From Random to Pre-trained ViT Backbones for Improved Feature Extraction: Application to Diagnosis with Sensor Images

Luan B. Guerra, Ana Beatriz S. Zerati, Wallace Casaca, and Lucas C. Ribas(✉)

São Paulo State University, Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, Brazil
{luan.bonizi,ab.zerati,wallace.casaca,lucas.ribas}@unesp.br

Abstract. The development of robust and low-cost diagnostic tools is essential for the early and fast detection of infectious and chronic diseases, such as COVID-19 and prostate cancer. Biosensors based on molecular recognition have emerged as promising alternatives to conventional laboratory techniques. Recent studies indicate that microscopy images acquired from these materials contain discriminative morphological patterns associated with analyte presence, enabling the development of machine learning models capable of automatically performing medical diagnoses based on these images. In this work, we propose a feature fusion strategy that combines representations extracted from random and pretrained Vision Transformers (ViTs) to improve image embeddings from genosensors and immunosensors and enrich domain-specific descriptors. The proposed approach is built upon three token-based feature extraction strategies, which are explored to capture complementary information from the models. To systematically assess the effectiveness of the method, we evaluate the approach using ten different ViT backbones. Experiments conducted on SARS-CoV-2 immunosensor and PCA3 genosensor microscopy image datasets demonstrate that feature fusion consistently improves performance and that randomly initialized ViTs frequently achieve results comparable to, or even surpassing, pretrained models. Furthermore, the proposed method significantly outperformed various baseline Convolutional Neural Networks (CNNs) across all datasets. These findings highlight the strong intrinsic representational capacity of Transformer architectures and their suitability for small-scale biosensor image analysis.

Keywords: Biosensors · Vision Transformers · Feature fusion · Random models · Transfer Learning

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16767, pp. 510–528, 2027.
https://doi.org/10.1007/978-3-032-30491-9_32

1 Introduction

The rapid and accurate detection of infectious and chronic diseases is a critical challenge for healthcare systems worldwide [2]. The COVID-19 pandemic reinforced the urgent demand for sensitive, low-cost, and easily deployable diagnostic technologies [8,22,34], while prostate cancer continues to rank among the leading causes of cancer-related mortality in men [16,28]. In this context, biosensors based on molecular recognition have emerged as promising alternatives to conventional laboratory techniques, enabling early detection, portability, and reduced operational costs [18,25,34].

Two important classes of biosensors are genosensors and immunosensors. Genosensors rely on the specific recognition between nucleic acid sequences, enabling the detection of viral genetic material or molecular biomarkers associated with diseases [4,20]. Immunosensors, in turn, are based on antigen–antibody interactions and are widely used for detecting tumor-related biomarkers [13,22]. Recent studies [12,22,34,42] have demonstrated that, beyond traditional electrochemical measurements, microscopy images of functionalized substrates contain morphological and textural patterns associated with analyte presence. Such patterns can generate informative visual features that can be directly analyzed and classified for diagnostic purposes. This observation opens new opportunities for computer vision and machine learning approaches aimed at the automatic interpretation of biosensing signals.

Deep learning models have been employed as automatic feature extractors in image analysis [1,5,29,32]. More specifically, in the field of texture analysis, which includes biosensor images, CNNs predominate as backbones for feature extraction [22,28,36]. CNNs achieve strong performance in tasks involving texture and biosensor images because their convolutional operations introduce an inductive bias that effectively captures the local patterns typical of these images [27]. However, these models have a limited receptive range, causing them to capture fewer long-range relationships [17]. As an alternative to CNNs, Vision Transformers (ViTs) [11] are capable of modeling global dependencies through attention mechanisms, which is particularly beneficial when the diagnostic information relies on long-range spatial relationships between different regions [40]. Nevertheless, there are no studies in the literature that explore ViTs as feature extractors for images obtained from genosensors and immunosensors, even though these models present state-of-the-art results in several applications [3,38], often surpassing CNN-based approaches [35].

Despite their strong global modeling capability, ViTs present limitations in capturing local patterns and typically require large-scale datasets and high computational resources to achieve competitive performance [44]. These challenges are especially relevant in biosensing applications, where datasets are often limited and subtle local morphological variations may be crucial for analyte detection [28]. To address the first ViT limitation, several studies have proposed complementary strategies such as feature fusion, which combines representations extracted at different layers or levels to enrich visual descriptions and recover fine-grained information [39,43]. Regarding data dependency, recent works have

explored training-free Transformer architecture search [44,45], which uses models with randomly initialized weights to investigate the intrinsic representational capacity of the architecture itself and to automatically design ViT architectures. These studies show that, under certain conditions, untrained ViTs can achieve competitive results, enabling architectural exploration with reduced computational cost.

This work investigates the use of Vision Transformers as feature extractors for images obtained from genosensors and immunosensors. In addition, both pretrained and randomly initialized models are explored to assess the intrinsic representational capacity of transformer architectures. We propose feature fusion strategies based on representations extracted from both pretrained and randomly initialized models. In particular, we explore token-based representations obtained from the CLS token and from global average pooling (GAP) over all tokens, which allows capturing richer textural and fine-grained patterns in order to improve feature extraction and produce more robust representations of biosensor images. The combination of these representations aims to produce more robust image descriptors and mitigate known limitations of Vision Transformers in modeling local patterns, making the approach particularly suitable for small-scale biosensing datasets. The main contributions of this work are summarized as follows:

- A feature fusion strategy that integrates representations extracted from random and pretrained Vision Transformers to obtain richer and more robust texture image descriptors for biosensor images.
- An analysis of token-based feature extraction strategies, evaluating three different approaches to capture complementary information from ViT representations.
- A systematic evaluation across ten ViT backbones using microscopy images obtained from genosensors and immunosensors, assessing the effectiveness of the proposed approach for biosensor image representation.

2 Proposed Methodology

This section describes the proposed methodology for extracting and analyzing representations from Vision Transformers applied to biosensor images. An overview of the proposed pipeline is illustrated in Fig. 1.

2.1 Feature Extraction from Vision Transformers

When used as a feature extractor, the backbone is employed without its classification head, and the internal representations generated by the network are used as image descriptors. In this work, Vision Transformers are employed to generate such representations from biosensor images.

In a ViT, an input image is first divided into fixed-size patches, which are linearly projected into embedding vectors, combined with positional encodings

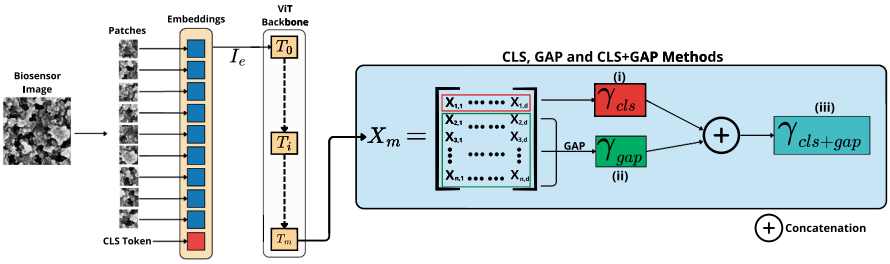

Fig. 1. Illustration of the methodology of this work. The Υ_{CLS} and Υ_{GAP} features are represented in (i) and (ii), respectively. (iii) represents the $\Upsilon_{CLS+GAP}$ feature, which is a combination of (i) and (ii).

and processed through a sequence of transformer blocks T. Each transformer block produces a matrix of token embeddings X that encode visual information extracted from different regions of the image. The matrix resulting of block T_i can be denoted as $X_i \in \mathbb{R}^{n \times d}$, where n is the number of tokens and d is the hidden dimension. Among these tokens, the first one corresponds to the classification token (CLS), while the remaining tokens correspond to spatial tokens associated with image patches.

The CLS token provides a compact global representation of the image, and is commonly used as a compact representation for image classification tasks. In contrast, spatial tokens encode spatial information from different image regions and, although they require additional processing to form a single representation, prior works [30,31] highlight their importance, showing that spatial tokens contribute significantly to robustness in complex visual scenarios. In order to investigate how these different sources of information contribute to the representation of biosensor images, we explore multiple strategies for extracting features from the final transformer layer. It is important to note that, since our work also explores the use of random ViTs, the representations were extracted using both pre-trained and random models.

CLS-Based Representation. The first strategy uses only the CLS token extracted from the final transformer block, shown in Fig. 1 (i). Formally, the CLS feature vector is defined as

$$\Upsilon^{\theta}_{CLS} = X_m[0], \tag{1}$$

where X_m denotes the output matrix of the last transformer block T_m and $\theta \in \{r, p\}$ denotes the model parameterization used to extract the representation, with r for randomly initialized and p for pretrained models. This representation captures global contextual information about the image.

Spatial-Token Representation. To obtain a compact representation from the spatial tokens, we apply global average pooling (GAP) over all tokens except the CLS token:

$$\Upsilon^{\theta}_{GAP} = GAP(X_m[1], X_m[2], ..., X_m[n-1]). \tag{2}$$

This operation aggregates the spatial information encoded by the transformer into a single feature vector while preserving contributions from all image regions. We refer to this method as the GAP method, which can be seen in Fig. 1 (ii).

Combined Representation. Since CLS and spatial tokens capture complementary information, we also investigate a combined representation obtained by concatenating both feature vectors:

$$\Upsilon^{\theta}_{CLS+GAP} = [\Upsilon^{\theta}_{CLS}; \Upsilon^{\theta}_{GAP}], \tag{3}$$

where $[;]$ denotes concatenation along the feature dimension, originating the CLS+GAP method (Fig. 1 (iii)). This strategy allows the representation to simultaneously encode global contextual information and aggregated spatial patterns.

All representations are extracted directly from the models without performing any fine-tuning or architectural modification. In addition, feature extraction is performed using both pretrained and randomly initialized Vision Transformers, enabling the analysis of the intrinsic representational capacity of the transformer architecture when applied to biosensor image recognition tasks.

2.2 Combination of Random and Pretrained Features

In addition to analyzing pretrained and randomly initialized models independently, we investigate whether combining their representations can yield more informative feature vectors. Pretrained ViTs are typically trained on large-scale natural image datasets such as ImageNet-21k, which may introduce biases related to natural image statistics. In contrast, randomly initialized models produce representations that are not influenced by such pretraining. As a result, the representations generated by these two types of models may capture different aspects of the visual patterns present in biosensor images.

To explore this hypothesis, we combine features extracted from pretrained and random models through feature concatenation. The fused representations are defined as

$$\Upsilon^{r+p}_{CLS} = [\Upsilon^{r}_{CLS}; \Upsilon^{p}_{CLS}], \tag{4}$$

$$\Upsilon^{r+p}_{GAP} = [\Upsilon^{r}_{GAP}; \Upsilon^{p}_{GAP}], \tag{5}$$

$$\Upsilon^{r+p}_{CLS+GAP} = [\Upsilon^{r}_{CLS+GAP}; \Upsilon^{p}_{CLS+GAP}], \tag{6}$$

where the features with p represent the vectors extracted from pretrained models, while the representations with r represent the vectors extracted from random models. Here the symbol [;] also represents a horizontal concatenation of the representations.

3 Experimental Setup

3.1 Selected Vision Transformer Architectures

To provide a comprehensive evaluation of transformer-based feature extractors, we considered ten ViT backbones that differ in architectural design, model size, internal modules and pre-training. As evaluating all existing models is impractical, we selected a representative set of widely used backbones that capture different design choices explored in the literature. The main characteristics of the selected architectures are summarized in Table 1. All models were obtained from the *timm* PyTorch library, which provides standardized and widely adopted implementations of modern vision architectures.

Table 1. Details of the ViT backbones evaluated in this work. The number of parameters refers only to the feature extractor, excluding the final classification layer. In the features column, [1] denotes the feature size for CLS and GAP, while [2] denotes the size for CLS+GAP.

Backbone	Pre-training	Feature extraction cost		Features	
		GFLOPs	param. (M)	[1]	[2]
ViT-B/16 [11]	IN-1k and IN-21k	16.9	85.8	768	1536
ViT-S/16 [11]	IN-1k and IN-21k	4.25	22	384	768
ViT-Betwix [10]	IN-1k	11.7	60.3	640	1280
ViT-RelPos [41]	IN-1k	16.9	86.4	768	1536
ViT-B/14 dinov2 [23]	IN-1k	16.9	85.8	768	1536
DeiT-B/16 [37]	IN-1k	16.9	85.8	768	1536
BeiTv2-B [24]	IN-1k	12.7	85.8	768	1536
BeiT-B [6]	IN-22k	12.7	85.8	768	1536
DeiT3-B [38]	IN-1k or IN-22k	16.9	85.8	768	1536
FlexiViT [7]	IN-1k or IN-21k	16.9	86.6	768	1536

3.2 Datasets

– **PCA3** [28]: The PCA3 dataset consists of 32 grayscale (8-bit) scanning electron microscopy (SEM) images of genosensors developed to detect the prostate cancer biomarker PCA3. The sensors were fabricated using layer-by-layer films with gold nanoparticles and an immobilized complementary DNA

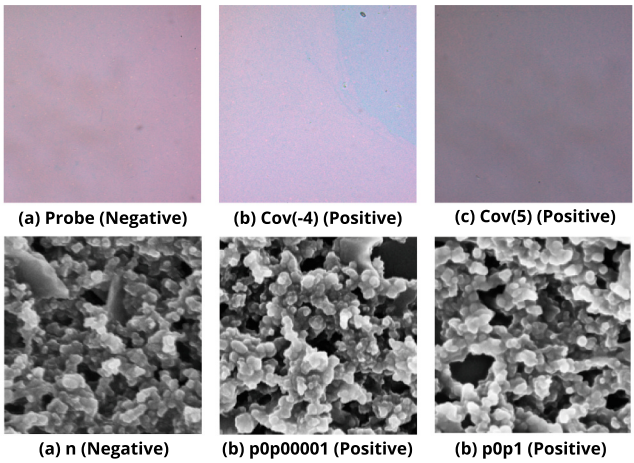

Fig. 2. Some samples of the datasets used in this work. The first row shows images from different classes of dataset SARS-CoV-2 Immunosensor [22], while the second row shows images from dataset PCA3 [28].

probe on carbon electrodes, and imaged after exposure to different PCA3 concentrations to capture morphology changes caused by DNA hybridization, covering eight experimental conditions: one negative control, one blank control, and six positive classes with concentrations ranging from 10^{-5} to 1 $\mu \cdot \text{mol L}^{-1}$. The pre-processed images have 300×300 pixels, following a similar protocol to that observed in [28]. We extracted three 300×300 pixels windows from each image in the dataset in order to increase the total number of samples, resulting in 96 images. Thus, the dataset supports both binary (presence vs. absence of PCA3) and multiclass classification settings. Some samples of this dataset can be seen in the second row of Fig. 2 and the number of samples per class for each classification configuration is reported in [28].

– **SARS-CoV-2 Immunosensor** [22]: The immunosensor dataset consists of 894 RGB optical microscopy images acquired from plasmonic substrates functionalized for the detection of SARS-CoV-2 antibodies. The sensing platform was based on gold-coated substrates modified with specific biological recognition elements, and images were captured after exposure to samples containing different antibody concentrations to record surface alterations induced by antigen–antibody interactions. The dataset comprises optical images obtained under controlled illumination and magnification conditions, representing distinct experimental groups including negative controls and positive samples with varying antibody concentrations ranging from 10 $\text{ng} \cdot \text{mL}^{-1}$ to 100 $\mu\text{g} \cdot \text{mL}^{-1}$. Similar to the PCA3 dataset, this dataset can be formulated either as a binary classification task (presence vs. absence of antibodies) or as a multiclass problem depending on the concentration levels considered. The per-class sample counts for the binary and multiclass problem are reported in

[22]. In the experiments, we rescaled all images to 1344×1024 pixels. Some samples of this dataset can be seen in the first row of Fig. 2.

3.3 Classification Setup

After extracting all feature vectors, a linear Support Vector Machine (SVM) classifier [9] was trained and used to classify the test features. The SVM was configured with a linear kernel and a penalty parameter $C = 1$. For the PCA3 dataset, a resampling procedure was adopted during the classification step using 100 random configurations (i.e., 100 trials). In each configuration, the same number of samples was selected from each class, resulting in a uniform class distribution. For each trial, a 10-fold cross-validation scheme was employed to split the data into training and test sets. The average accuracy and the standard deviation across all trials were then computed to evaluate the performance of the proposed method. For the SARS-CoV-2 dataset, a 5-fold cross-validation scheme was adopted and repeated three times. The final performance was reported as the mean accuracy and the corresponding standard deviation.

We also fine-tuned the ViT models to compare their performance with that of the proposed representations. In this case, the models were fine-tuned using the same cross-validation configuration as the respective datasets, where, for each configuration of the training and test folds, fine-tuning was performed over 40 epochs with a learning rate of 3×10^{-5}. From this, we calculated the mean and standard deviation of the accuracies.

4 Results and Discussion

As described in Sect. 3.2, the datasets used in this study support both binary and multiclass classification settings. To provide a more comprehensive evaluation of the proposed methodology, experiments were conducted under both scenarios.

4.1 Binary Classification Analysis

Table 2 presents the results of the binary classification on the SARS-CoV-2 dataset. It can be observed that there is no clear pattern among the results of the different models, with performance varying between CLS, GAP, and CLS+GAP methods. This variation is due to architectural differences among the models. Nevertheless, the GAP and CLS+GAP methods have higher average accuracy compared to the CLS method, indicating that spatial tokens provide valuable information about immunosensor images, especially when these tokens are combined with the CLS token.

Furthermore, an important observation is that combining random and pre-trained features leads to an increase in the average accuracy across all methods for the SARS-CoV-2 dataset. This result suggests that features extracted from randomly initialized models can provide complementary information to

Table 2. Accuracy of each proposed method considering the binary classification of the dataset SARS-CoV-2 Immunosensor. The results in bold represent the best result for a specific backbone.

Backbone	Pretrained			Random+Pretrained			Fine-tunning
	CLS	GAP	CLS+GAP	CLS	GAP	CLS+GAP	
ViT-B [11]	89.97 ± 2.15	89.11 ± 2.17	90.86 ± 1.72	96.05 ± 1.32	**96.42 ± 1.08**	96.23 ± 1.28	90.16 ± 4.90
ViT-S [11]	90.19 ± 2.31	88.29 ± 3.40	89.78 ± 2.59	95.12 ± 1.52	**95.53 ± 1.53**	95.45 ± 1.22	93.4 ± 0.65
ViT-Betwix [10]	92.73 ± 1.95	91.76 ± 2.34	92.47 ± 2.24	93.25 ± 2.02	92.50 ± 2.50	**94.44 ± 2.34**	93.28 ± 0.65
ViT-RelPos [41]	74.68 ± 1.99	89.22 ± 2.19	84.38 ± 1.22	77.52 ± 2.69	91.50 ± 2.07	87.58 ± 2.21	**95.19 ± 2.15**
ViT-B dinov2 [23]	83.60 ± 2.87	85.20 ± 2.78	85.31 ± 2.60	85.91 ± 2.43	87.36 ± 2.19	**87.77 ± 2.53**	84.45 ± 2.78
DeiT-B [37]	88.44 ± 2.33	89.23 ± 2.00	89.60 ± 2.22	97.20 ± 1.17	96.65 ± 1.31	**97.32 ± 1.28**	86.25 ± 9.29
BeiTv2-B [24]	89.15 ± 2.53	89.67 ± 2.59	89.78 ± 2.79	92.17 ± 1.58	89.41 ± 2.70	**93.44 ± 1.70**	74.85 ± 11.97
BeiT-B [6]	95.00 ± 1.94	94.11 ± 1.89	95.67 ± 1.88	97.69 ± 1.13	**97.72 ± 1.11**	97.13 ± 1.47	83.68 ± 6.53
DeiT3-B [38]	85.64 ± 3.05	84.94 ± 3.04	84.93 ± 3.00	88.03 ± 1.81	87.92 ± 3.00	88.74 ± 2.30	**89.03 ± 2.68**
FlexViT [7]	87.70 ± 2.74	87.92 ± 2.12	88.33 ± 2.08	96.12 ± 0.73	**96.94 ± 1.06**	96.53 ± 0.99	89.60 ± 2.36
Mean accuracy	87.71	88.95	89.11	91.91	93.20	**93.46**	78.98

the image representation. Unlike pretrained models, whose representations may reflect biases inherited from large-scale datasets such as ImageNet-21k, random models produce unbiased representations that can act as a form of regularization when combined with pretrained features. As a result, the fusion of both types of representations leads to more domain-specific and robust image descriptors for biosensor image recognition. It is also worth noting that the proposed method outperforms even results of fine-tunned models, where the best-performing method outperformed the average accuracy of the tuned models by 14.36%, demonstrating the effectiveness of our approach when applied to biosensor datasets, which often suffer from a lack of data for an efficient training of deep learning models. Among the evaluated methods, the CLS+GAP strategy combined with random and pretrained features achieved the highest mean accuracy.

Table 3. Accuracy of each proposed method considering the binary classification of the dataset PCA3. The results in bold represent the best result for a specific backbone.

Backbone	Pretrained			Random+Pretrained			Fine-tunning
	CLS	GAP	CLS+GAP	CLS	GAP	CLS+GAP	
ViT-B [11]	97.50 ± 7.50	97.22 ± 5.69	97.50 ± 7.50	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	99.6 ± 3.07
ViT-S [11]	93.89 ± 9.44	97.22 ± 5.69	95.00 ± 8.03	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.00**
ViT-Betwix [10]	95.83 ± 7.16	95.00 ± 8.03	95.00 ± 8.03	**100.0 ± 0.0**	96.51 ± 3.13	99.65 ± 1.31	93.23 ± 12.89
ViT-RelPos [41]	83.06 ± 14.40	95.83 ± 7.16	95.83 ± 7.16	81.96 ± 5.26	**99.30 ± 1.79**	97.23 ± 3.23	94.86 ± 11.62
ViT-B dinov2 [23]	94.17 ± 11.81	95.28 ± 9.46	95.28 ± 9.46	98.60 ± 2.33	**99.30 ± 1.79**	98.95 ± 2.11	92.13 ± 13.82
DeiT-B [37]	99.17 ± 2.50	97.78 ± 6.67	98.06 ± 3.94	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	99.65 ± 3.08
BeiTv2-B [24]	98.33 ± 5.00	98.33 ± 5.00	98.33 ± 5.00	98.95 ± 2.11	**99.30 ± 1.79**	**99.30 ± 1.79**	95.45 ± 9.04
BeiT-B [6]	94.17 ± 9.50	95.28 ± 8.06	95.28 ± 8.06	99.30 ± 1.79	97.58 ± 3.22	98.26 ± 2.46	**100.0 ± 0.0**
DeiT3-B [38]	96.94 ± 4.72	98.33 ± 5.00	99.17 ± 2.50	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	96.34 ± 8.51
FlexViT [7]	99.17 ± 2.50	96.94 ± 4.72	98.33 ± 5.00	**100.0 ± 0.0**	**100.0 ± 0.0**	**100.0 ± 0.0**	90.19 ± 15.38
Mean accuracy	95.28	96.19	96.78	97.88	99.20	**99.34**	96.14

Table 3, with the results from the binary classification of PCA3 database, shows something similar to the Table 2, where there is no discernible pattern among the results of the different ViTs and we have also a increase in the average accuracy of the GAP and CLS+GAP methods compared to the CLS method, which indicates that spatial tokens provide valuable information about genosensor images. Once again, it is very clear to verify an increase in the mean accuracy of all methods when combining random and pretrained features, again indicating that features extracted by random models contain important information and serve as a regularization, which reduces the biases of the representations and makes them much more robust, since, in this case, many methods have achieved 100% accuracy. In this case, our approach also outperformed the fine-tuned models, which achieved similar results to those of methods that rely solely on features from pre-trained ViTs, but were outperformed by methods that combine random and pre-trained features, resulting in a 3.2% increase in average accuracy. The CLS+GAP method with combined features also presented the best results for the PCA3 binary classification.

It is worth noting that, as shown in Table 3, several models achieve 100% accuracy on PCA3. However, due to the small size of this dataset, such performance may reflect the inherent simplicity of the binary task or potential overfitting, rather than clear superiority of a given method. In the original PCA3 study [28], a t-SNE analysis showed that the binary features form two well-defined and easily separable clusters, even in two dimensions. This supports the interpretation that the PCA3 binary classification task is intrinsically simple, corroborating the near-perfect supervised results reported in Table 3.

4.2 Multiclass Classification Analysis

From Table 4, which presents the results of the multiclass classification on the SARS-CoV-2 dataset, a noticeable decrease in accuracy can be observed across all methods when compared to the binary classification results. This drop in performance is expected, since multiclass classification requires models to recognize subtle differences in patterns among multiple classes, making the problem considerably more challenging. However, similarly to the binary classification scenario, the GAP and CLS+GAP strategies achieve higher average accuracies than the CLS method, demonstrating, once again, the importance of spatial tokens in detecting patterns of biosensor images. Moreover, when random and pretrained features are combined, an improvement in performance is again observed across all models. This indicates that random features, despite the increased difficulty of distinguishing interclass patterns, still contain valuable information and contribute to more robust image descriptors when fused with pretrained features. Once again, the proposed method outperformed the fine-tuned models in all cases without the need to train the ViTs, achieving an 8.11% increase in average accuracy, indicating the effectiveness of different feature extraction methods for biosensor images, particularly when random and pre-trained representations are combined. Unlike the previous cases, the method that achieved the highest aver-

age accuracy in the SARS-CoV-2 multiclass classification was the GAP strategy with combined random and pretrained features.

Table 4. Accuracy of each proposed method considering the multiclass classification of the dataset SARS-CoV-2 Immunosensor.

Backbone	Pretrained			Random+Pretrained			Fine-tunning
	CLS	GAP	CLS+GAP	CLS	GAP	CLS+GAP	
ViT-B [11]	73.19 ± 2.83	71.59 ± 3.52	72.78 ± 2.82	79.38 ± 2.19	**79.57 ± 2.56**	79.57 ± 2.50	69.57 ± 1.75
ViT-S [11]	69.20 ± 3.03	68.42 ± 3.16	69.20 ± 2.48	76.62 ± 3.22	77.63 ± 3.57	**77.67 ± 3.32**	69.23 ± 3.9
ViT-Betwix [10]	73.98 ± 3.13	73.12 ± 3.35	74.09 ± 3.03	78.97 ± 3.41	77.41 ± 3.04	**79.46 ± 2.75**	68.68 ± 5.06
ViT-RelPos [41]	56.71 ± 2.59	74.91 ± 2.75	68.09 ± 2.82	61.04 ± 1.77	**79.12 ± 2.44**	70.36 ± 3.01	76.85 ± 1.41
ViT-B dinov2 [23]	63.46 ± 3.95	65.62 ± 3.71	65.29 ± 4.01	71.59 ± 3.72	**73.41 ± 3.02**	72.37 ± 3.20	65.55 ± 1.81
DeiT-B [37]	72.34 ± 2.57	71.96 ± 2.89	72.75 ± 3.49	80.91 ± 2.55	80.95 ± 2.79	**80.98 ± 2.81**	70.25 ± 1.62
BeiTv2-B [24]	76.33 ± 2.57	76.81 ± 2.44	77.78 ± 2.71	81.66 ± 2.36	76.92 ± 2.37	**82.10 ± 2.94**	73.38 ± 1.87
BeiT-B [6]	80.09 ± 2.91	74.65 ± 2.95	78.41 ± 3.63	**83.56 ± 2.83**	81.24 ± 2.50	81.99 ± 2.65	72.60 ± 2.5
DeiT3-B [38]	66.93 ± 3.51	67.94 ± 2.53	67.75 ± 3.06	75.02 ± 3.51	**75.69 ± 2.76**	75.36 ± 3.56	64.54 ± 1.03
FlexViT [7]	70.88 ± 3.24	69.16 ± 2.11	70.66 ± 2.82	78.86 ± 2.51	**79.23 ± 2.86**	79.01 ± 3.31	69.35 ± 2.55
Mean accuracy	70.31	71.42	71.68	76.76	**78.12**	77.89	70.01

Table 5. Accuracy of each proposed method considering the multiclass classification of the dataset PCA3.

Backbone	Pretrained			Random+Pretrained			Fine-tunning
	CLS	GAP	CLS+GAP	CLS	GAP	CLS+GAP	
ViT-B [11]	41.11 ± 16.56	39.72 ± 15.81	43.06 ± 13.90	67.42 ± 9.50	65.95 ± 10.02	**68.46 ± 9.86**	49.3 ± 16.08
ViT-S [11]	37.78 ± 15.53	30.83 ± 20.20	34.17 ± 15.22	53.88 ± 8.44	**61.51 ± 8.14**	59.81 ± 8.00	48.56 ± 17.15
ViT-Betwix [10]	33.89 ± 16.52	30.83 ± 17.32	32.22 ± 13.56	**55.49 ± 9.54**	47.21 ± 10.52	51.68 ± 8.56	44.34 ± 16.08
ViT-RelPos [41]	13.33 ± 7.22	32.78 ± 8.77	28.61 ± 9.79	23.23 ± 10.32	**56.16 ± 9.36**	52.05 ± 8.29	40.20 ± 17.90
ViT-B dinov2 [23]	34.17 ± 20.03	34.17 ± 22.36	29.17 ± 18.64	48.65 ± 7.51	**52.11 ± 6.63**	51.05 ± 8.27	35.09 ± 23.07
DeiT-B [37]	40.83 ± 11.25	43.61 ± 15.52	45.00 ± 14.74	59.09 ± 8.80	**62.16 ± 7.87**	60.79 ± 8.48	37.45 ± 16.53
BeiTv2-B [24]	61.39 ± 16.59	61.39 ± 15.03	59.44 ± 14.50	**66.33 ± 8.61**	65.93 ± 8.88	64.58 ± 9.99	38.05 ± 22.87
BeiT-B [6]	39.44 ± 11.17	40.83 ± 13.15	36.39 ± 13.97	**64.61 ± 7.72**	56.26 ± 6.06	61.51 ± 5.81	44.51 ± 13.39
DeiT3-B [38]	29.44 ± 15.48	32.22 ± 11.12	30.56 ± 13.61	58.33 ± 7.84	**61.79 ± 6.26**	60.42 ± 7.25	45.43 ± 16.38
FlexViT [7]	53.89 ± 14.07	41.11 ± 15.41	50.83 ± 15.52	**70.53 ± 10.02**	68.75 ± 7.36	70.14 ± 8.30	49.57 ± 20.11
Mean accuracy	38.53	38.75	38.94	56.76	59.78	**60.05**	43.25

Table 5, which presents the results of the multiclass classification on the PCA3 dataset, shows an even greater drop in performance when compared to the binary classification results for this dataset. As expected, this is mainly due to the increased difficulty of multiclass classification. In addition, the PCA3 dataset poses a more challenging classification problem than the SARS-CoV-2 dataset, exhibiting even smaller variability in patterns among classes [28]. Nevertheless, results similar to those observed in Table 4 can still be identified. In particular, combining random and pretrained features consistently improves the performance of the evaluated methods, further highlighting the importance of

random features and feature fusion for constructing more robust image representations. In this scenario, fine-tunned models outperformed methods that rely solely on pre-trained features in terms of average accuracy, however, there is a significant improvement in performance compared to methods that combine random and pre-trained features, with an increase of 16.8% in the mean accuracy, which reaffirms the importance of features extracted by models without training. Finally, in this case, the CLS+GAP strategy with combined random and pretrained features achieved the highest average accuracy.

4.3 Impact of Transformers Architecture on Feature Extraction

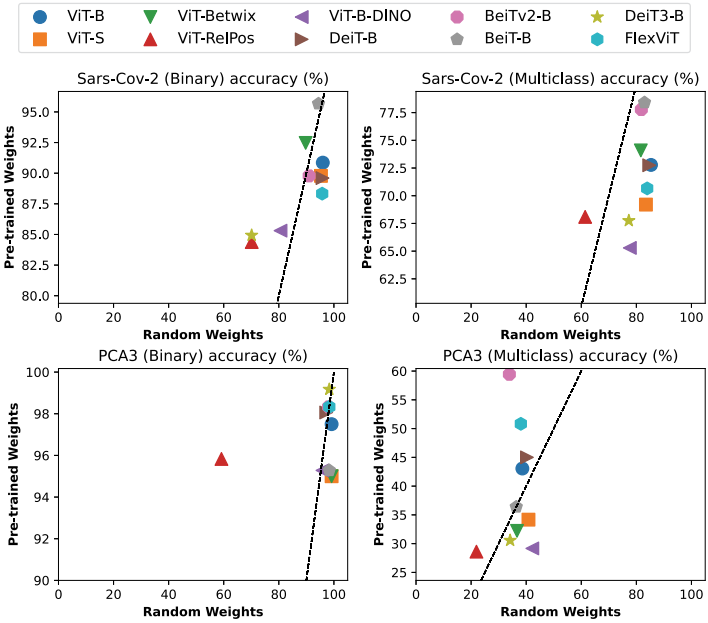

Fig. 3. Correlation between the accuracies of random and pretrained ViT models. The dotted line represents the diagonal of the plane, i.e., x = y.

In addition to exploring other feature extraction methods using ViTs, we also decided to analyze the influence of the Transformer architecture itself by comparing the robustness of features extracted by random and pretrained ViTs. Figure 3 shows the results of the SVM model used to classify the features extracted with the CLS+GAP method, since it was the method that yielded the best results in most cases. The graphs show the direct relationship between random ViT results (X-axis) and pretrained results (Y-axis), where it is also possible to verify the contribution of learned weights over random ones by observing the distance of

the points from the dashed line, which represents where the random results would be equal to the pretrained ones, i.e., $X = Y$.

For the SARS-CoV-2 dataset, in both binary and multiclass classification scenarios, the vast majority of models lie close to the dotted line or to its right. This indicates that, in general, randomly initialized models perform as well as, or even better than, their pretrained counterparts. These results suggest that, for the type of application explored in this work, Vision Transformers do not necessarily depend on pretrained weights to achieve strong performance, and that pretraining may even degrade performance in some cases.

In the binary classification of the PCA3 dataset, we observed the same case as before, where the models are close to or to the right of the dashed line, indicating that the random models achieved results as good as, or even better than, the pretrained ones, with the exception of the ViT-RelPos model, which showed a clear dependence on pretrained weights. In the multi-class classification of the PCA3 dataset, half of the models are to the left of the dashed line, indicating a certain dependence on pre-training for some models. However, these models still remain relatively close to this line, while many others lie above or to the right of it, once again indicating that ViT models are not strongly dependent on pretrained weights to perform feature extraction in biosensor images.

To provide a more rigorous analysis of why random ViTs can rival, and in some cases outperform, pretrained models, we conducted a comparative analysis of the attention maps produced by both settings. This analysis allows us to examine potential differences in attention patterns and in the way each model captures spatial and structural relationships within the images, offering more concrete insights into the factors that may influence their performance. Figure 4 shows representative samples from both datasets, together with the attention maps generated by the randomly initialized and pretrained models. A clear distinction can be observed between the two settings: randomly initialized models exhibit more spatially distributed attention, whereas pretrained models tend to concentrate attention on more localized regions. This behavior may be attributed to the biases acquired during ImageNet pretraining, whose images are predominantly object-centric. As a result, pretrained ViTs tend to focus on salient object regions learned from object-centric pretraining datasets [21], often assigning less attention to other image areas. In contrast, randomly initialized models are not constrained by these learned biases and therefore distribute attention more uniformly across the image. This property is particularly advantageous for texture-related tasks, where discriminative information is often homogeneous and spatially distributed, as observed in biosensor images.

This analysis further clarifies why concatenating random and pretrained features can improve performance, as reported in Sects. 4.1 and 4.2. The attention maps suggest that randomly initialized ViTs, even without training, can capture distributed textural information in biosensor images, likely because they are not constrained by the biases acquired during pretraining on generic object-centric datasets such as ImageNet [21]. In contrast, pretrained models benefit from large-scale training and can capture relevant structural patterns, but their learned

attention mechanisms may also lead them to focus on localized regions that are not always the most informative for texture-oriented biosensor data. Therefore, combining random and pretrained representations yields a more robust image descriptor by integrating complementary information captured by both model types. These findings also highlight the importance of the Transformer architecture itself for biosensor image feature extraction, since randomly initialized models achieved performance comparable to, and in several cases even better than, that of pretrained models.

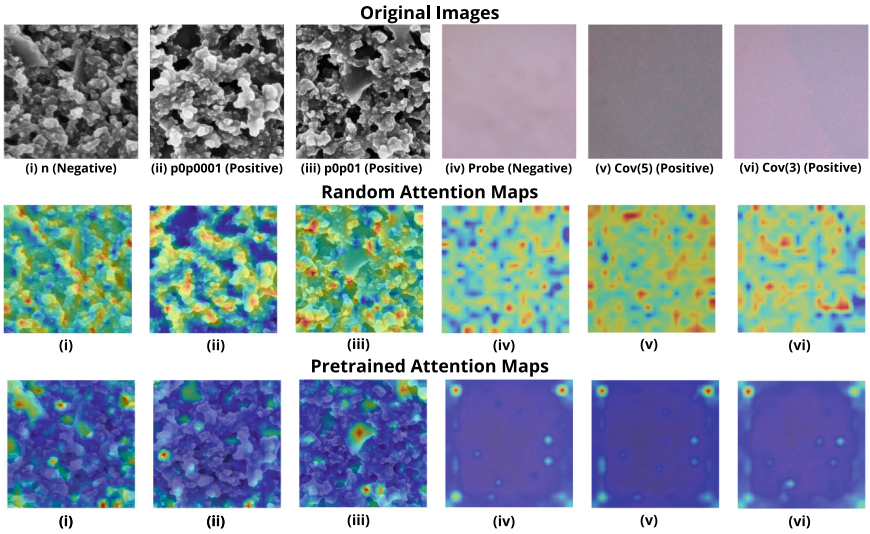

Fig. 4. Representative samples from the evaluated datasets with their corresponding attention maps obtained from random and pretrained ViT backbones.

4.4 CNN Baseline Comparison

Finally, in order to conduct a more in-depth analysis of our proposal, we decided to perform a brief comparison between the performance of different baseline CNNs and the proposed method. Table 6 shows the results of different CNNs used as feature extractors in the PCA3 and SARS-CoV-2 Immunosensor datasets, along with our own results. It is worth noting that, regarding CNNs, the features of the last layer were extracted and these features were used to train and test an SVM classifier, respecting the respective configurations of each dataset. In the case of ViTs, the results of the CLS+GAP method combining random and pretrained features were considered.

As we can see on Table 6, the features proposed in this study achieved overall better results compared to baseline CNNs. For the PCA3 dataset, the model

Table 6. Comparison of classification accuracies between baseline CNN-based features and the proposed representation (CLS+GAP strategy combining random and pretrained features) on the PCA3 and SARS-CoV-2 Immunosensor datasets. Bold values indicate the best performance.

Backbone	PCA3		SARS-CoV-2 Immunosensor	
	Binary	Multiclass	Binary	Multiclass
ConvNext-B [19]	96.11 ± 7.88	48.89 ± 9.31	90.45 ± 2.02	75.73 ± 3.08
ResNet-101 [14]	95.28 ± 8.06	37.78 ± 19.73	82.66 ± 2.64	63.05 ± 3.7
MobileNetV4 [26]	90.83 ± 10.1	20.28 ± 11.92	83.41 ± 2.47	62.6 ± 2.46
VGG 11 [33]	93.06 ± 9.4	30.56 ± 14.33	83.45 ± 3.36	66.45 ± 3.4
DenseNet-121 [15]	93.06 ± 11.74	42.22 ± 19.28	81.02 ± 1.88	63.8 ± 2.99
Proposed representation				
ViT-B	**100.0 ± 0.0**	68.46 ± 9.86	96.23 ± 1.28	79.57 ± 2.50
ViT-S	**100.0 ± 0.0**	59.81 ± 8.00	95.45 ± 1.22	77.67 ± 3.32
ViT-Betwix	99.65 ± 1.31	51.68 ± 8.56	94.44 ± 2.34	79.46 ± 2.75
ViT-RelPos	97.23 ± 3.23	52.05 ± 8.29	87.58 ± 2.21	70.36 ± 3.01
ViT-B dinov2	98.95 ± 2.11	51.05 ± 8.27	87.77 ± 2.53	72.37 ± 3.20
DeiT-B	**100.0 ± 0.0**	60.79 ± 8.48	**97.32 ± 1.28**	80.98 ± 2.81
BeiTv2-B	99.30 ± 1.79	64.58 ± 9.99	93.44 ± 1.70	**82.10 ± 2.94**
BeiT-B	98.26 ± 2.46	61.51 ± 5.81	97.13 ± 1.47	81.99 ± 2.65
DeiT3-B	**100.0 ± 0.0**	60.42 ± 7.25	88.74 ± 2.30	75.36 ± 3.56
FlexViT	**100.0 ± 0.0**	**70.14 ± 8.30**	96.53 ± 0.99	79.01 ± 3.31

FlexViT obtained the best overall results, Surpassing ConvNext results by 3.89% and 21.25% in binary and multi-class classifications, respectively. It is interesting to note that even though ConvNext-B obtained the best result among the CNNs, its performance was still worse than all ViTs, both in binary and multiclass classification in the PCA3 dataset.

In the case of the SARS-CoV-2 Immunosensor dataset, the model with the best performance in binary classification was DeiT-B, surpassing the accuracy of ConvNext-B by 6.87%, while in multiclass classification, BeiTV2-B showed the highest accuracy, surpassing the performance of ConvNext-B by 6.46%. Again, in the SARS-CoV-2 Immunosensor dataset, with the exception of the ViT-RelPos, ViT-B dinov2 and DeiT3-B models, the ViTs models significantly outperformed ConvNext, which was the CNN with the best result.

5 Conclusion

In this work, we conducted an investigation of Vision Transformers as feature extractors for biosensor image analysis, considering both genosensor and immunosensor datasets for the detection of COVID-19 and prostate cancer. Different token-based extraction and feature fusion strategies were explored, and

the use of both pretrained and random models allowed us to analyze the representational capacity of the Transformer architecture itself. This way, our work demonstrated that ViTs can serve as powerful feature extractors in biosensing applications without any type of end-to-end training or fine-tuning, making the approach viable for small-scale datasets. The results highlight the central role of random representations and feature fusion in improving classification performance. The combination of features extracted from random and pretrained models consistently increased the average accuracy across datasets and experimental settings, with random features acting as a form of regularization that mitigates biases inherited from large-scale pretraining datasets and yields more robust image descriptors. Moreover, in many scenarios, randomly initialized ViTs achieved performance comparable to, and sometimes surpassing, their pretrained versions, reinforcing the importance of the Transformer architecture itself in capturing relevant morphological and textural patterns from biosensor images, even without learned semantic priors from large-scale datasets.

For future work, it would be valuable to explore multi-layer feature aggregation in ViTs to investigate whether fusing tokens from different layers can further enrich image descriptors. Additionally, future studies could extend the approach to other biosensing applications, including genosensor-based imaging for COVID-19 detection and other infectious or oncological biomarkers, enabling a broader evaluation of Transformer-based representations in diverse biochemical and morphological contexts.

Acknowledgments. The authors acknowledge support from the São Paulo Research Foundation (FAPESP), Brazil (Process N. #2025/14029-8, #2025/27044-5, #2025/01057-3 #2024/04492-0 and #2024/04492-0) and Brazilian National Council for Scientific and Technological Development (CNPq) (grants #305206/2024-9 and #130337/2026-9).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, S., Sahu, S.P.: Image-based Parkinson disease detection using deep transfer learning and optimization algorithm. Int. J. Inf. Technol. (Singapore) **16**(2), 871–879 (2024)
2. Akbari Nakhjavani, S., Mirzajani, H., Carrara, S., Onbaşlı, M.C.: Advances in biosensor technologies for infectious diseases detection. TrAC - Trends Anal. Chem. **180** (2024)
3. Azad, R., et al.: Advances in medical image analysis with vision transformers: a comprehensive review. Medical Image Anal. **91** (2024)
4. Babaei, A., Pouremamali, A., Rafiee, N., Sohrabi, H., Mokhtarzadeh, A., de la Guardia, M.: Genosensors as an alternative diagnostic sensing approaches for specific detection of virus species: a review of common techniques and outcomes. TrAC, Trends Anal. Chem. **155**, 116686 (2022)

5. Bansal, M., Kumar, M., Sachdeva, M., Mittal, A.: Transfer learning for image classification using vgg19: Caltech-101 image data set. J. Ambient. Intell. Humaniz. Comput. **14**(4), 3609–3620 (2023)
6. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers (2022). https://arxiv.org/abs/2106.08254
7. Beyer, L., et al.: Flexivit: one model for all patch sizes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14496–14506 (2023)
8. Bhatt, G., Bhattacharya, S.: Biosensors on chip: a critical review from an aspect of micro/nanoscales. J. Micromanuf. **2**(2), 198–219 (2019)
9. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
10. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. In: International Conference on Learning Representations (ICLR) (2024)
11. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
12. Fam, M.K., Ismail, N.I., Zahari, A.I., Ghazali, N.A., Hassan, N.: Biosensor technologies in cancer: tools for early detection and prognostic monitoring. Biomarker Res. **14**(1) (2026)
13. Gomes, H.C., et al.: Laser-induced graphene-based fabry-pérot cavity label-free immunosensors for the quantification of cortisol. Sensors Actuators Rep. **7** (2024)
14. 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 (CVPR) (2016)
15. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
16. Ibau, C., Md Arshad, M., Gopinath, S.C.: Current advances and future visions on bioelectronic immunosensing for prostate-specific antigen. Biosens. Bioelectron. **98**, 267–284 (2017)
17. Kang, M., Song, H., Park, S., Yoo, D., Pereira, S.: Benchmarking self-supervised learning on diverse pathology datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3344–3354 (2023)
18. Kassal, P., Steinberg, M.D., Steinberg, I.M.: Wireless chemical sensors and biosensors: a review. Sens. Actuators, B Chem. **266**, 228–245 (2018)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
20. Nakowong, P., et al.: Detection of high-risk HPV 16 genotypes in cervical cancers using isothermal DNA amplification with electrochemical genosensor. Talanta **269** (2024)
21. Naseer, M., Ranasinghe, K., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Intriguing properties of vision transformers. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Red Hook, NY, USA. Curran Associates Inc. (2021)
22. Oiticica, P.R.A., et al.: Using machine learning and optical microscopy image analysis of immunosensors made on plasmonic substrates: Application to detect the sars-cov-2 virus. ACS Sensors **10**(2), 1407–1418 (2025). pMID: 39960416
23. Oquab, M., et al.: Dinov2: learning robust visual features without supervision (2024)

24. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: masked image modeling with vector-quantized visual tokenizers (2022). https://arxiv.org/abs/2208.06366
25. Qian, L., et al.: Biosensors for early diagnosis of pancreatic cancer: a review. Transl. Res. **213**, 67–89 (2019). https://doi.org/10.1016/j.trsl.2019.08.002
26. Qin, D., et al.: Mobilenetv4: Universal models for the mobile ecosystem. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV 2024. LNCS, vol. 15098, pp. 78–96. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73661-2_5
27. Ribas, L.C., Casaca, W., Fares, R.T.: Conditional generative adversarial networks and deep learning data augmentation: a multi-perspective data-driven survey across multiple application fields and classification architectures. AI **6**(2), 32 (2025)
28. Rodrigues, V.C., et al.: Electrochemical and optical detection and machine learning applied to images of genosensors for diagnosis of prostate cancer with the biomarker pca3. Talanta **222**, 121444 (2021)
29. Scabini, L., Zielinski, K.M., Ribas, L.C., Gonçalves, W.N., Baets, B.D., Bruno, O.M.: Radam: texture recognition through randomized aggregated encoding of deep activation maps. Pattern Recogn. **143**, 109802 (2023)
30. Scabini, L., Sacilotti, A., Zielinski, K.M., Ribas, L.C., De Baets, B., Bruno, O.M.: A comparative survey of vision transformers for feature extraction in texture analysis. J. Imaging **11**(9), 304 (2025)
31. Scabini, L., et al.: Vortex: challenging CNNs at texture recognition by using vision transformers with orderless and randomized token encodings. Neurocomputing 133852 (2026). https://doi.org/10.1016/j.neucom.2026.133852
32. Senan, E.M., Jadhav, M.E., Rassem, T.H., Aljaloud, A.S., Mohammed, B.A., Al-Mekhlafi, Z.G.: Early diagnosis of brain tumour MRI images using hybrid techniques between deep and machine learning. Comput. Math. Methods Med. **2022** (2022)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
34. Soares, J.C., et al.: Detection of a sars-cov-2 sequence with genosensors using data analysis based on information visualization and machine learning techniques. Mater. Chem. Front. **5**, 5658–5670 (2021)
35. Takahashi, S., et al.: Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. J. Med. Syst. **48**(1) (2024)
36. Togaçar, M.: Biotransnet: Detection of plant stress through the conversion of biosensor data into RGB channels and combination with transformer networks. J. Crop Health **77**(5) (2025)
37. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention (2021). https://arxiv.org/abs/2012.12877
38. Touvron, H., Cord, M., Jégou, H.: Deit III: Revenge of the VIT (2022). https://arxiv.org/abs/2204.07118
39. Wang, J., Yu, X., Gao, Y.: Feature fusion vision transformer for fine-grained visual categorization (2022)
40. Wang, X.O.: Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal. **79**, 102430 (2022)
41. Wu, K., Peng, H., Chen, M., Fu, J., Chao, H.: Rethinking and improving relative position encoding for vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)

42. Zerati, A.B.S., Guerra, L.B., Oliveira, O.N., Jr., Ribas, L.C.: Sars-cov-2 detection in genosensor images via deep learning and data augmentation. Revista de Informática Teórica e Aplicada **33**(2), 269–275 (2026)
43. Zhang, K., Cao, J., Li, H.s., Cai, Q.: Cross-layer feature fusion vision transformer for fine-grained visual classification. In: 2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS), pp. 1–6 (2023)
44. Zhou, H., Ma, K., Chen, H., Yao, R.: Training-free transformer architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9682–9691 (2022)
45. Zhou, Q., Zhu, Y.: When training-free NAS meets vision transformers: a neural tangent kernel perspective. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7405–7409. IEEE (2024)