Publicly Available Multimodal Biomedical Datasets: An Analysis of EEG, ECG, PPG, and fNIRS Signal Quality and Data Completeness

1. Executive Summary

The landscape of biomedical research is increasingly reliant on multimodal physiological data to unravel the complexities of human biology and cognition. This report provides a critical review of publicly available datasets encompassing electroencephalography (EEG), electrocardiogram (ECG), photoplethysmography (PPG), and functional near-infrared spectroscopy (fNIRS). Key repositories such as PhysioNet and OpenNeuro serve as foundational resources, offering diverse collections of these signals. A recurring theme in the analysis is the inherent challenge of maintaining high signal quality, particularly due to pervasive motion artifacts and physiological confounds, which are exacerbated in real-world data acquisition scenarios. Furthermore, data completeness, influenced by factors like sample size and the "missing modality problem," significantly impacts the generalizability and utility of these resources. The report highlights the critical role of "ground truth" data in validating signal processing techniques and underscores how evolving data sharing standards, coupled with advanced benchmarking frameworks, are instrumental in driving continuous improvements in data quality and reusability. Ultimately, while challenges persist, the synergistic combination of these modalities offers a profoundly more comprehensive understanding of physiological states, paving the way for more robust biomarkers and interventions.

2. Introduction to Multimodal Physiological Data in Biomedical Research

The integration of multiple physiological signals, known as multimodal physiological data, represents a pivotal advancement in contemporary biomedical research. This approach moves beyond the limitations of single-modality studies, offering a more holistic and nuanced perspective on complex biological and cognitive processes. By combining diverse data streams, researchers can gain a deeper understanding of phenomena such as brain states, cognitive workload, stress levels, and cardiovascular health.[1]

Each physiological modality brings unique strengths to this integrated framework. Electroencephalography (EEG) stands out for its exceptional temporal resolution, capturing millisecond-level changes in neuronal electrical activity with high precision.[2] In contrast, functional near-infrared spectroscopy (fNIRS) measures hemodynamic changes, specifically blood oxygenation, providing superior spatial resolution and

direct insights into brain responses, particularly within the prefrontal cortex.[1] The combination of EEG and fNIRS is particularly powerful, as it leverages the temporal strengths of EEG and the spatial insights of fNIRS, offering a more complete picture of neural activation than either method alone.[2]

For cardiovascular assessment, the electrocardiogram (ECG) is considered the gold standard, accurately measuring the heart's electrical activity and cardiac cycle.[7] Complementing this, photoplethysmography (PPG) offers a simple, non-invasive optical method for detecting changes in blood volume. PPG is highly suitable for continuous monitoring, especially through wearable devices, and its metrics often correlate strongly with those derived from ECG.[7] The increasing popularity of consumer wearable devices has significantly driven the collection of such physiological data, enabling studies in more naturalistic, real-world settings.[7] This technological progress, including the development of low-power, high-precision wearable sensors, is fundamentally reshaping biomedical research by expanding the scope of what can be monitored and understood outside traditional laboratory environments.[11]

However, the collection and analysis of multimodal data are not without their complexities. Challenges include ensuring precise signal synchronization across different devices, managing diverse sampling rates, and effectively addressing inherent noise and artifacts that can contaminate recordings.[2] The very desire to capture data in more ecologically valid settings often introduces additional noise and complexity, necessitating sophisticated data integration and artifact removal techniques.

To illustrate the distinct contributions and characteristics of these key modalities, Table 2 provides a comparative overview:

Table 2: Comparative Characteristics of EEG, ECG, PPG, and fNIRS

Feature	EEG (Electroencephalography)	ECG (Electrocardiogram)	PPG (Photoplethysmography)	fNIRS (functional Near-Infrared Spectroscopy)
Measured Process	Neuronal electrical activity	Heart's electrical activity	Blood volume changes in microvasculatur	Hemodynamic changes (HbO, HbR) in brain

			e	tissue
Temporal Resolution	Excellent (millisecond-level) [2]	Excellent (millisecond-level) [7]	Good (pulse waveform)	Slower (seconds, peak 4-6s post-stimulus) [4]
Spatial Resolution	Limited (scalp surface) [2]	Limited (cardiac source)	Limited (local blood flow)	Better (cortical regions) [1]
Invasiveness	Non-invasive	Non-invasive (surface electrodes)	Non-invasive (optical sensor)	Non-invasive
Common Artifacts/Noise	Eye blinks, muscle activity, motion, line noise [2]	Motion, baseline drift, power line noise [8]	Motion, ambient light, pressure variations [7]	Motion, respiration, superficial hemodynamics [16]
Continuous Monitoring	Possible (wearable EEG)	Challenging (requires stationary subject) [7]	Highly suitable (wearables) [7]	Possible (wearable fNIRS) [1]

The inherent limitations of single physiological signals necessitate a multimodal approach to achieve a comprehensive understanding of complex biological processes. For example, EEG's high temporal resolution is complemented by fNIRS's better spatial localization, allowing for a more complete characterization of brain activity.[2] Similarly, the accuracy of ECG is balanced by the practicality of PPG for continuous monitoring, enabling broader applications in daily life.[7] This combination of signals is not merely an additive increase in data volume; it represents a qualitative leap in data utility, enabling researchers to address questions that are intractable with single-modality data. This leads to the development of more robust biomarkers and effective interventions.[1]

3. Key Public Repositories for Multimodal Physiological Signals

Several prominent public repositories serve as vital resources for researchers seeking multimodal physiological datasets. These platforms play a crucial role in promoting data sharing, reusability, and collaborative research in the biomedical domain.

PhysioNet

PhysioNet is a highly established and widely recognized repository for physiological and clinical data.17 It hosts extensive collections of digital recordings and biomedical signals, encompassing cardiopulmonary and neural signals from both healthy individuals and patients with various significant public health conditions, including sudden cardiac death, congestive heart failure, epilepsy, and sleep apnea.17 The platform employs a tiered access policy—Open Access, Restricted Access, and Credentialed Access—to balance the imperative of data sharing with necessary privacy and ethical considerations.18 Notably, PhysioNet explicitly hosts relevant multimodal datasets, such as the "Multimodal Dataset for Investigating Working Memory in Presence of Music," which includes fNIRS, ECG, PPG, and other physiological signals.1 It also features specialized datasets like the "Brno University of Technology Smartphone PPG Database (BUT PPG)," which focuses on evaluating PPG signal quality.18

OpenNeuro

OpenNeuro is a free and open platform specifically designed for the validation and sharing of neuroimaging data that adheres to the Brain Imaging Data Structure (BIDS) standard.19 Its collection includes a wide array of modalities such as MRI, PET, MEG, EEG, iEEG, and NIRS. As per available information, OpenNeuro hosts a substantial number of public EEG datasets (301) and NIRS datasets (11), indicating its significant contribution to brain-related signal data.19 The platform is a designated data archive under the BRAIN Initiative, further solidifying its role in neuroscience data sharing.19 Crucially, OpenNeuro hosts specific multimodal datasets directly relevant to the user query, including ds003838, which contains EEG, ECG, PPG, and pupillometry data 20, and ds004022, featuring EEG and fNIRS.21

NIH Data Repositories

The National Institutes of Health (NIH) plays a foundational role in fostering a robust biomedical data ecosystem. The NIH differentiates between data repositories and knowledgebases, acknowledging their distinct functions and emphasizing the importance of sound data management practices for their sustainability and scientific impact.22 While the NIH provides a comprehensive list of supported data sharing resources, categorized into domain-specific and generalist repositories 17, the provided information does not explicitly detail specific multimodal physiological datasets within this broader listing. However, the mention of "Pennsieve: Impactful Multimodal Data Sharing for Epilepsy Research" as a funded project suggests the NIH's active support for initiatives involving multimodal data, even if the precise physiological modalities are not always specified in high-level descriptions.22

Pennsieve

Pennsieve is described as a scalable, cloud-based platform dedicated to scientific data management, analysis, and publication.24 A core tenet of Pennsieve is its emphasis on adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable) for data publishing, which is crucial for maximizing data utility and reproducibility.25 The platform supports a wide array of scientific file formats and modalities, enabling users to explore public datasets.24 Pennsieve also functions as a backend for other public repositories, including the NIH SPARC Portal and Epilepsy.Science.26

The increasing emphasis on "good data management practices" and the alignment with "FAIR and TRUST principles" across major funding bodies and platforms like NIH,

Pennsieve, and OpenNeuro (which is BIDS-compliant) signifies a significant and ongoing trend toward standardization and quality assurance in biomedical data sharing.[19] This concerted effort improves the overall reusability, interpretability, and reliability of publicly available datasets. For researchers, datasets adhering to these standards are generally more valuable, as they are more likely to have comprehensive metadata, clear access protocols, and well-structured data, which directly impacts the ease of assessing signal quality and understanding patterns of missing data. This also implies a future where data integration across different studies will become significantly more streamlined.

Despite the growth in the number and size of data repositories, a notable challenge persists in the efficient discoverability of specific multimodal datasets. Several queries for information on how to search for or identify prominent multimodal physiological datasets (EEG, fNIRS, ECG, PPG) often yield responses indicating that such granular information is unavailable in the high-level descriptions provided by the documents.[19] Even when multimodal data sharing is mentioned, the precise physiological modalities included are frequently not detailed in the initial descriptions. This difficulty in efficiently querying for datasets that precisely match specific needs, such as a particular combination of physiological signals, creates a bottleneck in data reuse. Researchers often must manually explore individual dataset descriptions, which is a time-consuming and inefficient process. This highlights a critical need for more sophisticated, standardized metadata tagging and advanced semantic search functionalities within these platforms to fully unlock the potential of multimodal data.

4. Detailed Analysis of Prominent Multimodal Datasets

This section provides a detailed examination of key multimodal physiological datasets, assessing their modalities, acquisition protocols, signal quality, and data completeness. To facilitate comparison, Table 1 offers a concise overview of these datasets.

Table 1: Overview of Key Multimodal Physiological Datasets

Dataset Name	Primary Modalities	Participants	Key Acquisition Parameters	Noted Signal Quality Aspects	Missing Data Info	Repository	Task/Context
Multimo	SC, ECG,	Small	Biopac,	fNIRS	Small	PhysioN	Working

dal n-back Music	PPG, fNIRS, EMG, Behavior al	sample size [1]	2 kHz samplin g rate for raw data [1]	for "high-qu ality data on hemody namic variation s" [1]	sample size noted [1]	et	memory (n-back) with music [1]
OpenNe uro ds0038 38	EEG, ECG, PPG, Pupillom etry, Behavior al	86 (initial), 65 (latest) [20]	EEG: 64-ch, 1000 Hz; ECG/PP G: aux inputs; Pupillom etry: 120 Hz [32]	Detailed acquisiti on protocol s suggest quality [32]	Participa nt count discrepa ncy [20]	OpenNe uro	Digit span task, resting state [20]
EEG-fNI RS WG Dataset (Sensor s 2024)	EEG, fNIRS	26 [33]	fNIRS: 72-ch, 10 Hz; EEG: 30-ch, 1000 Hz; simultan eous acquisiti on [33]	Preproc essing & artifact filtering improve d perform ance [34]	Balance d classes [33]	MDPI (via original pub.)	Mental state recogniti on (Word Generati on/Basel ine) [33]
fNIRS Resting State (Synthe tic HRF)	fNIRS (+ Accel, PPG)	14 per subset [16]	5-min/1 0-min resting state [16]	Provides "ground truth" for validatio n [16]	Not specifie d	Researc hGate/O penNeur o	Resting state, method validatio n [16]
OpenNe uro ds0040 22	EEG, fNIRS	7 (orthope dic impairm	EEG: 18-ch; fNIRS: raw [21]	Used for EEG denoisin g, SNR improve	Not specifie d	OpenNe uro	Motor imagery tasks [21]

		ent) [21]		ments reported [37]			
DREAMER dataset	EEG, ECG	Not specified	EEG: 14-ch, 128 Hz; ECG: 2-ch, 256 Hz [40]	High classification accuracy implies quality [40]	Not specified	PMC	Emotion recognition [40]
CAN-STRESS	Wearable Physiological, Self-reported	82 [12]	E4 wristband, full day [12]	Collected in "real-world settings" (implies noise) [12]	Not specified	arXiv	Cannabis use, stress, physiological responses [12]
OpenDriver dataset	ECG, 6-axis Motion	81 vehicles/drivers [11]	Non-intrusive ECG on steering wheel [11]	Benchmarks for ECG quality assessment, realistic noise [11]	Large scale, long-term [11]	arXiv	Real-world driving scenarios [11]

Dataset 1: PhysioNet's Multimodal n-back Music Dataset
This dataset is notable for its comprehensive collection of physiological signals, aiming to provide a rich understanding of human responses during cognitive tasks. It includes skin conductance (SC), electrocardiogram (ECG), skin surface temperature (SKT), respiration (RESP), photoplethysmography (PPG), functional near-infrared spectroscopy (fNIRS), electromyogram (EMG), de-identified facial expression scores, and behavioral metrics such as correct/incorrect responses and reaction time.1 The acquisition protocols involved recording raw physiological signals using a Biopac configuration, with a high sampling frequency of 2 kHz for the raw data.1 Data for each signal type (EDA, ECG, PPG, RESP, EMG) are organized into separate CSV folders, and crucial timing triggers for experimental blocks and trials are also provided, facilitating precise synchronization and analysis.1 The experimental context involves a working memory n-back task performed with background music, allowing for the study of cognitive load and its interaction with environmental factors.1

Regarding signal quality, the dataset's description highlights the utility of fNIRS for direct evaluation of brain responses, noting its superior spatial resolution compared to EEG for certain applications.[1] The portability and ease of use of fNIRS head caps are cited as factors contributing to the acquisition of high-quality data on hemodynamic variations.[1] The multimodal nature of the dataset is emphasized as a means to capture complementary aspects of neural activity and physiological changes, thereby offering a more comprehensive picture of brain responses.[1] While the dataset is rich in modalities, a stated limitation is its "small sample size" [1], which can affect the generalizability of findings. The provided information does not detail specific quantitative signal quality metrics (e.g., Signal-to-Noise Ratio, SNR) or explicit strategies for handling missing data within modalities.

Dataset 2: OpenNeuro's ds003838 (EEG, ECG, PPG, Pupillometry)
This dataset offers a robust collection for studying cognitive load and working memory. It comprises raw 64-channel EEG, cardiovascular data (ECG and PPG), pupillometry, and behavioral data (correctness of recall, reaction time).20 Initially collected from 86 human participants, the dataset's latest version includes data from 65 participants.20 Data acquisition took place during a 4-minute eyes-closed resting state and a classic working memory task, specifically a digit span task with serial recall.20 The EEG data were acquired using a 64-channel ActiCHamp system (Brain Products, Germany) with active electrodes, positioned according to the extended 10–20 system. The online reference was at FCz, and the ground electrode at Fpz, with impedance maintained below 25 kOm. The sampling rate for EEG was 1000 Hz, with no online digital filters applied.32 ECG and PPG signals were acquired using the same amplifier from auxiliary inputs, with specific electrode placements for ECG and PPG from the left index finger.32 Pupillometry was recorded with a Pupil Labs wearable eye-tracker at a 120 Hz sampling rate, with one-point calibration preceding each recording.32 The data is stored in BIDS format, with EEG and ECG in EEGLAB (.set) format and pupillometry in.tsv format.32
While explicit quantitative signal quality metrics for all modalities are not provided in the snippets for this specific dataset, the detailed acquisition parameters, such as maintaining low impedance for EEG electrodes, indicate a commitment to high-quality data collection.[32] The dataset's utility as a resource for the BrainBeats toolbox, which emphasizes signal visualization at various processing steps, further suggests a focus on signal integrity.[9] The reduction in participant count from 86 to 65 between initial publication and later versions could suggest data exclusion due to quality issues or re-curation, though the specific reasons are not detailed in the provided information.[20] No explicit mention of missing data points or trials within the provided snippets for ds003838 is present, but the broader literature acknowledges the "missing modality problem" in multimodal learning.[14]

Dataset 3: EEG-fNIRS WG Dataset (Sensors 2024)

This dataset is an open-access resource featuring simultaneous recordings of EEG and fNIRS signals.33 It is specifically designed for mental state recognition tasks, including Word Generation (WG) and Baseline (BL) trials.33 The dataset includes data from 26 healthy subjects, with each participant completing 60 trials across three sessions.33 The fNIRS data were captured using 72 channels at a sampling rate of 10 Hz, while the EEG data were recorded from 30 channels at a higher sampling rate of 1000 Hz.33 A key aspect of its acquisition protocol is that both fNIRS optodes and EEG electrodes were mounted on the same cap, ensuring precise spatial alignment and facilitating simultaneous data acquisition.33 Regarding signal quality, a related publication indicates that "fNIRS signal preprocessing and artifact noise filtering" were implemented, which significantly improved performance.[34] This suggests that the raw data likely contained artifacts requiring careful handling. Review papers on EEG-fNIRS systems also note that these techniques can be "de-artifacted," implying inherent noise characteristics that demand attention.[4] The dataset has been utilized to demonstrate that multimodal fusion can enhance classification accuracy and versatility compared to single-modality approaches.[3] The experimental design is balanced between the two task classes (WG and BL trials).[33] No explicit details on missing data points or trials within this specific dataset are provided in the snippets; the emphasis is on the successful simultaneous acquisition and the benefits derived from fusing the modalities.

Other Relevant Multimodal Datasets

Beyond these detailed examples, several other datasets contribute significantly to the multimodal physiological data landscape:

- **Open Access Multimodal fNIRS Resting State Dataset With and Without Synthetic Hemodynamic Responses** [16]: This dataset is primarily fNIRS, but includes additional physiological signals like accelerometer or PPG.[16] A crucial feature for signal quality assessment is its provision of "realistic fNIRS ground truth data by modeling a hemodynamic response function (HRF) on top of real resting state data".[16] This allows for objective validation of noise removal and signal processing methods. It includes 5-minute and 10-minute resting state data from 14 participants each.[16] The explicit inclusion of "ground truth" or "reference" signals in this and other datasets (e.g., the PhysioNet EEG/fNIRS motion artifact dataset utilizing "reference ground truth" signals from an unimpacted channel [47]) highlights a critical methodological development in signal quality assessment. This growing recognition of the need for a known clean signal to compare against allows for objective evaluation and benchmarking of noise reduction techniques and overall signal fidelity. Without such a comparison, assessing the true quality of a noisy physiological signal or the effectiveness of a denoising algorithm remains subjective and challenging. Therefore, datasets that explicitly include or

simulate "ground truth" are particularly valuable for method development, validation, and establishing reliable performance metrics. This indicates a shift towards more rigorous and quantifiable approaches to signal processing.

- **Multimodal EEG and fNIRS Biosignal Acquisition during Motor Imagery Tasks in Patients with Orthopedic Impairment (OpenNeuro ds004022)** [4]: This dataset contains raw 18-channel EEG and fNIRS signals from 7 participants with orthopedic impairment during motor imagery tasks.[21] It is specifically referenced in studies on EEG denoising using adversarial learning (GANs), with reported SNR improvements (up to 14.47 dB).[37] This suggests that the dataset likely contains significant noise or artifacts, making it a valuable resource for testing and developing robust denoising algorithms.
- **DREAMER dataset** [40]: This is a multimodal physiological signal dataset designed for emotion recognition research, incorporating EEG (14 electrodes, 128 Hz sampling rate) and ECG (2-channel, 256 Hz sampling rate) data.[40] Its use in validating deep learning models, which achieved high classification accuracy (e.g., 95.95% for the 'value' dimension), implies sufficient signal quality for robust feature extraction.[40]
- **CAN-STRESS** [12]: This dataset comprises multimodal physiological data collected via E4 wearable wristbands, combined with self-reported questionnaires, from 82 participants over a full day of daily activities.[12] Its collection in "real-world settings" inherently implies higher noise levels, yet it is described as a large resource suitable for developing advanced signal processing algorithms.[12]
- **OpenDriver dataset** [11]: This large-scale dataset includes six-axis motion data and ECG signals collected using non-intrusive methods (ECG sensors on the steering wheel cover) in real-world driving scenarios.[11] It addresses limitations of existing datasets, such as poor signal quality and intrusive measurement methods, and provides benchmarks for ECG signal quality assessment, including a noisy augmentation strategy for realistic noise simulation.[11] This dataset is also notable for its extensive sample size and long-term data collection from 81 vehicles and 81 drivers, addressing limitations of small sample sizes and short data collection periods that can hinder generalizability due to inter-individual variability in physiological signals.[11]

The consistent identification of motion artifacts as a significant challenge across all physiological modalities, particularly when data is collected in less controlled environments or with wearable devices, highlights a pervasive factor affecting signal quality. PPG is notably susceptible to movement and environmental factors [7], fNIRS signals can be obscured by motion and breathing [16], and even EEG is prone to muscle activity and motion artifacts.[2] This challenge is amplified by the growing desire for

more ecologically valid data, which inherently introduces more movement and environmental interference. Thus, the need for robust artifact identification and correction is a primary hurdle for data utility.[2] Datasets that explicitly include or simulate motion artifacts, or provide "ground truth" for artifact removal, are particularly valuable for developing robust and generalizable preprocessing techniques that can make real-world data usable.

A notable trade-off exists between the richness and depth of multimodal data (i.e., collecting many different signals) and the scale or completeness of the datasets (number of participants, duration of recording, absence of missing data). Collecting comprehensive multimodal data is often resource-intensive, which contributes to smaller sample sizes or shorter recording durations.[1] This, in turn, creates challenges for the generalizability and robustness of models trained on such data, as physiological signals exhibit significant inter-individual variability.[11] The "missing modality problem" is a recognized challenge in multimodal learning, occurring due to factors such as sensor limitations, cost constraints, privacy concerns, or data loss during collection or transmission.[14] This problem necessitates the development of specific computational approaches that can inherently handle incomplete data, rather than simply discarding incomplete samples. This implies that researchers must either seek larger, more diverse, and complete datasets (which are rare) or focus on developing advanced methods to handle the inherent incompleteness that often characterizes real-world multimodal data.

5. Comparative Observations: Signal Quality and Data Completeness Across Modalities and Datasets

The review of publicly available multimodal physiological datasets reveals several overarching observations regarding signal quality and data completeness. These observations highlight both inherent challenges and strategic approaches within the field.

Inherent Challenges of Multimodal Physiological Signal Acquisition

A primary challenge across all modalities (EEG, ECG, PPG, fNIRS) is the pervasive presence of motion artifacts. PPG, for instance, is highly susceptible to corruption by movements and environmental factors.7 Similarly, fNIRS signals can be masked by physiological noise such as motion and breathing 16, and EEG, despite its high temporal resolution, is vulnerable to muscle activity and general motion artifacts.2 This problem is particularly pronounced when data is collected in less controlled, real-world environments or using wearable devices.10 Beyond motion, other physiological processes introduce confounding noise. For fNIRS, superficial (scalp) blood flow and low-frequency oscillations, like Mayer waves, are significant confounds.[16] ECG and PPG signals can also be affected by ambient

light interference and pressure variations.[8] These intrinsic noise sources necessitate meticulous preprocessing.

A consistent theme in multimodal brain imaging is the complementary nature of EEG and fNIRS. EEG offers superior temporal resolution, capturing millisecond-level changes in electrical brain activity, but its spatial resolution is limited, making precise localization of brain regions challenging. Conversely, fNIRS provides better spatial localization of hemodynamic changes but exhibits a slower temporal response, with a typical 1-2 second delay and a peak response 4-6 seconds after a stimulus.[2] This inherent trade-off is the primary driver for their combined use, as their strengths compensate for each other's weaknesses.

For cardiovascular monitoring, a trade-off exists between practicality and accuracy, particularly when comparing ECG and PPG. ECG remains the clinical gold standard for measuring cardiac electrical activity, offering high accuracy. However, its requirement for subjects to remain stationary makes it impractical for continuous, daily monitoring. PPG, being non-invasive and easily integrated into wearables, offers the advantage of continuous monitoring but is more prone to noise, complicating the accurate inference of ECG-like waveforms.[7]

Approaches to Signal Quality Enhancement and Artifact Removal
To address these challenges, various strategies are employed to enhance signal quality and remove artifacts. Standard preprocessing pipelines typically involve steps such as removing power line noise (e.g., using a 50 Hz notch filter) and correcting for baseline drift, which is crucial for ensuring local stationarity of the mean value in physiological responses.5 Multimodal fusion itself can contribute to denoising and artifact correction. For example, simultaneous fNIRS recording can assist in identifying and correcting artifacts in EEG data.[2] The combination of EEG and fNIRS signals has the potential to significantly compensate for each other's limitations, thereby improving the overall signal-to-noise ratio.[15]

Advanced denoising techniques are also a significant area of research. Machine learning methods, such as adversarial learning (GANs), are being explored for EEG denoising, with studies demonstrating that models like WGAN-GP can substantially improve EEG signal fidelity and achieve higher SNRs.[37]

The presence of "ground truth" or "reference" signals within datasets is increasingly recognized as critical for objectively evaluating the effectiveness of denoising and preprocessing techniques. Datasets that provide such references, like the "Open Access Multimodal fNIRS Resting State Dataset" with its synthetic HRF ground truth or the PhysioNet EEG/fNIRS motion artifact dataset with its "reference ground truth"

channel, are invaluable for validating new algorithms and establishing reliable performance metrics.[16]

Strategies for Managing Missing Data in Multimodal Contexts

Data completeness is another significant aspect. The "missing modality problem" is a recognized and substantial challenge in multimodal learning. Data modalities can be absent due to various factors, including sensor limitations, cost constraints, privacy concerns, or simple data loss during collection or transmission.14 This highlights that the presence of missing data in multimodal physiological datasets is not just a minor inconvenience but a fundamental characteristic inherent to their real-world acquisition. This necessitates a paradigm shift in how researchers approach data analysis: instead of solely focusing on data cleaning to achieve complete datasets (which may not be feasible or desirable), there is a critical need for computational approaches that can inherently and robustly handle incomplete data. This implies that future research and dataset design should explicitly account for the "missing modality problem," making datasets with realistic patterns of missing data particularly valuable for developing and testing these robust, real-world-ready models. The field is actively developing specific computational approaches, known as Multimodal Learning with Missing Modality (MLMM) techniques, to ensure model robustness even when some modalities are unavailable during training or testing.[14] While simply removing missing-modality samples is a common preprocessing strategy, it can lead to significant data loss and reduced generalizability.[14]

Furthermore, the impact of sample size and recording duration on generalizability is a crucial consideration. Many existing physiological datasets suffer from "small sample sizes" and "short data collection periods".[11] This limitation is critical because inter-individual variability in physiological signals can significantly impact the generalizability and robustness of algorithms trained on such limited data.[11]

Benchmarking Frameworks and Their Role in Dataset Evaluation

Benchmarking frameworks play a crucial role in standardizing evaluation and driving improvements in data quality. The BenchNIRS framework, for example, is an open-source tool specifically for fNIRS data that aims to establish best practices for machine learning methodology. It utilizes five open-access datasets and robust techniques like nested cross-validation to optimize and evaluate models without bias, providing standardized metrics for comparison.6

Another significant initiative is CLIMB (Clinical Large-scale Integrative Multimodal Benchmark), which unifies diverse clinical data, including EEG and ECG, from multiple medical institutions. A key feature of CLIMB is its novel data collection and preprocessing pipeline that standardizes data formats while importantly preserving the natural patterns of missing data.[50] This framework emphasizes multitask pretraining to improve performance across various clinical tasks, even on

understudied modalities.[50] The existence and design of such frameworks are not merely passive evaluation tools; they act as catalysts for improving the quality and utility of biomedical datasets. By providing standardized, objective means of evaluation, they encourage researchers to focus their efforts on addressing known data limitations (such as noise and missingness) in a quantifiable and comparable manner. This, in turn, drives the development of more robust signal processing techniques, more effective artifact removal algorithms, and ultimately, the creation of higher-quality, more reliable datasets. This implies that datasets included in or validated by such benchmarks are likely to be of higher intrinsic value for the research community, as their quality has been rigorously assessed against established standards.

A fundamental tension exists between the desire for high data quality and the pursuit of real-world applicability. The increasing drive to collect data in "real-world settings" [11] and with "wearable devices" [7] is a clear trend aimed at developing practical applications. However, these environments are inherently less controlled than laboratories, and signals collected in such contexts are explicitly noted to be "easily corrupted by movements" [7] or to suffer from "poor signal quality".[11] Reviews on EEG-based multimodal Human-Computer Interfaces also highlight "difficulties in signal synchronization" and "limited data availability" as challenges for real-time online systems.[3] This inherent lack of control in real-world environments directly contributes to a degradation in raw signal quality. Therefore, for multimodal physiological data to be truly useful in real-world applications, there is an urgent and continuous need for sophisticated, robust preprocessing and denoising techniques that can effectively handle this increased noise without sacrificing valuable physiological information. This suggests a continuous feedback loop where real-world data informs the development of advanced processing methods, and improved methods, in turn, enable more reliable and widespread real-world data collection.

6. Recommendations for Researchers Utilizing Multimodal Physiological Datasets

Based on the comprehensive review of publicly available multimodal physiological datasets, the following recommendations are provided for researchers seeking to leverage these valuable resources:

- **Prioritize Established Repositories and Standards:** Researchers should actively seek datasets from well-regarded repositories such as PhysioNet and OpenNeuro, which are known for their curated, high-quality data collections.[17] Furthermore, favoring datasets that adhere to established data sharing principles like FAIR (Findable, Accessible, Interoperable, Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, Technology), as well as

specific data structures like BIDS (Brain Imaging Data Structure), is crucial. These standards are instrumental in promoting data quality, enhancing reusability, and ensuring comprehensive metadata, which are all vital for robust research.[19]

- **Thoroughly Scrutinize Acquisition Protocols and Metadata:** A critical step before utilizing any dataset is to meticulously review its acquisition parameters. This includes understanding the sampling rates (e.g., 2 kHz for Biopac data, 1000 Hz for EEG), channel configurations, and the types of sensors used.[1] Such detailed information is fundamental for comprehending the raw signal characteristics and for applying appropriate preprocessing steps. Equally important is assessing the completeness and quality of associated contextual metadata, which encompasses task descriptions, participant demographics, and any recorded behavioral data. Rich and well-structured contextual information significantly enhances the utility and interpretability of physiological signals, allowing for deeper analysis and more meaningful conclusions.[1] The utility and scientific value of raw physiological signals are profoundly enhanced by the presence of rich, contextual metadata. This metadata, including behavioral responses, task conditions, environmental factors, and self-reports, provides the necessary framework to interpret the physiological changes. This leads to a more complete understanding of the underlying biological and cognitive processes, enabling more nuanced interpretations, robust model development, and the ability to test complex hypotheses. Therefore, researchers should prioritize datasets that offer comprehensive and well-structured metadata, as this directly impacts the depth of analysis and the ability to draw meaningful, generalizable conclusions, moving beyond mere signal processing to true biomedical insight.
- **Seek Datasets with "Ground Truth" or Reference Signals:** For research specifically focused on signal processing, artifact removal, or denoising, prioritizing datasets that provide "ground truth" or "reference" signals is highly recommended. The availability of a known clean signal allows for objective evaluation and rigorous benchmarking of new algorithms, which is essential for advancing signal processing methodologies.[16]
- **Be Prepared for Artifact Management:** Researchers must acknowledge that motion artifacts and other physiological confounds are inherent challenges in physiological data, particularly in real-world or wearable data collection scenarios. Consequently, it is imperative to be prepared to implement robust artifact identification and removal techniques.[2] Datasets that explicitly address or demonstrate preprocessing steps for artifact reduction can also provide valuable guidance.
- **Account for Data Completeness and Generalizability:** Awareness of dataset sample sizes and recording durations is crucial, as these factors directly influence

the generalizability of research findings due to significant inter-individual variability in physiological signals.[1] For datasets that inherently contain missing modalities, it is advisable to explore or develop Multimodal Learning with Missing Modality (MLMM) techniques. This approach, rather than simply discarding incomplete data, maximizes data utility and enhances model robustness in real-world applications.[14]

- **Leverage Benchmarking Studies:** Consulting established benchmarking frameworks, such as BenchNIRS and CLIMB, and relevant studies that evaluate models on diverse datasets can provide invaluable insights. These resources offer a standardized perspective on dataset quality and the performance of various analytical approaches, guiding researchers toward effective methodologies and reliable data sources.[6]

7. Conclusion and Future Directions

This report has provided a critical review of publicly available multimodal biomedical datasets, focusing on EEG, ECG, PPG, and fNIRS, with an assessment of their signal quality and data completeness. Key repositories like PhysioNet and OpenNeuro, alongside initiatives from NIH and platforms such as Pennsieve, were identified as crucial resources.

The analysis of specific datasets underscored the intrinsic challenges in multimodal physiological data acquisition, predominantly characterized by pervasive motion artifacts and physiological confounds. The complementary strengths of modalities—EEG providing high temporal resolution and fNIRS offering better spatial resolution—drive their combined use, while the practicality of PPG for continuous monitoring often comes with increased susceptibility to noise. The critical need for "ground truth" data to objectively evaluate signal quality and denoising techniques was also highlighted. Furthermore, the "missing modality problem" and the limitations imposed by small sample sizes in many datasets present significant hurdles to data completeness and generalizability.

Despite these challenges, the field is actively progressing. Benchmarking frameworks like BenchNIRS and CLIMB are emerging to standardize evaluation and drive continuous improvements in data quality and processing methodologies. The increasing focus on collecting data in real-world settings, while introducing more noise, simultaneously propels the development of more robust analytical tools capable of handling such complexities.

The accelerating trend in biomedical research towards increasing the ecological validity of data collection, by moving from highly controlled laboratory settings to

"real-world driving scenarios" or "naturalistic conditions," is a positive development for translational research.[4] However, this shift inherently leads to an increase in data complexity, noise, and potential for missingness, as real-world environments are far less controlled than laboratory settings. Therefore, a critical future direction is to balance the scientific rigor achievable in controlled experiments with the practical utility of real-world data. This necessitates a continuous feedback loop where challenges encountered in real-world data inform the development of more robust computational methods, and these improved methods, in turn, enable more reliable and widespread real-world data collection and analysis. This suggests that the future of multimodal physiological data research lies in embracing, rather than avoiding, the complexities of naturalistic environments.

Future Directions:

- **Larger, More Diverse, and Longitudinal Datasets:** There is a pressing need for the collection and public release of larger, more diverse, and longitudinally acquired multimodal datasets, particularly those captured in naturalistic, real-world settings. This will be instrumental in addressing issues of generalizability and inter-individual variability, which are significant limitations in many current datasets.
- **Advanced AI/ML for Imperfect Data:** Continued research and development of sophisticated artificial intelligence and machine learning techniques are paramount. These techniques must be specifically designed to robustly handle noisy, artifact-ridden, and inherently incomplete multimodal data, including advanced Multimodal Learning with Missing Modality (MLMM) techniques and more effective denoising algorithms.
- **Enhanced Data Interoperability and Standardization:** Further adoption and rigorous enforcement of standardized data formats (e.g., BIDS) and data sharing principles (FAIR, TRUST) will be crucial. This will facilitate seamless data reuse, enable more effective integration across disparate studies, and promote more efficient collaborative research efforts.
- **Integration of Broader Physiological and Contextual Signals:** Future datasets should aim to integrate an even wider array of physiological signals, environmental factors, and rich behavioral or self-reported data. This comprehensive approach will provide a truly holistic and ecologically valid understanding of human states and conditions, moving beyond isolated physiological measurements to a more integrated view of human health and behavior.

Works cited

1. A Multimodal Dataset for Investigating Working Memory in Presence of Music - PhysioNet, accessed June 8, 2025, https://physionet.org/content/multimodal-nback-music/
2. Fusion of fNIRS and EEG: a step further in brain activity research - Bitbrain, accessed June 8, 2025, https://www.bitbrain.com/blog/fusion-fnirs-eeg-brain-activity-research
3. A review of hybrid EEG-based multimodal human–computer interfaces using deep learning: applications, advances, and challenges | Request PDF - ResearchGate, accessed June 8, 2025, https://www.researchgate.net/publication/390118310_A_review_of_hybrid_EEG-based_multimodal_human-computer_interfaces_using_deep_learning_applications_advances_and_challenges
4. Strategic Integration: A Cross-Disciplinary Review of the fNIRS-EEG Dual-Modality Imaging System for Delivering Multimodal Neuroimaging to Applications - MDPI, accessed June 8, 2025, https://www.mdpi.com/2076-3425/14/10/1022
5. Full article: Emerging Neuroimaging Approach of Hybrid EEG-fNIRS Recordings: Data Collection and Analysis Challenges, accessed June 8, 2025, https://www.tandfonline.com/doi/full/10.1080/26941899.2024.2426785
6. Benchmarking framework for machine learning ... - Frontiers, accessed June 8, 2025, https://www.frontiersin.org/journals/neuroergonomics/articles/10.3389/fnrgo.2023.994969/full
7. Improving Atrial Fibrillation Detection Using a Shared Latent Space for ECG and PPG Signals - Harvard Data Science Review, accessed June 8, 2025, https://hdsr.mitpress.mit.edu/pub/vifgdvlv
8. Inferring ECG Waveforms from PPG Signals with a Modified U-Net Neural Network - MDPI, accessed June 8, 2025, https://www.mdpi.com/1424-8220/24/18/6046
9. BrainBeats as an Open-Source EEGLAB Plugin to Jointly Analyze EEG and Cardiovascular Signals - JoVE, accessed June 8, 2025, https://app-jove-com.remotexs.ntu.edu.sg/t/65829/brainbeats-as-an-open-source-eeglab-plugin-to-jointly-analyze-eeg
10. Translating Emotions to Annotations: A Participant's Perspective of Physiological Emotion Data Collection - arXiv, accessed June 8, 2025, https://arxiv.org/html/2503.19636
11. OpenDriver: An Open-Road Driver State Detection Dataset - arXiv, accessed June 8, 2025, https://arxiv.org/html/2304.04203v2
12. CAN-STRESS: A Real-World Multimodal Dataset for Understanding Cannabis Use, Stress, and Physiological Responses - arXiv, accessed June 8, 2025, https://arxiv.org/html/2503.19935v1
13. Multimodal Dataset Construction and Validation for Driving-Related Anger: A Wearable Physiological Conduction and Vehicle Driving Data Approach - MDPI,

accessed June 8, 2025, https://www.mdpi.com/2079-9292/13/19/3904
14. Deep Multimodal Learning with Missing Modality: A Survey - arXiv, accessed June 8, 2025, https://arxiv.org/html/2409.07825v3
15. Bimodal EEG-fNIRS in Neuroergonomics. Current Evidence and Prospects for Future Research - PubMed Central, accessed June 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10790898/
16. Open Access Multimodal fNIRS Resting State Dataset ... - Frontiers, accessed June 8, 2025, https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2020.579353/full
17. AHA Approved Data Repositories - Professional Heart Daily ..., accessed June 8, 2025, https://professional.heart.org/en/research-programs/awardee-resources/aha-approved-data-repositories
18. Databases - PhysioNet, accessed June 8, 2025, https://physionet.org/about/database/
19. OpenNeuro, accessed June 8, 2025, https://openneuro.org/
20. EEG, pupillometry, ECG and photoplethysmography, and behavioral ..., accessed June 8, 2025, https://openneuro.org/datasets/ds003838
21. Multimodal EEG and fNIRS Biosignal Acquisition during Motor Imagery Tasks in Patients with Orthopedic Impairment - OpenNeuro, accessed June 8, 2025, https://openneuro.org/datasets/ds004022/versions/1.0.0
22. Biomedical Data Repositories and Knowledgebases | Data Science ..., accessed June 8, 2025, https://datascience.nih.gov/data-ecosystem/biomedical-data-repositories-and-knowledgebases
23. Open Domain-Specific Data Sharing Repositories, accessed June 8, 2025, https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
24. Pennsieve - Public and Private Data Management and Analysis, accessed June 8, 2025, https://discover.pennsieve.io/
25. Pennsieve, accessed June 8, 2025, https://sparc.science/resources/2j9IC0YFI5P34wGIkJOb49
26. Accessing previous versions of public datasets - Pennsieve, accessed June 8, 2025, https://docs.pennsieve.io/docs/accessing-previous-versions-of-public-datasets
27. Pennsieve Discover - Find and access public scientific datasets, accessed June 8, 2025, https://discover.pennsieve.io/datasets
28. Pennsieve Discover - Re3data.org, accessed June 8, 2025, https://www.re3data.org/repository/r3d100013148
29. Datasets | Computational Medicine Laboratory, accessed June 8, 2025, https://wp.nyu.edu/cml/datasets/
30. EEG, pupillometry, ECG and photoplethysmography, and behavioral data in the digit span task and rest - OpenNeuro, accessed June 8, 2025, https://openneuro.org/datasets/ds003838/versions/1.0.6/metadata
31. Pupillometry and electroencephalography in the digit span task - bioRxiv,

accessed June 8, 2025, https://www.biorxiv.org/content/10.1101/2021.10.21.465288v2.full-text
32. (PDF) Pupillometry and electroencephalography in the digit span task, accessed June 8, 2025, https://www.researchgate.net/publication/361369142_Pupillometry_and_electroencephalography_in_the_digit_span_task
33. EF-Net: Mental State Recognition by Analyzing Multimodal EEG ..., accessed June 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10974548/
34. Multimodal Model for Automated Pain Assessment: Leveraging Video and fNIRS, accessed June 8, 2025, https://www.researchgate.net/publication/391512202_Multimodal_Model_for_Automated_Pain_Assessment_Leveraging_Video_and_fNIRS
35. EF-Net: Mental State Recognition by Analyzing Multimodal EEG-fNIRS via CNN - GitHub, accessed June 8, 2025, https://github.com/DL4mHealth/EF-Net
36. Adversarial denoising of EEG signals: a comparative analysis of standard GAN and WGAN-GP approaches - PMC, accessed June 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12089060/
37. Adversarial denoising of EEG signals: a comparative analysis of standard GAN and WGAN-GP approaches - Frontiers, accessed June 8, 2025, https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2025.1583342/full
38. Adversarial denoising of EEG signals: a comparative analysis of standard GAN and WGAN-GP approaches - Frontiers, accessed June 8, 2025, https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2025.1583342/pdf
39. Adversarial denoising of EEG signals: a comparative analysis of standard GAN and WGAN-GP approaches - ResearchGate, accessed June 8, 2025, https://www.researchgate.net/publication/391880720_Adversarial_denoising_of_EEG_signals_a_comparative_analysis_of_standard_GAN_and_WGAN-GP_approaches
40. Emotion recognition based on multimodal physiological electrical signals - PMC, accessed June 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11919864/
41. BrainBeats, an Open-Source EEGLAB Plugin to Jointly Analyze EEG and Cardiovascular Signals - bioRxiv, accessed June 8, 2025, https://www.biorxiv.org/content/10.1101/2023.06.01.543272v3.full.pdf
42. A simultaneous EEG-fNIRS dataset of the visual cognitive motivation study in healthy adults, accessed June 8, 2025, https://pubmed.ncbi.nlm.nih.gov/38533112/
43. EF-Net: Mental State Recognition by Analyzing Multimodal EEG-fNIRS via CNN - MDPI, accessed June 8, 2025, https://www.mdpi.com/1424-8220/24/6/1889
44. Simultaneous EEG-fNIRS Data Classification Through Selective Channel Representation and Spectrogram Imaging - PubMed Central, accessed June 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11379445/
45. (PDF) fNIRS Dataset During Complex Scene Analysis - ResearchGate, accessed June 8, 2025, https://www.researchgate.net/publication/377651283_fNIRS_Dataset_During_Com

plex_Scene_Analysis
46. Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset - Bohrium, accessed June 8, 2025, https://bohrium.dp.tech/paper/arxiv/8130594684713041 92
47. Motion Artifacts Correction from Single-Channel EEG and fNIRS ..., accessed June 8, 2025, https://www.mdpi.com/1424-8220/22/9/3169
48. Early-stage fusion of EEG and fNIRS improves classification of motor imagery - PMC, accessed June 8, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC9869134/
49. www.nature.com, accessed June 8, 2025, https://www.nature.com/articles/s41597-022-01414-2
50. Data Foundations for Large Scale Multimodal Clinical Foundation Models - arXiv, accessed June 8, 2025, https://arxiv.org/html/2503.07667v1