Accurate estimation of prediction models for operator-induced defects in assembly manufacturing processes

applied sciences

MDPI

Article

Enhancing Cross-Dataset Mental Workload Detection Using Electrodermal Activity and Domain Adaptation

Luis Sigcha [1], Eduarda Pereira [1], Luigi Borzì [2], Diego Gachet [3], Paulo Cardoso [1] and Nélson Costa [1,*]

[1] Centro ALGORITMI/LASI, School of Engineering, University of Minho, 4800-058 Guimarães, Portugal; d8610@algoritmi.uminho.pt (L.S.); id8976@alunos.uminho.pt (E.P.); paulo.cardoso@dei.uminho.pt (P.C.)
[2] Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy; luigi.borzi@polito.it
[3] Advanced Artificial Intelligence Group (A2IG), Escuela Politécnica Superior, Universidad Francisco de Vitoria, 28223 Madrid, Spain; diegogabriel.gachet@ufv.es
* Correspondence: ncosta@dps.uminho.pt

Abstract

Mental workload assessment using physiological signals has gained increasing attention for applications in human–computer interaction and occupational monitoring. Among these signals, electrodermal activity (EDA) is widely recognised as a reliable indicator of sympathetic activation associated with cognitive effort. However, most existing machine learning- based approaches are evaluated within a single dataset, limiting their generalisability across different populations and experimental conditions. This study investigates the cross-dataset performance of machine learning models for mental workload detection using EDA features. Two independent datasets were employed, and a cross-dataset evaluation framework was adopted to simulate realistic deployment scenarios under domain shift. Three classifiers (Random Forest, XGBoost, and Support Vector Classifier (SVC)) were evaluated, together with two domain adaptation techniques: Correlation Alignment (CORAL) and Subspace Alignment (SA). The results show that model performance is strongly dependent on the direction of transfer, with a notable performance drop when generalising across datasets. Domain adaptation improved performance in several configurations, particularly for SVC with CORAL, achieving the best overall F1-score (0.815). However, improvements were not consistent across all models and target domains. Overall, this study highlights the challenges of cross-dataset generalisation in EDA-based workload detection and demonstrates the potential, yet limited robustness, of domain adaptation techniques in mitigating distribution shifts.

Keywords: Human Factors; Mental Arithmetic; machine learning; wearable sensors; explainable artificial intelligence; nested cross-validation

Academic Editor: Jing Jin

Received: 9 April 2026
Revised: 3 May 2026
Accepted: 6 May 2026
Published: 8 May 2026

1. Introduction

Globalisation has significantly transformed the nature of work. The widespread adoption of computer-based systems and digital technologies has driven a shift from physically demanding occupations to cognitively intensive tasks [1–3]. As a result, mental workload has become a central factor in many professional environments [4].

Mental work primarily involves cognitive processes such as problem solving, decision making, and sustained attention. These activities are typically performed over extended periods, often in sedentary settings such as office environments [5]. Occupations in information technology, finance, research, and administrative domains exemplify this trend,

where high cognitive demands highlight the importance of effectively managing mental workload to preserve both performance and well-being [6,7].

Recent reports have emphasised the growing prevalence of mental health problems associated with excessive cognitive demands on the job [4,8]. Elevated mental workload has been linked to reduced concentration and energy levels [9], increased stress [10], and long-term mental fatigue [11]. Importantly, such effects are frequently observed even in non-pathological populations exposed to prolonged cognitive demands [12].

Mental workload is inherently complex, dynamic, and highly individual-dependent [6]. It reflects the amount of cognitive resources required to perform a task and is influenced by task characteristics, environmental conditions, and individual differences [13,14], as well as the nature of secondary tasks and interaction modalities, which can significantly affect cognitive demand and performance [15]. Consequently, accurate monitoring of mental workload remains a challenging but essential objective.

In this context, Human Factors research has increasingly focused on developing methods to objectively assess and manage workload [8,16]. Automated monitoring systems, often based on wearable and Internet of Things (IoT) technologies, have emerged as promising solutions for continuous and unobtrusive workload assessment [17]. These systems rely on physiological signals to provide objective indicators of cognitive states.

Among laboratory-based approaches, several standardised protocols have been widely used to induce mental workload under controlled conditions, including the N-Back task, the Stroop test, and the Mental Arithmetic Test (MAT) [14,18–20]. In particular, MAT is one of the most commonly used paradigms due to its effectiveness in eliciting cognitive load and sympathetic activation [21,22].

Despite significant progress, a key limitation in the field lies in the lack of generalisation across datasets. Most existing studies evaluate models within a single dataset, often resulting in performance degradation when applied to unseen data due to distribution shifts [23]. This limitation restricts the practical deployment of workload monitoring systems in real-world scenarios.

To address this challenge, this study investigates cross-dataset mental workload detection using electrodermal activity (EDA) signals. In this work, cross-dataset refers to the setting in which models are trained on data collected under one experimental protocol (source domain) and evaluated on data acquired under different conditions (target domain). This setting involves variability in experimental protocols, acquisition conditions, and measurement systems, which can affect both the magnitude and statistical properties of physiological signals, introducing domain shift, defined as differences in data distributions between source and target domains. In addition, it evaluates the effectiveness of domain adaptation techniques, specifically Correlation Alignment (CORAL) and Subspace Alignment (SA), to mitigate distribution mismatches between datasets. Furthermore, several machine learning algorithms, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Support Vector Classifier (SVC), are compared under cross-domain conditions.

The main contributions of this work are as follows:

- A cross-dataset evaluation framework for mental workload detection using electrodermal activity (EDA) enables the assessment of model generalisation under domain shift conditions.
- A systematic comparison of classical machine learning models (RF, XGBoost, and SVC) in cross-dataset scenarios using subject-independent validation.
- The application and evaluation of two feature-based domain adaptation techniques (CORAL and SA) to mitigate distribution differences between datasets.

- An analysis of the impact of domain adaptation on classification performance, highlighting that improvements depend on the combination of model and adaptation method.

The remainder of this paper is structured as follows. Section 2 reviews the related literature on workload assessment using electrodermal activity and cross-domain learning approaches. Section 3 outlines the materials and methods. The results are presented in Section 4, followed by the discussion in Section 5, and the conclusions in Section 6.

2. Related Work

2.1. Workload Assessment Using Electrodermal Activity

EDA, also referred to as galvanic skin response, has been widely adopted as a physiological indicator for assessing mental workload [24,25]. Due to its non-invasive nature, low cost, and ease of acquisition, EDA is particularly suitable for real-world and wearable-based applications [22,25]. Its physiological origin lies in the activity of eccrine sweat glands, which are exclusively innervated by the sympathetic nervous system. As a result, EDA provides a direct and relatively uncontaminated measure of sympathetic activation associated with mental effort and stress [21,25,26].

EDA signals are typically analysed through skin conductance measurements, which can be decomposed into tonic (slow-varying) and phasic (event-related) components. In particular, skin conductance responses (SCR) have been shown to reflect transient changes in mental demand and are commonly used in workload analysis [27,28]. These properties have motivated a wide range of studies investigating the use of EDA for mental workload detection [29].

Early work by Setz et al. [30] demonstrated that EDA features such as peak amplitude and peak rate can effectively distinguish between stress and cognitive load, achieving classification accuracies of up to 82.8%. Subsequent studies have explored the integration of EDA into adaptive systems. For example, Zhou et al. [31] proposed a dynamic workload adaptation framework that adjusts task difficulty in real time based on EDA measurements. Similarly, Androutsou et al. [5] developed an unobtrusive monitoring system using EDA sensors embedded in a computer mouse, highlighting the feasibility of continuous workload assessment in office environments.

More recently, machine learning approaches have been increasingly applied to EDA-based workload classification. Saha et al. [25] reported that random forest classifiers can achieve high accuracy when using EDA features, emphasising their discriminative capability. In parallel, Ding et al. [14] highlighted the importance of physiological signals for improving classification performance in cognitively demanding tasks. These studies collectively demonstrate the potential of EDA as a reliable signal for workload assessment.

However, most existing work evaluates models under within-dataset conditions, where training and testing data are drawn from the same distribution. While this setup facilitates controlled experimentation, it does not reflect real-world deployment scenarios, where models are expected to generalises across different populations, devices, and experimental protocols. As a result, performance often degrades when models are applied to unseen datasets, highlighting the need for cross-dataset evaluation strategies.

2.2. Domain Adaptation for Physiological Signal Analysis

To address the challenges associated with distribution shifts across datasets, domain adaptation techniques have been proposed as a means to improve model generalisation. Domain adaptation aims to reduce discrepancies between source and target data distributions, allowing models trained on one domain to perform effectively on another [32,33].

These approaches have been successfully applied in various fields, including computer vision, speech recognition, and biomedical signal processing.

Among feature-based domain adaptation methods, Correlation Alignment (CORAL) and Subspace Alignment (SA) have gained attention due to their simplicity and effectiveness. CORAL aligns the second-order statistics of the source and target feature distributions by transforming the covariance of the source data to match that of the target domain [34]. This method does not require labelled target data and can be easily integrated into existing machine learning pipelines. Subspace Alignment (SA), on the other hand, projects source and target data into lower-dimensional subspaces and learns a transformation matrix that aligns these representations [35]. By operating in a shared latent space, SA aims to reduce domain discrepancy while preserving discriminative information.

Although these methods have shown promising results in general pattern recognition tasks, their application to physiological signals remains relatively limited. In recent years, more advanced domain adaptation approaches based on deep learning, such as domain-adversarial learning and deep transfer learning, have been proposed to learn domain-invariant representations [33]. These methods rely on end-to-end optimisation and have demonstrated strong performance across several domains. However, their performance is often influenced by data availability and computational resources. In the context of physiological signal analysis, where datasets are typically limited and heterogeneous, these factors may affect their applicability and performance. In particular, several studies have investigated transfer learning methodologies within the context of biosignal analysis, including applications in emotion recognition and brain–computer interface (BCI) systems [36,37]. However, systematic evaluations of classical domain adaptation methods in EDA-based mental workload detection are still scarce.

In particular, there is a lack of studies that combine (i) cross-dataset evaluation, (ii) EDA-based feature representations, and (iii) classical machine learning models within a unified framework. Addressing this gap is essential for understanding the practical applicability of workload detection systems in real-world scenarios.

In this context, the present work investigates the use of CORAL and SA in combination with three widely used machine learning algorithms: RF [38], XGBoost [39], and SVC [40]. This provides a systematic evaluation of their effectiveness in improving cross-dataset generalisation for EDA-based mental workload detection.

3. Materials and Methods

3.1. Datasets

This study employed two independent datasets to evaluate cross-dataset generalisation in mental workload detection using EDA: the Worker H&P dataset [41] and the UNobtrusIVE measuRement of mental workload and stress in uncontrolled environments (UNIVERSE) dataset [42].

The Worker H&P dataset comprises physiological recordings collected in a controlled office-like environment designed to simulate real-world working conditions. Data were acquired using a BIOPAC MP160 system (Biopac Systems Inc., Goleta, CA, USA), with EDA recorded via EL507 electrodes placed on the palmar surface of the hand. Signals were sampled at 250 Hz. In this study, only the second cohort was considered, consisting of 41 healthy participants (20 males, 21 females; mean age: 32.5 $\pm$ 10.8 years). Participants were seated during the experiment and performed the Mental Arithmetic Test (MAT), with different difficulty levels corresponding to distinct cognitive load conditions.

The UNIVERSE dataset consists of multimodal physiological recordings collected from 24 participants in both controlled laboratory and real-world environments. EDA signals were acquired using a wrist-worn Empatica E4 device (Empatica Inc., Cambridge, MA,

USA), with sensors positioned on the wrist and sampled at 4 Hz. In the controlled setting, participants were seated and performed cognitive tasks including Mental Arithmetic, Stroop, N-back, and Sudoku at varying difficulty levels.

To ensure comparability between datasets and minimise bias due to differences in data amount and task variability, the Mental Arithmetic task (MAT) was selected, as it is common to both datasets and effective in eliciting cognitive load and sympathetic activation [21,22]. Accordingly, only data corresponding to the MAT from the first laboratory session was retained from the UNIVERSE dataset. Furthermore, data from three participants (IDs: 102, 103, and 114) were excluded due to a significant amount of missing or null EDA data. After preprocessing and filtering, the final UNIVERSE subset included 21 participants.

Both datasets were harmonised in terms of preprocessing and feature extraction to enable a fair cross-dataset evaluation.

3.2. Classification Strategy

In this study, a binary classification framework was adopted to distinguish between the absence and presence of mental workload. Specifically, the classification task aimed to differentiate between a no mental workload condition (baseline) and a mental workload condition induced during task execution.

For both datasets, the original labels describing different levels of cognitive load (e.g., low and high workload) were merged into a single mental workload class. This unified class was then contrasted against the corresponding baseline condition. In the Worker H&P dataset, the baseline condition corresponds to periods in which participants were instructed to remain relaxed while observing a neutral grey screen. In the UNIVERSE dataset, the baseline condition corresponds to relaxation periods recorded prior to task execution. Although both datasets include a baseline condition, these are not strictly equivalent. Differences in experimental context may lead to variations in physiological signals, as baseline EDA levels are known to depend on prior activity and recording conditions. Consequently, these discrepancies may affect data comparability and contribute to the observed domain shift between datasets.

This label harmonisation was performed to ensure consistency across datasets and to facilitate a robust cross dataset evaluation. By reducing the problem to a binary classification task, the study focuses on detecting the presence of cognitive load independently of its intensity. This formulation simplifies the cross-dataset comparison by reducing label variability and enables a consistent evaluation framework across datasets. However, this simplification reduces the granularity of the problem by merging different workload levels into a single class. As a result, the model is expected to capture dominant differences between baseline and task conditions, while being less sensitive to variations associated with different levels of cognitive demand. This trade-off reflects the need to balance label consistency with representational detail and defines the scope of the analysis, which is focused on workload detection rather than fine-grained workload differentiation. Table 1 summarises the distribution of the data for both datasets after merging the original labels into a binary classification problem.

Table 1. Duration of data (in minutes) for the UNIVERSE and Worker H&P datasets after relabelling into a binary classification task.

Dataset	Baseline	Mental Workload (MAT)	Total
UNIVERSE	190.8	370.8	560.8
Worker H&P	246.8	262.8	508.8

As shown in Table 1, the UNIVERSE dataset exhibits a higher proportion of cognitive load data relative to baseline, whereas the Worker H&P dataset presents a more balanced distribution between the two classes. In terms of total duration, both datasets are comparable, with the UNIVERSE dataset containing a slightly larger amount of data than the Worker H&P dataset.

3.3. Signal Preprocessing

EDA signals from both datasets were resampled to a common sampling frequency of 4 Hz, which is sufficient to capture the slow dynamics of skin conductance signals [42,43]. This harmonisation was performed to ensure consistency across datasets, as EDA signals are typically band-limited to low frequencies (generally below 2 Hz) and are therefore commonly analysed at reduced sampling rates [44,45]. Such sampling frequencies are also consistent with those adopted in many consumer grade wearable devices for continuous skin conductance monitoring [46], and support consistent feature extraction across datasets. However, reducing the sampling frequency may limit the temporal resolution required to accurately capture fast phasic components, particularly for SCR peak detection and related features, for which higher sampling rates (4–10 Hz) are generally recommended [46]. As a result, peak-based and spectral features may be less precisely characterised under this sampling configuration, reflecting a trade-off between cross-dataset harmonisation and temporal resolution

The preprocessing pipeline was implemented and applied uniformly across datasets using an environment with Python (version 3.9.21) and the libraries Scipy (version 1.9.3) and Biobss (version 0.1.1). First, a fourth-order zero-phase Butterworth low-pass filter with a cutoff frequency of 1 Hz was applied to remove high-frequency noise. Then, a Gaussian smoothing filter of 40-point window and sigma of 400 ms was applied to reduce motion artefacts, as proposed in Campanella et al. [47].

Subsequently, signals were normalised using z-score normalisation computed on a per-subject basis to reduce inter-subject variability. Finally, the processed signals were segmented into fixed-length windows using a sliding window approach.

3.4. Segmentation and Feature Extraction

EDA signals were segmented into 60-s windows with 80% overlap. High-overlap sliding windows are commonly used in physiological signal analysis to preserve temporal continuity and increase the number of training samples [42]. Recent EDA-based studies have adopted near-maximum overlap strategies, further supporting the use of high overlap in this context [43]. However, the use of high overlap may introduce correlations between adjacent segments from the same recording. To avoid this affecting model evaluation, a subject-independent splitting strategy was used, assigning all windows from each participant to a single data subset (e.g., train, test, or align). This ensures that correlated segments remain within the same subject and do not appear across training and evaluation sets. As a result, dependencies are limited to within-subject data, while independence across subjects is preserved due to the inherent variability in EDA signals [26,48].

Given the differences in data acquisition between datasets, distinct labelling strategies were applied. For the Worker H&P dataset, where data were collected continuously, a majority voting approach was used to assign a label to each window based on the predominant class within the window. In contrast, the UNIVERSE dataset provides labelled data in discrete chunks corresponding to specific experimental conditions. Therefore, a direct labelling strategy was applied, assigning each window the label associated with its corresponding annotated segment.

From each window, a set of 30 features was extracted using the Biobss library (version 0.1.1). These features capture both tonic (SCL) and phasic (SCR) characteristics of the EDA signal, including statistical descriptors, peak-related features, and temporal dynamics. Although the majority correspond to statistical descriptors, they are grounded in standard EDA analysis through the decomposition into tonic and phasic components, which reflect underlying sympathetic activity [26,49]. A summary of the extracted features is presented in Table 2.

Table 2. Summary of extracted EDA features (30 features).

Component	Feature Group	Description	# Features
SCR	Peak-based features	Number of detected skin conductance responses (SCR peaks)	1
	Statistical features	Statistical descriptors (mean, standard deviation, max, min, range)	5
	Distribution features	Higher-order statistics (skewness, kurtosis)	2
	Hjorth parameters	Activity, mobility, and complexity of the signal	3
	Signal properties	Momentum, RMS, arc length, and integral of the signal	4
	Spectral features	Frequency-domain descriptors (F1, F2, F3, max frequency)	4
	Energy features	Energy and average power measures	2
	Entropy features	Entropy-based measure of signal complexity	1
SCL	Statistical features	Statistical descriptors (mean, standard deviation, max, min, range)	5
	Distribution features	Higher-order statistics (skewness, kurtosis)	2
	Signal properties	Momentum of the tonic component	1

To assess potential redundancy among features, a correlation analysis using the Pearson correlation coefficient (r) was conducted across all extracted descriptors, including those derived from SCR and SCL components. Correlation analysis was performed on the combined dataset.

Furthermore, to characterise the domain shift between datasets, a statistical comparison of feature distributions was performed. Feature-wise comparisons between the UNIVERSE and Worker H&P datasets were conducted using the Mann–Whitney U test, and effect sizes were quantified using Cliff's delta.

3.5. Machine Learning Models

To evaluate the capability of EDA features to discriminate between mental workload and the baseline, the following machine learning algorithms were considered: RF, XGBoost, and SVC. These algorithms were selected due to their robustness, complementary learning strategies, and proven effectiveness in physiological signal classification tasks [10,29,50].

RF is an ensemble learning method based on decision trees, which improves generalisation performance by aggregating multiple weak learners [38]. XGBoost is a gradient boosting framework that iteratively minimises prediction errors through additive model optimisation, often achieving high predictive performance in structured data [39]. The SVC, on the other hand, is a margin-based method that seeks optimal decision boundaries in high-dimensional feature spaces [40].

For each model, hyperparameters were optimised using a grid search strategy embedded within a nested cross-validation framework. The optimisation aimed to maximise the (macro) F1-score, ensuring balanced performance across classes. The hyperparameter search space for the three machine learning algorithms was defined based on previous studies evaluating multiple algorithms in cross-dataset settings [51,52] and is summarised in Appendix A (Table A1). The optimal hyperparameters for each model and experimental configuration are reported in Appendix A, Table A2 for the UNIVERSE dataset as target, and Table A3 for the Worker H&P dataset as target.

Deep learning approaches were not considered in this study, as the study focuses on domain adaptation and the interpretability of the resulting models under heterogeneous data conditions. Classical machine learning algorithms provide greater transparency and facilitate explainability analyses. In addition, deep learning methods typically rely on larger amounts of data to achieve optimal performance, which is not aligned with the cross-dataset setting considered in this work.

All the machine learning experiments were conducted using an environment with Python (version 3.11.14) and the Scikit learn library (version 1.8).

3.6. Domain Adaptation

Given the distribution differences between datasets collected under different experimental conditions, domain adaptation techniques were employed to improve cross-dataset generalisation. In this study, two feature-based unsupervised methods were considered: CORAL and SA.

CORAL performs alignment in the original feature space by matching the covariance structure of the source data to that of the target domain. In contrast, SA projects source and target data into a lower-dimensional latent space and aligns these representations through a linear transformation.

These methods were selected as well-established approaches that enable a controlled evaluation of cross-dataset generalisation within a classical machine learning framework. However, both CORAL and SA rely on linear transformations, which may limit their ability to capture more complex, non-linear discrepancies between domains. Therefore, this study can be interpreted as a first step towards cross-dataset mental workload detection, providing a baseline for future work exploring more advanced adaptation strategies.

Both methods were integrated into the training pipeline as feature transformation steps, using only the training data from the source domain and the alignment subset of the target domain. This strategy enables the model to learn domain-invariant representations without requiring labelled data from the target domain. The methods were implemented within the nested cross-validation framework using the Awesome Domain Adaptation Python Toolbox (ADAPT, version 0.4.5) [53].

3.7. Experimental Design and Validation

A cross-dataset evaluation framework was adopted to assess the generalisation capability of the proposed approach under domain shift conditions. In this setup, models were trained on one dataset (source domain) and evaluated on a different dataset (target domain), simulating realistic deployment scenarios. The overall evaluation pipeline is illustrated in Figure 1.

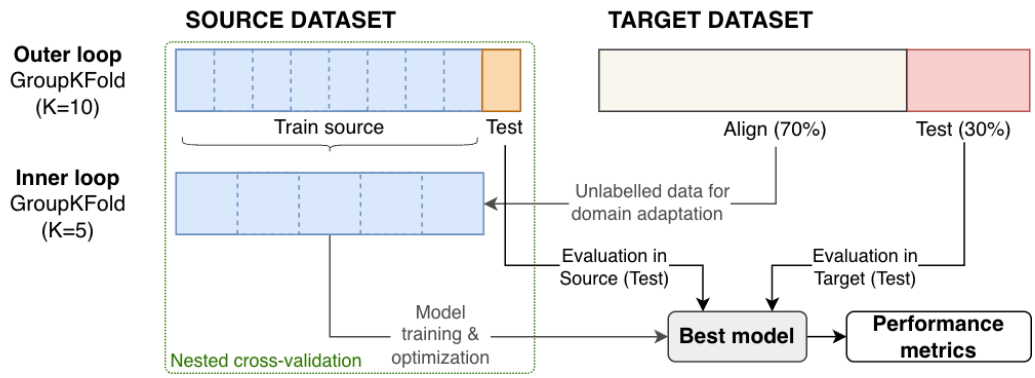

Figure 1. Cross-dataset evaluation framework with nested cross-validation and domain adaptation.

To ensure subject independence and prevent data leakage, both source and target datasets were handled using subject-wise splits. The source dataset was split within a nested cross-validation framework using subject-wise grouping, while the target dataset was partitioned into an alignment subset (70%) and a test subset (30%), ensuring that no participant appeared in both subsets. This target split was performed at the subject level and kept fixed throughout the evaluation. The alignment subset was used exclusively for domain adaptation using unlabelled data, whereas the test subset remained fully independent and was not involved in any stage of model training, hyperparameter optimisation, or model selection. This design ensures that the evaluation on the target domain reflects a realistic deployment scenario, where no target labels are available and no information from the test subset is used during model development.

Following the subject-wise splitting strategy, model development on the source dataset was conducted using a nested cross-validation framework. Specifically, a 10-fold outer GroupKFold was employed to estimate generalisation performance across subjects. Within each outer fold, a 5-fold inner GroupKFold was used for hyperparameter optimisation via grid search. In both cases, subject identifiers were used as grouping variables to ensure that samples from the same participant were not shared across folds. The optimisation criterion for model selection was the macro F1-score.

Within each outer fold, the model was trained on the source training data obtained from the outer split. When domain adaptation was applied, the model was subsequently aligned using unlabelled data from the target alignment subset. The resulting model was then evaluated on two independent sets: (i) the held-out source fold (source test) to assess in-domain performance, and (ii) the target test subset to assess cross-domain generalisation.

The final performance metrics are reported as mean (standard deviation) across the 10 outer folds. Baseline results include both source and target test evaluations, while domain adaptation results are reported only on the target test set.

To evaluate whether the observed differences in performance between baseline and domain adaptation methods were statistically significant, a Wilcoxon signed-rank test was applied to the F1-scores obtained across the outer cross-validation folds, enabling paired comparisons between methods. Additionally, 95% confidence intervals of the performance differences were computed to quantify the variability of the observed effects. Statistical significance was assessed at a level of $p < 0.05$.

3.8. Performance Evaluation

The performance of the proposed approaches was evaluated using three standard classification metrics: accuracy, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC) [54]. These metrics were selected to provide a balanced assessment of classification performance in the binary workload detection task.

In this context, true positives (TP) correspond to correctly identified cognitive load samples, true negatives (TN) to correctly identified baseline samples, false positives (FP) to baseline samples incorrectly classified as cognitive load, and false negatives (FN) to cognitive load samples incorrectly classified as baseline.

Accuracy measures the proportion of correctly classified samples over the total number of instances and is defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

The F1-score, defined as the harmonic mean of precision and recall, is particularly suitable in scenarios where class distributions may be imbalanced. It is computed as:

$$F1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (2)$$

where precision and recall are defined as:

$$Precision = \frac{TP}{TP + FP} \quad Recall = \frac{TP}{TP + FN} \quad (3)$$

The AUC metric evaluates the ability of the classifier to distinguish between classes across all possible decision thresholds and is commonly used as a threshold-independent performance measure. It summarises the Receiver Operating Characteristic (ROC) curve and provides insight into the discriminative capability of the model [55].

For the computation of classification metrics, a decision threshold was applied to the probability output of the binary classifiers. This threshold was determined using the Equal Error Rate (EER) computed on the training data and subsequently applied to the test data, ensuring a consistent evaluation across all experiments.

3.9. Model Explainability Analysis

To better understand the contribution of individual features to the model predictions, a post-hoc explainability analysis was conducted using SHapley Additive exPlanations (SHAP) [56]. This method provides a unified framework to estimate the contribution of each feature to the model output.

The explainability analysis was performed exclusively on the baseline models (i.e., without domain adaptation). The application of domain adaptation techniques modifies the original feature space through transformations such as CORAL or SA, altering the direct correspondence between input features and their physiological meaning. As a result, interpreting feature contributions after adaptation becomes less straightforward.

To provide a concise interpretation, SHAP values were aggregated across classifiers (RF, XGBoost, and SVC) by averaging normalised mean absolute SHAP values. This aggregation highlights feature importance patterns that are consistent across models while reducing model-specific variability. In addition to the aggregated analysis, a subject-level SHAP evaluation was conducted using the baseline model. For each participant, mean absolute SHAP values were computed to assess variability in feature importance across individuals.

4. Results

4.1. Baseline Performance

The baseline performance of the evaluated classifiers, obtained without applying any domain adaptation technique, is summarised in Tables 3 and 4 for both cross-dataset configurations.

When the UNIVERSE dataset was used as the target domain (Table 3), all models achieved comparable performance across evaluation domains. RF and XGBoost showed similar results, with target accuracies of 0.647 and 0.649, respectively. The SVC achieved the highest performance among all models, reaching an accuracy of 0.720, an F1-score of 0.773, and an AUC of 0.779 on the target dataset.

In contrast, when the Worker H&P dataset was used as the target domain (Table 4), a more pronounced decrease in performance was observed. Although all models achieved high performance on the source domain (accuracy above 0.81), their performance decreased when evaluated on the target dataset. RF achieved an accuracy of 0.601, XGBoost reached

0.618, and SVC obtained 0.596. Similar trends were observed for F1-score and AUC, indicating a stronger domain shift in this configuration.

Table 3. Baseline performance of the classifiers without domain adaptation, trained on the Worker H&P dataset (source domain) and evaluated on the UNIVERSE dataset (target domain). Results are reported as mean (standard deviation) across folds.

Model	Evaluation Domain	Accuracy	F1-Score	AUC
RF	Source	0.655 (0.075)	0.667 (0.072)	0.721 (0.079)
RF	Target	0.647 (0.020)	0.710 (0.018)	0.725 (0.023)
XGBoost	Source	0.657 (0.065)	0.669 (0.062)	0.730 (0.082)
XGBoost	Target	0.649 (0.020)	0.711 (0.019)	0.718 (0.022)
SVC	Source	0.650 (0.064)	0.662 (0.061)	0.707 (0.072)
SVC	Target	0.720 (0.072)	0.773 (0.063)	0.779 (0.072)

Table 4. Baseline performance of the classifiers without domain adaptation, trained on the UNIVERSE dataset (source domain) and evaluated on the Worker H&P dataset (target domain). Results are reported as mean (standard deviation) across folds.

Model	Evaluation Domain	Accuracy	F1-Score	AUC
RF	Source	0.820 (0.127)	0.854 (0.105)	0.843 (0.146)
RF	Target	0.601 (0.017)	0.614 (0.017)	0.634 (0.010)
XGBoost	Source	0.813 (0.137)	0.848 (0.114)	0.845 (0.155)
XGBoost	Target	0.618 (0.014)	0.632 (0.014)	0.648 (0.014)
SVC	Source	0.816 (0.137)	0.851 (0.115)	0.871 (0.114)
SVC	Target	0.596 (0.008)	0.610 (0.008)	0.636 (0.013)

Overall, these baseline results, obtained without any domain adaptation, indicate that cross-dataset generalisation is dependent on the direction of transfer, with better performance observed when using the UNIVERSE dataset as the target domain.

4.2. Domain Adaptation Results

The impact of domain adaptation techniques on cross-dataset performance is presented in Table 5, considering the UNIVERSE dataset as the target domain, and in Table 6, considering the Worker H&P dataset as the target domain. In both tables, the baseline corresponds to models trained and evaluated without applying any domain adaptation.

Table 5. Cross-dataset performance on the Universe data set as target. Results are reported as mean (standard deviation) across folds. The baseline corresponds to models without domain adaptation. Δ indicates the variation with respect to the baseline. The en dash (–) indicates that no comparison is applicable for the baseline model.

Classifier	Method	Accuracy	Δ Accuracy	F1-Score	Δ F1-Score	AUC	Δ AUC
RF	Baseline	0.647 (0.020)	–	0.710 (0.018)	–	0.725 (0.023)	–
	CORAL	0.658 (0.049)	+0.011	0.719 (0.045)	+0.009	0.713 (0.048)	−0.012
	SA	0.718 (0.019)	+0.071	0.772 (0.016)	+0.062	0.749 (0.020)	+0.024
XGBoost	Baseline	0.649 (0.020)	–	0.711 (0.019)	–	0.718 (0.022)	–
	CORAL	0.664 (0.039)	+0.015	0.724 (0.036)	+0.013	0.725 (0.043)	+0.007
	SA	0.677 (0.062)	+0.028	0.736 (0.056)	+0.025	0.711 (0.050)	−0.007
SVC	Baseline	0.720 (0.072)	–	0.773 (0.063)	–	0.779 (0.072)	–
	CORAL	0.767 (0.011)	+0.047	0.815 (0.009)	+0.042	0.832 (0.011)	+0.053
	SA	0.712 (0.014)	−0.008	0.767 (0.013)	−0.006	0.748 (0.024)	−0.031

Table 6. Cross-dataset performance on the Worker H&P dataset as target. Results are reported as mean (standard deviation) across folds. The baseline corresponds to models without domain adaptation. Δ indicates the variation with respect to the baseline. The en dash (–) indicates that no comparison is applicable for the baseline model.

Classifier	Method	Accuracy	Δ Accuracy	F1-Score	Δ F1-Score	AUC	Δ AUC
RF	Baseline	0.601 (0.017)	–	0.614 (0.017)	–	0.634 (0.010)	–
	CORAL	0.621 (0.014)	+0.020	0.634 (0.014)	+0.020	0.658 (0.013)	+0.024
	SA	0.503 (0.033)	−0.098	0.517 (0.033)	−0.097	0.502 (0.043)	−0.132
XGBoost	Baseline	0.618 (0.014)	–	0.632 (0.014)	–	0.648 (0.014)	–
	CORAL	0.587 (0.033)	−0.031	0.600 (0.033)	−0.032	0.624 (0.033)	−0.024
	SA	0.509 (0.032)	−0.109	0.523 (0.032)	−0.109	0.507 (0.037)	−0.141
SVC	Baseline	0.596 (0.008)	–	0.610 (0.008)	–	0.636 (0.013)	–
	CORAL	0.587 (0.017)	−0.009	0.601 (0.017)	−0.009	0.623 (0.016)	−0.013
	SA	0.515 (0.030)	−0.081	0.529 (0.030)	−0.081	0.524 (0.033)	−0.112

When considering the Worker H&P dataset as the target domain, the effect of domain adaptation differs markedly from that observed for the UNIVERSE dataset. For the RF classifier, the application of CORAL resulted in consistent improvements across all metrics, increasing the F1-score from 0.614 to 0.634 (+0.020). This improvement was statistically significant (p = 0.002, CI [0.0099, 0.0307]), indicating a consistent gain across folds. In contrast, SA led to a substantial decrease in performance (F1-score = 0.517, Δ = −0.097), also statistically significant (p = 0.005, CI [−0.150, −0.044]) indicating that this method was not suitable for this transfer scenario.

For XGBoost, both domain adaptation methods resulted in a reduction in performance compared to the baseline. CORAL decreased the F1-score from 0.632 to 0.600 (−0.032), with a statistically significant degradation (p = 0.014, CI [−0.0497, −0.0125]), while SA produced a larger degradation (F1-score = 0.523, Δ = −0.109), also significant (p = 0.002, CI [−0.1328, −0.0848]). A similar pattern was observed for the SVC model, where CORAL led to a slight decrease in F1-score (−0.009), whereas SA again resulted in a more pronounced reduction (−0.081), which was statistically significant (p = 0.020, CI [−0.1002, −0.0608]).

When comparing across datasets, the highest performance for the UNIVERSE target was achieved using SVC with CORAL (F1-score = 0.815), while for the Worker H&P target, the best-performing configuration corresponded to RF with CORAL (F1-score = 0.634). This represents a reduction of approximately 0.18 in F1-score between the two target domains, highlighting the increased difficulty associated with transferring to the Worker H&P dataset.

A statistical comparison of feature distributions confirmed the presence of a substantial domain shift between datasets. As detailed in Appendix A (Table A6), the largest discrepancies were observed in amplitude- and energy-related SCR features, including *scr_integral* ($\delta = -0.784$), *scr_RMS* ($\delta = -0.767$), and *scr_average_power* ($\delta = -0.767$), all indicating large effect sizes ($|\delta| > 0.474$) (effect size interpretation followed standard thresholds for Cliff's delta). These features consistently exhibited higher values in the Worker H&P dataset compared to the UNIVERSE dataset. In contrast, other descriptors, such as *scl_mean*, *scr_skew*, and *scr_max_frequency*, showed negligible or non-significant differences, indicating the presence of relatively stable features across datasets.

Furthermore, correlation analysis revealed strong redundancy within feature groups (Appendix A, Table A4), particularly among SCR amplitude- and energy-related descriptors (e.g., correlations up to $r \approx 1.00$ for *scr_std* and *scr_rms*). In contrast, cross-correlations between SCR and SCL features were moderate (Appendix A, Table A5), with a maximum of $|r| = 0.77$, suggesting that these components capture complementary aspects of the EDA

signal. In line with the data-driven design of this study, all extracted features were retained to evaluate domain adaptation methods under a comprehensive and potentially redundant feature space.

Although CORAL yielded improvements for the RF classifier, its effect was not consistent across models, and SA consistently resulted in performance degradation across all classifiers. Overall, these results indicate that the effectiveness of domain adaptation is highly dependent on both the classifier and the direction of transfer, with more limited benefits observed when adapting to the Worker H&P dataset. These observations are further supported by a statistical significance analysis based on the Wilcoxon signed-rank test across folds, which confirms that only selected configurations (e.g., RF with CORAL) yield statistically significant improvements, while others exhibit non-significant differences or significant performance degradation. For completeness, the full set of pairwise comparisons, including confidence intervals, is provided in Appendix A (Table A7).

4.3. Explainability Results

The SHAP-based analysis revealed that features associated with the tonic component (SCL) consistently showed the highest contribution to the classification task across both datasets (see Appendix A, Figures A1 and A2). In particular, *scl_mean* emerged as the most relevant feature across classifiers in both configurations, indicating a stable role in distinguishing between baseline and cognitive load conditions.

Differences in feature importance were observed between the two target domains. When the UNIVERSE dataset was used as the target (Figure A1), the most relevant features included *scl_mean*, *scl_max*, and *scr_integral*, with relatively balanced contributions across classifiers. In contrast, when the Worker H&P dataset was used as the target (Figure A2), *scl_min* and *scl_range* became more prominent, alongside *scl_mean*, indicating a shift in the relative importance of tonic-related features.

Phasic-related features, such as *scr_integral*, also contributed to the classification task in both scenarios, although their relative importance varied between datasets. This suggests that phasic dynamics provide complementary information but may be more sensitive to dataset-specific characteristics.

Despite these differences, a high degree of consistency was observed in the contribution patterns across classifiers within each dataset, as reflected by the aggregated SHAP representations. This indicates that the models rely on similar physiological markers, supporting the robustness of the feature representation. Overall, the SHAP analysis confirms that the most informative features are meaningful and highlights both shared and dataset-specific patterns in feature importance.

At the subject level, distinct variability patterns were observed across target domains when analysing the top-ranked features (Appendix A, Tables A8 and A9). For the UNIVERSE target, tonic-related features such as *scl_mean* and *scl_min* showed relatively low variability across subjects, whereas features including *scl_max* and *scr_integral* exhibited higher dispersion, with variability comparable to their average contribution. For the Worker H&P target, several features, including *scr_mobility*, *scl_std*, and *scr_std*, presented consistently low variability, indicating stable contributions across individuals. In contrast, features such as *scl_min* and *scl_max* displayed substantially higher variability, suggesting increased sensitivity to inter-subject differences under this transfer scenario. Overall, the results indicate that feature relevance is not uniform across subjects and varies depending on the target domain, particularly for features associated with signal amplitude. This variability is consistent with the known inter-subject differences in physiological responses to cognitive workload [26].

5. Discussion

This study evaluated the cross-dataset generalisation of mental workload detection models using EDA features, as well as the potential benefits of domain adaptation techniques. The results provide insight into both the baseline transferability of the models and the extent to which domain adaptation can mitigate distribution shifts between datasets.

When analysing the baseline results (i.e., without domain adaptation), performance was strongly dependent on the direction of transfer. When the UNIVERSE dataset was used as the target domain, the models achieved moderate performance, with SVC obtaining the highest F1-score (0.773). In contrast, when the Worker H&P dataset was used as the target, performance decreased across all models, with the best F1-score reaching 0.632 (XGBoost). This difference of approximately 0.14 in F1-score reflects a stronger domain shift when transferring to the Worker H&P dataset.

The application of domain adaptation techniques revealed different behaviours depending on the target dataset. When adapting to the UNIVERSE dataset, consistent improvements over the baseline were observed. For example, SVC combined with CORAL increased the F1-score from 0.773 to 0.815 (+0.042), representing the highest performance across all configurations. Similarly, RF with SA improved from 0.710 to 0.772 (+0.062), indicating that both CORAL and SA can enhance performance in this transfer direction.

In contrast, when adapting to the Worker H&P dataset, the impact of domain adaptation was limited and less consistent. The only improvement over the baseline was observed for RF with CORAL, where the F1-score increased from 0.614 to 0.634 (+0.020). For the remaining configurations, domain adaptation had a negligible effect or resulted in performance degradation. In particular, SA consistently reduced performance across all models, with decreases of up to -0.109 in F1-score for XGBoost.

Overall, these results indicate that domain adaptation can improve cross-dataset performance in some scenarios, but its effectiveness is not consistent across transfer directions. Importantly, the asymmetry observed in cross-dataset transfer can be interpreted in light of the observed differences between datasets. The large distribution shifts identified in amplitude- and energy-related features (Appendix A, Table A6), together with acquisition-related differences, suggest that the Worker H&P dataset represents a more challenging target domain. In contrast, the UNIVERSE dataset appears to provide feature representations that are more compatible with models trained on external data, resulting in improved transferability in that direction. This behaviour is consistent with the directional nature of domain adaptation, where transfer performance depends on the relative alignment between source and target domains [57].

The results also show that the effect of domain adaptation is model-dependent. SVC achieved the largest improvement when combined with CORAL in the UNIVERSE scenario, whereas RF was the only model that improved in the Worker H&P scenario. This suggests that the interaction between the classifier and the adaptation method influences the resulting performance.

Several dataset-related factors may contribute to the observed differences. First, although the UNIVERSE dataset was restricted to the first laboratory session to match the total duration of the Worker H&P dataset, differences remain in class proportions. The UNIVERSE dataset contains a higher proportion of cognitive load data, whereas the Worker H&P dataset is more balanced. These differences may contribute to variations in classification behaviour.

Differences in data acquisition protocols may also play a role. The Worker H&P dataset was collected using a laboratory-grade BIOPAC system (250 Hz, hand electrodes), whereas the UNIVERSE dataset was acquired using a wrist-worn wearable device (4 Hz). These differences may affect signal characteristics and the resulting feature distributions. Although

preprocessing steps such as low-pass filtering (1 Hz cutoff) and Gaussian smoothing were applied, residual differences between datasets may still influence the results.

Furthermore, resampling the signals to 4 Hz reduces the temporal resolution available to capture fast phasic components. However, this limitation appears to have a limited impact on the main findings. The statistical comparison of feature distributions showed that the largest differences were associated with amplitude- and energy-related features, whereas features related to temporal dynamics exhibited smaller or negligible differences. Consistently, the SHAP analysis indicated that tonic-related features contributed most to the classification. Together, these findings suggest that the dominant discriminative information is largely associated with slower components of the signal and is therefore less sensitive to the reduced temporal resolution.

In this context, the observed cross-dataset performance reflects the combined influence of task-related factors (cognitive workload) and instrumental differences, including sensor type, electrode placement, and acquisition conditions. These sources of variability are not fully separable within the current study design, and their relative contribution cannot be quantified.

In addition, the binary formulation of the problem (baseline vs. workload) simplifies the underlying phenomenon and emphasises general rest–task differences rather than finer-grained workload levels. This choice was driven by the need to ensure label consistency across datasets with heterogeneous annotations. However, it introduces the possibility that the models capture contextual or experimental factors in addition to cognitive workload. Future work should explore multi-level workload representations under more controlled conditions, enabling a clearer separation between different levels of cognitive demand and reducing the influence of contextual or acquisition-related factors.

Finally, domain adaptation methods such as CORAL and SA assume a certain level of similarity between source and target feature distributions. When this assumption is not fully met, the effectiveness of the alignment may be limited, which is consistent with the variability in performance observed across datasets and methods.

Overall, these findings indicate that domain adaptation can partially mitigate domain shift under certain conditions, but its effectiveness varies depending on dataset characteristics, including class distribution, acquisition conditions, and feature representation. This highlights the importance of evaluating both baseline and adapted models when assessing cross-dataset generalisation.

Several limitations should be acknowledged. First, the study focuses on a binary classification setting, which simplifies the problem but may not capture the full variability of cognitive load levels. Second, only EDA signals were considered, which may limit the generalisability of the findings to multimodal settings. In addition, the extracted features were not explicitly derived from physiological models, and future work could explore model-based approaches (e.g., SCR point-process models) to further enhance interpretability. Finally, although the evaluation was conducted on two independent datasets with comparable task structures, enabling a meaningful assessment of cross-dataset generalisation, further validation across additional datasets and experimental conditions would be beneficial to strengthen the evidence supporting the proposed approaches.

6. Conclusions

This study investigated the cross-dataset performance of machine learning models for mental workload detection using EDA signals under domain shift conditions.

The results demonstrate that cross-dataset performance is strongly dependent on the direction of transfer, revealing a marked asymmetry between configurations. This asymmetry is associated with distributional differences in amplitude- and energy-related

features, as well as acquisition-related variability, leading to different levels of domain compatibility across datasets.

Domain adaptation improved performance in specific scenarios, particularly when source and target feature distributions were sufficiently aligned. However, its effectiveness was not consistent across models or datasets, and remained limited under pronounced domain mismatch. These results indicate that domain adaptation does not uniformly compensate for distribution shifts, but depends on the degree of alignment between feature spaces.

Overall, the findings show that cross-dataset performance is constrained by feature-level distribution shifts and acquisition differences. In particular, the analysis indicates that discriminative information is primarily driven by amplitude- and tonic-related components, while variability across subjects reflects inter-individual differences in EDA responses. These results highlight that performance depends not only on model selection, but also on the compatibility of feature representations, emphasising the need for harmonised acquisition protocols or more robust domain-invariant representations.

Future work will focus on extending the analysis to additional datasets, exploring multimodal approaches, and investigating alternative domain adaptation techniques to further improve robustness and generalisation.

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

Funding: This research was funded by FCT—Fundação para a Ciência e Tecnologia within the R&D Unit Project Scope UID/00319/2025—Centro ALGORITMI (ALGORITMI/UM) https://doi.org/10.54499/UID/00319/2025.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the University of Minho (reference CEICSH 005/2023, dated 5 January 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the Worker H&P study prior the experimentation phase.

Data Availability Statement: The datasets used in this study are publicly available on Zenodo repositories for research purposes. The Worker H&P dataset is available at https://zenodo.org/records/10688787 (accessed on 7 January 2026), and the UNIVERSE dataset is available at https://zenodo.org/records/10371068 (accessed on 7 January 2026).

Acknowledgments: The authors acknowledge the WORKER H&P project (NORTE-01-0247-FEDER-070146) and the Sepri Group for their support and contributions. During the preparation of this manuscript, the authors used ChatGPT (version 5.3) for the purposes of refining the language and improve clarity of the manuscript. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Receiver Operating Characteristic Curve
CORAL	Correlation Alignment

EDA	Electrodermal Activity
EER	Equal Error Rate
FN	False Negative
FP	False Positive
MAT	Mental Arithmetic Test
RF	Random Forest
ROC	Receiver Operating Characteristic
SA	Subspace Alignment
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SHAP	SHapley Additive exPlanations
SVC	Support Vector Classifier
TN	True Negative
TP	True Positive
XGBoost	Extreme Gradient Boosting

Appendix A

Table A1. List of parameters used for optimisation of the machine learning algorithms.

Model	Parameter	Range	Step
RF	number of trees	[100, 150, 200]	1
	max depth	3 to 5	
	max features	[20, 25, 30]	
XGBoost	number of trees	[100, 150, 200]	1
	learning rate	[0.01, 0.05, 0.1]	
	max depth	3 to 5	
	subsample	[0.6]	
	subsample ratio of columns	[0.6, 0.8]	
	L1 regularization (alpha)	[0.1, 0.5]	
	L2 regularization (lambda)	[1, 2, 5]	
SVC	kernel	[rbf, linear]	
	regularization parameter (C)	[0.1, 1, 10]	
	kernel coefficient (gamma)	[0.01, 0.1, 1]	

Table A2. Best hyperparameter configurations for each model when the UNIVERSE dataset is used as the target domain. Results correspond to the target F1-score.

Model (Experiment)	Best Parameters	F1-Score
RF (Baseline)	number of trees: 150, max depth: 5, min samples split: 20	0.698
RF (CORAL)	number of trees: 100, max depth: 5, min samples split: 25	0.738
RF (SA)	number of trees: 150, max depth: 5, min samples split: 20	0.761
XGBoost (Baseline)	number of trees: 200, learning rate: 0.01, subsample ratio: 0.8, max depth: 5, alpha: 0.5, lambda: 5	0.724
XGBoost (CORAL)	number of trees: 150, learning rate: 0.05, subsample ratio: 0.8, max depth: 5, alpha: 0.5, lambda: 5	0.644
XGBoost (SA)	number of trees: 150, learning rate: 0.01, subsample ratio: 0.8, max depth: 3, alpha: 0.5, lambda: 5	0.755
SVC (Baseline)	C: 0.1, gamma: 0.01, kernel: linear	0.837
SVC (CORAL)	C: 0.1, gamma: 0.01, kernel: linear	0.818
SVC (SA)	C: 0.1, gamma: 0.01, kernel: linear	0.755

Table A3. Best hyperparameter configurations for each model when the Worker H&P dataset is used as the target domain. Results correspond to the target F1-score.

Model (Experiment)	Best Parameters	F1-Score
RF (Baseline)	number of trees: 200, max depth: 5, min samples split: 20,	0.629
RF (CORAL)	number of trees: 150, max depth: 5, min samples split: 30,	0.633
RF (SA)	number of trees: 150, max depth: 5, min samples split: 30,	0.550
XGBoost (Baseline)	number of trees: 150, learning rate: 0.05, subsample ratio: 0.8, max depth: 3, alpha: 0.1, lambda: 1	0.635
XGBoost (CORAL)	number of trees: 200, learning rate: 0.1, subsample ratio: 0.8, max depth: 3, alpha: 0.1, lambda: 1	0.558
XGBoost (SA)	number of trees: 200, learning rate: 0.1, subsample ratio: 0.6, max depth: 5, alpha: 0.1, lambda: 2	0.554
SVC (Baseline)	C: 10, gamma: 0.01, kernel: rbf	0.606
SVC (CORAL)	C: 0.1, gamma: 0.1, kernel: rbf	0.604
SVC (SA)	C: 0.1, gamma: 0.01, kernel: linear	0.580

Table A4 reports the strongest correlations among features, while Table A5 summarises cross-correlations between SCR and SCL descriptors.

Table A4. Top correlated feature pairs (Pearson $|r| > 0.9$).

Feature 1	Feature 2	Correlation
scr_momentum	scr_activity	1.00
scr_std	scr_rms	1.00
scr_momentum	scr_average_power	1.00
scr_activity	scr_average_power	1.00
scr_Energy	scr_Entropy	0.98
scl_std	scl_range	0.98
scr_activity	scr_acr_length	0.96
scr_momentum	scr_acr_length	0.96
scr_acr_length	scr_average_power	0.96
scr_std	scr_range	0.95
scr_range	scr_rms	0.95
scr_rms	scr_integral	0.95
scr_std	scr_integral	0.95
scr_max	scr_range	0.94
scl_mean	scl_min	0.93

Table A5. Top ten correlations (r) between SCR- and SCL-derived features.

SCR Feature	SCL Feature	Correlation
scr_integral	scl_range	0.77
scr_rms	scl_range	0.75
scr_std	scl_range	0.75
scr_integral	scl_std	0.73
scr_std	scl_std	0.71
scr_rms	scl_std	0.71
scr_range	scl_range	0.66
scr_range	scl_std	0.64
scr_min	scl_range	−0.61
scr_max	scl_range	0.61

The correlation analysis confirms substantial redundancy within feature groups, particularly among SCR amplitude- and energy-related descriptors (Table A4). Similarly, strong correlations were observed within SCL features. In contrast, cross-correlations between SCR and SCL descriptors remain moderate (Table A5), supporting their complementary role.

Table A6. Top features exhibiting the largest domain shift between the UNIVERSE and Worker H&P datasets, ranked by absolute Cliff's delta. Effect sizes are interpreted following standard thresholds, with large effects defined as $|\delta| > 0.474$.

Feature	Mean (UNIVERSE)	Mean (Worker H&P)	δ	Effect	Direction
scr_integral	6.712	23.700	−0.784	large	Worker H&P > UNIVERSE
scr_RMS	0.047	0.137	−0.767	large	Worker H&P > UNIVERSE
scr_average_power	0.014	0.026	−0.767	large	Worker H&P > UNIVERSE
scr_momentum	0.014	0.026	−0.767	large	Worker H&P > UNIVERSE
scr_activity	0.014	0.026	−0.767	large	Worker H&P > UNIVERSE
scr_std	0.047	0.137	−0.767	large	Worker H&P > UNIVERSE
scl_range	0.290	0.949	−0.751	large	Worker H&P > UNIVERSE
scl_std	0.087	0.287	−0.744	large	Worker H&P > UNIVERSE
scl_momentum	0.035	0.131	−0.744	large	Worker H&P > UNIVERSE
scr_max	0.172	0.420	−0.734	large	Worker H&P > UNIVERSE

Table A7. Statistical significance analysis (Wilcoxon signed-rank test) for F1-score across folds.

Target	Model	Comparison	Δ F1	CI (95%)	*p*-Value	Significant
UNIVERSE	RF	Baseline vs. CORAL	+0.009	[−0.0256, 0.0446]	0.557	No
	RF	Baseline vs. SA	+0.062	[0.0483, 0.0760]	0.002	Yes
	XGBoost	Baseline vs. CORAL	+0.013	[−0.0016, 0.0280]	0.106	No
	XGBoost	Baseline vs. SA	+0.025	[−0.0160, 0.0659]	0.322	No
	SVC	Baseline vs. CORAL	+0.042	[0.0002, 0.0837]	0.232	Yes
	SVC	Baseline vs. SA	−0.006	[−0.0500, 0.0382]	0.492	No
Worker H&P	RF	Baseline vs. CORAL	+0.020	[0.0099, 0.0307]	0.002	Yes
	RF	Baseline vs. SA	−0.097	[−0.1500, −0.0440]	0.005	Yes
	XGBoost	Baseline vs. CORAL	−0.032	[−0.0497, −0.0125]	0.014	Yes
	XGBoost	Baseline vs. SA	−0.109	[−0.1328, −0.0848]	0.002	Yes
	SVC	Baseline vs. CORAL	−0.009	[−0.0180, −0.0009]	0.084	No
	SVC	Baseline vs. SA	−0.081	[−0.1002, −0.0608]	0.020	Yes

Figures A1 and A2 show the SHAP summary plots and corresponding feature importance rankings for the RF, XGBoost, and SVC classifiers.

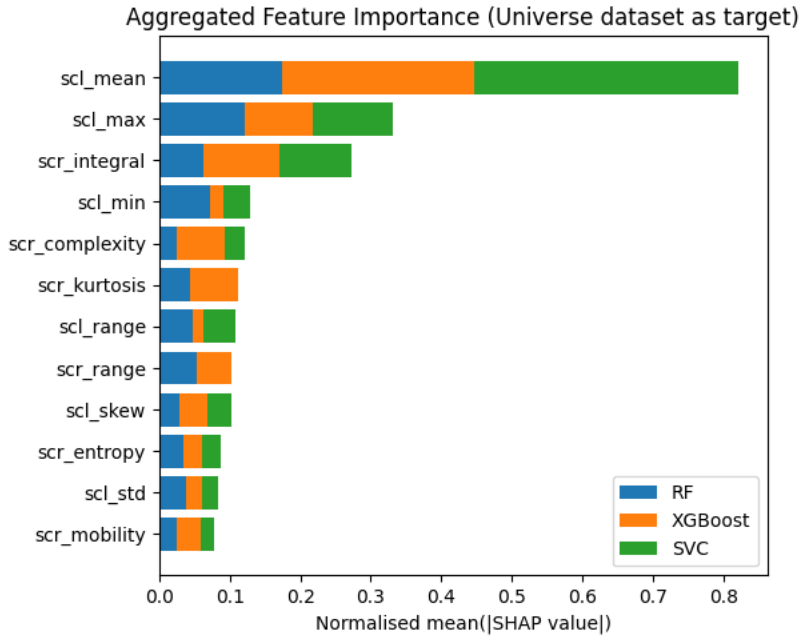

Figure A1. Aggregated SHAP-based feature importance for the UNIVERSE dataset as target domain. Bars represent the normalised mean absolute SHAP values, stacked to show the relative contribution of the RF, XGBoost, and SVC classifiers.

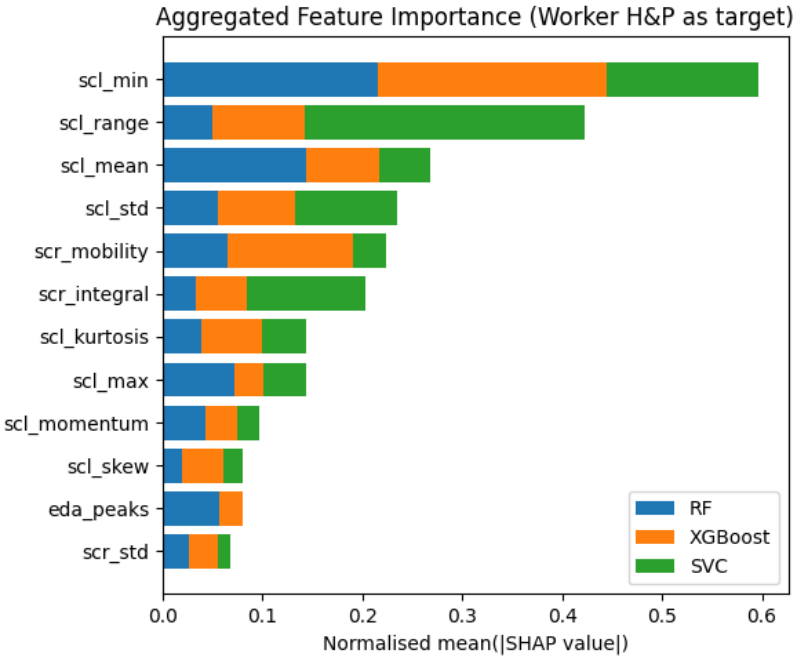

Figure A2. Aggregated SHAP-based feature importance for the Worker H&P dataset as target domain. Bars represent the normalised mean absolute SHAP values, stacked to show the relative contribution of the RF, XGBoost, and SVC classifiers.

Tables A8 and A9 present the subject-level SHAP analysis for the UNIVERSE and Worker H&P target scenarios, respectively, using the Random Forest baseline. The tables report the top ten features ranked by mean SHAP value, along with their standard deviation across subjects to characterise variability in feature importance.

Table A8. Subject-level SHAP analysis for the UNIVERSE dataset as target (Random Forest baseline). Top ten features ranked by mean SHAP value.

Feature	Mean SHAP	Std SHAP
scl_max	0.0138	0.0140
scl_mean	0.0110	0.0056
scl_min	0.0106	0.0053
scr_Entropy	0.0086	0.0039
scr_range	0.0074	0.0032
scr_integral	0.0072	0.0048
scr_momentum	0.0062	0.0029
scr_Energy	0.0053	0.0032
scr_kurtosis	0.0039	0.0031
scr_rms	0.0039	0.0024

Table A9. Subject-level SHAP analysis for the Worker H&P dataset as target (Random Forest baseline). Top ten features ranked by mean SHAP value.

Feature	Mean SHAP	Std SHAP
scl_min	0.0304	0.0136
scr_mobility	0.0237	0.0012
scl_std	0.0229	0.0022
scl_range	0.0221	0.0027
eda_peaks	0.0200	0.0030
scl_momentum	0.0179	0.0013
scl_max	0.0170	0.0072
scr_std	0.0120	0.0006
scr_average_power	0.0086	0.0007
scr_min	0.0084	0.0008

References

1. Brownson, R.C.; Boehmer, T.K.; Luke, D.A. Declining rates of physical activity in the United States: What are the contributors? *Annu. Rev. Public Health* **2005**, *26*, 421. [CrossRef] [PubMed]
2. Owen, N.; Sparling, P.B.; Healy, G.N.; Dunstan, D.W.; Matthews, C.E. Sedentary behavior: Emerging evidence for a new health risk. In *Mayo Clinic Proceedings*; Elsevier: Amsterdam, The Netherlands, 2010; Volume 85, pp. 1138–1141.
3. Bennie, J.A.; Pedisic, Z.; Timperio, A.; Crawford, D.; Dunstan, D.; Bauman, A.; Van Uffelen, J.; Salmon, J. Total and domain-specific sitting time among employees in desk-based work settings in Australia. *Aust. N. Z. J. Public Health* **2015**, *39*, 237–242. [CrossRef]
4. Cinaz, B.; Arnrich, B.; La Marca, R.; Tröster, G. Monitoring of mental workload levels during an everyday life office-work scenario. *Pers. Ubiquitous Comput.* **2013**, *17*, 229–239. [CrossRef]
5. Androutsou, T.; Angelopoulos, S.; Hristoforou, E.; Matsopoulos, G.K.; Koutsouris, D.D. A Multisensor System Embedded in a Computer Mouse for Occupational Stress Detection. *Biosensors* **2022**, *13*, 10. [CrossRef] [PubMed]
6. Longo, L.; Wickens, C.D.; Hancock, G.; Hancock, P.A. Human Mental Workload: A Survey and a Novel Inclusive Definition. *Front. Psychol.* **2022**, *13*, 883321. [CrossRef]
7. Zahmat Doost, E.; Zhang, W. Mental workload variations during different cognitive office tasks with social media interruptions. *Ergonomics* **2023**, *66*, 592–608. [CrossRef]
8. Muñoz, J.E.; Pereira, F.; Karapanos, E. Workload management through glanceable feedback: The role of heart rate variability. In *Proceedings of the 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*; IEEE: New York, NY, USA, 2016; pp. 1–6.
9. Mehta, R.K.; Agnew, M.J. Effects of physical and mental demands on shoulder muscle fatigue. *Work* **2012**, *41*, 2897–2901. [CrossRef]
10. Masi, G.; Amprimo, G.; Ferraris, C.; Priano, L. Stress and workload assessment in aviation—A narrative review. *Sensors* **2023**, *23*, 3556. [CrossRef]
11. Ahmed, S.; Babski-Reeves, K.; DuBien, J.; Webb, H.; Strawderman, L. Fatigue differences between Asian and Western populations in prolonged mentally demanding work-tasks. *Int. J. Ind. Ergon.* **2016**, *54*, 103–112. [CrossRef]
12. Ramírez-Moreno, M.A.; Carrillo-Tijerina, P.; Candela-Leal, M.O.; Alanis-Espinosa, M.; Tudón-Martínez, J.C.; Roman-Flores, A.; Ramírez-Mendoza, R.A.; Lozoya-Santos, J.d.J. Evaluation of a fast test based on biometric signals to assess mental fatigue at the workplace—A pilot study. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11891. [CrossRef]
13. Costa, N.; Costa, S.; Pereira, E.; Arezes, P.M. Workload measures—Recent trends in the driving context. *Occup. Environ. Saf. Health* **2019**, *202* , 419–430.
14. Ding, Y.; Cao, Y.; Duffy, V.G.; Wang, Y.; Zhang, X. Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics* **2020**, *63*, 896–908. [CrossRef]
15. Jin, L.; Liu, X.; Guo, B.; Han, Z.; Wang, Y.; Cao, Y.; Yang, X.; Shi, J. Impact of non-driving related task types, request modalities, and automation on driver takeover: A meta-analysis. *Saf. Sci.* **2025**, *181*, 106704. [CrossRef]
16. Alberdi, A.; Aztiria, A.; Basarab, A.; Cook, D.J. Using smart offices to predict occupational stress. *Int. J. Ind. Ergon.* **2018**, *67*, 13–26. [CrossRef] [PubMed]
17. Pütz, S.; Rick, V.; Mertens, A.; Nitsch, V. Using IoT devices for sensor-based monitoring of employees' mental workload: Investigating managers' expectations and concerns. *Appl. Ergon.* **2022**, *102*, 103739. [CrossRef]
18. Giorgi, A.; Ronca, V.; Vozzi, A.; Sciaraffa, N.; Di Florio, A.; Tamborra, L.; Simonetti, I.; Aricò, P.; Di Flumeri, G.; Rossi, D.; et al. Wearable technologies for mental workload, stress, and emotional state assessment during working-like tasks: A comparison with laboratory technologies. *Sensors* **2021**, *21*, 2332. [CrossRef]
19. Markova, V.; Ganchev, T.; Kalinkov, K.; Markov, M. Detection of acute stress caused by cognitive tasks based on physiological signals. *Bull. Electr. Eng. Inform.* **2021**, *10*, 2539–2547. [CrossRef]
20. Kaji, H.; Iizuka, H.; Sugiyama, M. ECG-based concentration recognition with multi-task regression. *IEEE Trans. Biomed. Eng.* **2018**, *66*, 101–110. [CrossRef]
21. Anusha, A.; Jose, J.; Preejith, S.; Jayaraj, J.; Mohanasankar, S. Physiological signal based work stress detection using unobtrusive sensors. *Biomed. Phys. Eng. Express* **2018**, *4*, 065001. [CrossRef]
22. Ghaderyan, P.; Abbasi, A. An efficient automatic workload estimation method based on electrodermal activity using pattern classifier combinations. *Int. J. Psychophysiol.* **2016**, *110*, 91–101. [CrossRef] [PubMed]
23. Mihirette, S.; De la Cal, E.A.; Tan, Q.; Sedano, J. Cross-contextual stress prediction: Simple methodology for comparing features and sample domain adaptation techniques in vital sign analysis. *Appl. Intell.* **2025**, *55*, 420. [CrossRef]
24. Aricò, P.; Reynal, M.; Di Flumeri, G.; Borghini, G.; Sciaraffa, N.; Imbert, J.P.; Hurter, C.; Terenzi, M.; Ferreira, A.; Pozzi, S.; et al. How neurophysiological measures can be used to enhance the evaluation of remote tower solutions. *Front. Hum. Neurosci.* **2019**, *13*, 303. [CrossRef] [PubMed]

25. Saha, S.; Jindal, K.; Shakti, D.; Tewary, S.; Sardana, V. Chirplet transform-based machine-learning approach towards classification of cognitive state change using galvanic skin response and photoplethysmography signals. *Expert Syst.* **2022**, *39*, e12958. [CrossRef]
26. Posada-Quintero, H.F.; Chon, K.H. Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors* **2020**, *20*, 479. [CrossRef]
27. Mehler, B.; Reimer, B.; Coughlin, J.F. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups. *Hum. Factors* **2012**, *54*, 396–412. [CrossRef]
28. Boucsein, W. *Electrodermal Activity*; Springer Science & Business Media: Berlin, Germany, 2012.
29. Pereira, E.; Sigcha, L.; Silva, E.; Sampaio, A.; Costa, N.; Costa, N. Capturing mental workload through physiological sensors in human–robot collaboration: A systematic literature review. *Appl. Sci.* **2025**, *15*, 3317. [CrossRef]
30. Setz, C.; Arnrich, B.; Schumm, J.; La Marca, R.; Tröster, G.; Ehlert, U. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *14*, 410–417. [CrossRef]
31. Zhou, J.; Jung, J.Y.; Chen, F. Dynamic workload adjustments in human-machine systems based on GSR features. In *Proceedings of the Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, 14–18 September 2015*; Proceedings, Part I 15; Springer: Berlin/Heidelberg, Germany, 2015; pp. 550–558.
32. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [CrossRef]
33. Wang, M.; Deng, W. Deep visual domain adaptation: A survey. *Neurocomputing* **2018**, *312*, 135–153. [CrossRef]
34. Sun, B.; Feng, J.; Saenko, K. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2016; Volume 30.
35. Fernando, B.; Habrard, A.; Sebban, M.; Tuytelaars, T. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: New York, NY, USA, 2013; pp. 2960–2967.
36. Wu, D.; Xu, Y.; Lu, B.L. Transfer learning for EEG-based brain–computer interfaces: A review of progress made since 2016. *IEEE Trans. Cogn. Dev. Syst.* **2020**, *14*, 4–19. [CrossRef]
37. Zanini, P.; Congedo, M.; Jutten, C.; Said, S.; Berthoumieu, Y. Transfer learning: A Riemannian geometry framework with applications to brain–computer interfaces. *IEEE Trans. Biomed. Eng.* **2017**, *65*, 1107–1116. [CrossRef] [PubMed]
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.
40. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
41. Sigcha, L.; Pereira, E.; Lima, A.; Cardoso, P.; Costa, N. *Worker HP—Worker Health and Performance: Biosignals of Induced Mental Workload in Office Environment*; Zenodo: Geneva, Switzerland, 2024. [CrossRef]
42. Anders, C.; Moontaha, S.; Real, S.; Arnrich, B. Unobtrusive measurement of cognitive load and physiological signals in uncontrolled environments. *Sci. Data* **2024**, *11*, 1000. [CrossRef]
43. Alchieri, L.; Garzon, M.; Alecci, L.; De Bona, F.B.; Gjoreski, M.; De Felice, G.; Santini, S. A foundation model for electrodermal activity data. *arXiv* **2026**, arXiv:2603.16878. [CrossRef]
44. Ishchenko, A.; Shev'ev, P. Automated complex for multiparameter analysis of the galvanic skin response signal. *Biomed. Eng.* **1989**, *23*, 113–117. [CrossRef]
45. Nardelli, M.; Greco, A.; Sebastiani, L.; Scilingo, E.P. ComEDA: A new tool for stress assessment based on electrodermal activity. *Comput. Biol. Med.* **2022**, *150*, 106144. [CrossRef]
46. Horvers, A.; Tombeng, N.; Bosse, T.; Lazonder, A.W.; Molenaar, I. Detecting emotions through electrodermal activity in learning contexts: A systematic review. *Sensors* **2021**, *21*, 7869. [CrossRef] [PubMed]
47. Campanella, S.; Altaleb, A.; Belli, A.; Pierleoni, P.; Palma, L. A method for stress detection using empatica E4 bracelet and machine-learning techniques. *Sensors* **2023**, *23*, 3565. [CrossRef]
48. Veeranki, Y.R.; Rao, N.K.; Posada-Quintero, H.F.; Swaminathan, R. Recent trends in electrodermal activity signal processing and deep learning methods for emotion recognition. *Neuroscience* **2026**, *602*, 12–30. [CrossRef] [PubMed]
49. Föll, S.; Maritsch, M.; Spinola, F.; Mishra, V.; Barata, F.; Kowatsch, T.; Fleisch, E.; Wortmann, F. FLIRT: A feature generation toolkit for wearable data. *Comput. Methods Programs Biomed.* **2021**, *212*, 106461. [CrossRef] [PubMed]
50. Gedam, S.; Paul, S. A review on mental stress detection using wearable sensors and machine learning techniques. *IEEE Access* **2021**, *9*, 84045–84066. [CrossRef]
51. Borzì, L.; Sigcha, L.; Firouzi, F.; Olmo, G.; Demrozi, F. Edge-based freezing of gait recognition in Parkinson's disease. *Comput. Electr. Eng.* **2025**, *127*, 110530. [CrossRef]
52. Sigcha, L.; Borzì, L.; Olmo, G. Deep learning algorithms for detecting freezing of gait in Parkinson's disease: A cross-dataset study. *Expert Syst. Appl.* **2024**, *255*, 124522. [CrossRef]
53. de Mathelin, A.; Atiq, M.; Richard, G.; de la Concha, A.; Yachouti, M.; Deheeger, F.; Mougeot, M.; Vayatis, N. Adapt: Awesome domain adaptation python toolbox. *arXiv* **2021**, arXiv:2107.03049.

54. Japkowicz, N.; Shah, M. *Evaluating Learning Algorithms: A Classification Perspective*; Cambridge University Press: Cambridge, UK, 2011.
55. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef]
56. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
57. Wilson, G.; Cook, D.J. A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol. (TIST)* **2020**, *11*, 1–46. [CrossRef]