SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Verbatim Duplication in a Public Urban Exposome Corpus: A Data Audit and Corrected Evaluation of Wellbeing Classification

Rabinarayan Panda[1], Thalaimalaichamy M[2], I Hameem Shanavas[3], Jeyakumar S[4], A. Sivakumar[5] Madhawa Surendar[6*]

[1] Department of Computer Science and Engineering, Sapthagiri NPS University, Bengaluru, India.
[2]Department of Electronics and Communication Engineering, SASTRA Deemed University, Tamil Nadu, India.
[3,6] Department of Computer Science and Engineering, Garden City University, Bengaluru, India.
[4]School of Health Sciences, Garden City University, Bengaluru, India.
[5]Varuvan Vadivelan Institute of Technology, Tamil Nadu, India.
*Corresponding Author: Madhawa Surendar, E-mail: m.surendar@gmail.com

Abstract
Public sensor corpora underpin an expanding body of work on environmental exposure and mental wellbeing, yet the integrity of those corpora is rarely audited before models are trained on them. This paper reports a duplication audit of the Digital Exposome dataset, a multimodal corpus of seven environmental and four physiological channels recorded from forty participants walking predefined urban routes, distributed through Mendeley Data and redistributed on IEEE Data-Port. The audit establishes that 14,352 of the 42,436 released rows, or 33.8 percent, are verbatim copies of rows appearing earlier in the same file, arranged in seven contiguous blocks, and that the label is identical within 99.9 percent of copy pairs. The consequence for evaluation is severe. Under the contiguous eighty-twenty partition natural to a time-ordered corpus, every one of the 8,488 held-out rows is present verbatim in the training partition; under the random partition adopted in the corpus's own benchmark, 51.9 percent are. A gated recurrent classifier trained on the released file attains a macro-averaged F1-score of 0.941, whereas the identical architecture trained on the de-duplicated corpus under a purged split with a training-fitted scaler attains 0.299, below the 0.535 accuracy of constant majority-class prediction. Further, an oracle persistence predictor that simply repeats the preceding label attains a macro-F1 of 0.980, and only 49 of 5,558 held-out windows, or 0.88 percent, exhibit a label change at the prediction step. We therefore report a negative result: momentary wellbeing is not recoverable from the released exposome channels above trivial temporal baselines, and previously published performance on this corpus is not interpretable as predictive capability. A reusable audit procedure and a reporting protocol for time-ordered sensor corpora are supplied.

Keywords - Data Leakage, Verbatim Duplication, Urban Exposome, Wellbeing Classification, Benchmark Integrity, Negative Results, Wearable Sensing, Evaluation Protocol.

Introduction

Environmental exposure and physiological stress are widely held to act jointly on mental wellbeing, and the availability of wearable sensing has made it feasible to record both continuously and at the level of the individual participant [1], [2]. Public corpora that pair calibrated environmental measurements with wearable physiological streams and momentary self-report are correspondingly scarce and correspondingly valuable. The Digital Exposome dataset is one such corpus, comprising forty participants who walked predefined urban routes in Nottingham, United Kingdom, while seven environmental channels and four physiological channels were recorded at a common one-hertz reference alongside a five-point self-reported wellbeing scale [3]. The corpus was released through Mendeley Data [4] and subsequently redistributed on IEEE Data-Port in derivative form [5], and it has supported predictive modeling of wellbeing from fused environmental and physiological signals [6].

Machine learning results reported on public corpora carry an implicit warranty, namely that the held-out partition is disjoint from the training partition in the sense that matters, which is informational rather than merely positional. That warranty is easy to lose. Leakage arising from preprocessing applied before partitioning, from overlapping windows in time-series construction, or from near-duplicate records distributed across the split has been documented as a recurring cause of irreproducible findings across scientific machine learning [7], and specific corpora in vision [8] and in clinical imaging [9] have been shown to contain duplicated or near-duplicated records sufficient to invalidate the benchmarks built upon them. Nevertheless, the audit of a corpus for exact duplication is neither routine nor commonly reported, and a published dataset is generally treated as trustworthy input.

This paper reports the outcome of such an audit, undertaken not as an end in itself but in the course of building a federated classifier of momentary wellbeing on the DigitalExposome corpus. The classifier initially attained a macro-averaged F1-score of 0.941 on a held-out partition, a figure sufficiently high, given the subjectivity of the target, to warrant verification rather than celebration. The verification procedure that follows established that the released file contains extensive verbatim duplication, that the duplication places the entire held-out partition inside the training partition under a contiguous split, and that the reported performance is therefore a measurement of memorization. We report the audit, the corrected evaluation, and the negative result that the corrected evaluation yields.

Three contributions are advanced. First, a duplication audit of the DigitalExposome corpus is presented that localizes the copied regions to seven contiguous blocks, quantifies contamination of the held-out partition under two standard partitioning protocols, and establishes that the defect is present in the primary Mendeley release rather than introduced by the IEEE DataPort derivative or by local preprocessing. Second, a corrected evaluation is reported in which the same architecture is retrained on the de-duplicated corpus under a purged contiguous split with a scaler fitted on the training partition alone, together with three non-learned reference baselines and an oracle persistence predictor. Third, we demonstrate that the pure-versus-transition stratification commonly used to argue that a classifier exceeds label autocorrelation does not, on this corpus, isolate the windows at which prediction is actually required, and we supply the corrected stratification criterion.

Although the immediate object of the audit is a single corpus, the procedure and the reporting protocol generalize to any time-ordered sensor dataset in which records are dense, continuous, and partitioned positionally. The remainder of this paper is organized as follows. The next section reviews the corpus, the modeling work built upon it, and prior treatments of leakage and duplication in benchmark data. Section 3 formalizes the audit procedure, the corrected evaluation protocol, and the baselines. Section 4 reports the results. Section 5 concludes and states the implications for practitioners using this corpus and for the maintenance of public sensor repositories.

Literature Review

The DigitalExposome corpus was introduced as a data descriptor documenting a protocol in which each participant traversed a fixed urban route of approximately forty minutes duration, carrying a custom environmental monitoring unit and wearing a wrist-worn physiological sensor, while reporting momentary wellbeing through a smartphone application on a five-point emoji-based scale [3]. The environmental instrumentation was independently validated against a reference station of the national Automatic Urban and Rural Network, reporting calibration accuracies between ninety-seven and ninety-eight percent for particulate matter and nitrogen dioxide, which establishes the metrological credibility of the environmental channel [10]. The descriptor is limited to describing the collection protocol and the released channel structure, and proposes no predictive model. Moreover, the descriptor states that heart rate variability is among the released physiological measures, whereas the file distributed at the corresponding repository record contains an inter-beat-interval column and no heart rate variability column, a discrepancy that we return to in Section 4.

Predictive modeling on this collection protocol was first reported by the corpus authors, who fused environmental and physiological streams from a preliminary cohort and classified wellbeing using a deep belief network [6]. Although that study established the feasibility of the task and constitutes the most directly comparable prior art, it centralized all recordings for training, evaluated no robustness condition, and partitioned the data by unstructured random sampling over rows. The choice of partitioning protocol is consequential in the light of the audit reported here, and Section 4 quantifies its effect. Subsequent work has treated the corpus as a benchmark without reporting an integrity check, which is unremarkable, since such checks are seldom reported for any corpus.

A parallel literature has established the wider feasibility of physiological inference outside the laboratory. Wearable stress and affect recognition benchmarks assembled under controlled protocols demonstrate that electrodermal activity, blood volume pulse, and cardiac timing carry recoverable signal about acute affective state [11], and the ecological momentary assessment paradigm supplies the methodological basis for pairing such signals with repeated self-report in situ [12]. Federated formulations of physiological inference have since been reported for emotion recognition from multimodal wearable sensors [13] and for seizure detection on low-power wearable hardware [14], and systematic reviews covering studies to 2025 confirm both the growth of the field and its continued reliance on centralized training [15], [24]. In comparison with these, the exposome setting is distinguished by the fusion of ambient environmental measurements with physiological channels, and the predictive contribution of the environmental channels specifically has not been isolated in prior work.

The methodological literature on leakage is more directly pertinent. A systematic survey of machine-learning-based science identified leakage as the dominant cause of irreproducible results across seventeen fields, cataloguing eight distinct leakage types and finding that the most damaging arise from the training

and test partitions failing to be independent in the informational sense [7]. In vision, the CIFAR benchmarks were shown to contain substantial numbers of near-duplicate images spanning the train and test partitions, with measurable inflation of reported accuracy once purged [8]. In clinical machine learning, a systematic review of models for COVID-19 diagnosis from chest imaging found that a majority of the surveyed models were trained on publicly assembled datasets containing duplicated and improperly sourced images, and concluded that none of the models surveyed was of potential clinical use [9]. Further, the test partitions of several canonical benchmarks have been shown to carry pervasive label errors sufficient to reorder model rankings [18], and replication of the ImageNet evaluation on freshly collected data revealed accuracy drops of up to fifteen percentage points, which indicates that reported figures had been partly specific to the released partition rather than to the task [19]. The common structure of these findings is that a defect in the corpus, not in the model, produced the reported performance. Practical guidance for avoiding such outcomes exists and is explicit on the point that data integrity must be established before modelling begins [20].

Evaluation protocols for temporally ordered data are a related concern. Random cross-validation applied to an autocorrelated series places temporally adjacent and therefore statistically dependent observations on both sides of the partition, and blocked or rolling-origin schemes have been shown to yield materially different and more defensible estimates [16], [21]. Further, the construction of overlapping sliding windows introduces a second dependence, since adjacent windows differ by a single timestep. Hence the recommended practice is to partition the raw sequence once, construct windows within each region separately, and interpose a purge interval at the boundary. Although this practice addresses positional dependence, it does not address verbatim duplication distributed non-locally through the file, which is the defect reported here, and which no purely positional partitioning scheme can remedy.

Finally, the interpretation of classifier performance on slowly varying labels requires reference baselines that exploit temporal continuity. A predictor that repeats the previously observed label attains high accuracy whenever the label is persistent, and comparison against such a predictor is the standard means of establishing that a model contributes beyond autocorrelation. Metrics that correct for marginal distribution, notably the Matthews correlation coefficient, are recommended over the F1-score for imbalanced multiclass problems and are reported alongside the conventional averages throughout Section 4 [17].

Proposed Work

3.1 Corpus Under Audit

The corpus audited in this study is the DigitalExposome dataset as distributed at the Mendeley Data record [4]. The released file contains 42,436 rows and twelve columns: seven environmental channels, namely particulate matter at one, two-point-five, and ten micrometres, carbon monoxide, ammonia, nitrogen dioxide, and ambient noise; four physiological channels, namely heart rate, inter-beat interval, electrodermal activity, and blood volume pulse; and a single wellbeing label taking integer values on a five-point scale. The file carries neither a participant identifier nor an explicit timestamp column. All channel values are distributed in normalized form.

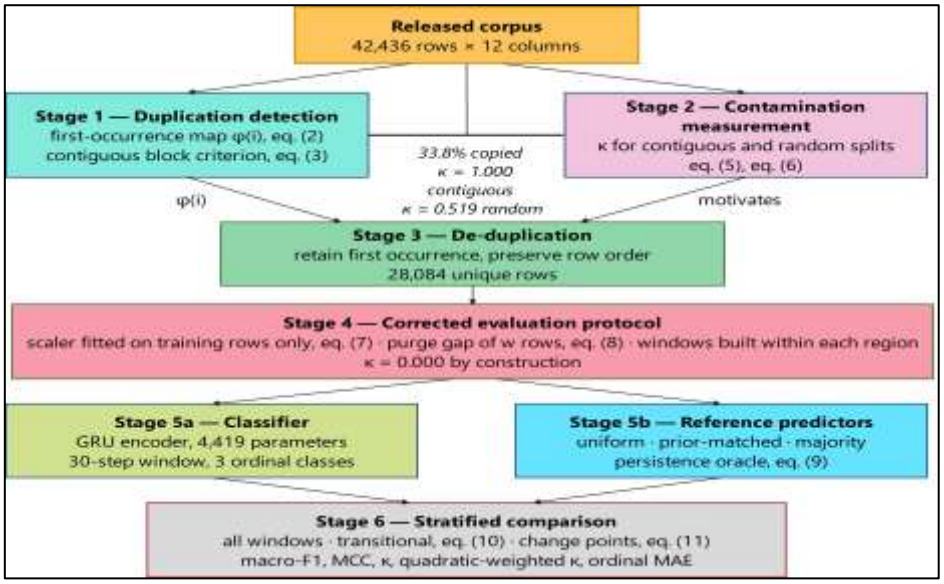

Fig 1. Audit and corrected-evaluation pipeline, from the released corpus through duplication detection and contamination measurement.

A derivative of this corpus is distributed on IEEE DataPort under a different title [5]. The derivative contains 42,436 rows and twenty-three columns, comprising the twelve original columns together with eleven appended categorical columns encoding thresholded risk annotations of the individual channels. We verified that the eleven raw channels and the label are identical row-for-row between the two files. Hence any defect present in the derivative is inherited from the primary release rather than introduced by redistribution, and any defect present in the primary release propagates to the derivative.

Throughout the audit, the eleven raw channels are treated as the feature vector and the appended risk annotations are excluded. The exclusion is deliberate rather than incidental: an annotation derived by thresholding a channel carries no information beyond that channel, and supplying both to a classifier would introduce a circularity independent of the duplication defect under examination.

The audit and the corrected evaluation are organized as a six-stage pipeline, shown in Fig. 1, in which the first two stages characterize the defect, the third removes it, and the remaining three establish what the corpus supports once it is removed. Each stage is specified formally in the subsections that follow.

3.2 Duplication Detection

Let X = (x_1, x_2, …, x_n) denote the released sequence of feature vectors, each an eleven-dimensional real vector, with n = 42,436, and let y_i denote the label associated with row i. A row is designated a copy when its feature vector has occurred at any earlier position in the file, which is expressed by the indicator in equation (1),

$$\mathcal{C}(i) = \mathbb{1}\left[\exists\, j < i : x_j = x_i\right] \tag{1}$$

and the first-occurrence map φ(i), which returns the earliest position at which the feature vector of row i appears, is defined in equation (2),

$$\phi(i) = \min\{j : x_j = x_i\} \tag{2}$$

so that C(i) = 1 precisely when φ(i) < i. Comparison is performed on exact floating-point equality across all eleven channels simultaneously. We adopt exact rather than approximate matching deliberately, since exact matching cannot produce false positives and therefore yields a conservative lower bound on the duplication present. Near-duplicate detection under a tolerance would report a larger figure, and the figures reported in Section 4 should be read accordingly.

Duplication is localized by identifying maximal contiguous runs of copied rows whose first-occurrence indices also advance contiguously. A block is recorded when the conditions of equation (3) hold over a run of length at least fifty,

$$\mathcal{C}(i) = 1 \text{ and } \phi(i+1) = \phi(i) + 1, \quad \forall i \in [a, b) \tag{3}$$

which identifies a segment [a, b] of the file as a verbatim replay of the segment [φ(a), φ(b)]. The threshold of fifty rows excludes incidental coincidences between rows of a quantized sensor stream and retains only structural replication.

Label agreement within copy pairs is measured by equation (4),

$$\alpha = \frac{1}{|\mathcal{K}|}\sum_{i \in \mathcal{K}} \mathbb{1}\left[y_i = y_{\phi(i)}\right], \quad \mathcal{K} = \{i : \mathcal{C}(i) = 1\} \tag{4}$$

and α value of α near unity indicates that the replication carries the target variable along with the features, which is the condition under which duplication becomes leakage rather than mere redundancy.

3.3 Contamination of the Held-Out Partition

The quantity of practical interest is not the duplication rate itself but the fraction of the held-out partition that is recoverable from the training partition by lookup. Let $\mathcal{T}$ and $\mathcal{H}$ denote the index sets of the training and held-out partitions under a given protocol. Contamination is defined in equation (5),

$$\kappa = \frac{1}{|\mathcal{H}|}\left|\left\{i \in \mathcal{H} : \exists\, j \in \mathcal{T},\ x_j = x_i\right\}\right| \tag{5}$$

and the accompanying label-agreement rate among contaminated rows is given by equation (6),

$$\alpha_\kappa = \Pr\left[\, y_i \in \{y_j : j \in \mathcal{T},\ x_j = x_i\} \;\middle|\; i \in \mathcal{H},\ \text{contaminated}\right] \tag{6}$$

Contamination is measured under two partitioning protocols. The first is the contiguous protocol, under which the first eighty percent of the file by position forms T and the remainder forms H; this protocol is the one appropriate to a time-ordered corpus and the one adopted in the modeling work that prompted the audit. The second is the unstructured random protocol, under which rows are assigned to partitions by uniform random sampling without regard to order; this protocol was adopted in the corpus's own benchmark evaluation [6]. Reporting both is necessary because the two protocols interact differently with block-structured duplication, and because published results on this corpus have used the second.

3.4 Corrected Evaluation Protocol

De-duplication retains the first occurrence of each distinct feature vector and discards subsequent copies, preserving the original row order among retained rows. The corrected corpus therefore contains one row per distinct feature vector, and every feature vector within it is unique by construction, so contamination under equation (5) is identically zero for any partitioning protocol.

Three further corrections are applied to the evaluation pipeline, each addressing a distinct source of optimism. Features are standardized using a scaler whose location and scale parameters are estimated on the training rows alone and then applied to the held-out rows, rather than estimated over the pooled corpus. A purge interval of w rows, where w is the window length, is discarded immediately after the partition point, so that no held-out window shares a raw timestep with any training window and residual autocorrelation across the boundary is excluded. Windows are constructed separately within each region after partitioning rather than before it. The corrected protocol is summarized in equations (7) and (8), in which c denotes the partition point at eighty percent of the de-duplicated sequence length,

$$\mu, \sigma \leftarrow \text{fit}(X'_{1:c}), \qquad \tilde{x}_i = (x'_i - \mu)/\sigma \tag{7}$$

$$\mathcal{T} = [1, c], \qquad \mathcal{H} = [c + w + 1,\ n'] \tag{8}$$

The contrast between the uncorrected and the corrected protocol is shown in Fig. 2. Under the uncorrected protocol the held-out region is drawn from replicated material and the scaler is estimated over the pooled corpus, whereas under the corrected protocol the scaler is estimated on training rows alone, a purge interval separates the regions, and contamination is zero by construction.

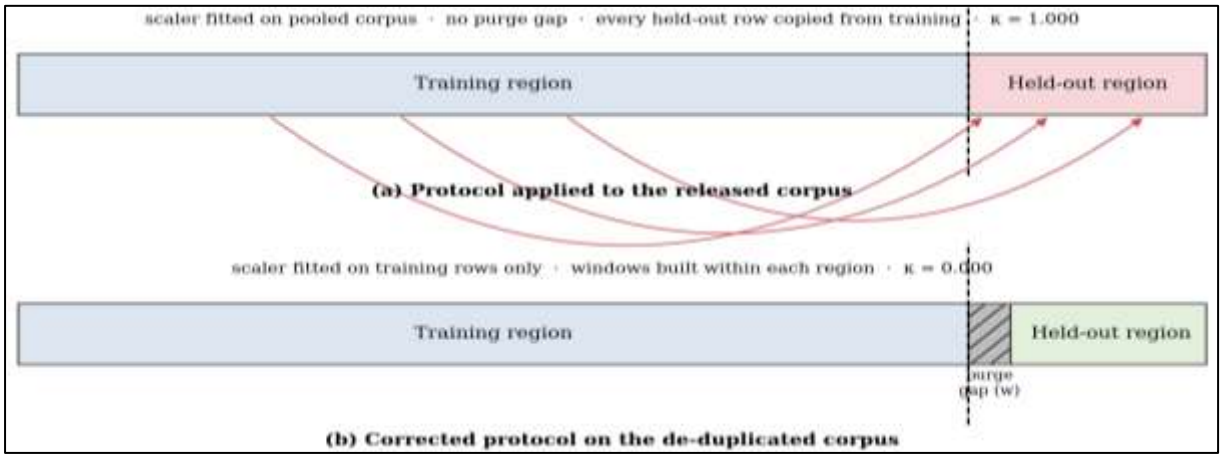

Fig 2. Partitioning and preprocessing under (a) the uncorrected protocol applied to the released corpus and (b) the corrected protocol applied to the de-duplicated corpus. Arcs in (a) indicate held-out rows recoverable verbatim from the training region.

3.5 Reference Baselines and Stratification

Four reference predictors are evaluated on the corrected corpus under the identical partition, the identical windowing, and the identical metric implementation as the classifier, so that the comparison is not confounded by differences of protocol. The uniform random predictor assigns classes with equal probability. The prior-matched random predictor samples from the class priors of the held-out partition. The majority-class predictor emits the modal training class unconditionally. These three exploit no temporal information and establish a floor.

The fourth predictor is the persistence oracle defined in equation (9),

$$\hat{y}^{\,\text{per}}(t) = y(t-1) \tag{9}$$

which repeats the label observed at the immediately preceding timestep. We emphasize that this predictor is an oracle and not a deployable method, since the wellbeing label is obtained by self-report and would not be available at inference time. Its purpose is to bound from above the accuracy attainable by exploiting label autocorrelation alone. Hence a classifier that fails to exceed the persistence oracle on a given subset has not been shown to contribute predictive information on that subset.

Stratification of the held-out windows requires care, and the criterion in common use is inadequate. A window is conventionally designated pure when all labels falling within its span are identical, and transitional otherwise, as expressed in equation (10),

$$\pi(t) = \mathbb{1}[\,|\{\, y(s) : s \in [t - w + 1,\ t] \,\}| = 1\,] \tag{10}$$

Although the transitional subset so defined excludes windows of constant label, it does not exclude windows whose label is constant at the prediction step, because a label change occurring early within the window leaves y(t) equal to y(t−1). We therefore introduce the change-point criterion of equation (11),

$$\chi(t) = \mathbb{1}[\, y(t) \neq y(t-1)\,] \tag{11}$$

which selects precisely those windows at which the persistence oracle must fail and at which genuine prediction is consequently required. Results are reported under both criteria in Section 4, and the discrepancy between them is itself a finding.

3.6 Classifier and Metrics

The classifier held constant across the contaminated and corrected evaluations is a single-layer gated recurrent encoder [22] of hidden dimension thirty-two operating on windows of thirty timesteps, followed by dropout [23] at rate 0.3 and a linear classification head over three ordinal wellbeing categories [26] obtained by coarsening the five-point scale, with the two lowest levels mapped to the low category, the midpoint retained, and the two highest levels mapped to the high category. The encoder and head together comprise 4,419 trainable parameters. Optimization uses the Adam rule [25] at a learning rate of 0.001 under a class-weighted cross-entropy objective. The architecture is deliberately unchanged between the two evaluations, since the object of comparison is the corpus rather than the model.

Macro-averaged F1 is treated as the primary metric, since it weights the minority category equally with the two majority categories, whereas weighted averaging is dominated by the majority categories and can conceal minority-class failure. Accuracy, the Matthews correlation coefficient, Cohen's kappa, its quadratic-weighted variant, and the mean absolute error in ordinal category units are reported alongside it. All metrics are computed by a single shared implementation invoked identically for every predictor reported.

4. Results

4.1 Duplication Structure of the Released Corpus

The audit establishes extensive verbatim duplication in the released file. Of the 42,436 rows, 28,084 carry distinct feature vectors and 14,352, or 33.8 percent, are exact copies of rows appearing earlier in the file. Label agreement within copy pairs, computed by equation (4), is 99.9 percent, which confirms that the replication carries the target variable along with the features.

The duplication is not diffuse. Seven contiguous blocks satisfying equation (3) account for all 14,352 copied rows, and their extents are reported in Table 1 and mapped in Fig. 3. Red segments mark rows that copy earlier material; blue segments mark the source regions; arcs link each copied block to its source. The dashed line marks the eighty-percent partition point, beyond which every row lies within a copied block. The blocks vary in length from 633 to 6,528 rows and their sources are scattered across the file, which is a signature consistent with a concatenation or assembly fault during dataset preparation rather than with any property of the underlying sensor recording. Two of the seven blocks, spanning rows 15,306 to 17,873, are duplicate copies of the same source segment, so that the segment beginning at row 10,089 appears three times in total.

Table 1. Contiguous Verbatim Duplication Blocks Identified in the Released DigitalExposome File

Block	Copied rows	Source rows	Length
1	15,306 – 16,589	10,089 – 11,372	1,284
2	16,590 – 17,873	10,089 – 11,372	1,284
3	30,651 – 32,135	29,166 – 30,650	1,485
4	32,137 – 32,769	1 – 633	633
5	32,770 – 34,278	5,923 – 7,431	1,509
6	34,279 – 40,806	8,778 – 15,305	6,528
7	40,807 – 42,435	19,526 – 21,154	1,629

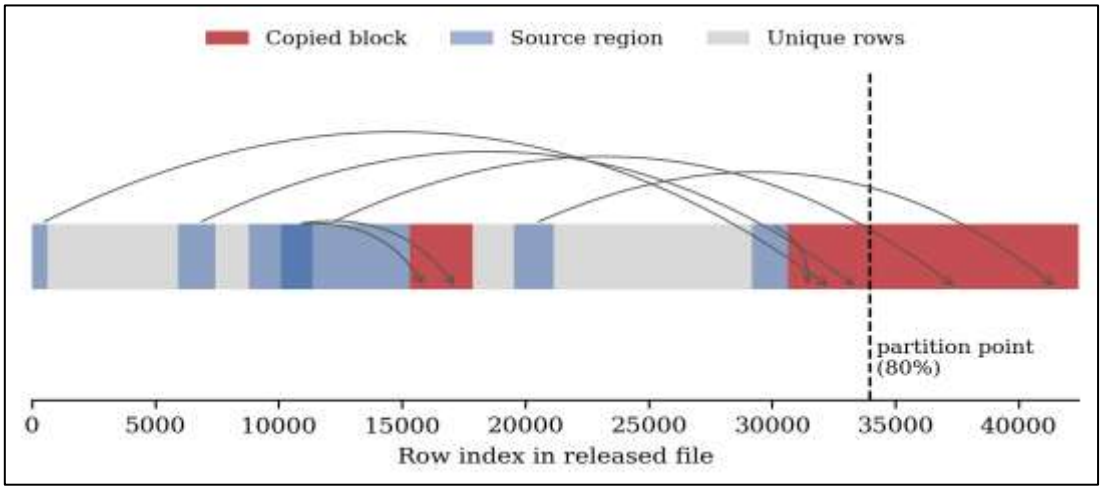

Fig 3. Duplication map of the released DigitalExposome file.

The final three blocks are decisive for evaluation, since together they span rows 32,770 to 42,435 and therefore cover the whole of the final quarter of the file. Any positional partition placing the tail of the file in the held-out set draws that set entirely from replicated material.

4.2 Contamination Under Two Partitioning Protocols

Contamination measured by equation (5) is reported in Table 2 and in Fig. 4. The contiguous partition appropriate to a time-ordered corpus yields total contamination, whereas de-duplication removes it entirely. Under the contiguous eighty-twenty protocol, all 8,488 held-out rows are present verbatim in the training partition, giving $\kappa = 1.000$. Under the unstructured random protocol, 4,408 of 8,488 held-out rows are so present, giving $\kappa = 0.519$. Label agreement among contaminated rows, computed by equation (6), is 99.9 percent under both protocols.

Table 2. Held-Out Partition Contamination by Partitioning Protocol

Protocol	Held-out rows	Contaminated	κ	α(κ)
Contiguous 80/20	8,488	8,488	1.000	0.999
Unstructured random 80/20	8,488	4,408	0.519	0.999
Contiguous 80/20, de-duplicated	5,617	0	0.000	—

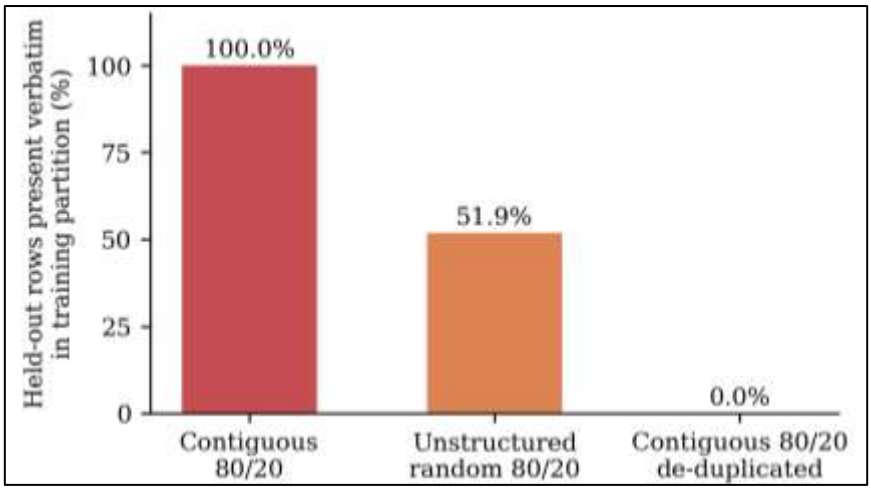

Fig 4. Held-out partition contamination under three protocols.

The asymmetry between the two protocols merits comment. The contiguous protocol is the more defensible choice for a time-ordered corpus and is the protocol recommended by the leakage literature [16], yet on this corpus it produces total contamination, because the duplicated blocks happen to occupy the tail of the file. The random protocol, which is ordinarily the weaker choice, produces roughly half the contamination simply because it distributes the copies across both partitions. Hence the defect inverts the usual ordering of protocol quality, and a practitioner following best practice on this corpus is penalized more heavily than one who does not. Prior published results using the random protocol [6] are nonetheless affected, since a contamination rate above one half with near-perfect label agreement is more than sufficient to dominate a reported metric.

4.3 Effect of Duplication on Reported Performance

The magnitude of the inflation is quantified in Table 3 and in Fig. 5. A random forest [27] trained on single rows of the released file, without any temporal window, attains 99.88 percent accuracy on the contiguous held-out partition, which is the signature of lookup rather than of learning. The gated recurrent classifier attains an accuracy of 95.8 percent and a macro-F1 of 0.941 on the same partition. That a windowed recurrent model performs measurably worse than a row-level random forest is itself diagnostic, since a genuine temporal relationship would be expected to favour the former.

Table 3. Held-Out Performance on the Released Versus the De-Duplicated Corpus

Corpus and protocol	Model	Accuracy	Macro-F1
Released, contiguous split	Random forest, single row	0.9988	0.9990
Released, contiguous split	GRU, 30-step window	0.9583	0.9413
De-duplicated, purged split	Random forest, single row	0.5476	0.2789
De-duplicated, purged split	GRU, 30-step window	0.4852	0.2987
De-duplicated, purged split	Majority class	0.5353	0.2324

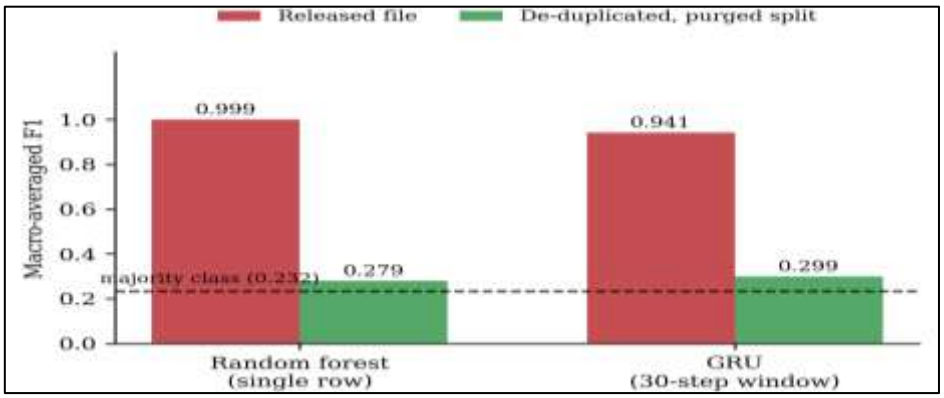

Fig 5. Macro-averaged F1 on the released file against the de-duplicated corpus under a purged split, for a row-level random forest and the 30-step gated recurrent classifier. The dashed line marks constant majority-class prediction.

Once the corpus is de-duplicated and the corrections of Section 3.4 are applied, macro-F1 for the recurrent classifier falls from 0.941 to 0.299, and accuracy falls from 0.958 to 0.485. Moreover, the corrected accuracy lies below the 0.535 attained by constant majority-class prediction. The corrected macro-F1 exceeds the majority-class macro-F1 of 0.232 by a small margin, which is attributable to the class-weighted objective producing some minority-class predictions rather than to recovered signal, and which we do not interpret as evidence of predictive capability.

4.4 Corrected Performance Against Reference Baselines

Table 4 and Fig. 4 report every predictor of Section 3.5 on the de-duplicated corpus under the identical protocol, stratified by the two criteria of equations (10) and (11). The held-out partition yields 5,558 windows, of which 4,546 are pure and 1,012 transitional under equation (10).

Table 4. Corrected Held-Out Performance by Predictor and Stratification Criterion

Predictor	All windows (n = 5,558)	Transitional, eq. (10) (n = 1,012)	Change points, eq. (11) (n = 49)
Uniform random	0.304	—	—
Prior-matched random	0.333	—	—
Majority class	0.232	0.209	0.179
GRU, 30-step window	0.299	0.282	0.290
Persistence oracle, eq. (9)	0.980	0.944	0.000

Values are macro-averaged F1. The persistence oracle attains 0.980 across all held-out windows and 0.944 on the transitional subset, against 0.299 and 0.282 respectively for the trained classifier. Further, the persistence oracle attains a macro-F1 of exactly 1.000 on the pure subset, which is a tautology rather than a result, since a pure window has constant label by definition and the preceding label therefore agrees with the target by construction.

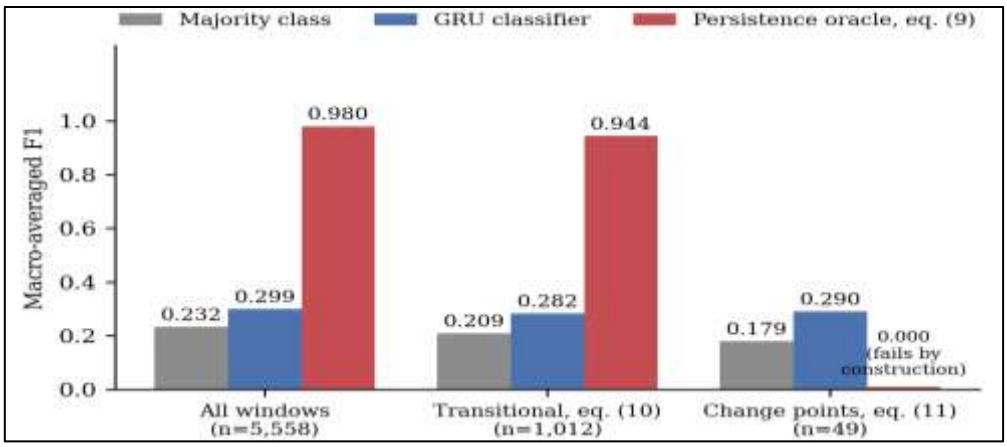

Fig 6. Corrected held-out performance by predictor and stratification criterion. The persistence oracle exceeds the trained classifier on every subset except the 49 change points, where it fails by construction.

The interpretation is unambiguous. The wellbeing label on this corpus is almost entirely predicted by its own immediately preceding value, and the eleven environmental and physiological channels add nothing detectable above that. Because the persistence oracle exceeds the trained classifier by a margin of 0.68 in macro-F1 on the full held-out partition, no claim of predictive contribution from the sensor channels is supportable on the corrected corpus.

4.5 Collapse of the Pure-Versus-Transition Stratification

The stratification of equation (10) is widely used to argue that a classifier exceeds label autocorrelation, and it fails on this corpus. Of the 1,012 windows designated transitional, 963 nevertheless satisfy $y(t) = y(t-1)$, because the label change falls earlier within the thirty-step span and not at the prediction step. Only 49 windows, or 0.88 percent of the held-out partition, satisfy the change-point criterion of equation (11). Table 5 and Fig. 5 report the decomposition.

Table 5. Decomposition of Held-Out Windows by Stratification Criterion

Window category	Count	Share of held-out partition
Pure, eq. (10)	4,546	81.8%
Transitional, eq. (10)	1,012	18.2%
Transitional but $y(t) = y(t-1)$	963	17.3%
Change points, eq. (11)	49	0.88%

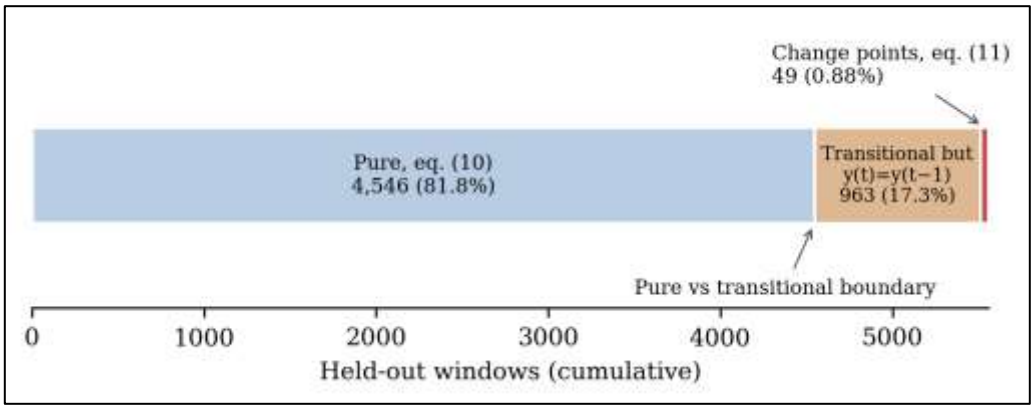

Fig 7. Decomposition of the 5,558 held-out windows. The transitional subset defined by equation (10) is dominated by windows whose label is nonetheless unchanged at the prediction step; only 0.88 percent of the partition satisfies the change-point criterion of equation (11).

Consequently, an evaluation restricted to the transitional subset still permits the answer to be read from the preceding timestep in ninety-five percent of cases, and the apparent difficulty of that subset is largely illusory. On the 49 genuine change points the classifier attains an accuracy of 0.347 and a macro-F1 of 0.290, whereas the persistence oracle attains zero by construction. Although the classifier consequently exceeds persistence on this subset, the subset comprises 49 windows, which is far too small to support any inference, and the classifier's absolute performance on it does not exceed chance. The majority-class predictor attains 0.179 on the change-point subset, where the three classes are approximately balanced with supports of 17, 14, and 18, so the classifier's 0.290 on that subset exceeds the majority baseline by a margin that 49 windows cannot render significant. We therefore report the change-point count as a property of the corpus rather than as an evaluation result: the corpus contains too few label transitions to support supervised learning of the transition itself.

4.: any published result on this corpus, obtained by any group, is affected.

A second and independent discrepancy concerns documentation. The data descriptor lists heart rate variability among the released physiological measures [3], whereas the released file contains an inter-beat-interval column and no heart rate variability column. Although inter-beat interval is the raw signal from which heart rate variability is conventionally derived, so that the substitution is recoverable by a careful user, the discrepancy may cause a reader to believe that a channel is available when it is not. We report it here for completeness rather than as a defect of comparable severity to the duplication.

5. Conclusion

This paper reports a duplication audit of the Digital Exposome corpus and a corrected evaluation of wellbeing classification upon it. The audit establishes that 33.8 percent of the released rows are verbatim copies arranged in seven contiguous blocks, that the label accompanies the copy in 99.9 percent of cases, and that the copied blocks occupy the tail of the file so completely that a contiguous eighty-twenty partition

places every held-out row inside the training partition. Under the unstructured random partition adopted in the corpus's own benchmark, contamination remains at 51.9 percent. The defect was confirmed present in the primary repository release and is therefore inherited by every derivative and every study built upon the corpus. The corrected evaluation yields a negative result, and we state it without qualification. On the de-duplicated corpus, under a purged contiguous partition with a scaler fitted on training rows alone, the same gated recurrent classifier that attained a macro-averaged F1 of 0.941 on the released file attains 0.299, with an accuracy of 0.485 that falls below the 0.535 obtained by constant majority-class prediction. An oracle predictor that repeats the preceding label attains 0.980. Momentary wellbeing, as labelled in this corpus, is therefore recoverable from its own recent history and is not recoverable from the eleven fused environmental and physiological channels above that trivial baseline. We do not claim that the underlying scientific hypothesis is false; the exposome literature rests on population-level correlational evidence that this result does not address. We claim only that the released corpus does not support the individual-timestep predictive claim that has been advanced on it.

A third finding concerns evaluation practice more broadly. The pure-versus-transition stratification used to demonstrate that a classifier exceeds label autocorrelation does not, on this corpus, isolate the windows at which prediction is required, since 963 of the 1,012 windows it designates as transitional still carry an unchanged label at the prediction step. Hence the criterion should be stated on the prediction step itself, as in equation (11), and the count of qualifying windows should be reported alongside the metric. Where that count is small, as it is here at 0.88 percent of the held-out partition, the corpus should be understood as unable to support the claim rather than as furnishing weak evidence for it.

Two recommendations follow for practitioners and for repository maintainers. First, an exact-duplication audit costs a single pass over the data and should be reported as a matter of course for any time-ordered sensor corpus, with the contamination rate of equation (5) stated for the partitioning protocol actually used; positional partitioning schemes, however carefully purged, do not detect non-local replication. Second, published performance on a corpus should be accompanied by a persistence baseline whenever the label is slowly varying, since the margin over that baseline, and not the absolute metric, is what carries the predictive claim. The corpus authors and the repository have been notified of both the duplication and the documentation discrepancy. Should a corrected release be issued, the corrected evaluation reported here provides the reference point against which any subsequent modeling result on this corpus can be assessed. Future work will extend the audit procedure to a wider set of public physiological and environmental corpora, in order to establish whether the defect reported here is isolated or characteristic of a class of assembled multimodal datasets.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Data and Code Availability

The audit procedure, the corrected evaluation pipeline, and the de-duplication utility are available from the author on request. The corpus audited is publicly available at the repository record cited as reference [4].

References

1. World Health Organization, *WHO Global Air Quality Guidelines: Particulate Matter, Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. Geneva, Switzerland: World Health Organization, 2021.
2. World Health Organization Regional Office for Europe, *Environmental Noise Guidelines for the European Region*. Copenhagen, Denmark: WHO Regional Office for Europe, 2018.
3. T. Johnson, E. Kanjo, and K. Woodward, "DigitalExposome: a dataset for wellbeing classification using environmental air quality and human physiological data," *Data in Brief*, vol. 59, art. 111373, 2025, doi: 10.1016/j.dib.2025.111373.
4. T. Johnson, "DigitalExposome: a dataset for wellbeing classification using environmental air quality and human physiological data," Mendeley Data, V2, 2025, doi: 10.17632/mbwxy48223.2.
5. M. A. I. Hemida, M. K. S. A. Altawil, and A. M. D. E. Hassanein, "Enviro-Health Cube: health and environmental monitoring dataset," IEEE DataPort, 2026, doi: 10.21227/y3r3-ht03.
6. T. Johnson, K. Woodward, E. Kanjo, and A. Bandera, "DigitalExposome: quantifying the urban environment influence on wellbeing based on real-time multi-sensor fusion and deep belief network," arXiv preprint, arXiv:2101.12615, 2021.
7. S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, art. 100804, 2023, doi: 10.1016/j.patter.2023.100804.
8. B. Barz and J. Denzler, "Do we train on test data? Purging CIFAR of near-duplicates," *Journal of Imaging*, vol. 6, no. 6, art. 41, 2020, doi: 10.3390/jimaging6060041.

9. M. Roberts, D. Driggs, M. Thorpe et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, pp. 199–217, 2021, doi: 10.1038/s42256-021-00307-0.
10. K. Woodward, E. Kanjo, D. J. Brown et al., "Beyond mobile apps: a survey of technologies for mental well-being," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1216–1235, 2022, doi: 10.1109/TAFFC.2020.3015018.
11. 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 Interaction (ICMI)*, 2018, pp. 400–408, doi: 10.1145/3242969.3242985.
12. S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annual Review of Clinical Psychology*, vol. 4, pp. 1–32, 2008, doi: 10.1146/annurev.clinpsy.3.022806.091415.
13. N. Gahlan and D. Sethia, "Federated learning inspired privacy sensitive emotion recognition based on multi-modal physiological sensors," *Cluster Computing*, vol. 27, pp. 3179–3201, 2024, doi: 10.1007/s10586-023-04133-4.
14. S. Baghersalimi, T. Teijeiro, A. Aminifar, and D. Atienza, "Decentralized federated learning for epileptic seizure detection in low-power wearable systems," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 6392–6407, 2024, doi: 10.1109/TMC.2023.3320862.
15. N. Gahlan and D. Sethia, "Federated learning in emotion recognition systems based on physiological signals for privacy preservation: a review," *Multimedia Tools and Applications*, vol. 84, pp. 12417–12485, 2025, doi: 10.1007/s11042-024-19467-3.
16. C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012, doi: 10.1016/j.ins.2011.12.028.
17. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, art. 6, 2020, doi: 10.1186/s12864-019-6413-7.
18. C. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in *Proc. 35th Conf. Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.
19. B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, PMLR 97, 2019, pp. 5389–5400.
20. M. A. Lones, "How to avoid machine learning pitfalls: a guide for academic researchers," arXiv preprint, arXiv:2108.02497, 2021.
21. L. J. Tashman, "Out-of-sample tests of forecasting accuracy: an analysis and review," *International Journal of Forecasting*, vol. 16, no. 4, pp. 437–450, 2000, doi: 10.1016/S0169-2070(00)00065-0.
22. K. Cho, B. van Merriënboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.
23. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
24. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, 2015.
25. A. O. Pataca, J. C. Ferreira, L. A. Silva et al., "Use of machine learning for predicting stress episodes based on wearable sensor data: a systematic review," *Computers in Biology and Medicine*, vol. 198, art. 111166, 2025, doi: 10.1016/j.compbiomed.2025.111166.
26. P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016, doi: 10.1109/TKDE.2015.2457911.
27. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.