Evaluating Vision-LSTM for Stress Detection

Mjellma Çitaku*, Silja Meyer-Nieberg, Marko Hofmann
Institute for Techincal Informatics, University of the Bundeswehr Munich,
Munich, Germany
mjellma.citaku@unibw.de*, silja.meyer-nieberg@unibw.de, marko.hofmann@unibw.de

***Abstract*—Vision-LSTM (ViL) has recently been introduced as a recurrent alternative to self-attention-based architectures, but its suitability for physiological data has not been studied. This paper provides the first systematic evaluation of ViL for stress detection from photoplethysmography-derived blood volume pulse (BVP) data. BVP segments are transformed into two dimensional image encodings, time-frequency scalograms and symmetric projection attractor reconstruction (SPAR), and ViL is compared against a Vision Transformer (ViT) under a strictly matched protocol: ImageNet-1K initialization, identical patchification, the same hyperparameter search space and tuning budget, and leave-one-subject-out (LOSO) evaluation. On WESAD dataset, ViL shows a consistent gain from SPAR across model sizes, while ViT achieves the strongest overall three-class performance and benefits more reliably from scalograms and increased capacity. In the binary stress vs. non-stress setting, ViL becomes competitive and slightly surpasses ViT in in-dataset LOSO. Cross-dataset experiments between the WESAD and UBFC-Phys datasets highlight a non-trivial domain shift and show that limited target-domain exposure can improve transfer. Overall, the results clarify when ViL is advantageous for physiological image encodings and highlight a strong architecture-representation interaction in BVP-based stress detection.**

***Index Terms*—Stress detection, blood volume pulse, image encodings, scalogram, SPAR, Vision-LSTM, Vision Transformer**

I. Introduction

Stress is a major aspect of everyday life. While acute stress can be adaptive and may even support performance in demanding situations, chronic stress can seriously affect both the physical and mental health. This can substantially reduce an individual's quality of life, making timely stress detection and early intervention essential [1].

As stress activates the autonomic nervous system, it alters the heart rate patterns, breathing rhythms, muscle tone, and sweat-related skin activity. These changes are captured by physiological signals, allowing stress-related dynamics to be monitored and analyzed [2]. These signals, raw or preprocessed, are commonly used as input to machine learning (ML) algorithms to determine stress states. Among these modalities, photoplethysmography (PPG) is particularly popular due to its non-invasive nature, low cost, and seamless integration into wrist-worn devices, from which the blood volume pulse (BVP) waveform can be derived.

Motivated by the strong performance of modern vision models on image recognition, one practical way to leverage pretrained backbones for physiological stress detection is to cast the task in an image-based form by transforming one dimensional (1D) time series into two dimensional (2D) image encodings (IEs) [3], [4]. This allows the model to learn discriminative patterns directly from the representation and reduces the reliance on handcrafted feature engineering [5]. A key question is which pretrained vision backbone transfers most robustly to subject-independent stress classification.

In this work, we focus on two IEs with strong performance for BVP data in previous studies: scalograms [6] and the symmetric projection attractor reconstruction (SPAR) [7], which has been recently adopted for physiological stress detection [8]. Scalograms emphasize non-stationary time-frequency structure via the magnitude of the continuous wavelet transform (CWT), while SPAR captures geometric aspects of the underlying signal dynamics via nonlinear state-space embeddings. Most existing IE-based stress detection approaches rely on convolutional neural networks (CNNs), whereas systematic evaluations of transformer-style vision backbones on benchmark physiological datasets remain limited.

Vision-LSTM (ViL) [9] was recently introduced as an alternative vision backbone that builds on xLSTM blocks and replaces self-attention with a recurrent token-mixing mechanism. Unlike the Vision Transformer's (ViT) global, context-aware attention, ViL introduces a directional, memory-based inductive bias through localized, sequence-like patch integration. While ViL has shown competitive or superior performance to ViT on several natural image benchmarks, its applicability to physiological image-based signal analysis remains unexplored. This is particularly relevant given the distinct visual structure of physiological IEs: scalograms exhibit dense, time-frequency texture, while SPAR attractors yield sparse, cyclic geometric forms derived from the underlying cardiovascular dynamics. These differences, both in data representation and architectural design, motivate a systematic comparison of ViL and ViT for stress detection across physiological IEs. To the best of our knowledge, this is the first work to evaluate ViL on physiological data. We investigate the following research questions:

- Under a strictly matched fine-tuning protocol, how do ViL and ViT compare for subject-independent BVP-based stress detection?
- How does the choice of IE interact with the inductive biases of ViT (global token mixing via self-attention) and ViL (directional recurrent token processing), and does this interaction change which architecture is preferable?

This paper is funded by dtec.bw - Digitalization and Technology Research Center of the Bundeswehr. dtec.bw is funded by the European Union - NextGenerationEU.

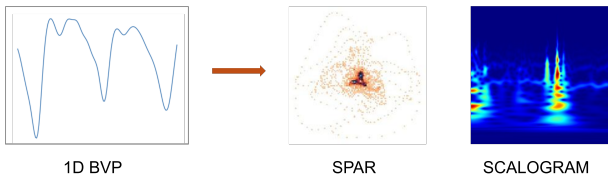

Fig. 1. Exemplary IEs of a converted 1D BVP segment.

- Which architecture generalizes better under domain shift, measured both by in-dataset subject-independent evaluation and by cross-dataset transfer between WESAD and UBFC-Phys?

This paper is organized as follows: Section II outlines the related work, followed by the methodology in Section III, which includes the datasets, architectures, and experimental design. Section IV presents the experimental results, and Section V concludes the paper with a summary and outlook.

II. Related Work

Physiological signals provide valuable insights into an individual's physical and emotional state and have long been used for stress detection. While early ML approaches relied on handcrafted features, more recent deep learning methods increasingly transform the 1D time series into 2D IEs to enable the use of computer vision architectures [10].

Time-frequency encodings such as scalograms have been widely applied to stress detection from PPG or electrocardiogram signals, predominantly in combination with CNNs [11], [12]. Beyond spectral representations, alternative encodings including Gramian Angular Fields and state-space representations such as SPAR have been proposed to capture geometric and temporal structure in physiological signals [8], [13], [4], [3]. More recently, transformer-based architectures have been explored for PPG-based analysis, including stress detection and related cardiovascular inference tasks, indicating their potential to model long-range dependencies in physiological image representations [14], [15], [16]. The ViL on the other hand, remains unexplored for physiological stress detection.

III. Methodology

In this section, the datasets used in this work are described first, followed by the IEs used to convert the BVP time series into 2D inputs. The ViT and ViL architectures are then summarized. Finally, the experimental protocol is detailed.

A. Datasets

All experiments were conducted on the provided BVP data from two public datasets: WESAD [17], and UBFC-Phys [18].

WESAD is a standard dataset in physiological stress detection and serves as the primary benchmark in this work. It contains multimodal physiological recordings from 15 subjects collected under a standardized laboratory protocol, including a baseline phase, a stress-induction phase based on the Trier Social Stress Test (TSST), and an amusement phase. In this paper, we considered the three-class setting (baseline / amusement / stress). For comparability with binary stress detection protocols, we additionally evaluated a binary mapping where baseline and amusement were merged into the non-stress class.

The **UBFC-Phys** dataset was used for the out-of-domain evaluation and cross-dataset transfer as it shares similar BVP acquisition characteristics with WESAD while differing in subjects and recording conditions. It includes recordings from 56 participants in a laboratory setting with a resting session and TSST-based stress induction. The protocol provides stress and non-stress phases and supports binary stress detection.

In line with common practice in physiological stress detection, the experimental phases provided by each dataset were used as class labels [3], [8], [17].

B. Image Encodings

Image encodings are 2D, image-like representations obtained by transforming 1D signals such as physiological time series, with each encoding highlighting particular properties of the underlying signal. Fig. 1 shows the two IEs used in this work: scalogram [6] and SPAR [7]. These were selected based on our prior work [8], [16] which showed consistent performance gains over other candidates, enabling a cleaner backbone comparison. A scalogram visualizes the magnitude of the CWT coefficients over time and frequency, thereby capturing transient, localized patterns that may be difficult to observe in either domain alone. SPAR is based on reconstructing a state-space attractor from the time series, with the goal of capturing aspects of the underlying signal dynamics. The resulting reconstruction is then mapped to a 2D image representation suitable for vision backbones.

Stress-related BVP changes are reflected in heart rate and beat-to-beat regularity. Scalograms capture these effects in the time-frequency domain, whereas SPAR offers a complementary dynamical view via reconstructed trajectory patterns.

C. Architectures

The ViL and ViT architectures both operate on patch-token sequences, however, they rely on different token-mixing principles and inductive biases. This motivates the controlled comparison conducted on physiological IEs.

Vision transformer adapts the Transformer encoder to vision applications by applying self-attention over a sequence of image patch tokens, allowing global context modelling and long-range interactions [19]. Instead of 1D token embeddings as in natural language processing, ViT splits an image into fixed-size 2D patches, which are then flattened and maped through a linear projection to obtain the patch embeddings. Positional embeddings are then added to the patch embeddings in order to preserve the spatial information.

Vision-LSTM [9] is a recently proposed vision backbone that replaces self-attention with a recurrent token-mixing mechanism built from xLSTM modules. Similar to ViT, ViL operates on a sequence of patch tokens, but integrates information via directional, sequence-like processing rather than global attention. In practice, patch tokens are processed in

alternating scan orders (e.g., forward and reverse traversals), enabling the information to propagate across the image while maintaining a recurrent inductive bias. While ViT's self-attention scales quadratically with the number of tokens, ViL's token-mixing mechanism scales linearly with sequence length.

As a convolutional reference, a ResNet50 [20] backbone pretrained on ImageNet-1K is evaluated. The original classification layer is replaced with a lightweight head consisting of dropout ($p = 0.5$), a 32-unit fully connected layer with ELU activation, and a final softmax layer with output dimensionality equal to the number of classes. ResNet50 uses a standard two-stage transfer-learning protocol, whereas ViT and ViL models are fine-tuned end-to-end. This reflects common practice for CNN transfer on small datasets and provides a strong, well-tuned convolutional reference rather than an intentionally constrained baseline. First, the head is trained with a frozen backbone (Adam, lr = 1.0e-3), then the full model is fine-tuned end-to-end (Adam, lr = 1.0e-5).

For ViT and ViL, a linear classification head whose output dimensionality matches the number of classes was used. For a fair comparison, their Tiny/ Small/ Base variants of similar sizes were compared.

D. Experimental Design

Data processing: The preparation of all datasets followed the same procedure. The raw BVP signals were split into 60 sec segments with a 0.75 window overlap, as this setting performed best in our previous evaluation [8] and is consistent with prior work [17]. Subsequently, the segments were transformed into scalograms and SPAR, resulting in 1120/610/316 (baseline/stress/amusement) images for WESAD and 504/1008 (baseline/stress) images for UBFC-Phys. The scalograms were generated via CWT using the ssqueezepy package [21], the generalized Morse mother wavelet, 16 voices per octave, and a JET colormap. For the SPAR images the embedding delay was set to one third of the mean NN interval estimated from NeuroKit2 [22] peak detection, and the attractor was rendered as a 100 x 100 density map using the icefire_r colormap.

Evaluation protocol: The subject-independent performance was evaluated using the leave-one-subject-out (LOSO) cross-validation on each dataset. In each fold, one subject was held out for testing, while the data of the remaining subjects were used for training and validation (80/20 split). The hyperparameter selection was performed exclusively on the WESAD dataset. Using the resulting best configuration, (i) in-dataset LOSO evaluation was conducted on WESAD and UBFC-Phys, and (ii) cross-dataset transfer under domain shift was evaluated by training on one dataset and testing on the other (WESAD $\leftrightarrow$ UBFC-Phys), both without and with 10% of the target-dataset labels used for tuning. To quantify the stochasticity, each experiment was repeated with 20 seeds. The same set of seeds was used for all model variants, ensuring identical data shuffling and train/validation splits across models within each LOSO fold. Unless stated otherwise, the accuracy and weighted F1-score were used as evaluation metrics due to the class imbalance, particularly in the three-class WESAD setting. Weighted F1 computes the per-class F1 scores and averages them using the class support as weights, yielding a single measure that accounts for both precision and recall while reflecting the relative prevalence of each class. For each seed, the performance across LOSO folds was aggregated by taking the median over folds (held-out subjects). The mean $\pm$ standard deviation across the 20 seeds were computed.

Statistical testing: The statistical significance was assessed using the two-sided Wilcoxon signed-rank tests [23] on paired per-seed LOSO scores, with the significance level set to $\mathrm{p} = 0.05$. The tests were limited to pre-specified ViL-ViT comparisons at matched model sizes and IEs.

Implementation details: We next describe the framework-specific implementation choices shared across all experiments. The ViT followed the JAX/Flax implementation [24] (with TensorFlow for data input), ViL used the PyTorch reference implementation [25], and ResNet50 the Keras implementation [26]. Despite using different frameworks, all models were trained on identical inputs and evaluated under the same protocol and metrics. The ViT data pipeline drops the final incomplete mini-batch by default. We instead padded the last batch and applied a validity mask so that padded entries were excluded from the loss and metrics computation, ensuring that no samples were discarded.

Hyperparameter tuning: To ensure a fair comparison between ViL and ViT, the fine-tuning protocol was standardized across all experiments. Both models were initialized from ImageNet-1K pretrained weights (ImageNet-21K weights are only available for ViT) and operated with identical patch tokenization (patch size 16×16), ensuring comparable input granularity and initialization. For hyperparameter selection, a matched procedure was applied: both models were tuned on WESAD using an identical image preprocessing, the same hyperparameter search space, the same tuning budget, and the same held-out tuning subjects. For tractability, the hyperparameter tuning was conducted using a partial LOSO procedure in which two subjects (S7 and S14) were reserved for tuning evaluation and held out in turn, while the remaining subjects were used for training. The hyperparameters were selected on WESAD by ranking configurations based on accuracy and weighted F1. Optimal hyperparameters were allowed to differ between ViL and ViT, since the objective was to compare the architectures under equal tuning opportunities rather than enforce a shared configuration that could favor one model.

A two-stage tuning strategy was employed. In Stage 1, a broad sweep was conducted on the Small variants of both architectures (Table I), covering the optimizer choice (AdamW/SGD), learning-rate and regularization ranges, and data augmentation options. All runs used linear warmup followed by cosine decay, early stopping, and a fixed input resolution of 224×224 with normalization to $[-1, 1]$. In Stage 2, the Tiny and Base variants were evaluated using a narrowed grid (Table II) from Stage 1 to reduce redundant exploration while keeping the evaluation protocol unchanged.

TABLE I
FINE-TUNING CONFIGURATION AND HYPERPARAMETER SWEEP SPACE.

Parameter	Value
Epochs	50
Batch size	64
Seeds	20
Optimizer	
AdamW	
Learning rate	{1e-3, 1e-4, 2.5e-4, 5e-4, 7.5e-4}
Weight decay	{0.0, 0.05}
Moments	$\beta_1 = 0.9,\ \beta_2 = 0.999$
SGD	
Learning rate	{1e-3, 3e-3, 1e-2, 3e-2}
Momentum	0.9
Learning-rate schedule	linear warmup $\rightarrow$ cosine decay
Warmup epochs	5
Early stopping	
Patience	7
Min. Δ	0.001
Grace period	15
Precision	float32
Data augmentation	
Horizontal flip	{true, false}(0.5 if true)
Random cropping	{true, false}
Crop area min	{0.5, 0.8} (only if cropping=true)
Crop area max	1.0 (fixed)
Resize	
Interpolation	bicubic
Image size	224×224
Normalize	$[-1, 1]$

TABLE II
STAGE-2 NARROWED HYPERPARAMETER GRIDS FOR TINY AND BASE VARIANTS DERIVED FROM THE STAGE-1 (SMALL) SWEEP IN TABLE I. ALL OTHER TRAINING SETTINGS FOLLOW TABLE I.

Variant	Optimizer	Learning-rate candidates
ViT-Ti/16	SGD	{1e-3, 3e-3, 5e-4, 7.5e-4}
ViT-B/16	SGD	{1e-3, 3e-3, 5e-4, 7.5e-4}
ViL-Tiny	AdamW	{2.5e-4, 5e-4, 7.5e-4}
ViL-Base	AdamW	{1e-4, 2.5e-4, 5e-4}

Random cropping removed in Stage 2; horizontal flip retained as optional.

IV. EXPERIMENTAL RESULTS AND EVALUATION

This section reports the LOSO and cross-dataset results obtained with the tuning protocol described in Section III-D. Unless stated otherwise, the best configuration for each model variant was selected on WESAD using the accuracy and weighted F1 metrics, and the same selected recipe was applied in all subsequent in-dataset and transfer experiments. Table III summarizes the LOSO main performance results, while Table IV shows the cross-dataset results. To keep the experimental matrix tractable, ResNet50 was included only for the primary in-dataset LOSO setting as a calibrated convolutional reference. All results were computed over the same set of 20 random seeds across models to ensure identical data shuffling and validation splits within each LOSO fold.

A. Outcome of Hyperparameter Selection

The Stage 1 sweeps on the Small variants revealed architecture-specific optimization preferences: ViT-S/16 performed best with SGD (best region around 1e-3) and with horizontal flipping disabled, whereas ViL-Small favored AdamW (best at 5e-4, WD=0.0) with horizontal flipping enabled. Random cropping was consistently detrimental for both architectures and was therefore removed from further consideration, while horizontal flipping was retained as optional due to its architecture-dependent effect. Guided by these observations, Stage 2 restricted each model to its preferred optimizer family and performed a localized search around the strongest learning-rate regions (Table II). The resulting best Tiny/Base configurations were then fixed for all subsequent experiments. For ViT, ViT-Ti/16 used SGD with 7.5e-4 (flip on) and ViT-B/16 used SGD with 5e-4 (flip off). For ViL, ViL-Tiny used AdamW with 2.5e-4 and WD=0.05, while ViL-Base used AdamW with 2.5e-4 and WD=0.0 (flip on in both cases).

B. LOSO Evaluation Results

Table III reports the LOSO performance on WESAD for both IEs. A clear interaction between architecture and representation is observed. For ViL, SPAR consistently improves the accuracy over scalograms across model sizes, whereas for ViT the strongest results are obtained with scalograms. A plausible interpretation is that the two IEs emphasize different structures: scalograms provide dense time-frequency patterns where discriminative cues can be distributed across distant regions, which may favor the ViT's global token interaction, whereas SPAR produces sparse, trajectory-like shapes dominated by the local geometry, which may be captured more effectively by ViL's directional, recurrent token mixing [6], [7]. This suggests that the benefit of an IE is not universal, but depends on how well its visual structure matches the inductive bias of the architecture.

Across model sizes, ViT exhibits a stronger scaling trend than ViL, with the Base variant achieving the best overall three-class performance on WESAD. Vision-LSTM remains competitive, particularly on SPAR where it closes part of the gap to ViT and surpasses its own scalogram-based performance. These trends are further reflected in Fig. 2, which shows the accuracy distributions across seeds. Furthermore, ViT exhibits tighter inter-seed variability than ViL, suggesting a slightly higher stability under the fixed protocol.

Table III also reports the binary stress detection results (stress /non-stress). The experiments were limited to scalograms and the Base variants to keep the computational budget manageable. On binary WESAD, ViL-Base outperforms ViT-B/16. A similar trend is observed on UBFC-Phys, where ViL-Base again yields higher accuracy and weighted F1 than ViT-B/16. Overall, the results show that the relative ranking can change between the three-class and binary formulations, and that ViL can match or slightly surpass ViT in the binary setting under the same tuning protocol.

Robustness to tuning subjects: For comparability with the literature, the LOSO results are reported using the standard

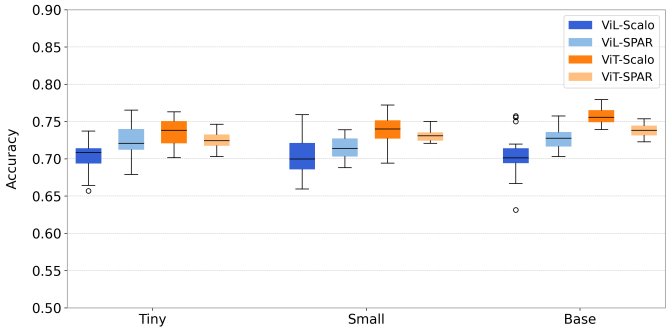

Fig. 2. Accuracy distributions for ViL and ViT models across different sizes (Tiny, Small, Base) for Scalograms and SPAR for the three-class WESAD classification task.

TABLE III
LOSO CLASSIFICATION PERFORMANCE (MEAN ± STD) ACROSS DATASETS FOR 3-CLASS AND BINARY SETTINGS, REPORTED IN TERMS OF ACCURACY AND WEIGHTED F1 ACROSS 20 SEEDS FOR SCALOGRAMS AND SPAR. RESNET50 SERVES AS A BASELINE.

Dataset	IE	Model	Accuracy	weighted F1
3-class classification				
WESAD	Scalo	ViL-Tiny	0.704±0.021	0.701±0.016
		ViL-Small	0.704±0.028	0.699±0.025
		ViL-Base	0.705±0.033	0.701±0.026
		ViT-Ti/16	0.735±0.018	0.712±0.010
		ViT-S/16	0.739±0.022	0.725±0.016
		ViT-B/16	**0.757±0.012**	**0.727±0.008**
		ResNet50	0.720±0.014	0.682±0.015
	SPAR	ViL-Tiny	0.726±0.022	0.706±0.013
		ViL-Small	0.714±0.014	0.695±0.012
		ViL-Base	0.728±0.015	0.711±0.015
		ViT-Ti/16	0.726±0.011	0.708±0.009
		ViT-S/16	0.732±0.009	0.703±0.011
		ViT-B/16	**0.738±0.009**	**0.709±0.009**
		ResNet50	0.711±0.017	0.698±0.017
Binary classification				
WESAD	Scalo	**ViL-Base**	**0.914±0.016**	**0.911±0.017**
		ViT-B/16	0.905±0.005	0.903±0.006
		ResNet50	0.894±0.020	0.894±0.022
UBFC-Phys	Scalo	**ViL-Base**	**0.833±0.021**	**0.829±0.023**
		ViT-B/16	0.826±0.013	0.823±0.017
		ResNet50	0.830±0.022	0.828±0.022

protocol over all 15 WESAD subjects. Because hyperparameters were selected using a partial LOSO procedure involving subjects S7 and S14, the LOSO performance on WESAD was additionally recomputed after excluding these two subjects. The scalogram results drop by roughly 1-3 percentage points, whereas the SPAR results changed only marginally (typically < 1 percentage point). Importantly, the architecture-encoding conclusions remained unchanged (ViT remains stronger on scalograms, while ViL benefits more from SPAR), whereas the binary ViL advantage did not persist under this exclusion.

Statistical comparison (ViT vs. ViL): The Wilcoxon signed-rank tests on paired per-seed LOSO scores ($n = 20$) confirmed the descriptive trends. On WESAD scalograms, ViT significantly outperforms ViL for all model sizes in both accuracy and weighted F1 ($p < 0.01$). On SPAR, the performance gap was reduced: the accuracy differences remained statistically significant for all model sizes, whereas the weighted F1 differences were significant only for the Small variant and not for Tiny or Base, indicating a weaker and metric-dependent advantage. In the binary WESAD setting (Base, scalograms), ViL showed a marginal accuracy advantage ($p \approx 0.044$), but the weighted F1 difference was not significant, suggesting that this effect is small and not robust across metrics.

TABLE IV
CROSS-DATASET TRANSFER ON SCALOGRAMS (MEAN ± STD, 20 SEEDS).

Transfer	Model	Accuracy	weighted F1
WESAD → UBFC-Phys	ViL-Base	0.694 ± 0.025	0.695 ± 0.027
	ViT-B/16	0.694 ± 0.014	0.686 ± 0.016
WESAD → UBFC-Phys*	ViL-Base	0.690 ± 0.025	0.692 ± 0.026
	ViT-B/16	0.702 ± 0.034	0.693 ± 0.035
WESAD* → UBFC-Phys*	ViL-Base	0.719 ± 0.036	0.716 ± 0.033
	ViT-B/16	0.724 ± 0.017	0.721 ± 0.015
UBFC-Phys → WESAD	ViT-B/16	0.571 ± 0.022	0.568 ± 0.025
	ViL-Base	0.530 ± 0.041	0.514 ± 0.049

[a] WESAD* includes subjects s1-s6 from UBFC-Phys additionally. UBFC-Phys* excludes s1-s6.

C. Cross-dataset Evaluation Results

Table IV reports the cross-dataset transfer results between WESAD and UBFC-Phys. When training on WESAD and testing on UBFC-Phys, both models transferred similarly in terms of accuracy, with ViL achieving a slightly higher weighted F1. In the reverse direction (UBFC-Phys → WESAD), the performance dropped substantially for both models, suggesting a stronger domain shift, but ViT generalized more robustly than ViL. Across both transfer directions, ViT exhibited lower inter-seed variability, whereas ViL showed a larger variance, particularly when trained on UBFC-Phys.

This asymmetric transfer can arise even under similar protocols and the same sensor type, because the learned decision boundary depends on the source data distribution. One contributing factor is the class composition: UBFC-Phys is stress-dominant (1008 stress vs. 504 baseline images), whereas WESAD contains a larger non-stress portion and an additional amusement condition. When mapped to binary labels, WESAD defines the non-stress class as baseline+amusement, while UBFC-Phys provides baseline only. This mismatch in the class priors and in what constitutes the non-stress class can bias training and reduce the transfer performance in the UBFC-Phys→WESAD direction, despite a similar acquisition setup.

To further probe the effect of limited target-domain exposure, Table V reports a partial adaptation setting in which 10% of UBFC-Phys subjects ($s_1 - s_6$) were added to the WESAD training set (denoted WESAD*), while validation remained restricted to WESAD. To ensure a fair comparison, the resulting models were evaluated on UBFC-Phys*, i.e., UBFC-

Phys with the same subjects ($s_1 - s_6$) removed. Compared to WESAD→UBFC-Phys*, adding these subjects during training improved the performance for both models: ViT increased from 0.70 to 0.72 accuracy (weighted F1: 0.69 to 0.72), and ViL increased from 0.69 to 0.72 accuracy (weighted F1: 0.69 to 0.72). These gains suggest that even a small amount of target-domain training data can improve transfer performance without using target-domain data for model selection.

V. Conclusion and Outlook

This work presented a controlled comparison of Vision-LSTM (ViL) and Vision Transformer (ViT) for BVP-based stress detection from physiological image encodings (IEs) under a matched fine-tuning protocol, thereby introducing ViL as a vision backbone for this setting. In terms of the relative performance of ViT and ViL under matched training conditions, ViT achieved the best overall results for the primary the three-class WESAD classification task and showed a clearer scaling benefit from Tiny to Base, particularly on scalograms. ViL was generally weaker in this setting but became competitive in the binary formulation, restricted to scalograms and Base variants for tractability, where it achieved comparable performance to ViT under in-dataset LOSO evaluation. With respect to the interaction between IE and architectural inductive bias, the choice of representation was found to interact strongly with the architecture. SPAR consistently improved ViL over scalograms, whereas ViT benefited more reliably from scalograms, indicating that no single encoding is universally optimal. When considering generalization under domain shift, the transfer performance between WESAD and UBFC-Phys was direction-dependent: WESAD→UBFC-Phys yielded comparable performance for both models, while UBFC-Phys→WESAD favored ViT, suggesting higher robustness in that direction. Finally, adding only 10% target subjects to the WESAD training set improved the performance on the remaining UBFC-Phys subjects, showing that limited target-domain exposure can reduce the transfer gap without using target-domain data for model selection. Overall, ViT is the stronger choice for three-class WESAD when using scalograms, while ViL is a viable alternative whose strengths emerge with SPAR and in the binary setting. Future work will explore additional IEs and state-space-based vision architectures to further improve robustness and cross-dataset generalization.

References

[1] World Health Organisation, "Stress," https://www.who.int/news-room/questions-and-answers/item/stress, 2023, accessed: 2026-01-28.

[2] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[3] M. Quadrini, S. Daberdaku, A. Blanda, A. Capuccio, L. Bellanova, and G. Gerard, *Stress Detection from Wearable Sensor Data Using Gramian Angular Fields and CNN*, 11 2022, pp. 173–183.

[4] S. Ghosh, S. Kim, M. F. Ijaz, P. K. Singh, and M. Mahmud, "Classification of Mental Stress from Wearable Physiological Sensors Using Image-Encoding-Based Deep Neural Network," *Biosensors*, vol. 12, no. 12, 2022.

[5] D. Altunkaya, F. Y. Okay, and S. Ozdemir, "Image transformation for IoT time-series data: A review," *arXiv preprint arXiv:2311.12742*, 2023.

[6] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, ser. Prentice Hall Signal Processing Series. Prentice Hall, 2013.

[7] P. J. Aston, M. I. Christie, Y. H. Huang, and M. Nandi, "Beyond HRV: attractor reconstruction using the entire cardiovascular waveform data for novel feature extraction," *Physiological Measurement*, vol. 39, no. 2, p. 024001, mar 2018.

[8] M. Çitaku, L. Zott, S. Meyer-Nieberg, C. Küsel, and M. Hofmann, "Images of Stress: Investigating Image Encodings and Transfer Learning for Stress Detection," in *2025 IEEE Symposium on Computational Intelligence in Health and Medicine*, 2025, pp. 1–7.

[9] B. Alkin, M. Beck, K. P oppel, S. Hochreiter, and J. Brandstetter, "Vision-lstm: xlstm as generic vision backbone," *arXiv preprint arXiv:2406.04303*, 2024.

[10] Z. Wang and T. Oates, "Encoding Time Series as Images for Visual Inspection and Classification Using Tiled Convolutional Neural Networks," *In Workshops at the 29th AAAI Conference on Artificial Intelligence.*, 01 2015.

[11] M. Amin, K. Ullah, M. Asif, A. Waheed, S. U. Haq, M. Zareei, and R. R. Biswal, "ECG-Based Driver's Stress Detection Using Deep Transfer Learning and Fuzzy Logic Approaches," *IEEE Access*, vol. 10, pp. 29 788–29 809, 2022.

[12] Y. Hasanpoor, B. Tarvirdizadeh, K. Alipour, and M. Ghamari, "Wavelet-Based Analysis of Photoplethysmogram for Stress Detection Using Convolutional Neural Networks," in *2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM)*, 2023, pp. 501–506.

[13] J. Venton, P. M. Harris, A. Sundar, N. A. S. Smith, and P. J. Aston, "Robustness of convolutional neural networks to physiological electrocardiogram noise," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2212, p. 20200262, 2021.

[14] H. Barki, S. H. Chung, R. Jafari, and W.-Y. Chung, "Stress-ViT: In-Ear Plethysmography for Mental Stress Classification with Vision Transformer," in *2023 IEEE Biomedical Circuits and Systems Conference*. IEEE, 2023, pp. 1–5.

[15] S. H. Darbhasayanam, Noorjahan, H. A. A. Mohammad, and A. R. Komalla, "Photoplethysmogram (PPG)-Based Blood Pressure Estimation using Vision Transformer Networks: A Deep Learning Approach," in *2024 3rd International Conference on Artificial Intelligence For Internet of Things*, 2024, pp. 1–6.

[16] M. Çitaku, S. Meyer-Nieberg, and M. Hofmann, "Attention or Convolution? Comparing Modern Architectures for Stress Detection on BVP Image Encodings," in *Artificial Intelligence XLII: 45th SGAI International Conference on Artificial Intelligence, AI 2025, Cambridge, UK, December 16-18, 2025, Proceedings, Part I*. Berlin, Heidelberg: Springer-Verlag, 2025, p. 320–326.

[17] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection." New York, NY, USA: Association for Computing Machinery, 2018.

[18] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappé, and F. Yang, "UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 622–636, 2023.

[19] A. Dosovitskiy *et al.*, "Vision Transformer and MLP-Mixer Architectures," https://github.com/google-research/vision_transformer, 2025, accessed: 2025-06-25.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE CVPR*, 2016, pp. 770–778.

[21] J. Muradeli, "ssqueezepy," *GitHub. Note: https://github.com/OverLordGoldDragon/ssqueezepy/*, 2020, accessed: 2026-01-30.

[22] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021. [Online]. Available: https://doi.org/10.3758%2Fs13428-020-01516-y

[23] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[24] A. Dosovitskiy *et al.*, "Vision Transformer and MLP-Mixer Architectures," https://github.com/google-research/vision_transformer, 2024, accessed: 2026-01-22.

[25] B. Alkin *et al.*, "Vision-LSTM (ViL)," https://github.com/NX-AI/vision-lstm, 2024, accessed: 2026-01-22.

[26] F. Chollet *et al.*, "Keras Applications," https://keras.io/api/applications/, 2026, accessed: 2026-01-30.