence continuing to rise [2]. Addressing this growing challenge requires significant efforts to advance diagnostic and therapeutic strategies tailored to the unique nature of skin cancer and its subtypes.

The most severe and potentially life-threatening form of skin cancer is melanoma [3]. Points out that in the United States, it is ranked as the fifth most common cancer in both men and women. It accounts for approximately 350,000 new cases and 57,000 deaths reported globally in 2020 [4]. The incidence increases with age, underscoring the importance of monitoring the population as it grows older. Survival rates for melanoma are closely linked to the stage of the disease at the time of diagnosis, making early detection a critical factor in improving patient outcomes and saving lives. While many melanomas are initially detected by patients themselves [5], clinician detection is often associated with thinner, more treatable tumours [6]. This highlights the value of professional screening in identifying melanomas at an early stage. For patients diagnosed with thin lesions and invasive melanomas (Breslow thickness ≤ 1 mm), treatment typically results in prolonged disease-free survival and, in most cases, a complete cure [7].

The current gold standard for melanoma diagnosis is histopathology, which analyses melanocytic neoplasms. These areas are tumours that originate from melanocytes, which are the cells responsible for producing the pigment in the skin. However, a subset of these neoplasms cannot be unequivocally classified as benign (nevus) or malignant (melanoma). These ambiguous cases are a significant source of diagnostic error, as evidenced by studies reporting discordance rates between expert dermatopathologists ranging from 14% to 38% using routine examination [8]. To reduce the risk of missing malignant lesions, the diagnostic criteria for melanoma have been adjusted to emphasise detecting as many potential cases as possible (sensitivity), even if this comes at the expense of a higher False Positive (FP) rate (reduced specificity). This trade-off highlights the urgent need for novel diagnostic tests that can increase accuracy by providing quantitative, objective information to minimise the inherent subjectivity of histopathological evaluation. Genetic analyses of melanocytic lesions have shown promise in distinguishing melanomas that harbour recurrent genetic aberrations absent in unequivocally benign lesions [9]. However, a “grey area” remains in histologically ambiguous melanocytic neoplasms with few genetic aberrations that continue to pose uncertainty regarding their biological behaviour.

Histopathologic diagnosis of melanocytic lesions is based on microscopic criteria including asymmetry, cytological atypia, deep maturation, pagetoid extension, mitosis and dusty pigmentation in melanocytes, which help the pathologist to classify these lesions as benign or malignant and to establish a specific diagnosis. Since these features may become ambiguous in some cases, expertise is most valuable in others. Doubtful cases are usually reviewed by several pathologists to arrive at the definitive diagnosis. This task is time-consuming and consumes significant resources due to the size and resolution of the images. Considering what has been mentioned previously, Artificial Intelligence (AI) techniques present a compelling solution to this problem.

By analysing big datasets, AI models can identify patterns in cases that are critical for human interpretation. These models can deliver diagnoses automatically and with remarkable speed, offering significant help in the accuracy and efficiency of melanoma diagnosis. AI is a discipline that aims to understand and replicate the mechanisms underlying intelligent behaviour in machines. This is achieved through various methods, which do

not necessarily mimic the original biological mechanisms. AI encompasses a wide range of approaches, with Machine Learning (ML) being the most prominent in recent years. As defined by [10], ML is a field of study that enables computers to learn autonomously without explicit programming. It includes numerous techniques, where Deep Learning (DL) has emerged as a groundbreaking advancement. DL, as described by [11], refers to models capable of learning hierarchical representations of data through multiple levels of abstraction. These models are inspired by artificial neural networks, designed to emulate the behaviour of biological neurons. They achieve this through interconnected layers arranged sequentially, enabling them to process complex patterns and relationships within data effectively.

The primary motivation of this study is to develop a fast and accurate approach for diagnosing melanoma using histopathological images, aiming to speed up the diagnostic process for physicians. The proposed methods address the challenges associated with the high dimensionality of histopathological images, which require significant computational resources and complex feature extraction. To tackle these issues, we leverage the advantages of Autoencoders for enhanced feature extraction by reducing the dimensionality, combined with classical ML models for classification. Furthermore, the study incorporates a subjective evaluation to better understand the model's performance and to identify common histopathological features that might confuse the classifier during diagnosis. This evaluation aims to bridge the gap between automated diagnostic tools and clinical practice, offering valuable insights to improve both the performance of the model and its interpretability for medical professionals.

The contribution of this work lies in the development of a workflow that first employs an Autoencoder for dimensionality reduction and feature extraction from histopathological images, followed by using these extracted features with various classifiers. This workflow produces a set of hybrid models that integrate novel DL techniques with classical ML algorithms. The performance of these hybrid combinations is evaluated both objectively, using standard mathematical metrics, and subjectively, through the insights of a medical expert in the field.

The innovation of the paper consists of the presented workflow and the different hybrid models, which, as far as we know, are the first time that these combinations are applied to diagnose melanoma using histopathological images. Moreover, the subjective evaluation lets us understand how the model works and what its limitations are.

The remaining sections of the paper are structured as follows. Section 2 provides a compilation of previous papers that pertain to the same field or share similarities with the proposed work. In Section 3, the data utilised for training the models and addressing the problems is described in detail. Section 4 presents the research findings and showcases the results obtained through the course of the study. Finally, Section 5 offers concluding remarks and highlights potential avenues for future research.

2. Related Works

As said above, in this work, different hybrid models using DL models alongside classical ML techniques have been proposed to diagnose melanoma with histopathologic images. Following, we compile different works using similar techniques for skin cancer diagnosis in general or melanoma.

First, we compile some works that only use classical ML techniques, for example [12] utilises Haar wavelet transformation in dermoscopy skin images to extract the main feature. These features are then introduced in different classifiers using the Random Forest (RF) algorithm, Support Vector Machines (SVM), or Naïve Bayes to discriminate between benign and malignant lesions. The same dataset is used in [13], where preprocessing techniques

are used to eliminate noise and occlusion, such as body hairs. Then, some characteristics were obtained by applying the ABCD rule and using them to feed different ML models like SVM, RFs, and K-Nearest Neighbours (KNN). Another type of image to diagnose skin cancer is Optical Coherence Tomography (OCT), which is used by [14] to acquire both intensity and birefringence images of healthy and cancerous mouse skin samples. Using an SVM-based classifier, the researchers aimed to automatically distinguish images indicative of basal cell carcinoma (BCC), the most common type of skin cancer. Vibrational OCT is used with logistic regression and SVM to use telemedicine to classify between basal cell carcinoma (BCC), squamous cell carcinoma (SCC), melanoma, and controls [15]. In [16], different preprocessing techniques, including the Gray Level Co-Occurrence Matrix (GLCM), Principal Component Analysis (PCA), and fuzzy C-means clustering, extract features from skin cancer images, and then utilise SVM for diagnosis. Moreover, [17] integrates Matrix-Assisted Laser Desorption/Ionisation Mass Spectrometry Imaging (MALDI-MSI) with Logistic Regression (LR) to characterise and delineate cutaneous squamous cell carcinoma, achieving a predictive accuracy of 92.3% in cross-validation. Investigation by [18] focused on the role of triaptopsis-related gene expression in melanoma prognosis by integrating single-cell and bulk RNA-seq data. A total of 101 ML algorithms and their combinations, including CoxBoost, Lasso, Random Survival Forest (RSF), and SurvivalSVM, were used to construct and validate prognostic models. The study presented by [19] combines 2D-IR spectroscopy and an ML algorithm called Partial Least Squares-Support Vector Machine (PLS-SVM) to classify melanoma patient serum samples, accurately predicting metastatic status and risk of relapse from protein fingerprint data. Finally, [20] applies Gaussian Naïve Bayes (GNB) and Logistic Regression (LR) to improve the histological classification of spitzoid tumours using Whole Slide Images (WSI) and clinicopathological features. To the best of our knowledge, the latter is the only work that utilises histopathological images in conjunction with classic machine-learning algorithms, although its primary aim is not melanoma diagnosis. In our case, we are not only using our dataset of histopathologies in melanoma, but we are also creating hybrid models that, apart from ML models, are using DL ones that face the problem of high dimensionality in this kind of image.

In the case of the DL models used to diagnose skin cancer, we have the following works where [21] uses different preprocessing techniques in images of skin lesions and evaluates how different CNN models (Resnet50, InceptionV3, and Inception Resnet) perform in diagnosing and classifying. Another work using CNN models is [22], where 11 different architectures were trained with a dataset that comprises 7 different skin lesions. After obtaining a model that performs well, it was also able to inform about the duration of sun exposure depending on the current UV radiation level, the individual's skin phototypes, and the level of protection provided by the sunscreen used. Another CNN model is applied in [23], adding a shape-adaptive method to explain the diagnosis of melanoma through a heatmap. In the case of DL, papers are using histopathologic images. In [24], different pre-trained models like EfficientNetB0, MobileNetv2, ResNet50, and VGG16 were used to find squamous cell carcinoma (SCC). Moreover, [25] uses histopathologic images and DL models. VGG-16 with gradient-weighted class activation mapping (Grad-CAM), Otsu's, and contour estimation, but in this case for the differentiation of skin melanocytes from keratinocytes. Another interesting work is [26], which evaluates DL melanoma classification using haematoxylin and eosin-stained WSI and immunohistochemistry-stained tissue slides using different ResNet models that were trained separately and jointly on these modalities. A study by [27] introduces a DL approach for melanoma prediction using Class-Agnostic Activation Maps (CAAMs), which enhance diagnostic reliability by addressing variability in lesion position and image transformations. The model was developed using ConvNeXt and ResNet backbones. Finally, [28] presents a multi-modal DL framework

integrating dermoscopic images, histopathological slides, and genomic data for melanoma diagnosis using Convolutional Neural Networks (CNNs) for image analysis and Graph Neural Networks (GNNs) for genomic data interpretation. In these works, there are more similarities because we could find works using histopathologic images, but compared with our proposal, we proposed different hybrid models using DL and classical ML models for diagnosing melanoma.

As said above, one of the highlights of the present work is the hybridisation of models. In this way, we have found different works in the field of skin cancer [28] developed and validated a DL model to distinguish small choroidal melanomas from benign nevi using standard and ultra-widefield fundus photographs. The approach combined a two-stage U-Net architecture, one for lesion segmentation and another for melanoma risk scoring, with a shallow RF classifier trained on pixel-wise probability outputs [29] proposes a hybrid DL model that integrates ResNet50, Long Short-Term Memory (LSTM), and transfer learning to enhance skin cancer classification by capturing both spatial and sequential lesion features [30] combining a particular type of CNN called faster region-based with fuzzy k-means clustering to detect melanoma, and segmenting images of skin cancer lesions. A three-level method that uses DL in the first level, SVM, MLP, RF, and KNN in the second, and LR in the third level for classifying benign and malignant images of skin cancers is described in [31,32] presents a hybrid model called SCDNet, combining VGG16 with a regular CNN to classify between 4 different classes.

In the case of skin cancer that use hybrid models, we found [33] that combines multiple public datasets to enhance model training and optimising CNN parameters using the Artificial Bee Colony (ABC) algorithm to improve classification accuracy. Also, [34] presents a DL-based method for automated skin disease diagnosis. By hybridising CNN and DenseNet architectures, the proposed model achieves 95.7% accuracy.

Our approach differs from previous works by proposing four distinct hybrid models that combine DL and classical ML techniques for the classification of histopathological images as melanoma. In our workflow, the classifiers operate at the feature vector level, using latent representations extracted from the bottleneck layer of the Autoencoder. In addition to the standard objective evaluation using DL performance metrics, we also provide an expert-based assessment, offering a more comprehensive understanding of the model's behaviour from a medical perspective. Furthermore, the two previously cited works reported lower accuracies of 93.04%, and 95.70% respectively, compared to our model, which achieves an accuracy of 97.95%.

3. Dataset Collection and Labelling

To train the proposed models, we collected a set of 506 histological raw images taken from 117 melanocytic lesions diagnosed between 2022 to 2023 at the 12 de Octubre University Hospital, including a wide spectrum of nevi and melanomas.

Whole-slide pictures were scanned at 200× with the Aperio Versa scanning system (Leica). Clinical data, including sex, age and location, were also collected. The age of the patient ranges from 10 to 89, with a percentage of 65.21% women, 30.43% men and 4.36% were not registered. Histological images were extracted from different parts of the body, including head and neck (33), trunk (35), Upper and lower extremities (40), metastases from deep tissues (6), and N/A (3 cases).

The labelling was performed by the three coauthors, histopathologists MGR, JLRP, and CPV, all of whom are expert dermatopathologists with extensive experience in the field. MGR and JLRP each have over 20 years of expertise, which is particularly valuable when diagnosing melanocytic lesions. All cases were independently reviewed by the three pathologists, using all available methods and clinical information.

The lesions included 31 melanomas (8 “in situ” melanomas, 12 superficial spreading melanomas, one acral lentiginous melanoma, 2 lentigo maligna melanomas, 7 melanoma metastases and one melanoma satellitosis) and 86 nevi (30 intradermal melanocytic nevi, 12 compound melanocytic nevi, 5 lentiginous nevi, 1 junctional nevus, 6 congenital nevi, 4 dysplastic nevi, 4 Spitz nevi, 9 Reed nevi, and 15 blue nevi). The hospital is officially categorised as a tertiary referral centre by the National Healthcare System. The present research has followed strict recommendations by the hospital Ethics Committee. In Table 1, we summarise the main features of the dataset.

Table 1. Summary of features of the collected dataset.

Dataset Feature	Values
Total raw images collected	499
Scanning magnification	200× (Aperio Versa scanner, Leica (Leica Microsystems, Wetzlar, Germany))
File format	Slide and Viewable Storage (.svs)
Bit depth/colour mode	24-bit RGB
Min., Max., Avg. image dimensions (pixels)	4583 × 6347, 91,683 × 170,804, 48,133 × 88,576
Min., Max., Avg. size dimensions (GB)	0.23, 9.41, 5.13

4. Methods

4.1. Preprocessing the Histological Images

The raw images obtained from the scanned biopsies needed some transformations due to their large size and other circumstances arising during the collection. First, not all the obtained images were valid to train the models, as we found some problems, such as defocused images, others containing some physical labels that bring noise and other issues. At first, the raw dataset contains 499 images. At this moment, a curation process was carried out to ensure that all the images were accomplished with a minimum quality to be used by the models. In this way, a manual revision was carried out, dividing the dataset into five groups. Regarding the images from these groups, the following three were discarded due to insufficient quality. The first group contained 64 images that were out of focus. The second group comprised 47 images that were not fully scanned, with parts of the epithelial slice missing. Lastly, the third group had 67 instances containing text or physical labels that could introduce noise.

Of the remaining images, 321 in total were considered valid. However, we identified two distinct groups within this set. The first group contains 144 images with multiple epithelial slices scanned correctly. The second group consists of 121 images, where all the information is contained within a single slice. This set of 313 images was used to train the Autoencoder and the classifiers by randomly splitting them into an 80–20 ratio for training and testing. The training set for the classifiers includes 194 nevus images and 56 melanoma images. For the test set, this proportion consists of 49 nevus images and 14 melanoma images.

To augment the dataset due to the limited number of images, we generated synthetic data by rotating the original images, effectively increasing the dataset size by a factor of four. Consequently, the training set expanded to include 776 nevus images and 224 melanoma images. Similarly, the test set grew to include 196 nevus images and 56 melanoma images. This augmentation step was crucial in enhancing the model’s ability to generalise effectively.

After setting the training and test datasets, we realised that the images were too big, with sizes ranging from 6134 to 91,683 in width and from 5880 to 170,804 in height. This was a problem at the time of being processed by the models in terms of computational capacity and feature extraction (extracting more particular features is more difficult in big pictures). Due to this, the first decision was to crop the images into small pieces of the same

size that were later fed to the model. In this case, it has been decided to divide the images into 256×256 pixel-coloured slices with no separation or union between them. For each image, a different number of crops has been obtained as the raw images have different sizes and because we have all the crops whose percentage of white pixels (normally pixels with no information) is greater than 60, which differs from image to image. In some cases, crops corresponding to the right and bottom border of the original image are discarded as crops of 256×256 cannot be created. At this stage, the number of cropped images is 612,397. After introducing them into the models, we normalised them by using the Whole Slide Images (WSI) normalisation algorithm, which is a particular way for managing histological images [35].

This algorithm starts by sampling a group of pixels from the tissue region of the histopathological image. Then, it converts the sampled pixels from the RGB (Red, Green and Blue) colour space to the Optical Density (OD) space. This conversion helps to separate the stain information from the image. After that, it utilises an ACD (Adaptive Colour Deconvolution) model to estimate the concentrations of different stains in the image. ACD models are commonly used for stain separation in digital pathology. In the following step, the algorithm calculates a stain-weight matrix based on the concentrations obtained from the ACD model. This matrix represents the importance or weight of each stain in the image. ACD matrix and stain-weight matrix are used to obtain some components, which are combined with the Stained Colour Augmentation (SCA) matrix that provides a reference for the desired colour appearance. The final output is a colour-normalised version of the input WSI, where the colours have been adjusted to avoid perception differences among images.

4.2. A Framework of Hybrid Models for Melanoma Diagnosis

Once the initial dataset has been preprocessed, 4 workflows have been proposed to classify histological images into melanoma and nevus. They all follow the same 3-stage scheme, where the third stage changes depending on the final classifier that has been used. In the first part of the model, dimensionality reduction and feature extraction of the images are performed using an Autoencoder, which is trained to reconstruct each input image as output (this is the differentiated characteristic of the model among Encoder-Decoder models). Then, in the second step, these features are obtained from the bottleneck of the Autoencoder and are combined in 5 ways to introduce the features in the classifiers. This approach aggregates all the crops extracted from a single image during the preprocessing stage to generate a unified input instance for the classifiers, effectively representing the entire image. These combinations are made by obtaining the maximum values, the minimum values, summing them up, calculating the average values and the median. Finally, these vectorised features are introduced into different classifiers, creating the 4 models evaluated in this work. These classifiers are KNN, SVM, RF, and MLP. The workflow for the 4 approaches is shown in Figure 1.

The decision to use an Autoencoder over other solutions like pretrained CNNs is not obvious, but in the present use case, the latter approach has several limitations. In general, as pointed out by [29], self-supervised learning approaches such as Autoencoders can maintain satisfactory performance even when trained with as little as 1% of the available target dataset, making them particularly well-suited for domains like medical imaging where labelled data is often scarce. Talking about the use of pretrained CNNs, in [36] it is demonstrated that in smaller and less similar datasets like ours, applying transfer learning from unrelated domains often risks poor results. Following, we formally describe all the algorithms and methods used in the proposed workflow.

Autoencoders. This is a particular type of Encoder-Decoder model where the input and output instances during each epoch of the training stage are the same instance. These two-part models employ a process to capture complex structures. Initially, a multilayer encoder network is utilised to represent high-dimensional structures in a reduced-dimensional space. Subsequently, a decoder network is employed to convert the data from this space back into high-dimensional structures while maintaining certain relationships to the initial representation [37]. The functioning of this architecture can be described as follows: the input data traverses various convolutional layers within the encoder, extracting essential features and condensing them into smaller data fragments. These fragments are then organised within a bottleneck, forming a representation known as the latent space. Finally, the feature representation is passed through the decoder, resulting in an output that closely resembles the input data.

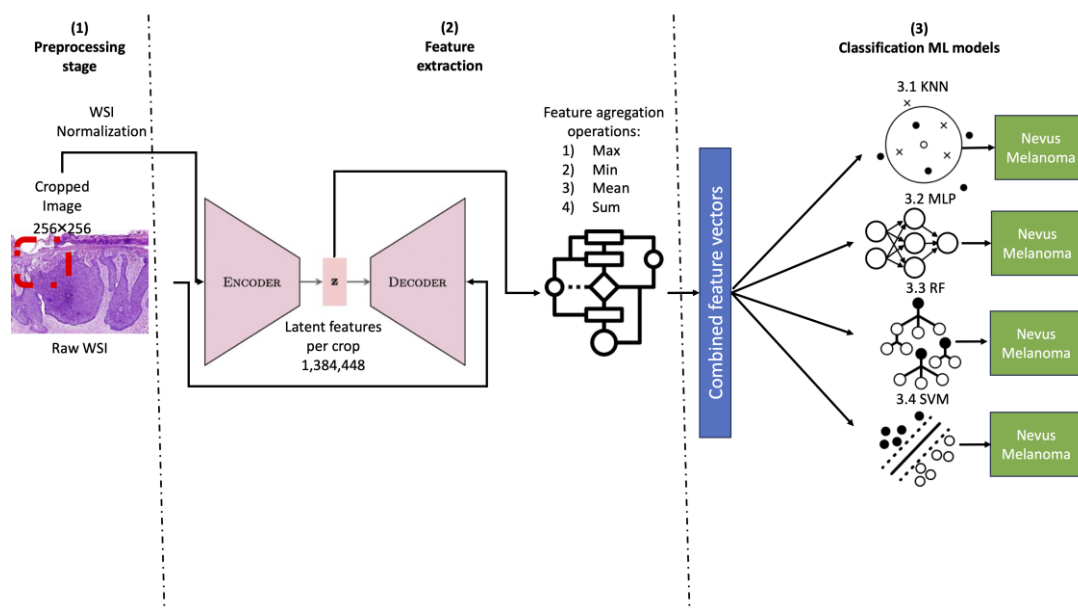


Figure 1. Workflow for the evaluated classifiers. The process starts with data preprocessing and feature extraction using an Autoencoder. The resulting feature vectors are then used to train and evaluate various classification algorithms, including both DL and classical ML models.

Autoencoders are formalised in [38]. During the encoding phase, an Autoencoder transforms an input vector X into a latent code vector Z through an encoding function f . In the decoding phase, the code vector Z is mapped back to a reconstructed output vector X' using a decoding function g , to accurately reconstruct the original input. The Autoencoder adjusts the network's weights W through fine-tuning, which is accomplished by minimising the reconstruction error L between the original input X and its reconstruction X' . This reconstruction error serves as the loss function used to optimise the network parameters. The objective function of the Autoencoder can be expressed as described in Equation (1).

$$\min_{\theta} J_{AE}(\theta) = \min_{\theta} \sum_{i=1}^n (x_i - x'_i) = \min_{\theta} \sum_{i=1}^n (x_i - g_{\theta}(f_{\theta}(x_i))) \quad (1)$$

where x_i denotes the i th element of the input sample, x'_i represents the corresponding element in the reconstructed output, and n is the total number of training samples. The term l refers to the reconstruction error between the original input and its reconstruction.

Feature combination. Due to the large size of the images and the bottleneck, we have proposed different ways to combine the sets of these particular features so they can be introduced in classifiers. Once all the slices of the different images have been processed

by the Autoencoder, the bottleneck will contain n slices (number of images, which is 96) of length m (number of features, 1,384,448, extracted by the Autoencoder). The length of these vectors is a consequence of the aggregation of latent features extracted from multiple 256×256 image crops. These vectors will be grouped depending on the image they belong to, and various operations will be performed to aggregate them. These operations could obtain the maximum, minimum, mean, sum, or median of each column of the vectors belonging to each image to create a single vector of length m per operation. Finally, we will have 5 vectors (one for each operation) of m characteristics for each image processed by the Autoencoder. These vectors will be the input data of the different classifiers. Feature aggregation operations are formalised in the following equations.

$$\max(x) = \max_{x=1}^n x_i \quad (2)$$

$$\min(x) = \min_{x=1}^n x_i \quad (3)$$

$$\text{mean}(x) = \frac{1}{n} \sum_{x=1}^n x_i \quad (4)$$

$$\text{sum}(x) = \sum_{x=1}^n x_i \quad (5)$$

$$\text{median}(x) = x_{(k)} \text{ where } x_{(k)} = \left\lceil \frac{n}{2} \right\rceil \text{ after sorting } \{x_i\} \quad (6)$$

In the equations above x_i are part of a set of values $\{x_1, x_2, \dots, x_n\}$ corresponding to a feature across n observations or sub-elements.

K-Nearest Neighbours. Introduced by [39], presents an algorithm that assigns a class to an instance by considering the K -nearest instances from a given dataset. This approach allows KNN to assign a label based on the local patterns and similarities observed in the dataset. This label is assigned based on the KNN search, which is formalised in [38] as follows.

Let D be a dataset consisting of n points in a d -dimensional space, and let q be a query point in the same space. Let k denote the number of nearest neighbours to retrieve, and let $\text{dist}(p, q)$ be a distance function that measures the distance between a data point $p \in D$ and the query point q . The KNN search problem aims to identify a set $R \subseteq D$, containing k points, such that for every $p \in R$, there is no point $p' \in D$ with $\text{dist}(p', q) < \text{dist}(p, q)$. In other words, R contains the k points in D that are closest to q .

Support Vector Machines. The current version of SVM was initially proposed by [40]. SVM can be conceptualised as a classifier that operates in an n -dimensional space, where instances are distributed. The primary goal of the algorithm is to locate a hyperplane that effectively separates individuals into distinct classes, maximising the margin of separation between them. The wider the margin, the better the classification performance of the SVM. By finding an optimal hyperplane, SVM enables the accurate and efficient classification of data points in higher-dimensional spaces. In [41], it is described as in Equation (7).

$$y(x) = w^T \phi(x) + b \quad (7)$$

In the equation above, $\phi(x)$ represents the kernel function responsible for the feature transformation; essentially, kernel functions map the original feature space into a higher-dimensional space. In this transformed space, the features become linearly separable. The parameter b denotes the bias term, and the vector w is perpendicular (normal) to the hyperplane. The training input feature vector is denoted by x , while the classification of test feature vectors is represented by $y(x)$.

Random Forest. As proposed by [42], ensemble learning is a method that leverages decision trees to improve predictive accuracy. This technique involves combining multiple decision trees and averaging their predictions, resulting in enhanced overall performance. One significant advantage of ensemble learning is its ability to mitigate overfitting, a common issue encountered when using individual decision trees. By aggregating the predictions of multiple models, ensemble learning reduces the likelihood of overfitting, leading to more robust and reliable results.

Formally, a RF is defined in [43] as a predictor composed of an ensemble of randomised base regression trees, denoted as $\{r_n(x, \Theta_m, D_n), m \geq 1\}$, where $\Theta_1, \Theta_2, \dots$ are independent and identically distributed (i.i.d.) realizations of a random variable Θ . These randomised trees are aggregated to produce the overall regression estimate represented in Equation (8).

$$\bar{r}_n(X, D_n) = \mathbb{E}_{\Theta}[r_n(X, \Theta, D_n)] \quad (8)$$

where \mathbb{E}_{Θ} denotes the expectation concerning the random parameter Θ , conditional on the input X and the training dataset D_n .

Multilayer perceptron. The Multilayer Perceptron (MLP) is a supervised neural network model that operates in a feedforward manner. It comprises an input layer and an output layer and can have any number of hidden layers that create a weight matrix. The fundamental MLP configuration typically includes a single hidden layer. Neurons within the network utilise nonlinear activation functions such as sigmoid, hyperbolic tangent, or Rectified Linear Unit (ReLU). The learning process is performed using backpropagation, employing the generalised delta rule to adjust the weight matrices. This iterative process updates the network's weights, allowing it to learn and make predictions or classifications based on the provided input.

The computations performed by each neuron in the hidden and output layers are given by the following equations extracted from [44].

$$hr = \Phi r = \rho v_1 gr + c_1 \quad (9)$$

$$gr = \varphi(v_1 h(r) + c_2) \quad (10)$$

where v_1 and v_2 are weight matrices, c_1 and c_2 are bias vectors, and ρ and φ denote the activation functions applied in the hidden and output layers, respectively. The set of parameters to be learned is given by $\theta = \{v_1, v_2, c_1, c_2\}$.

4.3. Training Stage

The training stage is used to obtain the best performance of the models. First, data needs to be split into a train and test set in a proportion of 80/20. Then, we are applying two typical ML strategies: K-fold validation and grid search. We should highlight that, as non-neural models do not involve iterative training, they do not perform cross-validation

The different training approaches are evaluated with the two metrics, a loss metric for the Autoencoders and an accuracy metric for the. For the Autoencoder, our goal is to identify the model that best reconstructs the images. For the classifier, we aim to select the one that most effectively discriminates between the two classes in the dataset: melanoma and non-melanoma.

Cross-validation has been applied only during the training of the neural classifiers. K-fold validation is a technique used to evaluate statistical analysis results, ensuring their independence from the partitioning between training and test data. In our approach, the training dataset is divided into k subsets or folds. We perform k -training processes, with each process utilising a different fold as the validation set while the remaining $k-1$ folds are used for training. During each iteration, a different fold is selected for validation until

all the folds have been used for this purpose. This process enables us to obtain multiple sets of results for the implemented metrics. To obtain more reliable estimates, we calculate the average values of the metrics over the k runs, along with the corresponding standard deviation. In our case, we divide the training set into 5 subsets, resulting in a 5-fold cross-validation. Each iteration involves training with 80% of the training set and validating with the remaining 20%. By following this methodology, we aim to achieve a reliable and unbiased evaluation of our model's performance while mitigating the risks of overfitting and chance-based results.

Grid search techniques have been applied to optimise the Autoencoder and the classifiers. This is a technique that involves exploring a range of values for hyperparameters to identify the model with the highest accuracy [45]. This approach systematically combines various parameter combinations to thoroughly search the parameter space and determine the optimal configuration for achieving optimal performance. In Table 2, we compile the hyperparameters used to train the Autoencoder, KNN, SVM, MLP, and RF.

Table 2. Hyperparameters for the models.

Model	Hyperparameter	Values
Autoencoder	Learning rate	0.001, 0.0001
	Depth	6, 7, 8
	Number of kernels	4, 8, 16
KNN	Number of neighbours	3, 5, 6, 10, 12
	Weights	Uniforms, Distance
SVM	Kernel	Poly, RBF, Sigmoid
	C	0.01, 0.1, 1.0, 10.0
MLP	Hidden layers	4, 6
	Neurons in the 1st layer	128, 256
	Epochs	50, 70
	Dropout	0.3, 0.5
	L2 reg.	1×10^{-4} , 1×10^{-3}
	Batch size	32, 64
RF	Number estimators	80, 100, 120
	Split criterion	Gini, Entropy, Log Loss
	Maximum depth	33, 66, 100

To summarise, the Autoencoders have been optimised via grid search, considering the best model to be the one that most faithfully reconstructs each input histopathological image, since the output is required to match the input. The grid search values in this step generate 64 different vectors (bottlenecks) that are combined according to the five functions of the feature combination algorithm. This operation generates 1189 input vectors fed into the six classifiers, which are fine-tuned according to different grid search strategies, generating 207 different test approaches. All of these combined generate 3105 accuracy values, which are evaluated by splitting the training and validation sets in an 80/20 ratio. In Figure 2, we schematically represent the previous process for the classifiers.

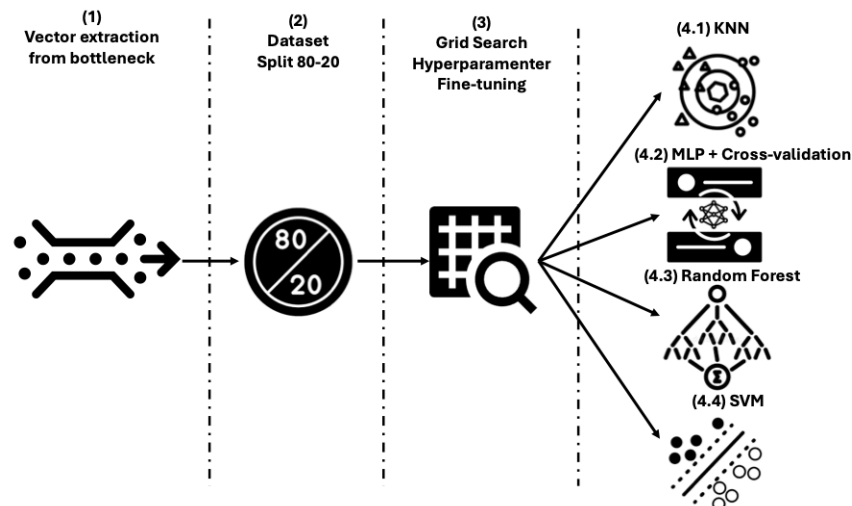


Figure 2. Description of the training process for the classifiers. The dataset is split into training and test sets (80/20). Model optimisation is performed using K-fold cross-validation and grid search, depending on the classifier.

5. Results

The proposed solution combines two types of models. In the first step, Autoencoders are used for dimensionality reduction and feature extraction of the histopathological images. As this is a regression problem, we are using an error metric like the Mean Squared Error (MSE) as its capacity to measure the performance in feature extraction [46]. MSE calculates the average of the squared differences between the predicted values (reconstructed pixel) and the actual values (pixel in the original image). Equation (6) formalised this metric.

$$\text{MSE} = \frac{1}{n} \sum_{k=0}^n (y_i - \hat{y}_i)^2 \quad (11)$$

In this context, n represents the number of data points (or examples) in the dataset. For each data point i , y_i denotes the true (actual) value, while \hat{y}_i represents the predicted value generated by the model for that same data point.

Regarding this, MSE has been used to decide which set of hyperparameters performed the best in reconstructing the images. The following are the values of this metric in training, validation and test: 0.015233, 0.015333 and 0.01474. Complementary to the metrics, we attach Figure 3 with the diagram of lines for training and validation for MSE. This indicates good generalisation, which means that the model is not overfitting and performs similarly on unseen validation data as indicated by the previous values of the metric.

The chosen Autoencoder begins with an input layer that feeds the cropped 256×256 images into the first convolutional block of the encoder, which reduces the dimensionality of the images. The encoder consists of 7 convolutional blocks whose objective is to extract a feature map from the image. Each of these blocks has a 3×3 convolutional filter with a pooling layer of size 2 and stride 1. All filters apply the ReLU function as the activation function.

The number of filters in each block starts at 8 and doubles with each block, reaching 16 in the second, 32 in the third, and culminating in 512 in the seventh block. This gradual increase leads to a continuous reduction in the dimensions of the input feature map. As this reduction progresses, the input data, originally sized 256×256 , is reduced to a size of 2×2 . During this process, the essential features of the input image are meticulously represented in this compact information, effectively capturing the essence of the histopathology.

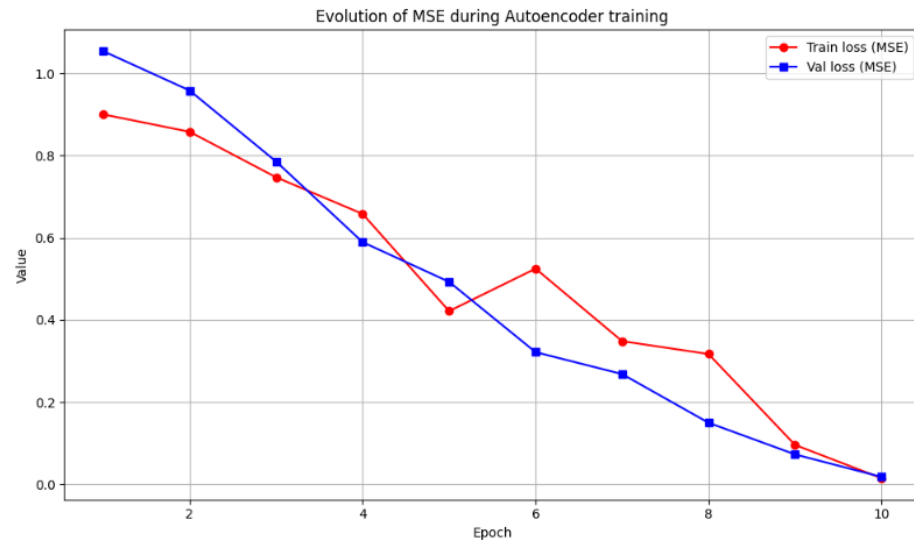


Figure 3. Evolution of Mean Squared Error (MSE) over training and test epochs for the Autoencoder model after 10 epochs.

The next part of the model is the bottleneck, which comprises a convolutional layer with 512 filters. Throughout this part, the size of the feature map remains constant.

The model then progresses to the dimensionality boosting or deconvolution stage. This phase consists of seven deconvolution blocks, reflecting the number of convolution blocks. Each block consists of one upsampling layer followed by two convolution layers. Again, the number of filters within layers of the same block remains constant but varies between each block. Specifically, there are 512 filters in each layer of the first block, 256 in the second, 128 in the third, and so on until reaching eight filters in the final block. This progress serves to amplify the dimensionality of the feature map. During this deconvolution process, starting from the compact feature map at the bottleneck stage, the initial image used as input to the model is reconstructed, using only the features extracted from the original image. Figure 4 shows a representation of the Autoencoder.

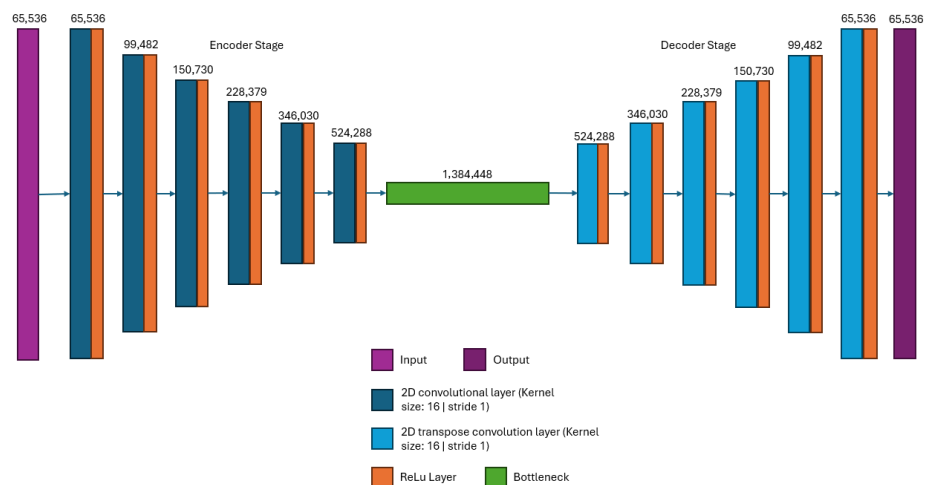


Figure 4. Autoencoder architecture for feature extraction. The model consists of an encoder that reduces the dimensionality of the input histopathological images into a compact latent representation, and a decoder that reconstructs the original image. The latent feature vectors obtained from the encoder are used as inputs for subsequent classifiers.

As said above, many combinations have been performed to obtain the best results for each of the 4 classifiers. Following, we compile the best metrics in a table by using the

accuracy, which measures the ratio of correctly classified instances to the total number of classified instances. It provides an initial indication of model performance and is often used as a starting point for evaluating the effectiveness of a model. To avoid the effects of randomness in the training and the separation between training and validation, we have obtained the mean and standard deviation of the 5-fold cross-validation. This information has been compiled in Table 3. In the case of the test stage, the value corresponds to the fold that performs best, so it does not reflect the mean or standard deviation. Non-neural methods such as KNN, SVM and RF do not report a value at the validation stage. This is because no separate validation set was used during the training of these models. These algorithms do not rely on progressive, epoch-based training and thus do not require iterative monitoring of validation performance. Results in bold indicate the best performance.

Table 3. Accuracies for the four approaches.

Approach	Training Accuracy	Validation Accuracy	Test Accuracy
KNN + mean	100.00% ± 0.00	N/A	97.95%
SVM + mean	84.96% ± 0.00	N/A	75.00%
MLP + minimum	90.17% ± 6.00	85.69% ± 14.40	76.53%
RF + sum	100.00% ± 0.00	N/A	97.95%

Following, we give the values for the best combinations of vector aggregation operations with the hyperparameters for KNN and RF.

First, we explored different methods for combining the extracted feature vectors. In this case, the mean operation is the best for KNN and the sum operation for the best performing with RF. The best-performing KNN model uses distance weights, meaning that weights point to the inverse of their distance. Additionally, it uses a K value of 3, representing the number of nearest neighbours considered during classification. For the RF approach, it uses 80 estimators or created trees in the forest, the criterion to measure the quality of the splits is Gini, and the maximum depth of the tree is set to 33.

For an in-depth analysis of the models' performance with FPs and False Negatives (FNs), we provide Figure 5. It contains the confusion matrices for the two best-performing models, KNN and RF, at the test stage.

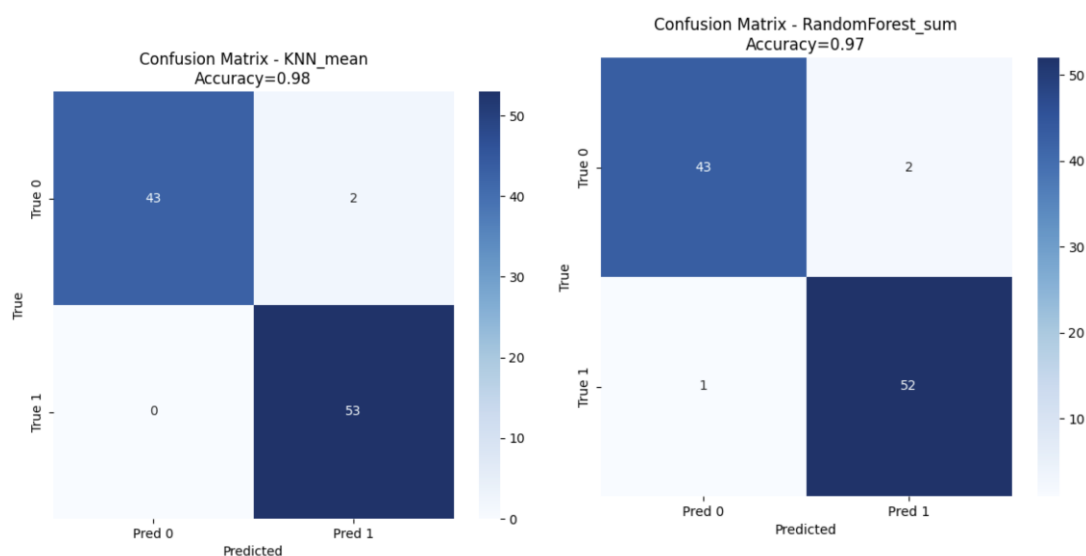


Figure 5. Confusion matrices for KNN (mean aggregation) and RF (sum aggregation) classifiers.

Accuracy is a good way to measure the performance of a model. Otherwise, in cases of an unbalanced dataset and fields like medicine, it is essential to study how the model fails, reporting FPs and FNs compiled in confusion matrices and additional metrics such as specificity and sensitivity.

Specificity represents the ratio between the number of True Negatives (TNs) (non-melanoma states correctly identified as non-melanoma) and the total number of instances predicted as negatives (including both TNs and FPs, where melanoma is misclassified as healthy). Specificity is particularly useful in avoiding situations where patients are not notified of a potential melanoma, ensuring that timely intervention and treatment can be provided when necessary.

Sensitivity measures the proportion of actual positive cases (melanomas) correctly identified by the model. It reflects the model's ability to minimise FNs, cases where melanoma is misclassified as nevi. This metric is particularly important in clinical settings to ensure that patients with melanoma are accurately diagnosed, allowing timely treatment and preventing missed diagnoses that could lead to severe consequences.

To further evaluate the balance between precision (accuracy of positive predictions) and sensitivity, we computed the F1-score, which is particularly relevant in the presence of imbalanced classes. It considers both the model's ability to correctly identify positive cases (melanomas) and its precision in avoiding misclassifying healthy cases (nevi) as melanomas. A high F1-score is critical in medical diagnosis to ensure that most melanoma cases are detected over FPs.

Considering these metrics in melanoma diagnosis provides a more comprehensive assessment of the diagnostic performance, enabling a better understanding of the model's ability to accurately identify melanoma cases while minimising FPs and FNs. The following Table shows these metrics for the best approach compiled in Table 4.

Table 4. Specificity, sensitivity and F1-score for the best approaches.

Model	Training Specificity	Test Specificity	Training Sensitivity	Test Sensitivity	Training F1-Score	Test F1-Score
KNN	100.00% ± 0.00	95.55%	100.00% ± 0.00	100.00%	100.00% ± 0.00	97.93%
RF	100.00% ± 0.00	97.77%	100.00% ± 0.00	100.00%	100.00% ± 0.00	91.75%

In addition, we calculated 95% confidence intervals (CIs) for sensitivity and specificity to quantify the uncertainty associated with our diagnostic performance. This is crucial in the context of medical imaging and AI-assisted diagnosis, where robust performance must generalise beyond the test sample. For the KNN model, the sensitivity was 100% with CI (93–100%), and the specificity was 96% with CI (85–99%), demonstrating that it consistently detects melanoma cases with very few FPs. The narrow CIs indicate high precision and reliability of these estimates. By comparison, the RF model achieved a sensitivity of 96% with CI (87–100%) but a lower specificity of 65% with CI (52–76%), highlighting a greater risk of FPs.

Finally, to complement these metrics, Receiver Operating Characteristic (ROC) curves and Precision-Recall (PR) curves are useful as they provide a threshold-independent evaluation.

The ROC Curve is a graphical representation that illustrates the diagnostic ability of a binary classifier across various threshold settings. It plots the sensitivity against the FP rate. The curve provides insight into the trade-off between correctly identifying positive cases (melanoma) and incorrectly labelling negative cases (nevi) as positive. In medical contexts, the ROC Curve is valuable for evaluating a model's performance independent of class distribution and decision threshold.

The Precision-Recall (PR) Curve is a performance evaluation tool particularly suited for binary classification tasks with imbalanced datasets. This curve plots precision against sensitivity, offering insight into the model's ability to identify positive cases (melanoma) while minimising FPs (misdiagnosed melanomas). It is especially useful when FPs can lead to unnecessary treatments, and FNs carry severe health risks.

To complete the analysis with other metrics, we provide Figures 6 and 7 that describe graphical metrics such as ROC-Curve and PR-Curve.

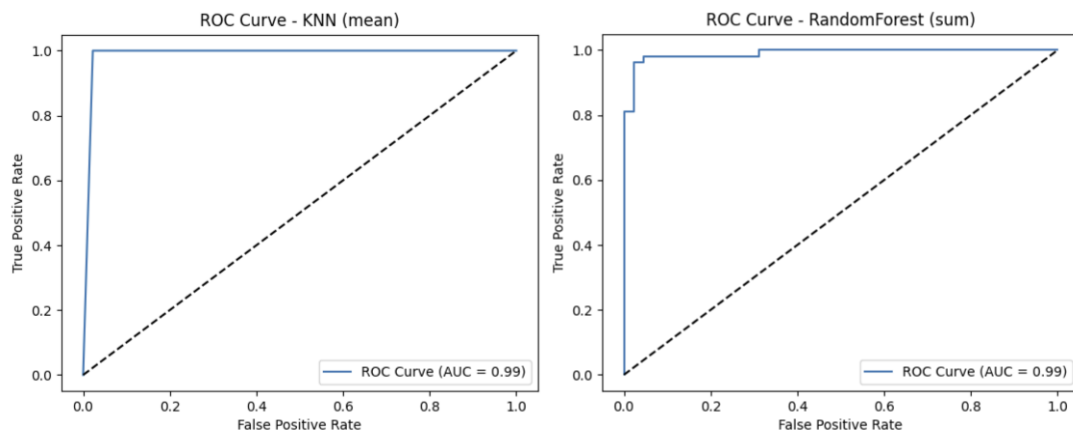


Figure 6. ROC curves for the KNN classifier using mean aggregation (left) and the RF classifier using sum aggregation (right).

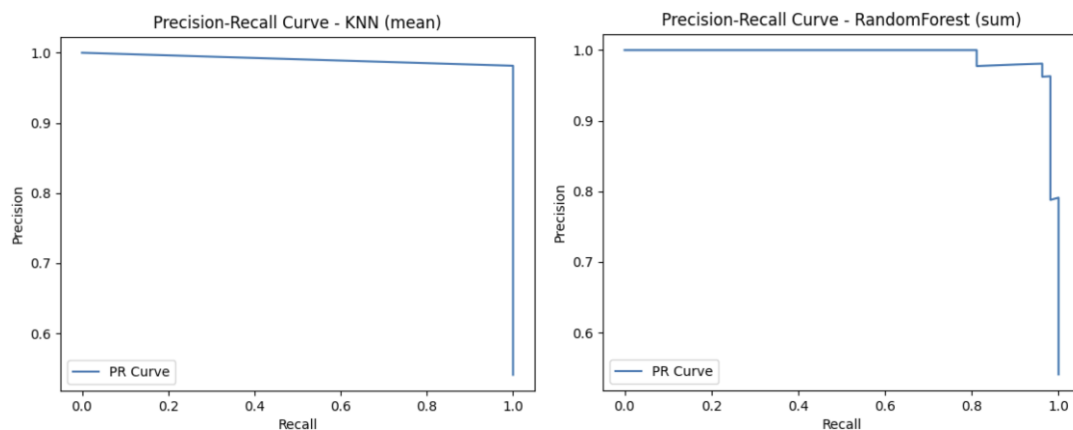


Figure 7. PR curves for the KNN classifier using mean aggregation (left) and the RF classifier using sum aggregation (right).

As mentioned above, one of the motivations for using hybrid models and extracting image features is the improved performance of Autoencoders for this task. To verify that we have compared our best approach to pretrained CNNs. This comparison has been made in terms of metric performance. We have compared our model to a well-known model used for image classification called ResNet50 [47]. Apart from that, three more baselines have been applied, which are the previous models using some of the hyperparameters obtained from the Autoencoder we have trained in this paper. This information is compiled in Table 5. Results in bold indicate that our method performs the best.

After demonstrating the good performance of our model, it is interesting to analyse its applicability in terms of diagnostic. In this way, regarding [48], the average time required to annotate WSIs that discriminate between melanoma and non-melanoma was approximately 15 min. By using our models, the average inference time per image was 3.44 min for the KNN classifier and 33.07 min for the RF model.

Table 5. Comparison with baseline.

Model	Test Accuracy
Our model	97.95%
EfficientNetB3	85.71%
DenseNet121	95.24%
ResNet50V2	84.13%
EfficientNetB3 + Autoencoder hyperparameters	63.84%
DenseNet121 + Autoencoder hyperparameters	71.31%
ResNet50V2 + Autoencoder hyperparameters	62.24%

Finally, to quantitatively evaluate the quality of our model classification, we tested the performance of our classifier by comparing it to the diagnosis made by two histopathologists (as determined with all methods and information at hand) with the test set. The physicians who performed the evaluation are MGR, CPV and JLRP, who are expert dermatopathologists in the field and assigned the class labels. Diagnosis of melanocytic lesions is based on histopathological criteria, including asymmetry, cytological atypia, maturation, pagetoid extension, mitosis and dusty pigmentation in melanocytes, which the pathologist uses to classify these lesions as benign or malignant and to establish a specific diagnosis. Since these features may become ambiguous in some cases, expertise is most valuable in others. However, our strikingly accurate models (KNN and RF models) yielded 100% sensitivity and 95.55% specificity for the KNN classifier and 97.7% for the RF model, which means two cases were FPs with the KNN model and only one with the RF model. Coincidentally, both model fails to classify the same case (a deep blue nevus) as benign. The histopathological features of the two false-positive cases are compiled in Table 6. The values range from 1 to 3, which means: 1 feature is fulfilled by nothing or very little, 2 features are fulfilled in some way, and 3 features are fulfilled by a lot or completely. In some cases, this value is NA (Not Available).

Table 6. Physicians' evaluation.

Case	Asymmetry	Atypia	Maturation	Pegatoid Extension	Mitosis	Pigment
Pigmented lentiginous nevus (FP)	0	2	NA	0	0	3
Deep blue nevus (FP)	3	0	2	0	0	3

If we describe these two cases in detail, the first was a deep blue nevus showing a proliferation of spindled pigmented melanocytes in the deep reticular dermis. In the second case, there was a basal proliferation of pigmented melanocytes. These cases are difficult to classify even for physicians, as they share common characteristics with both nevi and melanoma. Therefore, it is understandable that the model could misclassify them. A potential solution to this problem is to increase the number of similar cases in the dataset, which could help improve the model's ability to distinguish between these challenging examples.

To better illustrate how the model diagnoses, different pictures of the cases corresponding respectively to TNs, TPs and FPs are shown in Figures 8–10. Figure 8 shows several examples of TN cases, including a case of dermal melanocytic nevus (Figure 8A), a Reed nevus (Figure 8B), a Spitz nevus (Figure 8C) and a dysplastic nevus (Figure 8D). Figure 9 shows two examples of True Positive (TP) cases, including a spreading superfi-

cial melanoma of 3.79mm of Breslow thickness with balloon degeneration (case A) and a superficial spreading melanoma, ulcerated of 5.5 mm of Breslow thickness (case B).

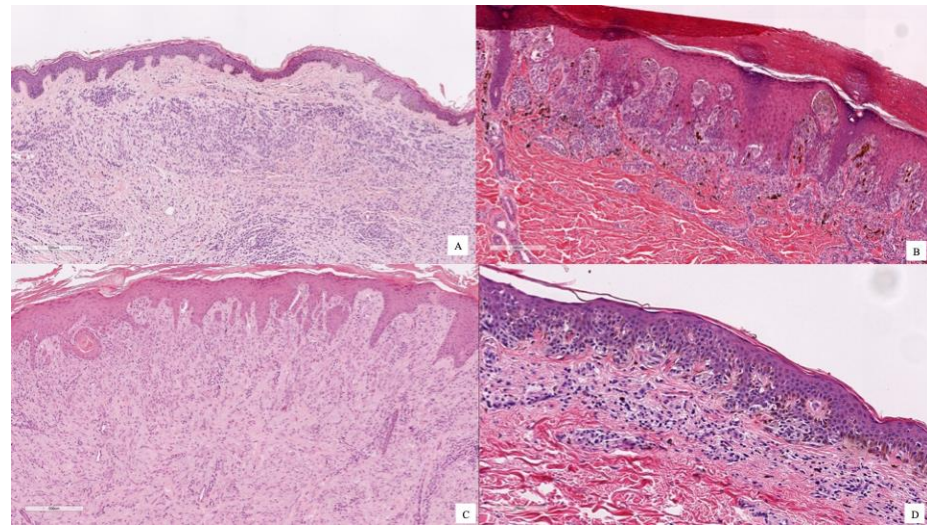


Figure 8. Representative TN cases illustrating different benign skin lesions: (A) dermal melanocytic nevus, (B) Reed nevus, (C) sclerosing Spitz nevus, and (D) dysplastic nevus. These examples highlight the variability in appearance among non-malignant lesions that may resemble melanoma.

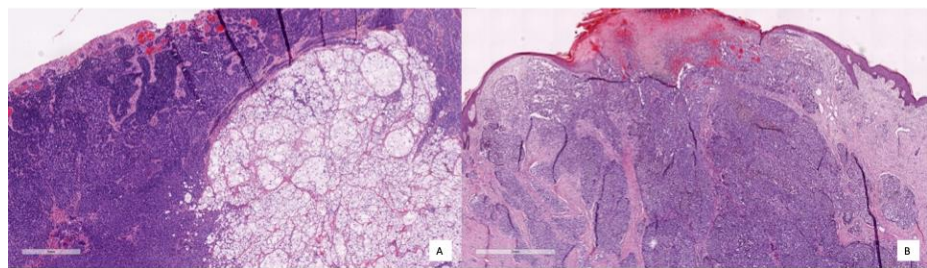


Figure 9. Representative TN cases: (A) superficial spreading melanoma with balloon degeneration, and (B) ulcerated superficial spreading melanoma. These examples illustrate the morphological diversity of malignant lesions.

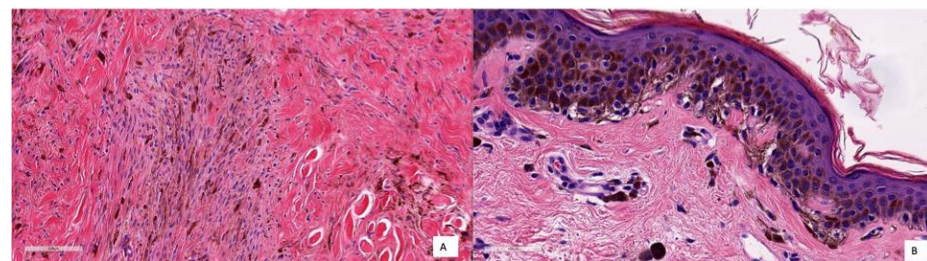


Figure 10. Representative FP cases: (A) deep cellular blue nevus, characterised by dermal proliferation of spindle, fusiform, or ovoid cells with associated melanin pigment (present both within melanocytic tumour cells and macrophages), and stromal sclerosis; (B) lentiginous nevus, exhibiting basal melanocytic proliferation of pigmented melanocytes.

As said earlier, only two cases were classified as FP with the KNN model and only one with the RF model. Coincidentally, both models failed to classify the same case (a deep blue nevus) as benign. The histopathological features of the two FP cases are compiled in Table 5, and the images are shown in Figure 10. In Case A, the deep infiltration of the pigmented melanocytes and the spindled shape may be misinterpreted; in Case B, there is a basal proliferation of pigmented melanocytes.

6. Discussion

As can be seen in Table 3, the MLP and SVM classifiers exhibited comparatively lower accuracy. This can be attributed to the nature of the extracted features and the dataset limitations. Although the Autoencoder reduced the dimensionality of the histopathological images, the resulting feature vectors remain relatively high-dimensional, while the number of instances is small, a scenario that often leads to overfitting, as reported respectively for SVM and MLP in [49,50]. Analysing the rest of the results, the best classifiers correspond to the KNN and RF approaches, which are the ones we are considering as our final model to do diagnosis.

Regarding the performance of the model, if we look at the accuracy metric, the model accomplishes the bias-variance trade-off [51]. Some papers establish the accuracy of professionals diagnosing melanoma with histopathologies between 59 and 80%, depending on the experience [52–54]. In our case, this value is about 97.95%, which means an improvement of 17 points in the worst case. In terms of variance, the values could be considered good enough less than 3%.

Although accuracy is a good metric to obtain a first evaluation of how the model performs, an in-depth analysis can be obtained from confusion matrices and the other metrics: specificity, sensitivity and F1-Score. In this way, in Figure 5, we can see that the KNN model has two FPs and no FNs, and the RF has one FN and two FPs, demonstrating balanced sensitivity and specificity.

To complement the performance of the models, we obtained Table 4. In this case, the values in the test are very good for sensitivity, with no errors, which is good in the case of medical diagnosis, as no patient with cancer is considered healthy. Another interpretation can be obtained with the lower value that corresponds to the specificity. These problems with specificity could lead to a situation in which a healthy person could be diagnosed as having melanoma. Although this problem is not the most serious, it entails spending money on treatments that should not be applied and some consequences on the physical and mental health of the person receiving the treatment. The high F1-scores align with the perfect test sensitivity (100%) but are slightly lower than 100% because the test specificity is not perfect (95.55% for KNN and 97.77% for RF). In practice, a near-perfect F1-score like this suggests the model is highly effective at minimising both missed diagnoses (FNs) and unnecessary interventions (FPs), which is critical for patient safety, early treatment, and reducing undue anxiety in melanoma screening.

Regarding the CIs, they provide a range of plausible values for each metric, enhancing the reliability and generalisability of the results and supporting the conclusion that KNN offers a more balanced and trustworthy approach for melanoma detection in this context.

In summary, our model achieved exceptionally good performance in distinguishing melanomas and nevi with our set of cases, with a sensitivity of 100%, a specificity of 95–97% and an F1-Score from 91 to 97. These results outperform other ancillary tools which are frequently used by dermatopathologists for that purpose, such as fluorescence in situ hybridisation with a sensitivity and specificity of 87% and 96%, respectively [55] or comparative genomes hybridisation [9].

A final measure of the performance of the models can be carried out with Figures 6 and 7, with the ROC and PR Curves. By analysing Figure 6, we can see that both models achieved an AUC curve of 0.99, indicating excellent discriminative performance. The KNN classifier exhibits a nearly perfect curve, suggesting minimal FPs and strong sensitivity. Similarly, the RF model maintains a very high TN rate and a low FP rate across thresholds. The proximity of both ROC curves to the top-left corner reflects a high ability to distinguish between classes (melanoma vs. nevi). An AUC of 0.99 implies that there is a 99% chance that the model will correctly rank a randomly chosen positive instance higher than a

randomly chosen negative one. Overall, these results demonstrate that both models perform robustly and reliably in terms of classification capability. In Figure 7, both models demonstrate strong performance, maintaining high precision across nearly the entire range of sensitivity values. The KNN model exhibits a nearly flat PR curve, indicating that precision remains consistently close to 1 even as recall increases, suggesting that FPs are minimal across different thresholds. Similarly, the RF model also maintains high precision, with a slightly more variable curve toward higher recall levels, yet still preserving strong discriminative ability.

Compared with the baselines in the table, our model offers the opportunity for better quantitative modelling of disease appearance with a lower amount of input data and outperforms other CNN models such as ResNet, EfficientNet and DenseNet in the classification of melanoma and nevi. Even tuning these baselines with the optimal hyperparameters of the Autoencoder used in this paper, the results of our model are much better.

Considering all the results described above, we can conclude that the KNN model is the best choice, as it performs better in all metrics (accuracy, sensitivity, and F1-score) except for specificity, which is less problematic in this context. Additionally, the analysis of confusion matrices and confidence intervals supports its trustworthiness. Finally, the inference time further reinforces KNN as the most suitable solution.

However, some small limitations remain: the model is currently binary (nevus vs. melanoma) and does not support broader differential diagnoses; it operates on cropped image sections rather than full WSI, which may omit contextual information; and its performance has yet to be validated on external datasets or through prospective clinical studies. Apart from that, the key limitation of this study is the relatively small size of 14 melanoma cases in the test set, which affects the robustness of the sensitivity estimates. Such a small sample size may limit the generalisability of the findings. Therefore, caution is warranted when interpreting these results, and further validation on larger, independent, and more diverse cohorts of histopathological melanoma images is needed to confirm the diagnostic accuracy and clinical applicability of the proposed solution.

Another key consideration for the clinical translation of this type of diagnostic model is the need to address regulatory hurdles, system integration, and clinician training. From a regulatory perspective, AI-based diagnostic tools must undergo thorough validation and certification before being adopted in routine clinical workflows. Ensuring compliance with relevant medical device regulations and obtaining approval from health authorities is a necessary step before clinical deployment. Additionally, the integration of these tools into hospital information systems poses significant technical challenges, as compatibility with electronic health records and existing pathology software must be ensured. Seamless integration is crucial to allow clinicians to access AI-supported diagnostic results within their established workflows. Finally, effective clinician training is essential to maximise the benefits of AI systems. Healthcare professionals must be familiar with the capabilities and limitations of these models, and training programs should be designed to help them interpret AI outputs and incorporate them into their decision-making process. Without adequate training, there is a risk of over-reliance or misuse of AI-generated results. Addressing these regulatory, technical, and educational aspects will be critical for the successful adoption and clinical impact of the proposed diagnostic framework.

7. Conclusions and Future Works

The main aim of this work is the implementation of different hybrid approaches that use DL models and classical ML techniques for the diagnosis of melanoma using histopathological images. Apart from obtaining a model that performs this task accurately

and improves the capacities of professionals in the field, we have provided a subjective evaluation that allows us to better understand model hits and misses in certain cases.

The workflow developed in this work comprises the following steps. First, images are cropped into smaller ones to avoid problems of computational capacity derived from the large size of the initial images. This stage also comprises the application of the WSI normalisation algorithm, a particular normalisation method developed particularly for managing histopathological images. After this, we applied a three-stage ML process for the diagnosis of the histopathologies. In the first stage, Autoencoders are used to obtain the most representative features of the images that allow for better discrimination between melanomas and healthy ones. Then, these feature vectors are aggregated by calculating the minimum of their values so they can be introduced into the ML algorithms. Even though different solutions were proposed, the two obtaining the best results were KNN and the RF algorithm. In this case, they obtain an accuracy of about 97.5%, which improves human performance by 9 points. This information has been complemented with other metrics, the calculation of CIs and the inference time. Regarding this, we can conclude that the best solution is the one using KNN

In future work, researchers aim to enhance the present diagnostic method. First, it is necessary to obtain a larger and more balanced dataset to mitigate the class imbalance issue. This will help determine whether current limitations are due to the dataset itself and may also offer insights for clinicians to improve disease diagnosis. Additionally, we propose implementing algorithms that replicate common tasks performed by dermatologists, such as counting mitoses and measuring the distance between the melanoma and the epithelial zone. Other complementary approaches—including the use of dual staining techniques or incorporating molecular analysis in complex cases—could also be explored to improve diagnostic accuracy. Lastly, to better understand how the proposed methods may fail in diagnosis, studies that provide a deeper analysis from a medical perspective are needed. In this way, the following research will include interobserver concordance comparing the performance of an AI classifying model with pathologists' performance to highlight how digital pathology has the potential to augment the pathologist's eye with objective information that cannot be gleaned by human examination.

Author Contributions: Conceptualization, A.N., M.C.G. and Á.J.G.-T.; methodology, A.N. and Á.J.G.-T.; software, D.D.-P. and A.G.; validation, A.N., M.C.G., C.P.V. and J.-L.R.-P.; formal analysis, A.N., M.C.G. and J.-L.R.-P.; investigation, A.N., M.C.G. and Á.J.G.-T.; resources, Á.J.G.-T.; data curation, M.C.G., C.P.V. and J.-L.R.-P.; writing—original draft preparation, A.N.; writing—review and editing, A.N. and M.C.G.; visualization, A.G.; supervision, A.N., M.C.G. and Á.J.G.-T.; project administration, Á.J.G.-T.; funding acquisition, A.N. and Á.J.G.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universidad Francisco de Vitoria internal research funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code for the project is publicly available at the following GitHub repository: <https://github.com/ufvceiec/melanoma>, accessed on 2 June 2025. The repository includes all scripts, notebooks, configuration files, and dependencies needed to replicate the experiments and results.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kaya, S.I.; Ozcelikay, G.; Mollarasouli, F.; Bakirhan, N.K.; Ozkan, S.A. Recent achievements and challenges on nanomaterial based electrochemical biosensors for the detection of colon and lung cancer biomarkers. *Sens. Actuators B Chem.* **2022**, *351*, 130856. [[CrossRef](#)]
2. Woo, Y.R.; Cho, S.H.; Lee, J.D.; Kim, H.S. The human microbiota and skin cancer. *Int. J. Mol. Sci.* **2022**, *23*, 1813. [[CrossRef](#)]
3. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2019. *CA Cancer J. Clin.* **2019**, *69*, 7–34. [[CrossRef](#)] [[PubMed](#)]
4. Arnold, M.; Singh, D.; Laversanne, M.; Vignat, J.; Vaccarella, S.; Meheus, F.; Cust, A.E.; de Vries, E.; Whitman, D.C.; Bray, F. Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatol.* **2022**, *158*, 495–503. [[CrossRef](#)] [[PubMed](#)]
5. Carli, P.; De Giorgi, V.; Palli, D.; Maurichi, A.; Mulas, P.; Orlandi, C.; Imberti, G.L.; Stanganelli, I.; Soma, P.; Dioguardi, D.; et al. Dermatologist detection and skin self-examination are associated with thinner melanomas: Results from a survey of the Italian Multidisciplinary Group on Melanoma. *Arch Dermatol.* **2003**, *139*, 607–612. [[CrossRef](#)]
6. Swetter, S.M.; Johnson, T.M.; Miller, D.R.; Layton, C.J.; Brooks, K.R.; Geller, A.C. Melanoma in middle-aged and older men: A multi-institutional survey study of factors related to tumor thickness. *Arch Dermatol.* **2009**, *145*, 397–404. [[CrossRef](#)]
7. Green, A.C.; Baade, P.; Coory, M.; Aitken, J.F.; Smithers, M. Population-based 20-year survival among people diagnosed with thin melanomas in Queensland, Australia. *J. Clin. Oncol.* **2012**, *30*, 1462–1467. [[CrossRef](#)]
8. Shoo, B.A.; Sagebiel, R.W.; Kashani-Sabet, M. Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center. *J. Am. Acad. Dermatol.* **2010**, *62*, 751–756. [[CrossRef](#)]
9. Bastian, B.C.; Olshen, A.B.; LeBoit, P.E.; Pinkel, D. Classifying melanocytic tumors based on DNA copy number changes. *Am. J. Pathol.* **2003**, *163*, 1765–1770. [[CrossRef](#)]
10. Samuel, A.L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* **1959**, *3*, 210–229. [[CrossRef](#)]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
12. Thepade, S.D.; Ramnani, G. Haar Wavelet Pyramid-Based Melanoma Skin Cancer Identification with Ensemble of Machine Learning Algorithms. *Int. J. Healthc. Inf. Syst. Inform.* **2021**, *16*, 1–15. [[CrossRef](#)]
13. Murugan, A.; Nair, S.A.H.; Kumar, K.P.S. Detection of skin cancer using SVM, random forest and kNN classifiers. *J. Med. Syst.* **2019**, *43*, 269. [[CrossRef](#)]
14. Liu, C.-H.; Qi, J.; Lu, J.; Wang, S.; Wu, C.; Shih, W.-C.; Larin, K.V. Improvement of tissue analysis and classification using optical coherence tomography combined with Raman spectroscopy. *J. Innov. Opt. Health Sci.* **2015**, *8*, 1550006. [[CrossRef](#)]
15. Silver, F.H.; Mesica, A.; Gonzalez-Mercedes, M.; Deshmukh, T. Identification of Cancerous Skin Lesions Using Vibrational Optical Coherence Tomography (VOCT): Use of VOCT in Conjunction with Machine Learning to Diagnose Skin Cancer Remotely Using Telemedicine. *Cancers* **2023**, *15*, 156. [[CrossRef](#)]
16. Mishra, R.; Meher, S.; Kustha, N.; Pradhan, T. A skin cancer image detection interface tool using vlf support vector machine classification. In *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2021*; Springer: Singapore, 2022; pp. 49–63.
17. Brorsen, L.F.; McKenzie, J.S.; Tullin, M.F.; Bendtsen, K.M.S.; Pinto, F.E.; Jensen, H.E.; Haedersdal, M.; Takats, Z.; Janfelt, C.; Lerche, C.M. Cutaneous squamous cell carcinoma characterized by MALDI mass spectrometry imaging in combination with machine learning. *Sci. Rep.* **2024**, *14*, 11091. [[CrossRef](#)]
18. Xie, J.; Zhang, M.; Qi, M. Integrating Machine Learning Algorithms to Construct a Triaptois-Related Prognostic Model in Melanoma. *Cancer Manag. Res.* **2025**, *17*, 1127–1141. [[CrossRef](#)]
19. Brown, K.; Farmer, A.; Gurung, S.; Baker, M.J.; Board, R.; Hunt, N.T. Machine-learning based classification of 2D-IR liquid biopsies enables stratification of melanoma relapse risk. *Chem. Sci.* **2025**, *16*, 8394–8404. [[CrossRef](#)]
20. Mosquera-Zamudio, A.; Launet, L.; Colomer, A.; Wiedemeyer, K.; López-Takegami, J.C.; Palma, L.F.; Undersrud, E.; Janssen, E.; Brenn, T.; Naranjo, V.; et al. Histological interpretation of spitzoid tumours: An extensive machine learning—Based concordance analysis for improving decision making. *Histopathology* **2024**, *85*, 155–170. [[CrossRef](#)]
21. Gouda, W.; Sama, N.U.; Al-Waakid, G.; Humayun, M.; Jhanjhi, N.Z. Detection of skin cancer based on skin lesion images using deep learning. *Healthcare* **2022**, *10*, 1183. [[CrossRef](#)] [[PubMed](#)]
22. Kousis, I.; Perikos, I.; Hatzilygeroudis, I.; Virvou, M. Deep learning methods for accurate skin cancer recognition and mobile application. *Electronics* **2022**, *11*, 1294. [[CrossRef](#)]
23. Shorfuzzaman, M. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimed. Syst.* **2022**, *28*, 1309–1323. [[CrossRef](#)]
24. Wako, B.D.; Dese, K.; Ulfata, R.E.; Nigatu, T.A.; Turunbedu, S.K.; Kwa, T. Squamous Cell Carcinoma of Skin Cancer Margin Classification from Digital Histopathology Images Using Deep Learning. *Cancer Control* **2022**, *29*, 10732748221132528. [[CrossRef](#)]
25. Parajuli, M.; Shaban, M.; Phung, T.L. Automated differentiation of skin melanocytes from keratinocytes in high-resolution histopathology images using a weakly-supervised deep-learning framework. *Int. J. Imaging Syst. Technol.* **2023**, *33*, 262–275. [[CrossRef](#)]
26. Kiran, A.; Narayanasamy, N.; Ramesh, J.V.N.; Ahmad, M.W. A novel deep learning framework for accurate melanoma diagnosis integrating imaging and genomic data for improved patient outcomes. *Ski. Res. Technol.* **2024**, *30*, e13770. [[CrossRef](#)] [[PubMed](#)]

27. Chu, Y.; Lee, S.; Oh, B.; Yang, S. Class-Agnostic Feature-Learning-Based Deep-Learning Model for Robust Melanoma Prediction. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 4946–4955. [[CrossRef](#)]
28. Sabazade, S.; Michalski, M.A.L.; Bartoszek, J.; Fili, M.; Holmström, M.; Stållhammar, G. Development and Validation of a Deep Learning Algorithm for Differentiation of Choroidal Nevi from Small Melanoma in Fundus Photographs. *Ophthalmol. Sci.* **2025**, *5*, 100613. [[CrossRef](#)] [[PubMed](#)]
29. Zhao, Z.; Alzubaidi, L.; Zhang, J.; Duan, Y.; Gu, Y. A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Syst. Appl.* **2024**, *242*, 122807. [[CrossRef](#)]
30. Nawaz, M.; Mehmood, Z.; Nazir, T.; Naqvi, R.A.; Rehman, A.; Iqbal, M.; Saba, T. Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. *Microsc. Res. Technol.* **2022**, *85*, 339–351. [[CrossRef](#)]
31. Bassel, A.; Abdulkareem, A.B.; Alyasseri, Z.A.A.; Sani, N.S.; Mohammed, H.J. Automatic malignant and benign skin cancer classification using a hybrid deep learning approach. *Diagnostics* **2022**, *12*, 2472. [[CrossRef](#)]
32. Naeem, A.; Anees, T.; Fiza, M.; Naqvi, R.A.; Lee, S.-W. SCDNet: A Deep Learning-Based Framework for the Multiclassification of Skin Cancer Using Dermoscopy Images. *Sensors* **2022**, *22*, 5652. [[CrossRef](#)] [[PubMed](#)]
33. Farea, E.; Saleh, R.A.; AbuAlkebash, H.; Farea, A.A.; Al-Antari, M.A. A hybrid deep learning skin cancer prediction framework. *Eng. Sci. Technol. Int. J.* **2024**, *57*, 101818. [[CrossRef](#)]
34. De, A.; Mishra, N.; Chang, H.-T. An approach to the dermatological classification of histopathological skin images using a hybridized CNN-DenseNet model. *PeerJ Comput. Sci.* **2024**, *10*, e1884. [[CrossRef](#)]
35. Zheng, Y.; Jiang, Z.; Zhang, H.; Xie, F.; Shi, J.; Xue, C. Adaptive color deconvolution for histological WSI normalization. *Comput. Methods Programs Biomed.* **2019**, *170*, 107–120. [[CrossRef](#)]
36. Berahmand, K.; Daneshfar, F.; Salehi, E.S.; Li, Y.; Xu, Y. Autoencoders and their applications in machine learning: A survey. *Artif. Intell. Rev.* **2024**, *57*, 28. [[CrossRef](#)]
37. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504–507. [[CrossRef](#)] [[PubMed](#)]
38. Halder, R.K.; Uddin, M.N.; Uddin, A.; Aryal, S.; Khraisat, A. Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *J. Big Data* **2024**, *11*, 113. [[CrossRef](#)]
39. Fix, E.; Hodges, J.L., Jr. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev.* **1989**, *57*, 238–247. [[CrossRef](#)]
40. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, New York, NY, USA, 27–29 July 1992; pp. 144–152.
41. Srivastava, S.K.; Singh, S.K.; Suri, J.S. Healthcare Text Classification System and its Performance Evaluation: A Source of Better Intelligence by Characterizing Healthcare Text. *J. Med. Syst.* **2018**, *42*, 97. [[CrossRef](#)]
42. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
43. Biau, G. Analysis of a Random Forests Model. *J. Mach. Learn. Res.* **2012**, *13*, 1063–1095.
44. Abinaya, S.; Devi, M.K.K. Enhancing crop productivity through autoencoder-based disease detection and context-aware remedy recommendation system. In *Application of Machine Learning in Agriculture*; Academic Press: Amsterdam, The Netherlands, 2022; pp. 239–262. [[CrossRef](#)]
45. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
46. Meng, Q.; Catchpoole, D.; Skillicom, D.; Kennedy, P.J. Relational autoencoder for feature extraction. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 364–371. [[CrossRef](#)]
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Li, M.; Abe, M.; Nakano, S.; Tsuneki, M. Deep Learning Approach to Classify Cutaneous Melanoma in a Whole Slide Image. *Cancers* **2023**, *15*, 1907. [[CrossRef](#)] [[PubMed](#)]
49. Chou, E.P.; Ko, T.-W. Dimension Reduction of High-Dimensional Datasets Based on Stepwise SVM. *arXiv* **2017**. [[CrossRef](#)]
50. Singh, T.P.; Gupta, S.; Garg, M. Machine Learning: A Review on Supervised Classification Algorithms and their Applications to Optical Character Recognition in Indic Scripts. *ECS Trans.* **2022**, *107*, 6233–6250. [[CrossRef](#)]
51. Belkin, M.; Hsu, D.; Ma, S.; Mandal, S. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15849–15854. [[CrossRef](#)] [[PubMed](#)]
52. Hekler, A.; Utikal, J.S.; Enk, A.H.; Solass, W.; Schmitt, M.; Klode, J.; Schadendorf, D.; Sondermann, W.; Franklin, C.; Bestvater, F.; et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur. J. Cancer* **2019**, *118*, 91–96. [[CrossRef](#)]
53. Phillips, M.; Marsden, H.; Jaffe, W.; Matin, R.N.; Wali, G.N.; Greenhalgh, J.; McGrath, E.; James, R.; Ladoyanni, E.; Bewley, A.; et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw. Open* **2019**, *2*, e1913436. [[CrossRef](#)] [[PubMed](#)]

54. Morton, C.A.; Mackie, R.M. Clinical accuracy of the diagnosis of cutaneous malignant melanoma. *Br. J. Dermatol.* **1998**, *138*, 283–287. [[CrossRef](#)]
55. North, J.P.; Garrido, M.C.; Kolaitis, N.A.; LeBoit, P.E.; McCalmont, T.H.; Bastian, B.C. Fluorescence In Situ Hybridization as an Ancillary Tool in the Diagnosis of Ambiguous Melanocytic Neoplasms. *Am. J. Surg. Pathol.* **2014**, *38*, 824–831. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.