Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 33, Num. 2 (2026) 193-201

RESEARCH ARTICLE

HEVA: Integrating ViT and ResNet50 with ABCD Clinical Features for Skin Cancer Diagnosis

HEVA: Integrando ViT e ResNet50 com recursos clínicos ABCD para diagnóstico de câncer de pele

Antônio Marcio Crepaldi Junior[1]*, Bianca Aparecida Andrade[2], Ademar Takeo Akabane[2]

Abstract: Melanoma is the deadliest skin cancer, causing over 90% of skin tumor deaths. Early detection ensures survival above 95%, but below 15% in late stages. We propose the HEVA architecture, which integrates ensemble deep learning with ABCD descriptors and a SegFormer-B5 model, for accurate lesion masks. HEVA achieved 93.2% accuracy and 84.2% F1-score on HAM-10000, and 87.7% accuracy with 82.01% macro F1 on ISIC 2018 (1,512 images). SegFormer achieved 82.6% mean IoU and 91.2% accuracy in segmentation.
Keywords: skin cancer — melanoma — vision transformers — ensemble learning — computer-aided diagnosis

Resumo: O melanoma é o câncer de pele mais letal, responsável por mais de 90% das mortes por tumores cutâneos. A detecção precoce garante sobrevida acima de 95%, mas cai para menos de 15% em estágios avançados. Este trabalho apresenta a arquitetura HEVA, que integra *ensemble* de deep learning com descritores ABCD, apoiada por um modelo SegFormer-B5 para geração de máscaras de lesões de alta qualidade. No HAM-10000, a HEVA alcançou 93,2% de acurácia e 84,2% de F1-score, enquanto no ISIC 2018 (1,512 imagens) obteve 87,7% de acurácia e 82,01% de F1 macro. O SegFormer, isoladamente, alcançou IoU médio de 82,6% e acurácia de 91,2% na segmentação.
Palavras-Chave: câncer de pele — melanoma — vision transformers — ensemble learning — diagnóstico auxiliado por computador

[1] *Centro de Pesquisa e Desenvolvimento em Telecomunicações (CPQD), Brazil*
[2] *Pontifícia Universidade Católica de Campinas (PUC-Campinas), Brazil*
***Corresponding author**: acrepaldi@cpqd.com.br
DOI: http://dx.doi.org/10.22456/2175-2745.150917 • **Received:** 14/10/2025 • **Accepted:** 28/10/2025

1. Introduction

According to the National Cancer Institute (INCA), an estimated 704,000 new cancer cases are expected in Brazil each year of the 2023-2025 triennium [1]. The majority of these cases are concentrated in the South and Southeast regions, which account for approximately 70% of the national incidence [2]. Skin cancer is the most common neoplasm in the country, representing about 30% of all cancer diagnoses. There are different types of skin cancer, including melanoma, basal cell carcinoma, and squamous cell carcinoma, among others. Among these, melanoma is considered the most aggressive because, although it represents less than 5% of skin tumors, it is responsible for more than 90% of deaths related to this type of cancer [3, 4]. Thus, early detection of melanoma is fundamental for treatment success, as the survival rate exceeds 95% when the disease is identified in the initial stages, but drops to less than 15% in advanced stages [5].

Typically, when skin cancer is suspected, the first diagnostic step is a visual inspection by a specialist. An important point to highlight in this stage is the difficulty of distinguishing between melanoma and melanocytic nevi (moles), especially in the early stages of the disease. The second step is a dermatoscopic analysis. In this stage, dermatoscopic images are analyzed to increase the chance of early detection of malignant melanomas. Studies indicate that the accuracy of unaided visual inspection is approximately 60% [6, 7], increasing to 75% with the aid of dermatoscopy [8, 7, 4].

Computational techniques based on Deep Learning (DL) models have shown great potential in the early detection of skin lesions [7, 4]. The use of DL allows for the development of interpretable models capable of analyzing medical information quickly and accurately, identifying relevant patterns and details. These computational advances contribute significantly to improving the diagnosis and treatment of skin lesions by healthcare professionals [5].

In this context, this work presents the HEVA (Hybrid

Ensemble Vision Architecture) for the classification of skin lesions. HEVA adopts an ensemble learning-based strategy, combining different deep learning (DL) approaches to increase the model's robustness and precision. Specifically, HEVA integrates (i) DL models, such as Vision Transformers and ResNet50, and (ii) ABCD descriptors of the skin (Asymmetry, Borders, Color, and Diameter), exploring the complementarity between these techniques. This strategy aims to improve the model's robustness, precision, and interpretability by exploring the complementarity between DL-based methods and traditional computer vision techniques. Additionally, we trained a semantic segmentation model (SegFormer-B5, pretrained) to produce high-quality lesion masks. These masks are crucial for our preprocessing pipeline and handcrafted descriptors, and they enabled a robust cross-dataset evaluation on ISIC 2018 test dataset.

The main contributions of this work are:

- New distribution of the HAM-10000 dataset classes
- Proposal of HEVA, a hybrid ensemble vision architecture for skin lesion classification;
- Integration of multiple DL models (Vision Transformers and ResNet50) within an ensemble learning strategy;
- Incorporation of handcrafted dermatological descriptors (ABCD: Asymmetry, Borders, Color, Diameter) to complement DL-based feature extraction.

The remainder of this paper is organized as follows: Section 2 presents a review of related works. Then, Section 3 presents the HEVA architecture and also describes the methodological steps carried out in this work, including lesion segmentation with SegFormer-B5. Section 4 presents the experimental results obtained and discussions, including the cross-dataset evaluation on ISIC 2018. Finally, Section 5 presents the final considerations and future work.

2. Related Work

This section summarizes recent DL-based approaches for skin lesion classification.

In [9], a CNN-based model is proposed with preprocessing (image resizing, data augmentation), global feature extraction, and a comparative evaluation of algorithms on HAM-10000, achieving 91.18% accuracy.

In [10], a hybrid feature-extraction strategy combines 17 pre-trained CNNs as feature extractors with 24 ML classifiers. On ISIC 2019 and PH2, the best results come from DenseNet201 features classified by Fine KNN and Cubic SVM, reaching 92.34% accuracy.

The work in [11] introduces a multimodal hybrid DL model that fuses dermatoscopic images, clinical images, and patient metadata with a soft attention module to improve accuracy and interpretability. Evaluated on the Seven-Point Criteria Evaluation dataset, it attains 83.04% mean accuracy for multi-label classification.

Distinct from multimodal-attention pipelines, our HEVA adopts an ensemble learning strategy that integrates Vision Transformers and ResNet50 with ABCD skin descriptors, leveraging the complementarity between learned representations and interpretable dermatological features to enhance robustness and accuracy. Additionally, we train a SegFormer-based segmentation model to obtain reliable lesion masks that generalize across datasets.

3. Methods

The HEVA (Hybrid Ensemble Vision Architecture), shown in Figure 1, proposes a hybrid approach for classifying skin lesions by combining two complementary deep learning paradigms: (i) deep learning models (Vision Transformers and ResNet50) and (ii) specialized handcrafted descriptors. This strategy aims to capture information at multiple levels of abstraction, enhancing the robustness and interpretability of the classification.

HEVA is based on two deep learning models for extracting discriminative representations. The first is the Vision Transformer (ViT), refined by fine-tuning on our dataset, used to capture global relationships in the image through the CLS (Class Token) token, resulting in a dense 768-dimensional representation. The second, ResNet50, also adjusted by partial fine-tuning, extracts relevant local patterns through its convolutional layers, complemented by a dense layer of 512 neurons. Both ViT and ResNet50 were chosen because they are well-established and widely validated models in the computer vision community, ensuring reliable feature extraction and strong generalization performance.

In addition to feature extraction by DL models, HEVA incorporates a rich set of specialized descriptors aligned with the clinical criteria of the ABCD method for analyzing skin lesions. These descriptors include:

- **Texture:** LBP (Local Binary Patterns), which analyzes the neighborhood of each pixel to identify local textures; GLCM (Gray-Level Co-occurrence Matrix), which captures statistical relationships between neighboring pixels; and Wavelet transforms, which decompose the image into different scales for multiresolution analysis—all essential for capturing irregularities and microstructural patterns;
- **Shape:** HOG (Histogram of Oriented Gradients), which characterizes the distribution and orientation of edges in the image, as well as morphological, fractal, and radiomic descriptors that quantify geometric complexity, boundary irregularities, and spatial heterogeneity—providing a deeper structural understanding of the lesion's morphology;
- **Color:** Histograms in the HSV (Hue-Saturation-Value) space, a color representation system that separates color

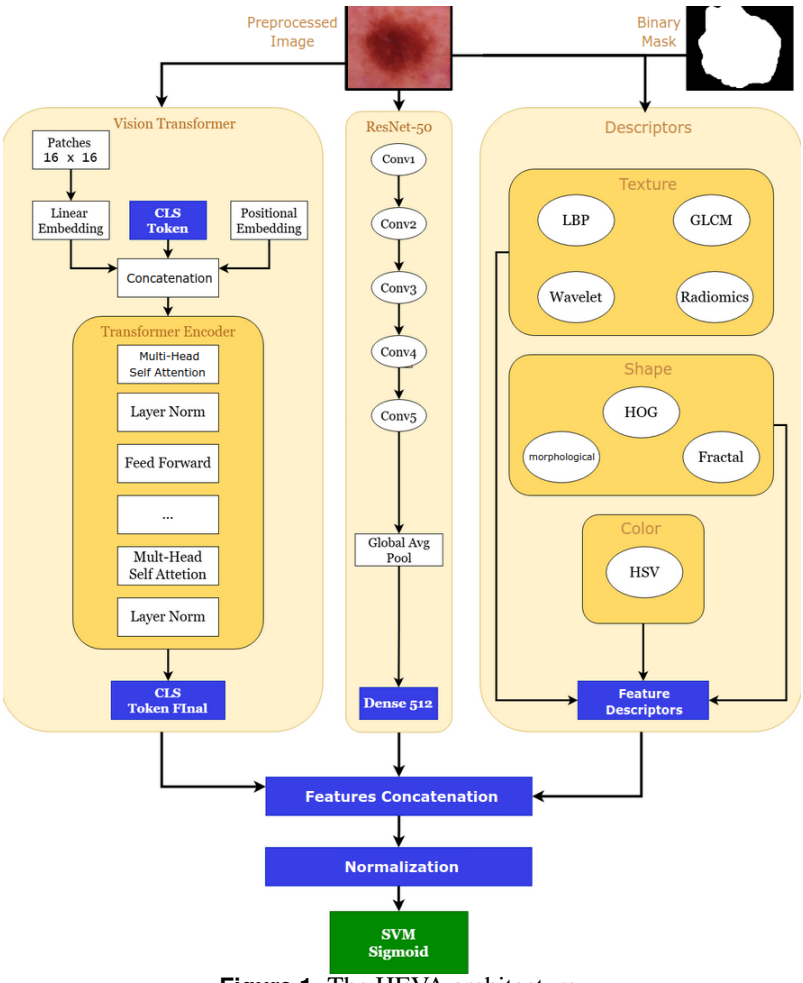

Figure 1. The HEVA architecture.

information from luminosity, facilitating the identification of chromatic patterns characteristic of different types of lesions.

The fusion of these representations occurs by concatenating the extracted features, forming a 1,395-dimensional vector (768 from ViT, 512 from ResNet50, and 115 from handcrafted descriptors). To ensure homogeneity in the feature scales, we apply normalization via Standard Scaler, and the resulting vector is used as input for an SVM classifier with a sigmoid kernel, chosen for its ability to model complex relationships between features and for providing a better balance between precision and recall, essential for medical applications.

3.1 Data Preparation

This section presents the dataset used in the experiments and the preprocessing applied to the data.

3.1.1 Dataset

The dataset used in this project was HAM-10000 [12]. The set consists of 10,015 dermatoscopic images, collected from diverse populations and acquired by different modalities, known for its diversity. HAM-10000 has seven available classes, each corresponding to a specific type of dermatological lesion. Some examples of cataloged lesions are melanoma, basal cell carcinoma, and melanocytic nevi, among others. This dataset also provides binary lesion segmentations for all images, called masks, which identify the lesion region in the image by highlighting which pixels belong to the lesion and which do not.

Due to the large number of classes, the dataset was binarized, grouping the seven original classes into two categories: Benign and Malignant (Table 1). This strategy simplifies the multi-class classification problem, facilitating the identification of critical patterns associated with potentially dangerous lesions.

Notably, the Vascular Lesions class (`vasc`), although generally considered harmless, was categorized as malignant in this study. This decision reflects a conservative approach, prioritizing the model's sensitivity in detecting possible atypical cases where such lesions might exhibit malignant characteristics [13]. In a clinical context, the model should maximize the detection of suspicious lesions, even if this leads to an increase in the false positive rate (benign lesions erroneously classified as malignant), rather than incurring false negatives

(malignant lesions not detected), which could result in serious consequences [13].

This conservative approach is supported by documented cases in the literature where vascular lesions have undergone malignant transformation [14], and by evidence that malignant tumors—particularly melanoma—may be misdiagnosed as benign vascular anomalies in real-world settings [13]. Furthermore, international classifications explicitly recognize vascular tumors across benign, borderline, and malignant spectra, reinforcing the rationale for a cautious grouping when aiming to protect against rare but high-impact outcomes [15].

This class grouping, particularly assigning `vasc` *to the malignant category, is an original choice of our work, grounded in clinical evidence and diagnostic safety principles*. A practical implication is that direct comparison with other studies becomes limited, as most prior work does not adopt this distribution. For this reason, we trained the base architectures (ResNet50 and ViT) under the same setting and then evaluated the hybrid ensemble (HEVA), enabling within-study ablations and fair assessment of the hybrid gains.

Table 1. Class division.

Benign	Malignant
Melanocytic Nevi (nv)	Melanoma (mel)
Dermatofibroma (df)	Basal Cell Carcinoma (bcc)
Benign Keratosis-like Lesions (bkl)	Actinic Keratosis (akiec)
	Vascular Lesions (vasc)

3.1.2 Preprocessing

The image preprocessing followed three sequential steps, inspired by [16]. Initially, all images were normalized to the 0–255 intensity range. Then, using the available segmentation masks, we computed the tight bounding box (bbox) of each lesion and cropped the image accordingly, delimiting the tumor area. We opted for a mask-derived bbox crop rather than using the full image or cropping strictly to the mask because it provides a practical balance: by excluding large portions of background, it removes external artifacts and noise (e.g., rulers, color charts, vignetting, and hair), while still preserving a narrow peri-lesional context (surrounding skin and border cues). In contrast, cropping exactly to the mask can discard edge/color transitions and contextual cues that help disambiguate lesions, whereas using the whole image increases background variability and the risk of spurious correlations. The resulting images were then resized to 100×100 pixels using nearest-neighbor interpolation, ensuring the preservation of the original intensity values. The resolution of 100×100 was chosen to significantly reduce the computational cost during training and inference, enabling faster experimentation without compromising the essential visual features.

Finally, a Gaussian filter was applied to reduce unwanted noise while preserving essential edges and morphological structures for classification. This pipeline was implemented on a total of 9,947 images, selected based on the availability of adequate segmentation masks, resulting in an optimized set for training the models. For external datasets (e.g., ISIC 2018 test), when ground-truth masks were not directly available or to ensure consistency, we used the masks predicted by our SegFormer-B5 model (Section 3.2). The complete process is illustrated in Figure 2.

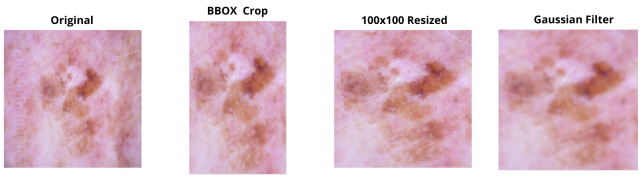

Figure 2. Preprocessing steps.

3.2 Lesion Segmentation with SegFormer-B5

To ensure robust and consistent lesion masks across datasets, we trained a semantic segmentation model based on SegFormer-B5. The model was trained on the original data prepared at 512×512 resolution for a binary segmentation task (classes: *background* and *lesion*). The predicted masks are essential for our preprocessing and for computing handcrafted descriptors over the lesion regions. Crucially, this component enabled cross-dataset generalization, allowing us to process the ISIC 2018 test set and evaluate HEVA in a cross-dataset scenario.

Segmentation test metrics. On the held-out segmentation test set, SegFormer-B5 achieved:

- Mean IoU (mIoU): 82.6%;
- Mean Accuracy: 91.2%.

These results indicate effective delineation of lesion boundaries, providing masks that are sufficiently accurate for downstream preprocessing and descriptor computation and that generalize to external datasets such as ISIC 2018.

3.3 Training of HEVA Component Models

This section details the training strategies applied to the main feature extraction components of the HEVA architecture. The training process was carefully designed for each model, considering their architectural particularities and potential contributions to the classification of skin lesions.

3.3.1 ViT Training

The ViT model used in this study was "*vit-base-patch16-224-in21k*", available from HuggingFace. This model was adapted for the binary classification task of skin lesions, preserving its fundamental architecture of processing 16x16 pixel patches and the attention mechanisms intrinsic to transformer models. The ViT receives preprocessed images, which are converted into PyTorch tensors, with their channels reorganized, resized to 224x224 pixels, and normalized to the range $[-1, 1]$, ensuring compliance with the required input format.

The training was conducted with carefully optimized hyperparameters, using batches of 32 images per iteration and a total of 10 epochs. The optimization strategy included a weight decay of 0.01 and a warmup phase of 500 steps, essential for stabilizing the initial learning process, which is particularly critical due to the sensitivity of the attention mechanisms in the early iterations. Additionally, the code was configured to extract the CLS tokens, which are used for the final classification.

To ensure the best possible performance, a systematic evaluation was implemented at the end of each epoch, saving the best model based on the F1-score metric. To guarantee the reproducibility of the experiments, the seed was fixed at 42. This rigorous methodological approach aims to ensure the reliability and consistency of the obtained results.

3.3.2 ResNet50 Training

The training of ResNet50 was carried out using a two-phase approach: (i) transfer learning and (ii) progressive fine-tuning to optimize feature extraction from the images. In the first phase, ResNet50 was loaded with weights pre-trained on ImageNet, keeping its base architecture frozen to preserve the generic features learned. The original fully connected layers were removed and replaced with a new structure composed of a Global Average Pooling layer, responsible for reducing the dimensionality of the feature maps, followed by a dense layer with 512 neurons and ReLU activation. To mitigate the risk of overfitting, a Dropout layer with a rate of 0.5 was added. Finally, an output layer with softmax activation was incorporated for multi-class classification. In this initial training phase, the model was trained for up to 30 epochs using the Adam optimizer, with an initial learning rate of 1e-3. Callbacks were used for dynamic process optimization, including ReduceLROnPlateau, which reduced the learning rate by a factor of 0.1 after five epochs without improvement in validation loss, and EarlyStopping, which interrupted training after ten epochs without progress, restoring the best weights obtained.

In the second phase, dedicated to fine-tuning, the last 50 layers of ResNet50 were unfrozen to allow for finer, more specific adjustments to the domain of skin lesions. For this stage, the learning rate was decreased to 1e-5 to avoid drastic changes in the already optimized weights. The training lasted for 50 epochs, with the callbacks readjusted for this phase—ReduceLROnPlateau was set to wait for 7 epochs and EarlyStopping for 12 epochs without improvement, allowing for a more detailed refinement of the model.

For feature extraction, a feature extractor was built to capture the output of the 512-dimensional dense layer. This layer was strategically chosen because it provides a compact and information-rich representation, where the features learned by the network have already undergone multiple nonlinear transformations and have been reduced to a dense and discriminative vector space.

3.4 Descriptor Implementation

The extraction procedure was conducted in three main steps. First, features were extracted for each image-mask pair, ensuring that the characteristics were captured accurately and contextualized concerning the regions of interest. Next, the resulting vectors were organized into a two-dimensional array in the format (n_samples × n_features), allowing for efficient data manipulation for subsequent analyses. Finally, normalization was applied using StandardScaler, ensuring that all features were standardized to a distribution with zero mean and unit standard deviation, optimizing compatibility with machine learning models. This process generates 115-dimensional vectors that represent consistent and reproducible manual features, regardless of the data source, as long as RGB images accompanied by corresponding binary masks are provided.

3.4.1 Implementation Parameters

The descriptors were implemented with the following technical parameters:

- **LBP**: Implemented with `local_binary_pattern` (radius=1, n_points=8, method='uniform'), generating a normalized 10-bin histogram;
- **GLCM**: Image quantized into 8 levels, with distances=[1] and angles=[0, π/4, π/2, 3π/4], extracting contrast, dissimilarity, homogeneity, energy, and correlation for each angle, totaling 20 features;
- **Wavelet**: `wavedec2` decomposition using Daubechies 2 ('db2') with 3 levels, generating statistics (mean, standard deviation, energy, skewness, kurtosis, and entropy) for approximation and detail coefficients in each direction;
- **HOG**: Configured with `pixels_per_cell`=(8, 8) and `cells_per_block`=(2, 2), extracting statistical features from the resulting descriptor;
- **Color**: Analysis in the HSV space with calculation of mean, standard deviation, and skewness for each channel;
- **Morphological**: Extraction via `regionprops` with calculation of circularity as $4\pi \cdot \text{area}/\text{perimeter}^2$ and asymmetry as the normalized difference between the mask and its mirrored version;
- **Fractal Dimension**: Box-counting algorithm implemented with decomposition into powers of 2, calculating the slope of the line on a logarithmic scale.

3.5 HEVA Classifier Training

For the training of the classifiers, a systematic experimentation approach was adopted, comparing different classifiers and data partitioning strategies. This section presents, in detail, the complete training workflow, from feature extraction and preparation to the evaluation of the models' performance.

3.5.1 Training Workflow

The complete training workflow consisted of the following steps:

1. **Feature extraction and integration:** Initially, features were extracted and integrated from three distinct sources: 115 handcrafted descriptors, 768-dimensional vectors from the ViT's CLS tokens, and 512-dimensional vectors extracted from the final dense layer of ResNet50. The concatenation of these features resulted in 1,395-dimensional vectors for each image;

2. **Normalization:** We applied StandardScaler to normalize the features, ensuring a distribution with zero mean and unit standard deviation. This step is essential to prevent features with different scales from disproportionately influencing the training process;

3. **Cross-validation:** We implemented K-Fold cross-validation with three partitions (3-Fold) to identify the best-performing subset, using the F1-Score as the selection criterion. This approach reduces the influence of variability in the train-validation split, ensuring a more robust model evaluation;

4. **Final model training:** The final model was trained using the data from the best fold;

5. **Evaluation on the test set:** Performance was evaluated on an independent test set, kept separate from the beginning of the process and not used in any model training phase.

3.5.2 Classifiers and Hyperparameters

Four distinct classifiers were evaluated as the final component of the HEVA architecture.

1. **MLP Neural Network:** A sequential feed-forward architecture composed of three dense layers (1024, 512, and 1 neuron). The intermediate layers used the ReLU activation function, while the output layer used the Sigmoid function. For regularization, a dropout of 0.3 was applied after the first layer and 0.2 after the second. The training was performed with the Adam optimizer, using a learning rate of 1e-4 and the Binary Cross-Entropy loss function. The model was trained for 10 epochs with a batch size of 32;

2. **SVM with RBF kernel:** The SVC (Support Vector Classifier) class from scikit-learn, configured with the Radial Basis Function (RBF) kernel, regularization parameter $C = 1.0$, gamma set to *scale*, and random state equal to 42 to ensure reproducibility;

3. **SVM with Sigmoid kernel:** Using the same SVC implementation, but with a sigmoidal kernel, keeping the other parameters identical to the SVM-RBF;

4. **SVM with Linear kernel:** Also based on the SVC implementation, but with a linear kernel, which projects the data into the original space without non-linear transformations.

3.5.3 Data Distribution Strategies

To analyze the influence of sampling strategies on classifier performance, we conducted experiments with three distinct approaches:

1. **Original:** The original distribution of the HAM-10000 dataset was used, consisting of approximately 80% benign and 20% malignant lesions, reflecting the prevalence observed in real clinical scenarios;

2. **Augmented Equalized:** Data augmentation techniques (rotation, zoom, translation, and mirroring) were applied, combined with oversampling of the minority class, resulting in a balanced training set with a 1:1 ratio between the classes. This strategy was adopted to reduce bias in favor of the majority class;

3. **Augmented Original:** Data augmentation techniques were applied to the training data, preserving the original proportion between the classes. This approach aimed to increase the diversity of the examples without compromising the statistical distribution of the dataset.

4. Results and Discussion

To evaluate the performance of the HEVA architecture, we used four standard metrics: **Precision**: the proportion of correct predictions among predicted positives, important when false positives undermine reliability. **Recall** (sensitivity): the ability to identify all positive cases, crucial in medical applications to avoid false negatives. **F1-Score**: the harmonic mean of precision and recall, useful on imbalanced datasets by balancing both error types. **Accuracy**: the overall proportion of correct classifications (positive and negative), a general indicator of effectiveness.

4.1 Comparative Analysis of Base Models

To establish a robust baseline for comparison, we initially analyzed the individual performance of the ResNet50 and ViT architectures under different data distribution settings. The results of these evaluations are presented in Table 2.

Table 2. Results of the base models with full metric names. Key: Aug. Eq. (Augmented Equalized), Aug. Orig. (Augmented Original).

Model & Dist.	Accuracy	Precision	Recall	F1-Score
ViT (Aug. Eq.)	91.26%	75.20%	88.55%	81.33%
ViT (Aug. Orig.)	92.56%	83.98%	80.84%	82.38%
ViT (Original)	91.26%	79.53%	79.91%	79.72%
ResNet50 (Aug. Eq.)	85.00%	78.00%	84.00%	80.00%
ResNet50 (Aug. Orig.)	88.70%	83.60%	79.80%	81.40%
ResNet50 (Original)	88.30%	84.20%	79.40%	81.40%

The results indicate that both models show competitive performance, with ViT achieving slightly higher metrics in configurations that use data augmentation. In particular, the "Augmented Original" configuration stood out for ViT, achieving the highest F1-Score (82.38%) and accuracy (92.56%) among all analyzed configurations. This result suggests that the data augmentation strategy, when preserving the original class distribution, favors the model's generalization ability. This effect may be associated with maintaining the natural statistical relationships between classes while simultaneously enriching the training set with useful variations.

4.2 Comparative Analysis of Classifiers and Data Distributions

The evaluation of the HEVA architecture was conducted considering various classifier configurations and data partitioning strategies. The obtained results are presented in Table 3.

The results show that the data augmentation strategy preserving the original class distribution (Augmented Original) provided the best performance among the evaluated approaches. This method allowed for the expansion of the training set without altering the natural proportion between benign and malignant lesions, resulting in better generalization capacity of the model and reducing the introduction of artificial biases. Compared to the `Original` and `Augmented Equalized` distributions, the `Augmented Original` showed consistent gains across all metrics, standing out as the most effective configuration for skin lesion detection.

In the analysis of the classifiers, the SVM-based models demonstrated competitive performance, with the SVM with a Sigmoid kernel achieving 84.81% recall and 84.22% F1-Score, indicating its high capacity to correctly detect malignant lesions. Although the SVM with a Linear kernel showed a slightly higher accuracy (93.72%) and the highest F1-Score (84.47%), the Sigmoid kernel was chosen for its better balance between precision (83.64%) and recall (84.81%) out of the box, ensuring a more stable relationship between minimizing false positives and maximizing the detection of malignant cases.

The comparative analysis also revealed that using artificially balanced data (`Augmented Equalized`) resulted in an increase in recall, indicating greater sensitivity for identifying malignant lesions. However, this approach negatively impacted precision, increasing the incidence of false positives, which can compromise the reliability of the diagnosis and lead to unnecessary investigations in healthy patients.

Based on the presented results, the final configuration adopted for HEVA consisted of an SVM with a Sigmoid kernel, trained on the `Augmented Original` distribution, achieving an accuracy of 93.17%, precision of 83.64%, recall of 84.81%, and an F1-Score of 84.22%. This configuration demonstrated a significant balance between the reliability of positive classifications (precision) and the ability to correctly identify malignant lesions (recall). This performance is particularly relevant for the computer-assisted diagnosis of skin cancer, where minimizing false negatives is a fundamental clinical priority.

4.3 Cross-Dataset Evaluation on ISIC 2018

To evaluate generalization across datasets, we applied the complete HEVA pipeline to the ISIC 2018 test set. Lesion masks were generated using our SegFormer-B5 model (Section 3.2), and the same preprocessing and feature extraction procedures were followed. On 1,512 successfully processed images, the classification results were:

Metric	Value
Accuracy	0.8770
Precision (Macro Avg)	0.8283
Recall (Macro Avg)	0.8129
F1-Score (Macro Avg)	0.8201

These outcomes demonstrate that the hybrid approach maintains strong performance under cross-dataset conditions, even when the class mapping differs from the majority of the literature (`vasc` considered malignant in our setup).

4.4 Comparison with the Literature

The results obtained by HEVA are promising when compared to other recent works in the literature. The model proposed by [9] achieved an accuracy of 91.18% using only a CNN, while our hybrid approach reached 93.17% with the Sigmoid SVM. Similarly, the model presented by [10], which combined the DenseNet201 architecture with Fine KNN and Cubic SVM classifiers, obtained an accuracy of 92.34% on the ISIC 2019 and PH2 datasets, which is lower than HEVA's results on the HAM-10000 dataset.

A in cross-paper comparisons is our conservative and *original* class mapping that assigns `vasc` to the malignant class. This choice enhances sensitivity for atypical vascular cases but reduces direct comparability with works that preserve the canonical class grouping. To mitigate this limitation and to quantify the hybrid gains reliably, we trained the base backbones (ResNet50 and ViT) and then compared them against the hybrid HEVA within our consistent setup.

This superior performance can be attributed to the hybrid nature of the HEVA architecture, which effectively combines the high-level representation capabilities of DL-based models (Vision Transformers and ResNet50) with interpretable clinical ABCD descriptors, enabling the capture of complementary features at various levels of abstraction, from textural and morphological aspects to complex semantic patterns. Furthermore, the SegFormer-B5 segmentation component provided reliable lesion masks that generalized to ISIC 2018, making cross-dataset evaluation feasible and robust.

5. Conclusion

This work presented the HEVA architecture, an innovative hybrid approach for the classification of skin lesions that integrates deep neural network models (Vision Transformers and

Table 3. Comparative results.

Distribution	Classifier	Accuracy	Precision	Recall	F1-Score
Original	Neural Network	91.40%	83.61%	74.95%	78.92%
	SVM (Kernel RBF)	91.16%	84.43%	72.20%	77.83%
	SVM (Kernel Sigmoid)	90.45%	79.17%	75.47%	77.27%
	SVM (Kernel Linear)	90.65%	81.68%	72.90%	77.04%
Augmented Equalized	Neural Network	91.84%	80.74%	81.68%	81.15%
	SVM (Kernel RBF)	91.06%	77.06%	83.18%	80.00%
	SVM (Kernel Sigmoid)	91.61%	79.06%	82.94%	80.96%
	SVM (Kernel Linear)	91.81%	80.46%	81.78%	81.11%
Augmented Original	Neural Network	93.53%	87.87%	81.17%	84.36%
	SVM (Kernel RBF)	93.57%	90.76%	78.04%	83.92%
	SVM (Kernel Sigmoid)	**93.17%**	83.64%	**84.81%**	**84.22%**
	SVM (Kernel Linear)	93.72%	90.19%	79.44%	84.47%

ResNet50) with clinical ABCD descriptors. The experiments conducted demonstrated the effectiveness of this approach, achieving competitive performance metrics with an accuracy of 93.17%, precision of 83.64%, recall of 84.81%, and an F1-Score of 84.22% in detecting malignant lesions. The combination of complementary representations extracted from different sources allows HEVA to exploit both the ability of deep neural networks to learn complex and abstract features and the clinically interpretable aspects provided by specialized descriptors. This synergy resulted in a more robust model with greater discriminative power, surpassing approaches based exclusively on CNNs or traditional feature extraction techniques.

A particularly relevant aspect of the obtained results was the demonstration that the data augmentation technique preserving the original class distribution provided significant performance gains compared to non-augmented or artificially balanced data. This result suggests that, for the specific domain of skin lesion classification, increasing the number of training samples while maintaining the natural proportion between benign and malignant lesions contributes to the development of more representative and generalizable models.

The choice of the SVM with a Sigmoid kernel as the final classifier was motivated by its excellent balance between precision and recall, with a special emphasis on maximizing recall. This characteristic is crucial in medical applications, where the non-detection of a malignant lesion (false negative) can have much more severe consequences than the incorrect classification of a benign lesion as malignant (false positive).

Additionally, we showed that a dedicated segmentation model (SegFormer-B5) is instrumental for generating reliable lesion masks that are essential to our preprocessing and descriptor pipeline and that enable strong cross-dataset generalization, as evidenced by the ISIC 2018 test results.

For future work, we suggest the exploration of feature selection methods and the optimization of the decision threshold to improve classification robustness and interpretability. Furthermore, we propose a detailed analysis of the computational cost involved in training the HEVA architecture, evaluating its feasibility for large-scale applications.

Author Contributions

Bianca Aparecida Andrade proposed and performed the experiments; Ademar Takeo Akabane proposed the research theme and assisted in the preparation of the article; Antonio Marcio Crepaldi Junior proposed and performed the experiments.

References

[1] MINISTÉRIO DA SAÚDE - INCA. *INCA estima 704 mil casos de câncer por ano no Brasil até 2025*. 2022. In Portuguese. Disponível em: ⟨https://www.gov.br/inca/pt-br/assuntos/noticias/2022/inca-estima-704-mil-casos-de-cancer-por-ano-no-brasil-ate-2025⟩.

[2] MINISTÉRIO DA SAÚDE - INCA. *Estimativa 2023 – Incidência de Câncer no Brasil*. 2023. In Portuguese. Disponível em: ⟨https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa⟩.

[3] ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, v. 542, n. 7639, p. 115–118, 2017.

[4] SHORFUZZAMAN, M. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. *Multimedia Systems*, v. 28, n. 4, p. 1309–1323, 2022.

[5] BHATT, H. et al. State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review. *Intelligent Medicine*, v. 3, n. 3, p. 180–190, 2023.

[6] KITTLER, H. et al. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, v. 3, n. 3, p. 159–165, 2002.

[7] JAVAID, A.; SADIQ, M.; AKRAM, F. Skin cancer classification using image processing and machine learning. In: IEEE. *2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. [S.l.], 2021. p. 439–444.

[8] VESTERGAARD, M. E. et al. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, v. 159, n. 3, p. 669–676, 2008.

[9] SHETTY, B. et al. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Scientific Reports*, v. 12, n. 1, p. 18134, 2022.

[10] BENYAHIA, S.; MEFTAH, B.; LÉZORAY, O. Multi-features extraction based on deep learning for skin lesion classification. *Tissue and Cell*, v. 74, p. 101701, 2022.

[11] OMEROGLU, A. N. et al. A novel soft attention-based multi-modal deep learning framework for multi-label skin lesion classification. *Engineering Applications of Artificial Intelligence*, v. 120, p. 105897, 2023.

[12] TSCHANDL, P.; ROSENDAHL, C.; KITTLER, H. The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, v. 5, p. 180161, 2018.

[13] SCORLETTI, F. et al. Malignant tumors misdiagnosed as benign vascular anomalies. *Pediatr Blood Cancer*, v. 65, n. 7, p. e27051, 2018.

[14] HEINZE, C. et al. Malignant conversion of eyelid capillary hemangioma to cutaneous angiosarcoma. *Ophthalmic Plast Reconstr Surg*, v. 37, n. 2, p. e54–e57, 2021.

[15] International Society for the Study of Vascular Anomalies. *ISSVA Classification for Vascular Anomalies*. 2018.

[16] DINIZ, J. O. B. et al. Efficientensemble: Diagnóstico de câncer de mama em imagens de ultrassom utilizando processamento de imagens e ensemble de efficientnets. In: *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2024)*. [S.l.]: SBC, 2024.