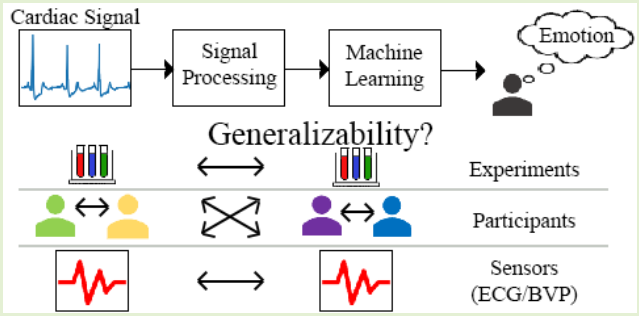
On the Robustness of Machine Learning Models for Stress and Anxiety Recognition From Heart Activity Signals

John Henry, Huw Lloyd, *Member, IEEE*, Martin Turner, and Connah Kendrick

***Abstract*—Many recent studies have addressed the detection of negative affective states such as stress and anxiety from physiological signals taken from body-worn sensors. Typically, machine learning classifiers are applied to features derived from sensor signals, and several authors have reported high accuracy results from a range of signals including cardiac, skin conductance, and skin temperature. However, the issue of how robust these models are for deployment in the field is rarely addressed. In this article, we use open data from two large experimental studies to evaluate the generalizability of models derived from cardiac signals, focusing on detection of stress and anxiety. We choose the cardiac signal since the commonly used heart rate variability features can be derived from multiple sensor modalities, allowing us to evaluate the robustness of models within, as well as between, experimental settings. We show that consistent classification outside the original experimental setting relies on high-quality training data with minimal artifacts, and that models may often train on proxies within the noise of lower quality data. Our results also underline the importance of including a wide range of emotional states in the training data to minimize erroneous classification from unseen regions of feature space.**



***Index Terms*— Emotion recognition, generalizability, machine learning, physiological signals.**

I. Introduction

MACHINE learning for emotion recognition has gained increasing interest from the scientific community, since the early 2000s. Cowie et al. [1] introduced their investigation into the use of emotion recognition in human-centered computing, outlining the potential for emotion recognition through multiple sources of information, powered by artificial intelligence. Since 2001, several studies have produced machine learning algorithms that classify positive and negative valence [2], [3]. Valence in relation to emotion describes an axis of emotion in a circumplex model of affect. The scale plots emotions against arousal as a means of depicting emotions in context with state. [4]. While substantial improvements are made available in classification accuracy of positive and negative valence [5], [6], [7], no available research focuses on the generalizability of machine learning models for negative valence, such as stress and anxiety. The generalizability of the related machine learning models improves impact and expands the potential application area. Consider the use of such machine learning algorithms in game-based solutions, where participants cannot stand still during data collection or observe other strict laboratory setting conditions.

This article investigates the generalizability of machine learning models for emotion recognition with data sourced from physiological signals, with rigid experiment processes, in particular using blood volume pulse (BVP) and electrocardiogram (ECG) data for predicting negative affect from different open emotion recognition datasets.

In this article, we consider "rigid" experiments to be those in which the data collection protocols require participants to follow specific rules that would be unlikely to be replicated outside the experimental setting without prior training or supervision. For example, having to keep a hand steady, having to avoid standing or sitting, or having to focus on a specific visual cue during the experiment. The generalizability in this context of the rigidity of the experimental conditions questions how replicable accuracies reported by literature would be in nonrigid settings, such as using these models

Manuscript received 17 March 2023; revised 28 April 2023; accepted 29 April 2023. Date of publication 23 May 2023; date of current version 29 June 2023. The associate editor coordinating the review of this article and approving it for publication was Dr. Theerawit Wilaiprasitporn. *(Corresponding author: John Henry.)*

John Henry, Huw Lloyd, and Connah Kendrick are with the Department of Computing and Mathematics, Manchester Metropolitan University, M15 6BH Manchester, U.K. (e-mail: john.henry@mmu.ac.uk).

Martin Turner is with the Department of Psychology, Manchester Metropolitan University, M15 6BH Manchester, U.K.

Digital Object Identifier 10.1109/JSEN.2023.3276413

in applications where participants can take part from home.

The use of rigid data collection techniques suits some scenarios, but neglect settings where users may not be constantly supervised to execute a strict protocol around the use and configuration of physiological sensors. Furthermore, our research explores the accuracy of pervasive sensors for a satisfactory accurate prediction of emotion. This article presents the following contributions.

1) We propose a novel methodology based on testing reasonable hypotheses on open datasets, which can be used to evaluate the generalizability of new machine learning models *outside* of their original experimental environment.
2) We evaluate the generalizability of BVP- and ECG-sourced data, focusing on prediction accuracy, using open datasets gathered under different laboratory conditions.
3) We show that the generalizability between sensor modalities of cardiac signals (BVP and ECG) depends on high-quality data, and that for lower quality data models will train on proxies within the noise.
4) We identify no meaningful improvement in classification between the machine learning and deep learning algorithms in the context of generalizability.

This article is structured as follows: Section II presents related background research, highlighting the motivation for our research. Section III details the methodology undertaken and explores the methods of testing prediction accuracy in terms of generalizability across multiple datasets. Section IV presents the results from several methods of evaluation. Finally, Section V concludes on our research, highlighting limitations and areas for future works.

II. Motivation and Background

The generalizability of the machine learning models is an underresearched field, with limited existing literature stating generalizability as a consideration for future research [7], [8]. The generalizability for the machine learning algorithms that consider positive and negative valence is restricted by complicated sensor setups that require rigid processes to be followed by participants. We consider how pervasive technology for data collection with proven prediction accuracy can support the generalization of the machine learning models in real-world settings. We focus on two data sources, widely regarded as accurate indicators of positive and negative valences; ECG and BVP sensors [9], [10], [11].

A recent investigation into stress updated the definition to consider the latest physiological understanding on the topic. Stress, therefore, describes a condition in which an organism's natural regulatory capacity cannot respond adequately to external environmental factors [12]. Anxiety relates to a negative emotion state that differs from healthier states of apprehension such as fear [13]. Though there are biological similarities between fear, stress, and anxiety, fear is an emotion state that occurs usually through external factors, whereas anxiety relates to the ability to cope with stress or fear inducing circumstances [14] and depends on an uncertain threat that has existential implications having to do with one's identity [15].

Multiple investigations into stress recognition [3], [7], [16], [17] and anxiety recognition [2], [5], [6], [9] show highly accurate predictions using several machine learning techniques. A recent review into wearable devices for stress, depression, and insomnia detection highlighted that low-cost devices coupled with machine learning can aid in mental health monitoring [18]. This article considers both stress and anxiety as examples of negative affective emotion states, considered as negative valence. A constant theme in this line of research is a rigid experimental setup that requires participants to remain in specific posture or conditions, varying greatly from circumstances in the field.

Conventionally, anxiety recognition is performed through questionnaires [19]. Over time, more research presented machine learning algorithms that recognize and classify anxiety. Research into anxiety classification using support vector machine (SVM) achieved high classification accuracy [2], [9], [20] with recent work including deep learning approaches [21].

Similarly, stress recognition practices revolve around self-reflective instruments [22]. The subjectivity of the self-reported questionnaires and the response bias that derives from their use generated research into machine learning stress recognition using physiological markers. An investigation into stress recognition using wearable sensors achieved a prediction accuracy of over 75% focusing on skin conductance, skin temperature, and accelerometer data [3]. Their research acknowledged the requirement of richer data to increase the robustness of their predictions in affective systems. Similar research explored the use of sourcing multiple physiological sources for affective stress detection [16]. Their research achieved 90% accuracy using cardiac and skin conductance features, predicted through SVM classifiers. Their findings, however, are limited to controlled settings, leaving room for further exploration on the robustness of such practices in the field.

A subset area of research in the field of machine learning emotion recognition is on-line or real-time recognition. Though there are similarities, predictions are required at a faster and more regular interval. Research into anxiety recognition for virtual reality exposure therapy (VRET) achieved a high accuracy of over 80% on four-point anxiety recognition [5]. Their research provides advances in the field of anxiety and stress recognition, but the validity of the experiment in field settings is not explored. Similarly, research into on-line anxiety recognition achieved high prediction accuracy using physiological data. A notable difference in their method was a training period which can improve the robustness of their results [6].

A review into the literature on machine learning emotion recognition discovered multiple sensors and physiological data can recognize a variety of emotion states, including stress and anxiety [7]. Their review found that further research is required in unconstrained settings to study validity, robustness, and generalizability. Our research aligns with their findings and investigates the generalizability of the machine learning techniques across multiple open-source

TABLE I
MODEL EVALUATION METHODS CITED IN THE LITERATURE

Reference	k-fold CV	LOSO CV	Subject hold-out
[5]			✓
[6]			✓
[7]	✓		
[8]	✓		
[10]	✓		
[11]	✓	✓	
[12]		✓	

datasets. Research into the robustness of machine learning stress recognition investigated the impact of several hardware sources on prediction accuracy [8]. Their research discovered that hardware and placement are crucial factors to performance metrics, supporting further research on the robustness of machine learning stress and anxiety recognition.

This article explores the problem of robustness and reliability and generalizability of the existing machine learning algorithms, including deep learning approaches for tabular data, for negative valence detection, focusing on stress and anxiety. We investigate whether the proposed models from previous authors can return accurate predictions on similar datasets. No research to our knowledge examines the claims of stress recognition accuracy for generalizability. Recent research highlighted the importance of new studies investigating the existing claims of machine learning accuracy for emotion classification, including effectiveness when the data collection methodology is generalized [7]. Furthermore, the evaluation methods in the literature vary, with some using leave one subject out (LOSO) cross-validation (CV) or subject hold-out to evaluate the generalization of classification performance between subjects, and others using k-fold CV across the entire dataset; the evaluation of generalizability even within the same experimental setting is therefore not universally followed. Table I illustrates the distribution of validation techniques in the included background literature. Finally, we present a hypothesis testing methodology on open datasets that evaluates the potential for the machine learning models to generalize *outside* of the original experimental environment, differing from the cited research which tests generalizability between subjects *within* an experiment.

III. METHODOLOGY

Our novel method for testing the generalizability of the machine learning techniques for stress recognition includes intraexperiment evaluation and interexperiment evaluation between two biosignal open datasets commonly used in related research. We detail our process of data sanitization and feature extraction required to perform this evaluation below.

A. Datasets

The *continuously annotated signals of emotion* (CASE) dataset [11] was gathered in an experiment on 30 participants, 15 each of male and female, with eight sensor modalities capturing physiological signals. For this work, we use the ECG and BVP signals, which are both captured using a 16-bit analog-to-digital converter. The dataset includes annotation data, gathered from a joystick device that the participants use to continuously self-report their emotional state during the experimental protocol. This two-axis joystick provides readings of *Arousal* and *Valence* and is captured at a low rate of 20 Hz. The joystick device is annotated with symbols from the *self-assessment manikin* (SAM) [23] to give the participants nonverbal cues to identify emotions. During the experiment, the participants watch emotion elicitation videos which are categorized as *amusing*, *boring*, *relaxing*, and *scary*.

The *wearable stress and affect detection* (WESAD) dataset [10] provides physiological sensor data on 15 participants who underwent an experimental protocol which includes *baseline*, *amusement*, *meditation*, and *stress* states. The stress state is induced using the *trier social stress test* [24]. The dataset also includes self-reported data from the participants based on questionnaires and SAM. For this work, as in [10] we use the experimental protocol itself to provide the ground truth. The sensor modalities include ECG data from a chest-worn device, which is sampled at 700 Hz and a BVP signal from a wrist-worn device which is sampled at 64 Hz.

B. Signal Processing

The features used in this study all relate to heart rate variability (HRV). The goal of the signal processing, therefore, is to produce for each participant and sensor modality a list of beat times. We first filter the signals to clean them as far as possible before running a peak detection process to find the beat times.

1) BVP Data: For both the datasets, we find issues in the BVP data which need to be addressed in the signal processing before beat detection. For the CASE data, we find occasional large spikes in the data, which are clearly unrelated to the underlying cardiac signal. Since the data are captured at a very high rate, we can effectively remove these using a *median filter* over 21 samples. We then apply a third-order Butterworth bandpass filter with low- and high-frequency cutoffs of 0.5 and 5 Hz, respectively, to remove low-frequency drift and high-frequency noise. Note that the Nyquist frequency corresponding to the median filter interval is higher than the cutoff of the bandpass filter. A typical sample of the unfiltered and filtered signals is shown in Fig. 1.

For the WESAD data, we find that the BVP signal is often of low quality and appears to be dominated at times by large amplitude variations apparently unrelated to the underlying cardiac signal. This may be due to sensor movement in the wrist-worn device; in the CASE protocol, the participants wore finger-mounted BVP sensors which would have limited movement of the hand, whereas in WESAD, participants would have more freedom of movement. This component is difficult to remove as it typically occurs at frequencies of interest in the cardiac data. For the WESAD data, a median filter was not used as there is no evidence of "spike" noise as in the CASE data. A third-order Butterworth bandpass filter with low- and high-frequency cutoffs of 0.32 and 3.2 Hz was used to de-noise the data. Example signals demonstrating typical issues in the data are shown in Fig. 1.

2) ECG Data: In both the cases, the ECG data are clean and free of artifacts and were treated with a third-order

Butterworth high-pass filter with a cutoff of 0.5 Hz to remove the dc component and low-frequency drift. Example signals are shown in Fig. 1.

3) Beat Detection: We apply a simple beat detection algorithm which is based on identification of peaks in the data within a sliding window. We move a window of width 1.0 s over the signal and take the time at which the maximum signal value occurs within the window as a peak, providing it is within the central 40% of the window. These peak time values are recorded in a hashed set data structure (a Python `set`) so that repeat detections are handled. The constraint on the maximum value occurring close to the center of the window is effective at preventing false detections of subsidiary peaks, such as the second wave following the dicrotic notch, which is clearly visible in the CASE BVP data (see Fig. 1). The raw peaks obtained using the windowing method are further filtered using the following scheme. Let the measured beat times be $t_i, i \in [1, N]$ where N is the number of detected beats. Then, for any beat time t_j, $j > 3$, we predict the time of the next beat, t_{j+1}, from the average interval over the preceding three beats, that is,

$$t_{j+1}^{\text{pred}} = t_j + \frac{1}{3}(t_{j-1} - t_{j-4}). \tag{1}$$

We then compare this predicted time to the following two beat detections t_{j+1} and t_{j+2}. If $|t_{j+2} - t_{j+1}^{\text{pred}}| < |t_{j+1} - t_{j+1}^{\text{pred}}|$, we remove t_{j+1} from the list as a spurious beat detection. The detected beat times are shown as vertical lines overlaying the plots of filtered BVP and ECG data in Fig. 1. Finally, the lists of beat times are converted into a list of *interbeat intervals*, given by $\Delta_i = t_{i+1} - t_i, i \in [1, N' - 1]$ where N' is the number of beat times remaining after removal of spurious beats. Note that if any group of the three peaks used for prediction includes one or more spurious peaks, the effect is to produce predicted peaks which occur early. Hence, the next "true" peak will always be accepted if there is not another intervening spurious peak. In this way, the method will recover from any accepted spurious beats as soon as it has been presented with three consecutive true beats.

C. Feature Extraction

We selected 11 commonly used HRV features and derived these from both the ECG and BVP data, for all the participants in both the experiments. The features are calculated at 6-s intervals for each participant, over a window covering the preceding 30 s of heart rate interval measurements. The features used are presented in detail in the following subsections.

1) Heart Rate Features: The two heart rate features, μ_{HR} and σ_{HR}, are the mean and standard deviation of the *heart rate* derived from intervals which start in the 30-s window. That is, given a set of intervals Δ_i which begin in the window, μ_{HR} is the mean value of Δ_i^{-1} and σ_{HR} is the standard deviation of Δ_i^{-1} in the interval.

2) HRV Features: HRV features are based on the differences in measured heart rate between successive intervals. The HRV in interval i is therefore given by $\Delta_i^{-1} - \Delta_{i-1}^{-1}$. Three features are calculated from these values, μ_{HRV}, σ_{HRV}, and rms_{HRV}, which are, respectively, the mean, standard deviation, and root mean square of the values of $\Delta_i^{-1} - \Delta_{i-1}^{-1}$ for which t_i lies within the interval.

We also include two features derived from th Poincaré plots [25]; for these features, the interbeat intervals are plotted on a scatter plot against the same data-shifted by one or more intervals. The principal axes of the data are then found using principal component analysis, and the ratio of the standard deviations of the data projected onto these axes (SD1/SD2) is calculated. We derive two features using this method, for a delay of one and two intervals, which we denote `poincare_1` and `poincare_2`.

Finally, two additional features are calculated from the HRV, *NN50* and *pNN50*. These are the number and percentage of interbeat intervals within the window which deviate by more than 50 ms from the previous interval.

3) Spectral Features: The frequency components of HRV are commonly divided into four spectral bands: ultralow frequency (ULF, 0.01–0.04 Hz), low frequency (LF, 0.04–0.15 Hz), high frequency (HF, 0.15–0.4 Hz), and ultrahigh frequency (UHF, 0.4–1.0 Hz) [26]. Note that these frequency bands refer to the Fourier components of the secular variation in the heart rate, not the BVP or ECG signals themselves. Here, we calculate the power in the LF and HF bands. Since the heart rate is unevenly sampled, Fourier analysis is not straightforward, and these features can be better estimated using the *Lomb–Scargle periodogram* [27]. We compute the periodogram of the sampled heart rate in a 30-s window (that is, the values of Δ_i^{-1} sampled at times t_i) for 36 evenly spaced frequencies between 0.04 and 0.4 Hz and sum the values in the relevant frequency ranges before normalizing with the total power to derive two features, `LFn` and `HFn` (low-frequency normalized power and high-frequency normalized power). These represent the normalized low- and high-frequency components of the HRV. Normalizing these features to the total power has been shown to be important in reducing the effect of variation in the overall power [28].

D. Ground Truth

The ground truth for the WESAD data is obtained directly from the experimental protocol; that is, the annotations of each data point as belonging to the *baseline*, *amusement*, *meditation*, or *stress* phases of the protocol are used to label the rows of training data as *stress* or *nonstress*. For the CASE data, we use the continuously annotated values of arousal and valence to derive a composite measure of negative affect. We define an *anxiety index*

$$\alpha = A(1 - V) \tag{2}$$

where $A and V \in [0, 1]$ are the arousal and valence, respectively. We then choose a cutoff of 0.5 above which we categorize a data point as belonging to the *anxiety* state. This cutoff value is arrived at from an examination of the distributions of α values recorded in the experiment (see Fig. 2); although values of less than 0.5 are recorded while participants are in the "scary" protocol, it is clear that this

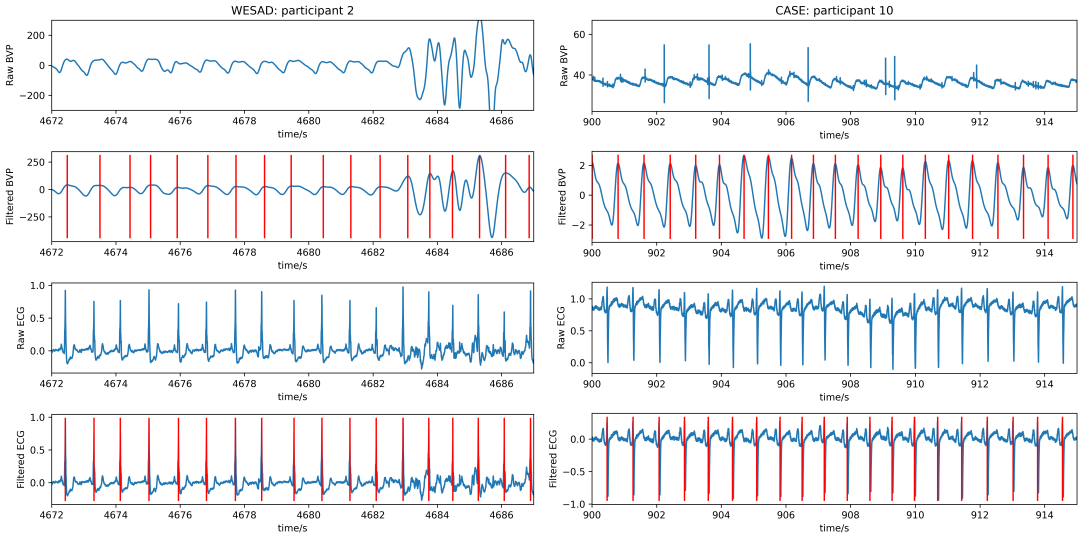


Fig. 1. Samples of raw and processed ECG and BVP signals from the CASE (right) and WESAD (left) datasets. From top to bottom, the plots show the raw BVP, filtered BVP, raw ECG and filtered ECG signals in a 15-s interval. The CASE raw BVP signal shows the characteristic "spikes" which are removed by median filtering. The effect of the filtering on the ECG signal is to remove the dc component and low-frequency drift. Detected beats are shown as vertical lines on the filtered data plots. Note the artifacts in the WESAD BVP signal in the last 5 s of the sample, where the signal varies at a large amplitude compared with the underlying cardiac signal. Detected beats are shown as vertical lines on the filtered data plots.

cutoff excludes the majority of data from the other protocols (bar the outliers for one participant). Note that this is a general measure of strongly negative affect, which is labeled as "anxiety" here but could include many other related states. Our aim here is to investigate the generalizability of emotion measures in general, rather than to distinguish between a range of closely related psychological states.

Fig. 2 shows histograms of α values for all the 30 participants in the CASE study, categorized by the four phases of the protocol, *amusing*, *boring*, *relaxing*, and *scary*. Clearly, the higher values of α predominantly occur during the *scary* protocol, justifying the use of this measure as a classifier for the ground truth.

E. Model Fitting and Evaluation

1) Training Data: The features for both ECG and BVP sensor modalities, and the target class derived from the ground truth, were sampled every 6 s during the active phases of the experiment protocols. The smaller of the two classes (*stress* and *no stress* for WESAD and $\alpha \geq 0.5$ and $\alpha < 0.5$ for CASE) for each experiment were retained in their entirety, and the other class was randomly sampled to produce a balanced dataset. This produced four datasets, the BVP and ECG datasets for the CASE experiment, each comprising 1188 examples, and the BVP and ECG datasets for the WESAD experiment, each comprising 3334 examples. Each feature set was normalized by removing the mean and scaling to unit variance. For each experiment and modality, the dataset comprises 11 features and a binary classification target. In summary, the dimensions of the datasets (features $\times$ time points) are (11 $\times$ 1188) for CASE (ECG and BVP), (11 $\times$ 3334) for WESAD (ECG and BVP). The time points are not evenly distributed between participants.

2) Classifiers and Hyperparameter Optimization: We used three "classical" machine learning classifiers: kernel SVM with a radial basis function kernel, random forests (RFs), and extreme gradient boosting (XGBoost). For the two tree ensemble classifiers, XGBoost and RF, we optimized the hyperparameters for each dataset using a random grid search of 100 trials, with a nested fivefold CV. For SVM, we used a grid search with nested fivefold CV over the two hyperparameters (the regularization parameter C and the inverse square width of the radial basis function kernel, γ). We selected these models since these are typical of the models used in the literature cited above.

We also include three neural network algorithms. Recent work on using deep learning for tabular data has shown promising results, including the use of transfer learning [29]. We train three deep learning networks developed for tabular data [30]: multilayer preceptron (MLP), ResNet [31], and a Feature-Tokenizer transformer (FTT) [32]. For training, we used MLP and ResNet without pretraining, and FTT with and without pretrained weights. We denote the pretrained FTT model as FTTP (FTT pretrained). We use a heart attack classification dataset for our pretraining [33]. We train each model using MSE loss optimized with Adam and a batch size of 128. We use patience of ten epochs for early stopping, based on a validation split of 10% of the training data. We used these models as they represent the state-of-the-art tabular data with deep learning. We train the networks as regressors, with a target of 0 or 1 for the two classes, and then find the optimal value of the regression target to split the two classes in the training data. This same value is then used for the validation data.

3) Evaluation: The models were evaluated using different variations in CV on the complete datasets, retraining at each fold using the best hyperparameters found in the optimization exercise. For the evaluation of the models on a given dataset, a tenfold CV was used. For evaluating how a model trained on one sensor modality generalizes to the other modality, we again used tenfold CV, but this time at each fold we evaluated the model on the left-out examples in the other

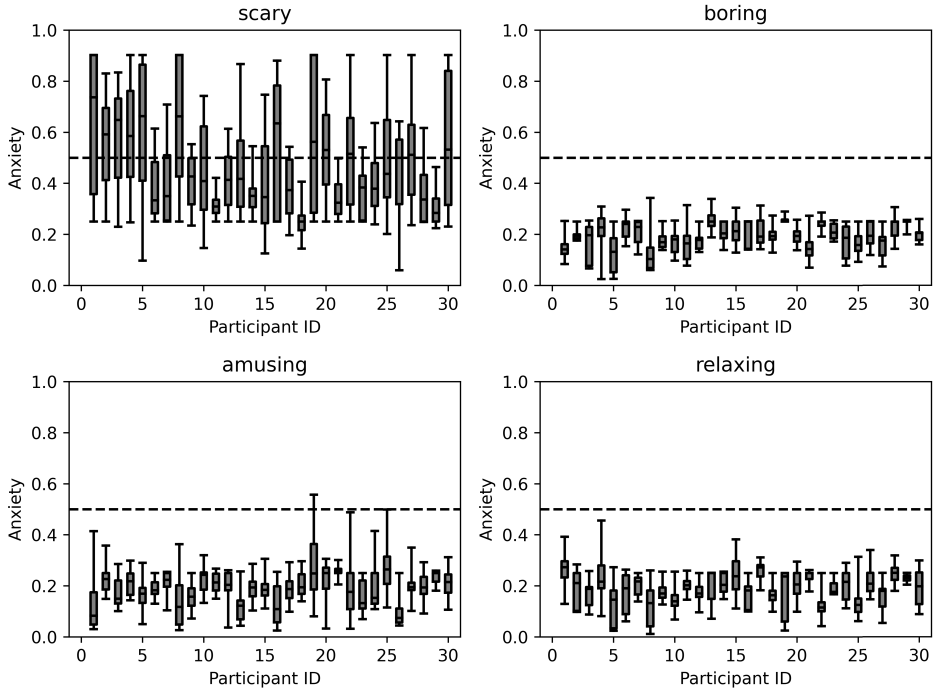


Fig. 2. Boxplots of the "anxiety index" α for each of the 30 participants in the CASE study, split by the current video category at the time of sampling the emotional state. Each panel shows the distribution of anxiety index values recorded for each participant for one of the video categories. The boxes show the interquartile range, with the whiskers showing the full range of values. The dotted line is drawn at an anxiety index of 0.5, which was chosen as the threshold to distinguish the negative affect state.

sensor modality. For example, in evaluating how a model trained on BVP data generalizes to the ECG data gathered in the same experiment, at each fold we train on the BVP data, but use the ECG data for the examples omitted from the training fold to evaluate.

For evaluating the generalization between participants, we used a LOSO CV. For n participants, we produce n folds, in which each participant is left out of the training data once and used to evaluate the models trained on the other $n-1$ participants. For interexperiment evaluation, we train models on the entire training data for one experiment and evaluate the models on the training data for the other experiment.

IV. Results

A. Intraexperiment Evaluation

In this section, we present the results of evaluating the machine learning models within each of the two experimental datasets. The models are evaluated using CVs; for each sensor modality, a tenfold CV over the whole dataset to determine the overall classification performance, a tenfold CV between sensor modalities in both the directions, and finally a LOSO CV to assess the generalizability to unseen subjects. The results are summarized in Table II which show the accuracy and F_1 score accumulated over the folds in the CVs.

For the CASE experiment, we see that the classification performance is roughly constant across the sensor modalities and algorithms, with typical accuracies and F_1 scores in the low 70%s. The generalization between sensor modalities is good, with only slight reductions in the metrics. Typically, the performance metrics shift from the low 70s to the mid 60s. The generalization between participants in the LOSO evaluation is relatively poor, however, with a much greater reduction in performance than observed in the intersensor evaluation. We suggest that this may be due to the self-reporting mechanism in this dataset. If the protocol induces consistent responses in the participants, but the reporting of the responses is inconsistent, we may expect the models to generalize less well between the subjects. The distributions of anxiety index (which are given in Fig. 2) show that the participants report quite different ranges; for example, participant 29 reports a very narrow range of α values, with none in the range > 0.5. Participants 17, 18, and 28 report generally low values, while other participants such as 1–5 report a wide range of values.

For the WESAD data, we see a difference in performance between the BVP and ECG modalities, with the ECG sensor showing the best classification performance in either experiment (around 80% accuracy and F_1 score), whereas the performance of the BVP data is below 70% on both the metrics. We have already noted in Section III-B.1 that the WESAD BVP data show many more artifacts than are seen in the CASE data. Comparing our results to those reported in [10], we see that with RF, the only common algorithm, we obtain worse classification performance on the BVP data, but very similar performance with the ECG data. This may be due to the differences in the beat detection algorithms, the size and frequency of the data windows, and the features chosen; [10] includes some HRV features not used here. The intermodality generalization in the WESAD data is considerably worse than that seen in the CASE data. We suggest that the BVP models are fitting features in the HRV data which are related to the artifacts although these may still relate to real signals which are significant for

TABLE II
RESULTS OF FITTING AND EVALUATING MACHINE LEARNING MODELS TO THE BVP AND ECG FEATURES ON THE CASE AND WESAD DATASETS. THIS TABLE SUMMARIZES THE RESULTS OF THE INTRAEXPERIMENT EVALUATION. BEST-PERFORMING ALGORITHMS ON ACCURACY OR F_1 SCORE ARE INDICATED IN BOLD

	Evaluation (train/test)	SVM		RF		XGBoost		MLP		ResNet		FTT		FTTP	
		acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1
CASE	BVP/BVP	0.733	0.740	**0.752**	**0.759**	0.735	0.736	0.704	0.713	0.704	0.729	0.659	0.620	0.455	0.453
	ECG/ECG	**0.763**	**0.770**	0.756	0.753	0.730	0.729	0.667	0.704	0.707	0.730	0.677	0.632	0.533	0.536
	BVP/ECG	**0.668**	**0.673**	0.637	0.618	0.646	0.633	0.601	0.583	0.624	0.626	0.606	0.581	0.522	0.463
	ECG/BVP	0.635	0.618	0.641	0.591	**0.653**	0.636	0.607	**0.643**	0.614	0.628	0.623	0.591	0.461	0.519
	BVP (LOSO)	0.560	0.517	0.580	0.533	0.584	0.541	**0.600**	0.569	0.583	**0.599**	0.543	0.510	0.455	0.453
	ECG (LOSO)	**0.590**	0.569	0.582	0.533	0.582	0.537	0.566	0.547	0.580	**0.582**	0.572	0.511	0.530	0.533
WESAD	BVP/BVP	**0.745**	**0.773**	0.732	0.757	0.737	0.761	0.729	0.763	0.731	0.762	0.718	0.759	0.555	0.599
	ECG/ECG	**0.811**	**0.818**	0.795	0.800	0.796	0.800	0.801	0.813	0.805	0.816	0.795	0.811	0.560	0.534
	BVP/ECG	0.586	0.313	0.568	0.253	0.561	0.236	**0.600**	0.354	0.596	0.348	0.582	0.305	0.596	**0.396**
	ECG/BVP	0.615	**0.711**	0.617	0.624	0.640	0.661	**0.642**	0.671	0.623	0.637	0.630	0.671	0.538	0.680
	BVP (LOSO)	0.713	0.741	0.708	0.733	0.708	0.730	0.708	0.740	**0.715**	**0.746**	0.703	0.744	0.555	0.599
	ECG (LOSO)	0.715	0.715	**0.720**	0.712	0.711	0.704	0.705	0.707	0.693	0.699	0.706	**0.722**	0.560	0.534

classification. For example, if the movement of the subjects is affecting the BVP signal, and the nature of the movement is related to the presence or absence of stress, the models trained on the BVP data may make use of this. This would not only explain the poor intermodality generalization (since the features are measuring different effects) but also explain the low performance of the BVP models here compared with [10]; our peak detection algorithm may reject many of the peaks induced by the artifacts, which may survive the signal processing in [10]. The interparticipant generalizability in the WESAD data is considerably better than for the CASE data. The ground truth in WESAD is derived from the protocol, rather than self-reporting; our comments in the previous paragraph relating to self-reporting in the CASE data may be relevant here to explain the greater consistency between subjects seen in WESAD.

Performance across all the algorithms is broadly similar, with the exception of the pretrained FTT model, which performs less well than the non-pretrained version. This is due to the small pretraining dataset where it is well known that transformers require large datasets to pretrain [34]. In addition, we used a heart-related dataset. However, due to the difference in data structure, it could be similar to out-of-domain transfer learning having a negative impact on the model, creating locally inductive bias. Overall, we do not see an advantage to using deep learning models on these data; the broadly similar performance across a diverse range of algorithms suggests that the choice of algorithm is relatively unimportant for this data.

B. Interexperiment Evaluation

For the interexperiment comparison, we evaluate trained on data from one of the experiments on the data from the other experiment. For example, we can train a model on the CASE BVP data and use the model to make predictions from the WESAD BVP data. Since the experimental protocols are different and the models are predicting different states, metrics such as accuracy or F_1 scores should be treated with caution. The results of this evaluation are shown in Table III. The models trained on the CASE ECG dataset generalize to some extent, but otherwise the performance on these metrics is relatively poor.

Although we would not expect models to predict different protocols across experiment, we should expect the results to show some consistency. For example, we do not expect to see a higher prevalence of *relaxed* states in the set of CASE samples for which the WESAD models predict *stress* than in the *no stress* class. Similarly, we might expect the $\alpha > 0.5$ class from the CASE models to show a lower prevalence of *meditation* and a higher prevalence of *stress* in the WESAD data. These results of these hypotheses are not predetermined. It is possible due to the nature of the generalizability experiments that we observe results beyond the expected. Our comparison is therefore largely qualitative; however, it still enables us to draw conclusions on the robustness or otherwise of the models especially in cases where we see behavior in the opposite sense to our reasonable expectations.

To carry out this analysis, we first propose a number of reasonable hypotheses which we would expect to be supported if the models generalize between the experiments, and then apply statistical tests to determine whether these are consistent with the data. In particular, we test the following hypotheses, labeled $H1$–$H5$.

- *H1:* The mean value of the anxiety index α for subjects predicting *stress* in models trained on the WESAD data is greater than the mean for subjects predicting *nonstress*.
- *H2:* Subjects in the CASE experiment indicating *stress* in models trained on the WESAD data are more likely to be in the *scary* protocol than subjects indicating *nonstress*.
- *H3:* Subjects in the CASE experiment indicating *stress* in models trained on the WESAD data are less likely to be in the *relaxing* protocol than subjects indicating *nonstress*.
- *H4:* Subjects in the WESAD experiment indicating $\alpha > 0.5$ from models trained on the CASE data are more likely to be in the *stress* protocol than subjects indicating $\alpha < 0.5$.
- *H5:* Subjects in the WESAD experiment indicating $\alpha > 0.5$ from models trained on the CASE data are less likely to be in the *meditation* protocol than subjects indicating $\alpha < 0.5$.

TABLE III

RESULTS OF EVALUATING MODELS ON THE SAME SENSOR MODALITY IN THE OTHER EXPERIMENT. MODELS ARE TRAINED ON THE ENTIRE TRAINING DATASET FOR ONE EXPERIMENT AND EVALUATED ON THE ENTIRE TRAINING DATASET FOR THE OTHER EXPERIMENT. BEST-PERFORMING ALGORITHMS ON ACCURACY OR F_1 SCORE ARE INDICATED IN BOLD

	SVM		RF		XGBoost		MLP		ResNet		FTT		FTTP	
Train/Test/Modality	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1	acc.	F_1
CASE/WESAD/BVP	0.564	**0.694**	0.458	0.278	0.516	0.494	**0.587**	0.686	0.577	0.650	0.500	0.586	0.536	0.682
CASE/WESAD/ECG	0.621	0.659	**0.657**	0.682	0.638	0.675	0.556	0.683	0.594	0.640	0.590	**0.694**	0.580	0.521
WESAD/CASE/BVP	**0.501**	0.003	**0.501**	0.003	**0.501**	0.003	**0.501**	0.003	**0.501**	0.003	0.500	0.007	0.480	**0.107**
WESAD/CASE/ECG	0.604	0.499	0.588	0.445	**0.605**	0.468	0.582	0.441	0.577	0.468	0.597	0.474	0.508	**0.554**

TABLE IV

RESULTS OF THE STATISTICAL TESTS ON HYPOTHESES $H1$–$H5$ FOR INTEREXPERIMENT COMPARISON USING BVP AND ECG DATA. p-VALUES ARE GIVEN FOR WELCH'S t-TEST ($H1$) OR PEARSON'S χ^2 TEST ($H2$–H5). p-VALUES LESS THAN 10^{-4} ARE SHOWN AS 0.0. ASTERISKS INDICATE THAT THE DATA SHOW THE OPPOSITE TREND TO THE HYPOTHESIS. FIGURES IN BOLD REPRESENT p-VALUES BELOW THE SIGNIFICANCE THRESHOLD OF 0.02

	SVM		RF		XGB		MLP		ResNet		FTT		FTTP	
	BVP	ECG	BVP	ECG	BVP	ECG	BVP	ECG	BVP	ECG	BVP	ECG	BVP	ECG
$H1$	0.029	**0.0**	0.029	**0.0**	**0.011**	**0.0**	0.220	**0.0**	0.529	**0.0**	0.524	**0.0**	**0.0**	0.686
$H2$	**0.016***	0.0	0.440	**0.0**	**0.005***	**0.0**	0.122	**0.0**	1.0	**0.0**	1.0*	**0.0**	**0.0***	**0.008***
$H3$	0.644*	0.986	**0.009**	0.211*	**0.019**	0.310*	1.0	0.823	1.0	0.492*	0.256	0.711	**0.0***	**0.0***
$H4$	**0.0**	**0.0**	**0.0***	**0.0**	**0.014**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	0.706*	**0.0**
$H5$	0.726*	**0.0***	0.700*	**0.0***	0.206*	**0.0***	0.073	**0.0***	0.299*	**0.0***	0.809	**0.0***	0.013	**0.0**

We evaluate $H1$ using Welch's t-test on the distributions of α values and the remaining hypotheses using Pearson's χ^2 test on a 2×2 contingency table. The results of the hypothesis testing are summarized in Table IV which gives the p-values derived from the statistical tests. Where the observation contradicts the hypothesis the p-value carries an asterisk superscript; for example, for $H2$ using the BVP data, the subjects indicating stress in the WESAD model are *less* likely to be in the *scary* protocol of the CASE experiment.

$H1$ is supported by the data at a statistically significant level for the ECG data for all the models apart from FTTP. For the same hypothesis evaluated on BVP data, we see poorer performance, with only the FTTP model giving a statistically significant result, although in all the cases the data agree with the hypothesis. $H2$ is supported by all the models, again apart from FTTP, for the ECG data. For the BVP data, we see mixed results including some cases with a statistically significant result in the sense opposite to the hypothesis; that is, participants indicating *stress* in the WESAD BVP models are *less* likely to be in the *scary* protocol of CASE than participants indicating *nonstress*. $H3$ shows weak results, with only the RF and XGB models trained on BVP data showing statistically significant support for the hypothesis. Overall, the WESAD model shows some generalizability to the CASE data; it is able to select the *scary* protocol in the CASE experiment, which qualitatively best matches the *stress* protocol in the original experiment, although the results are mixed, with much weaker effects seen for the second and third hypotheses.

Hypothesis $H4$ is strongly supported by the data in most cases, suggesting that the models largely generalize well to the WESAD data and are able to select the *stress* protocol based on the predicted values of α. However, there are exceptions in the case of the RF and FTTP models trained on BVP data. $H5$ gives a surprising result; in most cases, the CASE models are more likely to select subjects in the WESAD *meditation* protocol in its indications of high anxiety, with strong signals in many cases, particularly the ECG models. We suggest that this may be due to the absence of states in the CASE protocol sufficiently similar to the *meditation* state in WESAD; the models may therefore not have seen sufficient data similar to this state in training. The unseen regions of feature space may then give rise to counter-intuitive predictions. This underlines the importance of using models in a context in which the range of emotional states likely to be encountered is all well-covered in the training data.

In summary, we see strong indications that the models are able to produce consistent results in a different experimental setting. The best generalizability is seen for the higher quality data (ECG for both the experiments, and BVP for CASE) in cases where the experimental protocols are qualitatively similar. Weaker or inconsistent results are seen in cases where the training or evaluation data are of lower quality, or where the emotions induced by the experiment protocols differ widely.

V. CONCLUSION

This article presents an examination into the accuracy and effectiveness of machine learning emotion recognition when considering generalizability. It is crucial future works continue placing strong emphasis on generalizability to extend applications and algorithms from fixed-experiment conditions to widely adopted, multiconditional settings. After extracting standard features from cardiac signals, we evaluated ML classification algorithms using CV within and between sensor modalities and LOSO CV between subjects to determine how well ECG and BVP signals classify emotions in multiscenario settings and with unseen data. We show that the signal quality is a crucial factor to the success of generalizability. ECG data generally performed better than BVP across all the validation techniques. Weaker results were found with signals showing significant noise and artifacts, which we found to be

more prevalent in the BVP data, particularly in the WESAD dataset. Our study proposes all future research in the domain considers the quality of data before data capture occurs, to maximize the generalizability of the models. We also provide a methodology for assessing the generalizability of models by testing reasonable hypotheses on open datasets. Furthermore, we suggest future research considers multiple machine learning techniques to develop emotion recognition that is generalizable and appropriate for particular applications. For example, emotion trend calculations could aid in the discovery of new predictions for on-line emotion recognition.

REFERENCES

[1] R. Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui, "Anxiety recognition using relevant features from BVP signal: Application on phobic individuals," *Model., Meas., Control J., Model., Meas., Control*, vol. 75, pp. 131–141, Oct. 2014.

[3] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 671–676.

[4] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychol.*, vol. 39, no. 6, p. 1161, Dec. 1980.

[5] J. Šalkevicius, R. Damaševičius, R. Maskeliunas, and I. Laukien, "Anxiety level recognition for virtual reality therapy system using physiological signals," *Electronics*, vol. 8, no. 9, p. 1039, Sep. 2019.

[6] F. R. Ihmig, A. G. H., F. Neurohr-Parakenings, S. K. Schäfer, J. Lass-Hennemann, and T. Michael, "On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals," *PLoS ONE*, vol. 15, no. 6, Jun. 2020, Art. no. e0231517.

[7] P. J. Bota, C. Wang, A. L. N. Fred, and H. P. D. Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019.

[8] A. Simons, T. Doyle, D. Musson, and J. Reilly, "Impact of physiological sensor variance on machine learning algorithms," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2020, pp. 241–247.

[9] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui, "Short-term anxiety recognition from blood volume pulse signal," in *Proc. IEEE 11th Int. Multi-Conf. Syst., Signals Devices (SSD14)*, Feb. 2014, pp. 1–6.

[10] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 400–408.

[11] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Sci. Data*, vol. 6, no. 1, pp. 1–13, Oct. 2019.

[12] J. M. Koolhaas et al., "Stress revisited: A critical evaluation of the stress concept," *Neurosci. Biobehavioral Rev.*, vol. 35, no. 5, pp. 1291–1301, Apr. 2011.

[13] B. F. Chorpita and D. H. Barlow, "The development of anxiety: The role of control in the early environment," *Psychol. Bull.*, vol. 124, no. 1, p. 3, 1998.

[14] T. Steimer, "The biology of fear- and anxiety-related behaviors," *Dialogues Clin. Neurosci.*, vol. 4, no. 3, pp. 231–249, Sep. 2002.

[15] R. S. Lazarus, *Emotion and Adaptation*. London, U.K.: Oxford Univ. Press, 1991.

[16] J. Zhai and A. Barreto, "Stress recognition using non-invasive technology," in *Proc. FLAIRS Conf.*, 2006, pp. 395–401.

[17] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Anal. Appl.*, vol. 9, no. 1, pp. 58–69, 2006.

[18] K. Ueafuea et al., "Potential applications of mobile and wearable devices for psychological support during the COVID-19 pandemic: A review," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7162–7178, Mar. 2021.

[19] J. Ormel, M. W. Koeter, W. Van den Brink, and G. Van de Willige, "Recognition, management, and course of anxiety and depression in general practice," *Arch. Gen. Psychiatry*, vol. 48, no. 8, pp. 700–706, 1991.

[20] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui, "Objective model assessment for short-term anxiety recognition from blood volume pulse signal," *Biomed. Signal Process. Control*, vol. 14, pp. 217–227, Nov. 2014.

[21] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *Proc. 2nd Int. Conf. Inventive Res. Comput. Appl. (ICIRCA)*, Jul. 2020, pp. 51–57.

[22] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J. Health Social Behav.*, vol. 24, no. 4, pp. 385–396, 1983.

[23] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *J. Behav. Therapy Experim. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[24] K. P. C. Kirschbaum and H. D., "The trier social stress test—A tool for investigating psychobiological 1stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, nos. 1–2, pp. 76–81, 1993.

[25] G. J. K. K. Chandan, H. K. Ahsan, and P. Marimuthu, "Complex correlation measure: A novel descriptor for poincaré plot," *Biomed. Eng. OnLine*, vol. 8, no. 1, pp. 1–17, 2009.

[26] M. Malik, "Heart rate Variability.: Standards of measurement, physiological interpretation, and clinical use: Task force of the European society of cardiology and the North American Society for pacing and electrophysiology," *Ann. Noninvasive Electrocardiol.*, vol. 1, no. 2, pp. 151–181, Apr. 1996.

[27] D. S. Fonseca, A. D. Netto, R. B. Ferreira, and A. M. F. L. M. de Sa, "Lomb-scargle periodogram applied to heart rate variability study," in *Proc. ISSNIP Biosignals Biorobotics Conf., Biosignals Robot. Better Safer Living (BRC)*, Feb. 2013, pp. 1–4.

[28] W. Saengmolee, D. Cheaha, N. Sa-Ih, and E. Kumarnsit, "Exploring of cardiac autonomic activity with heart rate variability in long-term kratom (*Mitragyna speciosa* Korth.) users: A preliminary study," *PeerJ*, vol. 10, Oct. 2022, Art. no. e14280.

[29] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, *A Survey on Deep Transfer Learning in International Conference on Artificial Neural Networks*. Cham, Switzerland: Springer, 2018, pp. 270–279.

[30] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18932–18943.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[33] R. Detrano. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

[34] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.