This thesis by Dylan Green is accepted in its present form by the
Biomedical Engineering Program as satisfying the
thesis requirements for the degree of Master of Science

Date 04/30/2026 Marissa Gray

Dr. Marissa Gray, Advisor

Date 04/30/2026 Andrea Webb

Dr. Andrea Webb, Advisor

Date 4/30/2026 Christopher D. Hughes, PhD

Dr. Christopher Hughes, Reader

Approved by the Graduate Council

Date_____ _____

Professor David P. Lindstrom, Dean of the Graduate School

Multi-Limb Sensor Fusion for Actigraphy-Based Sleep/Wake Classification

By

Dylan Green, Draper Scholar

B.S., United States Military Academy, 2024

Thesis

Submitted in partial fulfillment of the requirements for the
Degree of Master of Science in Computer Science at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2026

Multi-Limb Sensor Fusion for Actigraphy-Based Sleep/Wake Classification

Dylan Green, Marissa Gray, Andrea Webb, Jeff Huang

Contributing authors: dylan_green@brown.edu; marissa_gray@brown.edu; awebb@draper.com ; jeff_huang@brown.edu ;

Abstract

Actigraphy-based sleep monitoring has gained popularity due to its affordability and unobtrusive nature compared to polysomnography (PSG), the clinical gold standard. However, existing single-sensor actigraphy methods suffer from a fundamental class imbalance; since the dominant class, sleep, constitutes **80 − 90%** of a night of study, algorithms that optimize for sleep detection accuracy consistently misclassify the minority wake class, inflating estimates of total sleep time while under-reporting wakefulness. We directly address this challenge with adaptations of the Cole-Kripke and Sadeh sleep detection algorithms, modified to support data fusion at the data, score, and decision levels from four Axivity AX6 actigraphy sensors, one on each wrist and ankle. Rather than optimizing for sleep detection, we report $\boldsymbol{F_1}$, Cohen's Kappa, and MCC, evaluating our algorithms' ability to detect both sleep and wake accurately. Validated against simultaneous PSG at the Brown University Health Sleep Lab, our results indicate that multi-sensor fusion eliminates placement-dependent $\boldsymbol{F_1}$ variance: the leading single-sensor algorithm had a **31%** relative $\boldsymbol{F_1}$ difference between wrists, a risk which every multi-sensor approach mitigates. As a pilot study with four participants, we identified trends rather than specific benchmarks: higher-level wake-biased fusion strategies consistently outperformed lower-level fusion, thresholds from the seminal algorithms need to be adjusted for multi-sensor approaches, and Cole-Kripke adaptations consistently outperformed Sadeh. These findings establish multi-sensor actigraphy as a robust, low-cost approach to wakefulness detection, with further implications for operational contexts requiring simple, cheap monitoring hardware.

Keywords: Actigraphy, Total Sleep Time, Wake After Sleep Onset, Sleep Onset Latency, Sensor Fusion, Class Imbalance, Polysomnography

1 Introduction

The importance of high-quality sleep for physical and mental health has been increasingly recognized over the past decade. In response, companies such as Apple, Amazon, Fitbit, and Whoop have developed wearable, 'nearable,' and 'airable' devices to monitor sleep. These devices provide insights into sleep duration, quality, and stages, but their accuracy depends on the sensors and their placement [1].

A common method employed in these devices is actigraphy, a non-invasive technique that estimates sleep stages by measuring limb movement through triaxial accelerometers. Previous actigraphy-based studies have estimated total sleep time with high reported accuracy rates between 88-90.7% depending on the algorithm used [2]. Though the accuracy seems high, upon further inspection, these algorithms tend to underestimate the amount of time it takes to fall asleep and underestimate the number of awakenings. This is because sleep is an overwhelmingly dominant class; an algorithm could predict sleep for every epoch and immediately have a relatively high accuracy (approximately 85-95%). The "Class Imbalance Problem" (CI) [3] is well-documented in machine learning, but interestingly has not formally entered the conversation of sleep classification; CI is a traditional problem in binary classification where models are biased towards the majority class, ultimately resulting in misclassifications of the minority or anomalous class. By optimizing sleep identification algorithms to predict sleep, the field has obfuscated the real area for improvement, wakefulness identification, which remains around 32% [4]. Our research team proposes that the critically necessary task in sleep identification relies on inverting this issue—accurately classifying wakefulness.

However, classification inversion alone does not present a way forward. Instead, we suggest the novel approach of using four actigraphy sensors, one on each limb, rather than the traditional approach of using one sensor, usually worn on the wrist. Studies have shown that fusing data streams from multiple sensors [5] dramatically reduces the mean average error from noise and anomalies when relying on an individual signal. Furthermore, using additional sensors with machine learning models [6] outperformed single sensors with the same model on all key metrics.

Acquiring even one research-grade sensor, such as the Empatica E4 or ActiGraph GT3X+ can be prohibitively expensive (with the Empatica E4 approaching $1700 and ActiGraph devices and software generally ranging from $600 to $2000) [7–9]. This is largely because current sleep trackers combine actigraphy with advanced sensors and complex proprietary software that transforms raw data into meaningful sleep metrics. By deliberately picking a less complex sensor without proprietary software, the Axivity AX6 (costing £199.00 or $252), we validate the efficacy of our algorithms while reducing cost. The core goal of this paper is to prove the efficacy of multi-sensor actigraphy algorithms, with the intention of supporting future, cheaper research and use cases.

The constraint of relying on cheaper actigraphy-only sensors stems not only from a desire to drive the cost as low as possible. This research aims to provide military operators in austere mission contexts with insight into their readiness through an accurate picture of their total sleep. In operational environments where sensors are lost, damaged, and can be potentially captured by adversarial forces, it is critical

that the hardware is cheap enough to be disposable and does not reveal any sensitive information about the original wearer. While our current proof of concept uses more expensive clinical sensors, following the framework of Intille et al. [10] could drive the cost per sensor as low as $63 in the future to fit this mission. Raw accelerometry data contains no positional, biological, or mission-relevant data, and traditional algorithms are efficient and reduce computational requirements compared to ML approaches. Every design choice in this thesis was made with this end state in mind, though building the hardware solution extends past the current scope.

Our study introduces a novel approach that addresses previous challenges in three key ways:

1. We attack the traditional class imbalance problem by optimizing for wakefulness identification rather than correctly identifying sleep, a significant ideological difference from previous papers.
2. We use a multi-sensor approach to capture whole-body positional transitions that single sensors miss, and fuse data to reduce the average error in single sensor signals.
3. We propose using a cost-effective and operationally safe actigraphy-based solution with open source resources (Neishabouri's Pipeline) to convert raw accelerometry data into epoch-level activity counts for sleep classification.

We hypothesize that fusing data from four limb-mounted sensors with our multi-sensor approach will allow us to improve epoch-level wake classification compared to single sensor approaches when validated against simultaneously recorded Polysomnography. This approach leverages the affordability and accessibility of actigraphy while potentially reducing the need for more expensive, complex sensors typically used in commercial wearables. Our long-term goal is to apply our results in a military context, where irregular sleep patterns are standard; a cost-effective, easily integrated method to quantify sleep could enhance operational readiness by providing personnel with accurate sleep data to inform planning.

2 Background & Related Works

To understand the novelty of our approach, we first explore existing literature around the fundamental sleep architecture, shift towards more actionable quantitative metrics, and describe the existing methodologies commonly used to track them.

2.1 Sleep Architecture and the Actionability Problem

Historically, sleep research has focused heavily on mapping the physiological architecture of sleep, which is divided into distinct stages: N1, a transitional phase between wakefulness and sleep, N2 deeper sleep which lasts roughly 45% of the night, N3, critically important restorative deep sleep, and Rapid Eye Movement (REM) sleep, which is similarly important for memory consolidation [11]. Because REM and deep sleep are heavily impacted by sleep deprivation—leading to well-documented declines in psychomotor vigilance, emotional stability, and cognitive function [12, 13], both clinical studies and modern consumer sleep trackers (CSTs) have become fixated on estimating the exact duration a user spends in each specific stage.

However, for the average user, or in high-stress operational environments where maintaining readiness is critical, granular sleep stage data is often unactionable. While it is true that a healthy adult requires four to five full sleep cycles per night [11], an individual cannot consciously force their brain into deep or REM sleep. Providing users with complex stage breakdowns often overwhelms them with data they do not know how to meaningfully address, resulting in stress or guilt regarding their "sleep performance," rather than providing clear guidance on how to adjust their schedules [14].

Instead, the most practical and reliable proxy for tracking the total time spent in each sleep phase is to ensure adequate Total Sleep Time (TST). The American Academy of Sleep Medicine indicates that adults should obtain at least seven hours of sleep per night [15]. Barring underlying medical conditions, if an individual achieves a TST of 7 to 9 hours, the required distribution of REM and deep sleep naturally follows. Therefore, our research team posits that generating a highly accurate TST, which depends on reliable detection of wakefulness, is arguably the most valuable function a sleep-tracking system can offer to inform daily planning and behavioral change.

2.2 Quantifying Total Sleep Time

Total Sleep Time (TST) is a simple metric that refers to the total duration of sleep achieved in a given sleep period, from sleep onset to awakening. This inherently includes the time spent in all sleep stages (N1-N3 and REM Sleep) [16]. While this metric may sound trivial, there are complexities that make it more challenging to track — namely, still wakefulness before falling asleep and wake-ups throughout the period of sleep.

Sleep Onset Latency (SOL) and Wake After Sleep Onset (WASO) are two additional metrics that precisely address this complexity. Sleep onset latency is the amount of time an individual takes to fall asleep after the night of sleep recording has begun. Wake after sleep onset represents the amount of time an individual spends awake after the determined sleep onset [16]. Other less studied metrics, such as Wake Time After Sleep Offset (WASF), have also been explored, and is the amount of time the individual spends awake after the determined point of sleep offset.

Some studies have linked sleep onset latency as a higher predictor of subjective sleep quality than even total sleep time [17]. Sleep onset latency, however, is one of the metrics that is hardest to track, specifically with consumer sleep trackers, due to the tendency for wrist-worn sensors to misclassify lying still while awake as being asleep. The incongruence between the subjective quality of sleep an individual experiences and the quality their devices report can be particularly frustrating and make it challenging to understand how to improve one's sleep. It simultaneously breeds distrust between the user and their tracking methodology [14].

Similarly, WASO is frequently under-reported on many actigraphy-based devices, as still-wakefulness after sleep onset is still classified as sleep. One such study indicated that only around two-thirds of WASO was correctly identified by actigraphy sensors when compared to PSG [4]. This represents another form of incongruency that makes understanding the sleep scores granted by consumer sleep tracking devices challenging for users.

While total sleep time does not paint a complete picture, as it does not provide an in-depth look into the individual's sleep architecture, it serves as an excellent baseline with which individuals can directly change their habits. Coupled with accurate metrics of SOL and WASO, individuals are provided with a clear, single value to optimize. The simplicity of this metric makes it uniquely applicable both for general consumers and high-stress austere environments, which demand actionable information.

2.3 What Has Been Tried So Far?

Sleep studies have been conducted for decades, with seminal papers such as those of Cole et al. and Sadeh et al. being written in the early nineties [18, 19]. We first discuss the clinical gold standard, polysomnography, before evaluating the various actigraphy-based studies that have been conducted over the years of research.

2.3.1 Polysomnography

Polysomnography (PSG) remains the gold standard for sleep studies, combining electroencephalogram (EEG), electromyography (EMG), electro-oculogram (EOG), electrocardiogram (ECG), and respiratory sensors [20] to paint a high-fidelity image of an individual's night of sleep. Experts are trained via the AASM Manual for the Scoring of Sleep and Associated Events to classify the different stages of sleep based on the readings from the different sensor feeds. This allows them not only to classify the different stages of sleep (N1-N3 and REM) but also collect a variety of other metrics such as TST, SOL, WASO, and sleep efficiency [20]. This is the most invasive methodology, requiring a full lab facility and trained specialists who are able to perform the appropriate sleep classification. This remains, by far, the most reliable form of sleep classification and is used as a baseline for most sleep studies.

2.4 Actigraphy-based Sleep Studies

Polysomnography, though reliable, requires medical professionals, a fully established laboratory environment, and a significant commitment from the participant. Furthermore, the general discomfort of sleeping in a new location and the sensation of having many new sensors attached to one's body is known to negatively impact sleep quality [21]; actigraphy has been explored as a lower-cost alternative. Computer scientists and sleep specialists alike have sought to develop algorithms that can accurately identify an individual's sleep state based on the raw data points provided by these sensors.

2.4.1 Single-Sensor Actigraphy and the Challenges of Specificity

Most actigraphy-based studies utilize a single wrist-worn sensor to estimate sleep-wake patterns. Cole et al. (1992) were one of the first to introduce such an algorithm and had to engage with the most appropriate way to match their sensor frequency with the polysomnography epochs. When validated against polysomnography, the algorithm achieved an accuracy of 88.25%, establishing both the feasibility and reliability of wrist actigraphy as a non-invasive tool for sleep monitoring [18]. To achieve such accuracy,

it would classify each minute epoch as wake or sleep depending on a weighted sum of the preceding four minutes and the following two minutes surrounding any given epoch. Even when rescored to adjust for the base algorithm's tendency to classify actual wake as sleep, it was two and a half times more likely to misclassify wakefulness than it was to misclassify sleep; despite their monumental step forward in the field of actigraphy, this demonstrated the class imbalance problem clearly.

A similar study conducted by Sadeh et al. (1994) used a large 11-minute window, along with the mean activity count (the average of the raw results from the actigraph), the standard deviation of activity, the number of highly active epochs, and the natural logarithm of the activity count. Rather than deliberately rescoring to address the class imbalance problem, they use the standard deviation and observed particularly active epochs to more accurately identify wakefulness. This algorithm demonstrated between 91-93% accuracy, furthering the appeal of actigraphy-based sleep assessment as a low-cost tool. Notably, this study also introduced the idea of using multiple actigraphy sensors. In this case, they used one for redundancy purposes and to assess if limb dominance played a role in accuracy. This is fundamentally different from our approach to use multiple data streams to feed into one algorithm [19].

Despite Cole and Sadeh's success with wrist-worn actigraphy in laboratory settings, there have been more studies measuring the applicability of actigraphy in various environments, such as the field or hospital, as well as attempts to change sensor positioning, placing them on the ankle instead of the wrist. Although actigraphy demonstrated high agreement with polysomnography regardless of the environment in which the study was performed, the findings regarding ankle versus wrist placement were less conclusive. One such study found that the ankle sensor and wrist sensor functioned interchangeably when testing individuals' sleep at high altitudes. In contrast, a hospital-based study found that the ankle and wrist sensors deviated dramatically, with a discrepancy of over 10 hours across two nights of sleep [22, 23]. While the individual sensors may have inconsistent agreement, we believe the broader movement patterns across limbs may provide positional insights that no individual sensor can capture, resulting in an improved ability to identify wakefulness. While this has not explicitly been tested with multiple sensors, studies have been able to infer sleep and wakefulness from positional changes in a bed with sensors attached to it [24].

Across more recent studies, such as in Marino et al. (2013), actigraphy has continued to demonstrate a relatively high accuracy and sleep sensitivity, 86.3% and 96.5% respectively, while struggling with specificity 32% [4]. Accuracy in this context is defined as the total amount of correctly identified epochs, sleep sensitivity is the number of sleep epochs correctly identified as sleep, and wake specificity is the number of wakefulness epochs correctly identified as wake. While they acknowledge that the high overall accuracy of actigraphy algorithms largely depends on their high sleep sensitivity, they do not identify the root cause of this symptom, which is the fundamental class imbalance.

Our research team addresses this class imbalance - with sleep being the dominant class and wake being the minority class - by identifying wakefulness as our target for sensitivity (true positives). However, we recognize that this simple inversion does

not actually improve the existing algorithmic space, and we opted to overcome this challenge with a relatively unexplored methodology: multi-sensor actigraphy.

2.4.2 Multi-Sensor Actigraphy and The Path Ahead

Currently, the body of literature exploring multi-limb actigraphy tends to focus on patients with various motor disorders such as Parkinson's Disease, Huntington's Disease, or restless leg syndrome. While not specifically sleep-focused, Adams et al. (2017) [25] demonstrated the ability to accurately predict the movements of their participants (standing, lying, walking, and sitting) based on five triaxial accelerometers, one chest-mounted and one on each limb, respectively.

Rather than directly improving the accuracy of sleep classification, some studies have shown that fusing data from multiple sensors can improve the actual underlying quality of the data by reducing error. Through the use of neural networks, Fuster-Garcia et al. [5] demonstrated a 40% reduction in errors due to noise, missing data, and artifacts, from independent datastreams via fusion. This accuracy increase alone may help improve traditional classification algorithms, such as Cole-Kripke and Sadeh, by providing a dataset that more accurately reflects the individual's night of sleep.

Studies such as SleepNet [26] and LTA2V [6] have utilized self-supervised methods on single and dual-sensor data, respectively. In the LTA2V paper, it was shown that self-supervised pretraining on dual-sensor data outperforms single-sensor data on all metrics (F_1-score, specificity, sleep sensitivity, AUC-ROC). There is a clear and marked advantage in synthesizing the data from multiple sensors. That being said, the F_1 score for both algorithms only remains at around 0.75. While direct comparison of F_1 scores across datasets is not possible, it is worth noting that computing an F_1 score on the published sleep sensitivity and wake specificity values of the Cole-Kripke algorithm in Marino's 2013 paper [4] resulted in a score of about 0.92. This suggests that machine learning methods may not have surpassed the extensively validated traditional algorithms, and further supports the need to test them with multi-sensor datastreams.

To the best of our knowledge, multi-sensor algorithms with more than two sensors have yet to be studied. Our research team conducted preliminary pilot testing with this methodology, using an Apple Watch Ultra 2 as an informal baseline while waiting for institutional approval to gather polysomnography data. This pilot actigraphy data was processed into ActiGraph counts, compatible with the Cole-Kripke algorithm, using the Brønd et al. [27] aggregation method. Even a naive fusion of the four actigraph data streams outperformed a single wrist-worn sensor regarding wake classification when processed through the traditional Cole-Kripke and Sadeh algorithms. These results were in no way generalizable, but they provided sufficient motivation to pursue a more formal multi-sensor validation study.

While the cost of four Axivity AX6 sensors at $252 represents a steep scaling cost, totaling $1008 for experimental setup, this remains significantly cheaper than the equipment needed for polysomnography and comparable to or less than a single research-grade device with appropriate software licensing. Further work can drive this price substantially lower without limiting the ability to emulate our methodology.

There are clear indications that the use of multiple sensors improves positional identification, and the synthesis of multiple sensors allows for a more accurate data signal. Specifically because SOL and WASO are currently underestimated by actigraphy, we believe additional sensors may help clarify the specific bodily behavior that occurs when someone is awake but lying still. For example, an individual's wrist may be still, but their ankle is twitching, indicating wakefulness. Similarly, the movement of all four limbs at once could represent someone turning in their sleep but may not indicate wakefulness.

Existing work in the multi-sensor actigraphy space has led our research team to believe there are improvements to traditional actigraphy-based sleep algorithms that could be made by increasing the number of data streams an algorithm can base its classification upon.

3 Methodology

To test our methodology, our study followed a single-night in-lab experimental design, conducted at the Brown University Health Sleep Lab, which was approved by Brown and Brown University Health's Institutional Review Boards. Participants were individuals aged 18 to 65 who were already scheduled for an overnight polysomnography (PSG) test. At the time of their sleep lab appointment, individuals were prompted to join the study and provided written informed consent to participate in the study extension, which involved additional actigraphy sensors.

Upon arrival, participants underwent a standard PSG setup, including EEG, EOG, EMG, ECG, and respiratory sensors. After licensed technicians completed standard sensor placement, the AX6 sensors were activated in the OmGui platform and secured in place. The wrist sensors used silicone bands, while the ankles required a larger mesh anklet with an interior pocket to seat the sensor in. Sensors were placed on the outside of the wrist like a wristwatch, and ankle sensors were affixed just above the ankle bone on the outside of the ankle. The participants then continued their normal overnight PSG study, which lasted approximately 12 hours.

The following morning, the AX6 devices were removed prior to PSG sensor removal. The resulting data were immediately downloaded to a secure laboratory workstation using OmGui and exported in CSV format. We exported the raw 100 Hz triaxial accelerometer data and timestamps without conducting any resampling. Fortunately, the actigraph timestamps contained a full date time group down to the millisecond. Thus we were able to temporally align the sensors with the polysomnography data post-night of study, allowing for simple recording procedures that could be easily administered by technicians without requiring special training.

The supervising sleep physician then performed the PSG classification according to the standard outlined in the AASM manual [28]. The physician then removed all identifying information from the PSG reports and data before transferring them to the Brown University research team. No personally identifying information was retained in the dataset accessed by researchers.

3.1 Data Preprocessing

While our preliminary analysis of pilot data used the approximated aggregation pipeline by Brønd et al. (2017) [27] to generate ActiGraph counts, our formal validation against PSG data relied on the Neishabouri et al. (2022) [29] agcounts python package and its "get_counts" method. Brønd's initial pipeline was a well-approximated reverse-engineered implementation of ActiLife's proprietary count aggregation prior to its open-sourcing. Neishabouri et al. published the exact signal processing methodology, sequence, and coefficients used in the ActiLife software. This provided a stronger, simpler, and more repeatable data processing pipeline for future cross-study comparison.

The pipeline begins by parsing timestamp strings from the raw CSV files, correcting for malformed time values (e.g., when seconds equal or exceed 60), and aligning all entries relative to a shared baseline for temporal synchronization based on the recorded start-time of the PSG study. We rounded down anomalous entries to 59.999 (occasional entries would have 60 seconds with some amount of milliseconds, rather than incrementing the minute), preemptively handling malformed data before feeding it into the validated agcounts pipeline.

We then passed the get_counts function a numpy array (N, 3) of our raw accelerometry data, the frequency of our AX6 sensors, 100 Hz, and an epoch time of 60, the size both Cole-Kripke and Sadeh expect. Internally, for our 100 Hz input, the get_counts function resamples our frequency to 30 Hz by first up-sampling it by 3, performing a low-pass filter, then down-sampling by 10. The function then runs a band-pass filter using the exact coefficients used in the ActiLife software. Finally, a noise suppression dead-band filter is applied, data is downsampled to 10 Hz, and epochs are consolidated into the desired time, in our case, 60 seconds.

The initial step of aligning our actigraphy data with the PSG sensors also helps address an issue other studies have encountered: aligning epoch lengths between PSG and Actigraphy data. PSG is traditionally recorded in 30-second intervals rather than the 60-second epochs used by sleep algorithms. Gao et al. [30] encountered this issue when aligning data from an actigraphy sensor that computed epochs internally. Their actigraphy epochs frequently fell in between three PSG epochs, as such, they determined the PSG epoch as "Sleep" or "Wake" if all three agreed, otherwise they would classify it as a transitional epoch. We used raw actigraphy data, allowing us to align our sensors directly with the PSG start and end time rather than using Gao's algorithm. We followed Cheung et al.'s [31] approach in which a 60-second epoch is classified as wake if any of the composing epochs are identified as wake.

This preprocessing pipeline was the most computationally intensive step in our study, taking between fifteen to twenty minutes to process our actigraph files of over two million lines. To avoid repeated computations, we saved five total comma-separated value (CSV) files: one file for each sensor, and one combined dataframe file that includes all four sensors. This pipeline was finalized prior to receiving any live data, thus we were able to run the preprocessing step each time we received new data and cache the results, enabling rapid iteration on our algorithms without reprocessing raw sensor data each time.

3.2 Building our Algorithms

We derived sleep-wake classifications from the processed actigraphy counts, using both single-sensor and multi-sensor adaptations of the Cole-Kripke and Sadeh algorithms. These seminal algorithms are offered in proprietary suites like ActiLife [32] and are the most robustly tested and commonly used sleep classification methodologies. Our Python-based implementation follows the formulation defined in the Actigraph Sleepr repository [33], a project which contains implementations of Cole-Kripke and Sadeh directly following the guidance of the ActiGraph manual. This repository was produced in R, and we adapted the necessary functions to match our Python-based preprocessing pipeline. This allowed us to run one effective Pytest suite, which gave our research team confidence that our underlying algorithms and preprocessing remained logically sound as we iterated upon our multi-sensor algorithms. Before discussing our multi-sensor algorithms, however, we will further explain how Cole-Kripke and Sadeh classify sleep epochs.

3.2.1 Cole-Kripke and Sadeh - Rediscovering the Optimal Single Sensor Approaches

As discussed in section 2.4.1, Cole-Kripke classifies each minute of sleep through a weighted combination of the preceding four minutes and the following two minutes surrounding any epoch. The basic equation is represented in equation (1) where D represents the classification—$D < 1$ is sleep and $D \geq 1$ is wake—P is a scale factor applied to the entire equation, $W_{-4}, W_{-3}, ..., W_2$ represent the weighting factor for each minute, and $A_{-4}, A_{-3}, ..., A_2$ represents the activity scores—based on the y-axis—for each minute [18].

$$\begin{aligned} D = P(&W_{-4}A_{-4} + W_{-3}A_{-3} + W_{-2}A_{-2} + W_{-1}A_{-1} \\ &+ W_0A_0 + W_{+1}A_{+1} + W_{+2}A_{+2}) \end{aligned} \tag{1}$$

Interestingly, our research team noticed that the open source implementation of the Cole-Kripke algorithm implemented neither the optimal activity score calculation methodology nor the coefficients that yielded the highest accuracy; initially, we believed there was room for further optimization. Cole et al. indicated that their most effective algorithm first relied on chunking data into 2-second epochs, sliding a 10-second window over it, and then retaining the maximum windowed sum as the activity count for that minute. With this counts they performed a regression analysis and determined their most accurate equation was as follows [18]:

$$\begin{aligned} D = 0.00001(&404A_{-4} + 598A_{-3} + 326A_{-2} + 441A_{-1} \\ &+ 1408A_0 + 508A_{+1} + 350A_{+2}) \end{aligned} \tag{2}$$

After implementing the 10-second sliding window algorithm and its optimized equations, it became clear this was not a mere convenience-based simplification by ActiGraph, but rather a deliberate choice due to the difference in the way modern

actigraphy devices calculate activity counts; using this "optimized" methodology with our data resulted in 100% of the night being classified as sleep. The original 1992 algorithm relied on a zero-crossing AMI Motionlogger, which produced fundamentally different data than the more sensitive GT3X+ device. The motion logger measured movement frequency whereas modern sensors measure acceleration amplitude. An independent study by Quante et al. [34] indicated that ActiGraph performed a side-by-side analysis with both devices, and empirically determined scaling and capping methodologies on the minute time quanta that allowed similar performance with the GT3X+ to the legacy motion logger sensor. They also discovered basic changes that needed to be made to thresholds in the Sadeh formulation. Understanding that these were well-researched and deliberate choices, we implemented the functions in the same capacity as those done in the Actigraph Sleepr repository [33] rather than the original 1992 papers. The final Cole-Kripke implementation followed this equation:

$$\begin{aligned} D = 0.001(&106A_{-4} + 54A_{-3} + 58A_{-2} + 76A_{-1} \\ &+ 230A_0 + 74A_{+1} + 67A_{+2}) \end{aligned} \tag{3}$$

Immediately upon adjusting to this equation and following the 60-second aggregation methodology, our accuracy and epoch classifications produced plausible sleep-wake patterns consistent with the original paper's reported behavior. We were, however, able to improve our pilot results by making one addition to the Actigraph Sleepr repository [33] by implementing the Webster Rescoring methodology discussed in the original paper [18]. These rules were applied in the initial paper to directly attack the wake sensitivity issue our paper aims to address; this was performed after the algorithmic classification, and thus could be built on top of the existing framework without being marred by the idiosyncrasies of specific devices.

The Webster Rescoring rules are as follows:

1. After at least four consecutive wake epochs, the next epoch would be classified as wake.
2. After at least ten consecutive wake epochs, the next three epochs would be classified as wake.
3. After at least fifteen minutes scored as wake, the next four minutes would be classified as wake.
4. Six epochs or fewer classified as sleep, surrounded by ten epochs of wake (on either side), are rescored as wake.
5. Ten epochs or fewer classified as sleep, surrounded by twenty epochs (on either side), are rescored as wake.

These five rules lowered the ratio of "false sleeps" to "false wakes" by nearly 30% [18]. Particularly given our efforts to accurately identify wakefulness, these were simple to implement, effective rules that directly improved wake sensitivity. Per their papers' conventions, we also mark the first four epochs and last two epochs as unscored as there are not sufficient preceding or trailing epochs to classify these accurately. This six-minute adjustment likely was not necessary for ActiGraph's general purposes, but is worthwhile in a rigorous performance analysis of the Kripke algorithm. Sadeh did

not have further optimizations to the base implementation, and the algorithm only required one minor thresholding change to match modern actigraphy measurements.

Sadeh used surrounding epochs, five minutes preceding and five minutes following, to perform sleep classification on y-axis actigraphy counts; however, they used an activity average across this time interval with additional variables, rather than weighting each minute independently. Their original algorithm is as follows:

$$\begin{aligned} \text{PS} = 7.601 - 0.065 \cdot \text{Mean-W-5-min} - 1.08 \cdot \text{NAT} \\ - 0.056 \cdot \text{SD-last 6 min} - 0.703 \cdot \text{LOG-Act} \end{aligned} \tag{4}$$

In this formulation, PS represents the probability of sleep—if $PS \geq 0$ the epoch is classified as sleep, if $PS < 0$ it is classified as wake—$Mean\text{-}W\text{-}5\text{-}min$ represents the average activity count over the eleven-minute window, NAT is the number of epochs with activity level greater-than or equal-to 50 but lower than 100 across the whole eleven-minute period, $SD\text{-}last\ 6\ min$ is the standard deviation of the current epoch and the five preceding minutes, and $LOG\text{-}Act$ is the natural logarithm of the number of activity counts in the current epoch plus one.

The thresholding change noted both in the Quante et al. paper [34] and in the Actigraph Sleepr repository [33] changes sleep classifications to $PS > -4$ and wake classifications to $PS \leq -4$. Interestingly, in our pilot, though the accuracy and epoch classification frequency closely mirrored that of the original paper with this exact implementation, it is worth mentioning that the NAT feature was almost always zero, with fewer than 1% of our epochs falling into the $50 - 100$ range found on the original frequency-based motion logger. This feature was likely kept by ActiGraph for algorithmic consistency, and because the zero value does not impact the remaining more-meaningful features.

3.2.2 Exploring Multi-Sensor Approaches

We followed data fusion strategies at three different levels outlined by Gravina et al. [35]: data-level, feature-level, and decision level. The feature-level is also commonly referred to as score-level fusion, as it combines the numerical outputs of models before making the final classification [36, 37]; the score level convention better applies to our "D" and "PS" sleep scores, and is used for the remainder of the paper. Data level fusion is the lowest level, combining multiple sources of raw sensor data with no information or performance loss. Score-level fusion happens as an intermediate step after raw data processing, aggregating these scores before making a final classification. Finally, decision-level fusion allows for independent classifications that are processed by a decision strategy like voting or summation.

The data level served as our baseline-performance algorithm for a multi-sensor approach. We averaged the data from the four sensors, per axis, for a combined signal. This leverages the reduction in mean average error when using redundant sensors as discussed by Fuster-Garcia et al. [5]. We expected more significant results at the score and decision levels, as our research team believed that leveraging higher-level score and classification results would better exploit our multi-sensor architecture. At this

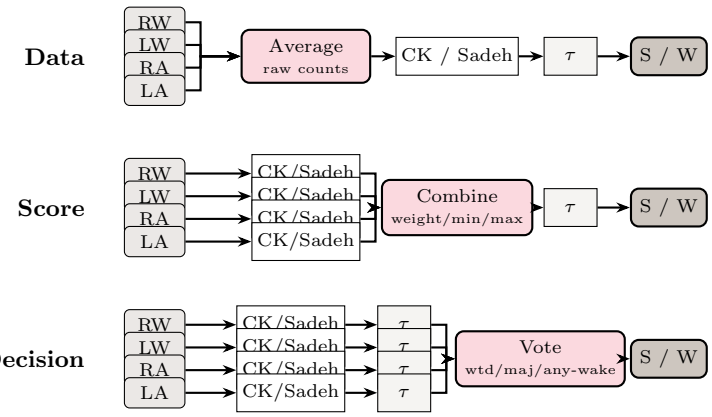

Fig. 1 Three distinct points in the pipeline where our four limbs' data can be combined — raw actigraph counts, raw algorithm scores, or at the sleep/wake decision level. Framework adapted from Gravina [35] and Kittler [36].

level, our algorithms were modeled from Kittler's seminal classifier combination paper [36] using strategies such as weighted combination, majority voting, and the max and min rules.

At the score-level, we implemented three algorithms: weighted combination, min, and max. To perform score-level transformations, we performed the single-sensor variant of Cole-Kripke and Sadeh, respectively, *without* making the final sleep or wake classification. To perform our weighted combination, we weighted the resulting numerical scores, 0.5 for ankles, given their tendency to be more inert during sleep [23], and 1.0 for wrists, given the documentation of stronger wrist-based motion [38], before performing our final classification with the standard thresholds for each algorithm. Our min and max algorithms took the minimum or maximum score from each epoch before applying the threshold. In Cole-Kripke, the max algorithm maximally favors wake-classification, and the min-algorithm maximally favors sleep classification whereas in Sadeh the opposite is true; this is due to thresholding conventions: Cole-Kripke $D < 1$ is sleep, whereas Sadeh $PS \leq -4$ is wake. With these three algorithms, we assessed whether the combination of these sensors provided a latent feature representation of information that single sensors couldn't capture; we aimed to produce a sleep-favored, wake-favored, and moderate approach for a balanced perspective on score-level strategies.

At the decision-level, we implemented a weighted vote, a majority vote (unweighted), and an any-wake algorithm (functionally similar to the Cole Max score-level approach). At this level, we biased our algorithms towards wake detection with our principal goal of improving wake-classification and the specificity of sleep algorithms; both our voting algorithms favored wakefulness in the event of ties. Our weighted vote provided half a vote to each ankle and a full vote to each wrist. Our majority vote provided each limb with a full vote. Finally, the any-wake algorithm was the upper bound of a wake-biased algorithm, classifying the epoch as wake if any sensor resulted in such a classification.

It is worth addressing a subtle similarity we uncovered between the decision-level any-wake variants and the score-level wake-biased variants. When we first ran the algorithms, we noticed, particularly for Sadeh, that the results for Any-Wake and Min Score were identical across all 6 values— this seemed impossible if the classifiers truly had underlying differences. Upon closer inspection, it became clear that the fundamental question both classifiers were asking was "did any sensor cross the wake threshold?" We considered removing this for simplicity, yet opted to keep it as Cole-Kripke had a subtle but key differentiator.

The Webster rescoring rules that we implemented when emulating optimal wake detecting Cole-Kripke variant [18] provide a subtle but impactful intermediate step that depends on the context of surrounding epochs rather than performing linear transformations on the raw sensor data. This non-linearity introduces real variance as the resulting classifications of individual data streams are changed prior to any decision-level consensus. Further exploration confirmed this; if Webster rescoring is applied to the already-fused label sequence, Cole-Kripke Max and Cole-Kripke any-wake produce identical results.

We opted to apply Webster rescoring to each of the sensors' labels before decision-level operations occur, as this maintains the most logical consistency with the semantic definition of decision-level fusion as described by Gravina et al. [35]. Each sensor first performs independent classification (rescoring is a part of the classification), and then performs actions on these component classifications.

We aimed to provide a wide array of multi-sensor algorithms at each level of data fusion, spanning the different levels of biases towards sleep and wake, and ultimately provide a high-fidelity picture of the various multi-sensor approaches.

Level	**Method**	**Fused Data**	**CK variants**	**Sadeh variants**
Single-sensor	Baseline	N/A	CK Single	Sadeh Single
Data level	Average raw counts	Raw axis1	CK Combined	Sadeh Combined
Score level	Fuse scores	D / PS Scores	Weighted, Min, Max	Weighted, Min, Max
Decision level	Fuse S/W labels	Per-limb votes	Weighted Vote, Majority, Any Wake	Weighted Vote, Majority, Any Wake

CK and Sadeh variants implemented at each fusion level. Single-sensor algorithms acted as a baseline. One multi-sensor variant per algorithm at the data level, and three variants per algorithm at the score and decision levels.

Table 1 Comparison of Fusion Levels and Variants.

3.3 Fine-Tuning our Methodologies

With our multi-sensor algorithms, there are two key sets of parameters that can be adjusted: the threshold value, and for weighted score-level and decision-level algorithms, their weights. Our initial thresholds, $D > 1$, being wake for Cole-Kripke and $PS \leq -4$, being wake for Sadeh, were based on the more-active wrist activity counts. By averaging four limbs, including ankles, we expected our algorithmic scores to trend more towards sleep with the current thresholds. To robustly test this factor, we assessed a range of threshold values for each function. For Cole-Kripke, we tested threshold values ranging from $[0.1, 2.5]$ at 0.1 increments, and for Sadeh we tested threshold values ranging from $[-8.0, 4.0]$ with an increment of 0.5. This resulted in 25 different thresholds to test per Cole-Kripke / Sadeh variant.

To adjust for limb weights, we limited our exploration space by assuming symmetry between wrist and ankle weights. This matched Sadeh's original finding [19], which indicated non-dominant and dominant wrists were similarly explanatory for the sleep classification. Our starting values of 1.0 and 0.5 for wrist and ankle weights, respectively, were based on the simple heuristic evaluation that ankles are typically less active than wrists [23], and thus would bias more towards sleep. We explored weights in the range of $[0, 2.0]$ for each limb at an increment of 0.25. While we expected wrists to be more indicative of wakefulness, this helped us understand *how indicative* limbs are. In the extreme case where wrists are weighted at 2.0, and ankles are weighted at 0.25, wrists would be eight times more impactful on the resulting classification of wakefulness. Excluding $(0, 0)$ weights, this resulted in 80 weight combinations.

Parameter	**Description**	**Starting Value**	**Search Range**	**Step**
CK threshold	D-score threshold	1.0	[0.1, 2.5]	0.1
Sadeh threshold	PS threshold	-4.0	$[-8.0, 4.0]$	0.5
Wrist sensor weight	Wrist weight for weighted fusion	1.0	[0.0, 2.0]	0.25
Ankle sensor weight	Ankle weight for weighted fusion	0.5	[0.0, 2.0]	0.25

Search ranges and step sizes for LOOCV parameter optimization. Starting threshold values based on published standards and starting weights based on heuristic evaluation. We deliberately used a wide search space to ensure we identified the global optima for multi-sensor variants.

Table 2 Parameter Configuration and Optimization Ranges.

The weighted algorithms dramatically increased the search space we had to optimize for. We expected each algorithm to require different thresholds to account for each fusion method's bias towards wake classification. For example, a maximally wake-favored algorithm likely would require a slightly higher threshold than a traditional or sleep-favored algorithm to maintain the same level of sleep-specificity. With this in mind, there were 2000 combinations to evaluate for each weighted sleep algorithm (25 thresholds $*$ 80 weights). With 4 weighted algorithms requiring

both threshold and weight optimizations and 10 algorithms requiring only threshold optimizations, we had to perform a total of 8250 evaluations per patient.

Because we knew our population would be extremely small, we opted for Leave-One-Out-Cross-Validation (LOOCV) [39]. This methodology would allow us to select the optimal coefficients and weights based on all but one member of the population; we then would apply this algorithm with optimized values to the test population, the member who was left out of the training set. We would record the highest performing combination's F_1 score, and rotate every member of the population through this process so each individual served as the testing population. Finally, we reported the average F_1 score for the optimal combination of metrics for each tested party. This process ensured a deterministic, exhaustive exploration of how well optimizations generalize to multiple parties.

4 Results

Given our team's unique emphasis on classifying wakefulness, it is worth remembering that the core function of our algorithms is still to properly classify an entire night of sleep. An algorithm that only detects wake for a given night would score remarkably on wake classification, but would fail as a sleep detection algorithm. Simply improving wake-identification at the expense of classifying sleep correctly is unacceptable. To accurately portray the impact of our algorithms, we first need to establish metrics that clearly demonstrate wake-classification improvement grounded in accurate sleep identification.

The metrics introduced in Section 2— TST, SOL, and WASO are all derived from epoch-level sleep/wake classification, yet our research team recognized these metrics contribute to masking the class imbalance problem. Fundamentally, TST is the count of sleep-classified epochs, SOL is the number of wake epochs before the first prolonged bout of sleep, and WASO is the total number of wake epochs between sleep onset and offset. However, misclassified wake epochs simply inflate TST while deflating WASO without providing further clarity on the underlying mechanics. We rather focus on the wake F_1 score, a harmonic mean of precision and recall that effectively encapsulates data about the number of correctly predicted sleep *and* wake epochs. Improving wake F_1 necessarily improves TST, SOL, and WASO.

4.1 F_1 Score

F_1 Score is the primary metric we report for our study, rather than accuracy, as it is a harmonic mean of precision and recall with the following equation:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5a}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5b}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5c}$$

Grounding this in the terminology we have used thus far in the paper, precision means: "Of all the epochs our algorithm classified as wake, how many epochs was the individual actually awake?" Recall means: "Of all the true wake epochs, what fraction did the algorithm detect?" By taking the harmonic mean of these two metrics, we penalize any algorithm that achieves high wake sensitivity (recall) at the expense of precision by only guessing "Wake," for example, while also penalizing the algorithm for being too conservative and only classifying wake when it is 100% certain, which would result in high precision but low recall. In our traditional class imbalance case, an algorithm that predicted only sleep could have an accuracy of over 85%, but would score an F_1 of 0.

4.2 Cohen's Kappa & Matthews Correlation Coefficient

While the F_1 score rigorously evaluates our algorithm's ability to identify our critical minority class (wakefulness), it possesses a significant blind spot: it entirely omits true negatives (correctly identified sleep epochs) from its calculation. In the context of actigraphy, sleep inherently constitutes the overwhelming majority of the dataset, frequently accounting for 80% to 90% of a given night. An algorithm optimized strictly for wake-sensitivity—such as our "Any-Wake" algorithm will inevitably generate more localized False Positives by flagging minor sleep disturbances as wakefulness, thus reducing true negatives. Here is an admittedly extreme but informative example:

Imagine a night with 100 epochs where an individual sleeps for 50 minutes, and is awake for the other 50. This algorithm is degenerate and only predicts wake. We have 50 True Positives, 0 False Negatives, 50 False Positives, and 0 True Negatives. Now imagine a second night with 1000 epochs where an individual slept for 950 minutes and was awake for 50. This algorithm is much better, and only classifies 50 epochs of sleep incorrectly (classifies 900 epochs as sleep). In this case, we have 50 True Positives, 0 False Negatives, 50 False Positives, and 900 True Negatives. In both cases, the F_1 score would be identical at about 67%. However, our broken algorithm had a 0% sleep specificity, whereas our much better algorithm had a 95% accuracy.

Because the F_1 does not "see" the size of the pool of correctly identified sleep epochs, solely optimizing for it can negatively impact sleep specificity, thus reducing the algorithm's functionality. To ensure our multi-sensor algorithms improve wake detection without silently degrading baseline sleep classification, we pair our primary F_1 metric with Cohen's Kappa and the Matthews Correlation Coefficient (MCC).

4.2.1 Cohen's Kappa

Cohen's Kappa measures the agreement between two classifiers while subtracting the probability that this agreement is occurring by chance [40]. This is represented by the following equation:

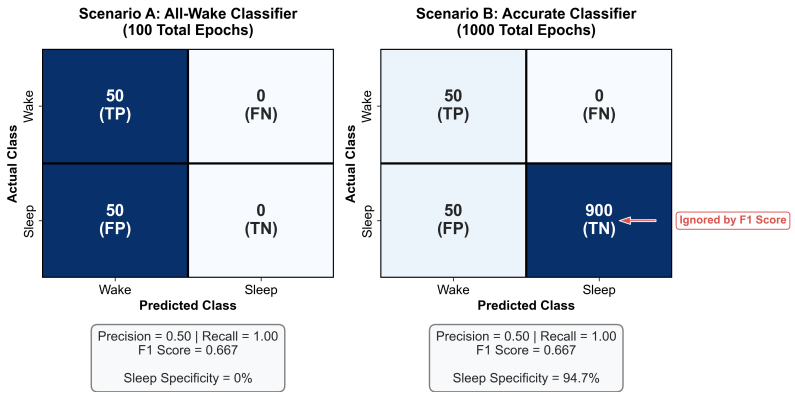

Fig. 2 Visualizing the F_1 score's blind spot: In both cases of high and low sleep specificity, the F_1 score remains identical because it does not account for True Negatives.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{6a}$$

$$N = TP + TN + FP + FN \tag{6b}$$

$$p_o = \frac{TP + TN}{N} \tag{6c}$$

$$p_e = \left(\frac{TP + FN}{N} \cdot \frac{TP + FP}{N}\right) + \left(\frac{FP + TN}{N} \cdot \frac{FN + TN}{N}\right) \tag{6d}$$

Kappa (κ) is a value that ranges from 0 to 1. Typically, < 0.20 is poor, $0.20 - 0.40$ is fair, $0.40 - 0.60$ is moderate, and > 0.60 is substantial. p_o is accuracy - total correctly classified epochs (sleep and wake). p_e is the expected agreement by chance, or the probability that both classifiers predict wake, plus the probability that both classifiers predict sleep. This is calculated as the sum of the probability of marginal totals. To further clarify this, let's return to our extreme example in Figure 2:

Analyzing our broken "All Wake" classifier, the observed accuracy is 50% as it got half of the epochs correct. Calculating the expected probability in plain English is the percentage of epochs where wake actually occurred (50/100 epochs) times the tendency for our algorithm to guess wake (100/100) plus the percentage of epochs where sleep actually occurred (50/100 epochs) times the probability of our algorithm predicting sleep (0/100). This results in p_e of 0.5. This would result in a Cohen's Kappa score of 0, highlighting that this is an incredibly poor classifier; it performed no better than random chance.

Performing the same calculation with our more accurate algorithm, the observed accuracy jumps up to 95% (50 identified wake and 900 identified sleep out of 1000 total epochs). The expected chance of agreement would be as follows: the percentage

of epochs where wake actually occurred (50/1000 epochs) times the tendency for our algorithm to guess wake (100/1000) plus the percentage of epochs where sleep actually occurred (950/1000 epochs) times the probability of our algorithm predicting sleep (900/1000). This results in a final p_e of 0.86 and a Kappa score of 0.643.

Cohen's Kappa directly shines a light on the blind spot present with our F_1 score. Particularly because we have a dominant class, sleep, where the actual chance of an epoch being sleep is high, it is important to have a metric that can clearly distinguish our classifier from chance.

4.2.2 Matthews Correlation Coefficient

Apart from proving that our algorithms are better classifiers than pure chance with Cohen's Kappa, Matthew's Correlation Coefficient (MCC) is a highly balanced metric that similarly compares predicted and actual classifications; it is the only widespread confusion matrix metric that requires high accuracy in all four quadrants of the matrix— a high MCC of 1 demands perfect classification over all four quadrants [41]. An MCC of 0 corresponds to a chance level prediction, and an MCC of -1 would be a perfect inversion. The equation looks as follows:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7a}$$

Reviewing this equation, it becomes immediately apparent that our bad algorithm from the previous example, with 0 False Negatives and True Negatives, immediately cancels out the numerator. Boundary cases are immediately caught and flagged with this formulation. When we apply the same equation to our accurate classifier, we get a resultant score of 0.68. The scale is similar to Cohen's Kappa, where < 0.30 is weak, $0.30 - 0.50$ is moderate, $0.50 - 0.70$ is strong, $0.70 - 0.90$ is very strong, and $0.90 - 1.00$ is near perfect. There is no way for an algorithm to "cheat" its way to a perfect MCC score.

The remainder of our results section reports these metrics for each of our classification functions.

4.3 Multi-Sensor Fusion Data Analysis

4.3.1 Patient Cohort Characterization

Due to the lengthy nature of both the study approval process and the time necessary to schedule and facilitate sleep studies, we had a total population size of 4 participants. While we are actively continuing to consent more individuals for polysomnography exams, the generalizability of our results is significantly impacted by our small population size.

Across our four patients, individuals ranged between a 6.5 to 7.9 hour sleep window with a range of 9% to 20.3% of that being wakefulness, just barely missing the recommended 7 to 9 hours of total sleep [15]. Interestingly, our patient who slept the least, $P001$, also demonstrated the least wakefulness. While this small population made

it challenging to robustly test coefficients for our algorithms, it provided us with the opportunity to examine the individual actigraphy scores across patients more closely.

Patient	Wake Epochs	Sleep Epochs	Total	Wake %	Hours
P001	35	355	390	9.0	6.5
P002	85	333	418	20.3	7.0
P003	90	384	474	19.0	7.9
P004	61	357	418	14.6	7.0
Mean	67.8	357.2	425.0	15.7	7.1

Per-participant 60-second epoch sleep classification. Wake % is the fraction of total epochs labeled wake by PSG; Hours is total recording duration within the lights-out to lights-on window. Note the severe class imbalance: across all participants, sleep accounts for 80%–91% of the recording period.

Table 3 PSG-derived class distribution for all four participants.

4.3.2 Per-Limb Activity Exploration

Before running our data through any classification algorithms, we sought to understand the range of activity per limb during epochs classified by polysomnography as wake and sleep, respectively. If any limbs had particularly low activity counts during wake, this likely meant they would be more subject to the poor WASO detection [4] that has been previously identified with actigraphy-based sleep algorithms. We then also took the average activity count during sleep. Looking at the ratio between the average sleep actigraphy counts and wake actigraphy counts painted a clear picture; if an individual had a score of 1, the amount of limb movement during sleep and wakefulness was nearly identical. Any algorithm would likely struggle to classify sleep and wakefulness appropriately as there isn't a clear signal to attach to.

Interestingly, the trends differentiated across patients. For Patients one and two, their left wrists were their most active limb during periods of wakefulness. For patients three and four, their right wrists were their strongest indicator. Across all four patients, ankles were never the most active limb during sleep or wakefulness, corroborating Kamdar et al.'s [23] findings.

Looking more closely at the W/S ratios presented in Table 4, patient one's left ankle is actually incredibly indicative of their sleep state. Their left ankle moves, on average, almost 15 times more when they are awake than when they are asleep. For other individuals, such as patients two and three, their ankles are actually more active when they are asleep. As such, ankles are challenging to rely on by themselves for sleep state detection, but provide interesting signals for algorithms to exploit when combined with more descriptive wrist data.

4.3.3 Identification of Patient Two as an Outlier

Looking at the activity data, patient two represents an outlier in multiple different ways. Their maximum activity score while awake, their left wrist, at 35.8 scores lower

than most other patients' wrist activity scores while they are asleep. Furthermore, their W/S ratios demonstrate incredibly poor discriminability at 1.13 and a perfect 1.0 respectively; their wrists essentially moved the same amount while awake as asleep.

Fundamentally, actigraphy classifies wakefulness on the volume and intensity of patient movement. In a patient who not only moves very little but also demonstrates similar movement patterns during sleep and wakefulness, we expect actigraphy algorithms to perform poorly. We performed our primary analysis with their data excluded. However, comparing a secondary analysis with their data included highlights unique trends which apply in the presence of anomalous individuals.

		R-Ankle	**R-Wrist**	**L-Wrist**	**L-Ankle**
P001	Wake Mean	19.8	75.0	171.1	123.9
	Sleep Mean	11.0	34.1	46.4	8.3
	W/S Ratio	1.80	2.20	3.69	14.96
P002	Wake Mean	16.8	28.8	35.8	6.8
	Sleep Mean	11.1	25.5	35.9	12.4
	W/S Ratio	1.52	1.13	1.00	0.55
P003	Wake Mean	50.7	136.1	109.4	48.1
	Sleep Mean	59.0	91.2	82.2	60.3
	W/S Ratio	0.86	1.49	1.33	0.80
P004	Wake Mean	15.6	73.3	54.5	21.1
	Sleep Mean	11.4	40.3	32.1	13.4
	W/S Ratio	1.37	1.82	1.69	1.57

Mean activity counts per 60-second epoch during PSG-labeled wake and sleep periods. W/S Ratio = Wake Mean / Sleep Mean; values near 1.0 indicate poor discriminability between states. P002 is anomalous in that both their activity counts are low, and their W/S ratios are much closer to 1.0 than their peers.

Table 4 Per-limb activity characteristics during wake and sleep epochs.

4.3.4 Primary Algorithmic Performance

First, looking at the results of our algorithms with their base heuristic level parameters ($D_\tau = 1$, $PS_\tau = -4$, $AW = 0.5$, $WW = 1.0$), nearly every Cole-Kripke variant outperforms its Sadeh counterpart across all three metrics, F_1, κ, and MCC— the only exception being the worst performing algorithms, CK Min and Sadeh Max. Overall, these results suggest weighted scores of surrounding epochs used by Cole-Kripke seem to generalize better to multi-sensor fusion strategies than the epoch averaging and statistical calculations of Sadeh.

In both sets of algorithms, our maximally sleep-favoring algorithms, CK Min Score and Sadeh Max Score, performed expectedly poorly with a wake F_1 of 0.045 and 0.129 respectively. Both their κ and MCC scores ranged between $0.024 - 0.090$, indicating these classifiers were barely better than chance and failed to mark any significant improvement in wake detection; this logically follows, as unless all four limbs were moving substantially, the epoch would be marked as sleep.

It is worth establishing the performance baseline for our multi-sensor variants as the Cole-Kripke Single and Sadeh Single Right Wrist variants. Our Cole-Kripke Single RW's accuracy was 0.847, which is about 4% lower than the agreement recorded in the original Cole-Kripke paper [18]. Similarly, the accuracy of our Sadeh Single RW was 0.799, almost $11 - 12\%$ lower than their reported 91% to 93% accuracy [19]. Sadeh's wake detection and performance on both wrists were consistent with the findings in their original paper; the right wrist (assumed to be dominant for most of our population, though not recorded) had slightly higher accuracy, but the results were generally similar. To our knowledge, there has not been a direct comparison between Cole-Kripke's performance per wrist— the differences between the left and right wrist performance were significant. Though the right wrist Cole-Kripke algorithm was one of our highest performing algorithms, significantly exceeding by (nearly 12.5%) the reported wake detection of approximately 36% in their initial findings [18], whereas our left-wrist variant was one of our worst performing algorithms, slightly undercutting their wake prediction.

Another factor stands out when looking at these algorithms. Across the Cole-Kripke algorithms, the top four performing algorithms evenly span across the score and decision levels. In Sadeh, however, the decision level is clustered near the top, with score algorithms clustering near the bottom (except the Any-Wake variant due to its logical equivalence with the decision level). Across the results, it is clear that algorithms that strongly bias towards the dominant class perform poorly in multi-sensor fusion; the likelihood that one limb meets the threshold necessary to be classified as sleep is very high. In contrast, more wake-biased approaches capture more of the minority class without significantly overwhelming the overall classification.

The most notable observation from the table is the fact that the Cole-Kripke baseline, CK Single RW with an F_1 of 0.459, κ of 0.371, and MCC of 0.374 outperforms every other algorithmic variant, seemingly undermining the multi-sensor approach outlined in this paper. However, the substantial difference between wrists— a 31% relative difference— indicates that this algorithm's efficacy is only as good as the experimenter's ability to predict the correct wrist prior to study. We did not have wrist dominance recorded in our cohort, and the implication of choosing the wrong wrist is substantial. Every multi-sensor variant, with the exception of the most-sleep biased algorithm, outperformed the worst-case single sensor placement. Thus, the true value of multi-sensor approaches lies not in achieving peak performance, but in robustness by significantly mitigating the downside risk of worst-case limb placement.

Ultimately, not only do our multi-sensor algorithms provide a marginal gain in wake F_1 performance, but it removes a significant downside risk of wrist-based variance. The 31% relative difference in wake detection between wrists could significantly impact future study results, and this alone promotes fused sensor setups.

4.3.5 Bringing Outliers into the Picture

Introducing patient two, our distinct outlier, helps identify trends that hold even with anomalous data. The most notable shift with the inclusion of their confounding data is that each algorithm's F_1 score decreased between $0.01 - 0.09$, with already poor-performing algorithms being less affected. A majority of algorithms lost between

Algorithm	Fusion	Acc	W-Sens	S-Spec	F_1	κ	MCC
Cole-Kripke Algorithms							
CK Single (RW)*	Single	0.847	0.486	0.906	**0.459**	0.371	0.374
CK Max Score	Score	0.786	0.621	0.814	0.420	0.300	0.333
CK Any Wake	Decision	0.793	0.565	0.830	0.407	0.289	0.313
CK Weighted Score	Score	0.835	0.349	0.915	0.363	0.270	0.272
CK Majority Vote	Decision	0.842	0.326	0.927	0.362	0.276	0.280
CK Weighted Vote	Decision	0.842	0.326	0.927	0.362	0.276	0.280
CK Combined	Data	0.844	0.315	0.931	0.353	0.270	0.273
CK Single (LW)*	Single	0.808	0.354	0.883	0.318	0.210	0.214
CK Min Score	Score	0.854	0.024	0.990	0.045	0.024	0.062
Sadeh Algorithms							
Sadeh Majority Vote	Decision	0.816	0.431	0.888	**0.391**	0.284	0.289
Sadeh Weighted Vote	Decision	0.818	0.423	0.891	0.388	0.282	0.287
Sadeh Any Wake†	Decision	0.729	0.609	0.756	0.372	0.222	0.258
Sadeh Combined	Data	0.800	0.393	0.867	0.360	0.243	0.247
Sadeh Single (RW)*	Single	0.799	0.374	0.873	0.345	0.226	0.228
Sadeh Single (LW)*	Single	0.775	0.352	0.849	0.298	0.166	0.171
Sadeh Weighted Score	Score	0.828	0.229	0.927	0.281	0.187	0.194
Sadeh Max Score	Score	0.849	0.075	0.981	0.129	0.090	0.137

*Single-sensor baselines; RW = Right Wrist, LW = Left Wrist. †Sadeh Any Wake is logically identical to Sadeh Min Score (score-level minimum) because published Sadeh has no per-sensor post-processing step; the two are reported once under Any Wake. All values are means across three participants (P001, P003, P004); P002 excluded due to anomalous activity profile. W-Sens = Wake Sensitivity; S-Spec = Sleep Specificity; κ = Cohen's Kappa. Bold = best F_1 per family. Sorted by F_1 within each family.

Table 5 Cross-patient mean performance of heuristic-parameter algorithms across all fusion levels, with the P002 outlier excluded ($N = 3$).

$0.04 - 0.09$, which is a significant and notable decrease in F_1 score. Their data was objectively harder to classify, and thus, the holistic performance of every algorithm decreased.

Another distinct indication of anomalous behavior is that their CK Majority Vote and CK Weighted Vote algorithmic performances differ. Referring to Table 5, most configurations result in identical outputs between these algorithms, which, without patient two, never occur. The only instance in which these algorithms deviate is when both ankles are classified as awake while both wrists are classified as asleep. This is unusual sleep behavior and is only noticeable when closely looking at the per-limb exploration Table 4.

However, the general trends of the tables remain the same, with Cole-Kripke outperforming Sadeh and with similar fusion clustering. Another point to note is that Cole-Kripke Max shifted to our highest performing algorithm, barely edging out the Single RW variant with F_1 scores of 0.387 and 0.372 respectively. This stands as a further implication that multisensor algorithms provide another layer of algorithmic robustness that single-sensor variants lack.

4.3.6 LOOCV Threshold and Weight Optimization

LOOCV parameter search of the surrounding threshold space revealed one core finding: the previously published thresholds for single sensor algorithms are not optimal for multi-sensor variants. When oriented to the graphs in Figure 3, the graphs in

Algorithm	Fusion	Acc	W-Sens	S-Spec	F_1	κ	MCC
Cole-Kripke Algorithms							
CK Max Score	Score	0.778	0.528	0.830	**0.387**	0.260	0.286
CK Single (RW)*	Single	0.826	0.382	0.914	0.372	0.282	0.285
CK Any Wake	Decision	0.782	0.471	0.845	0.363	0.240	0.259
CK Majority Vote	Decision	0.824	0.256	0.933	0.290	0.206	0.209
CK Weighted Vote	Decision	0.823	0.253	0.933	0.286	0.202	0.204
CK Weighted Score	Score	0.817	0.267	0.924	0.282	0.194	0.192
CK Single (LW)*	Single	0.793	0.292	0.890	0.275	0.162	0.166
CK Combined	Data	0.829	0.239	0.944	0.270	0.200	0.200
CK Min Score	Score	0.840	0.018	0.993	0.033	0.018	0.046
Sadeh Algorithms							
Sadeh Any Wake†	Decision	0.725	0.512	0.776	**0.339**	0.183	0.210
Sadeh Combined	Data	0.787	0.333	0.875	0.319	0.197	0.201
Sadeh Majority Vote	Decision	0.798	0.335	0.895	0.311	0.202	0.204
Sadeh Weighted Vote	Decision	0.799	0.329	0.898	0.308	0.200	0.202
Sadeh Single (RW)*	Single	0.786	0.319	0.879	0.308	0.185	0.187
Sadeh Single (LW)*	Single	0.762	0.285	0.859	0.251	0.116	0.119
Sadeh Weighted Score	Score	0.810	0.175	0.932	0.216	0.125	0.125
Sadeh Max Score	Score	0.834	0.056	0.984	0.097	0.064	0.092

*Single-sensor baselines; RW = Right Wrist, LW = Left Wrist. †Sadeh Any Wake is logically identical to Sadeh Min Score (score-level minimum) because published Sadeh has no per-sensor post-processing step; the two are reported once under Any Wake. All values are means across four participants (P001–P004). W-Sens = Wake Sensitivity; S-Spec = Sleep Specificity; κ = Cohen's Kappa. Bold = best F_1 per family. Sorted by F_1 within each family.

Table 6 Cross-patient mean performance of heuristic-parameter algorithms across all fusion levels ($N = 4$).

Cole-Kripke and Sadeh seem to have inverted trends. As a reminder, Cole-Kripke $D < 1$ is sleep, whereas Sadeh $PS > -4$ is sleep, and as such their behavior appears mirrored across the y-axis. In both cases, however, when looking at the average performance of each algorithmic variant ((a) and (b)), the optimal thresholds for F_1 performance need to be shifted to be more wake favored; Cole-Kripke's optimal threshold shifts from 1 to $0.1 - 0.5$ and Sadeh's threshold shifts from -4 to $0 - 4$. Fusing multiple data streams together dampens the most extreme behavior, which results in lower average thresholds that frequently fall below the accepted single-sensor thresholds. This essentially makes the algorithms more sleep-biased, which requires direct correction with thresholding. This is most apparent when looking at the maximally sleep-favored algorithmic variants ((c) and (d)).

In CK Min Score and Sadeh Max Score, algorithmic F_1 performance drops precipitously even with more generous thresholds than the published standards. These algorithms were more sleep-favored than the traditional approaches while also using a fused data stream, which naturally dampens wakefulness signals. While adjusting these degenerate algorithms improves their performance, it is worth recognizing that this only serves to improve these algorithms to the point that wake-favored variants started at. Thresholding serves as a meaningful optimization for wake-favored variants, whereas it is a necessity for functionality in sleep-biased setups.

It is also worth noting that patient two's behavior continues to deviate from the rest of the cohort, with their F_1 performance flat-lining where other patients' thresholds approach local optima at any given threshold. Even aggressive thresholding

adjustments barely allow their F_1 to match that of their cohort at the pre-optimized threshold. This behavior further highlights their unique and challenging movement profile for actigraphy-based sleep detection algorithms.

At a population size of $N = 3$, we don't have a sufficient sample size to publish rigorously explored and methodologically sound thresholds to be used for future studies. We can, however, provide strong indication for where thresholding can be effective and the direction which they need to be shifted.

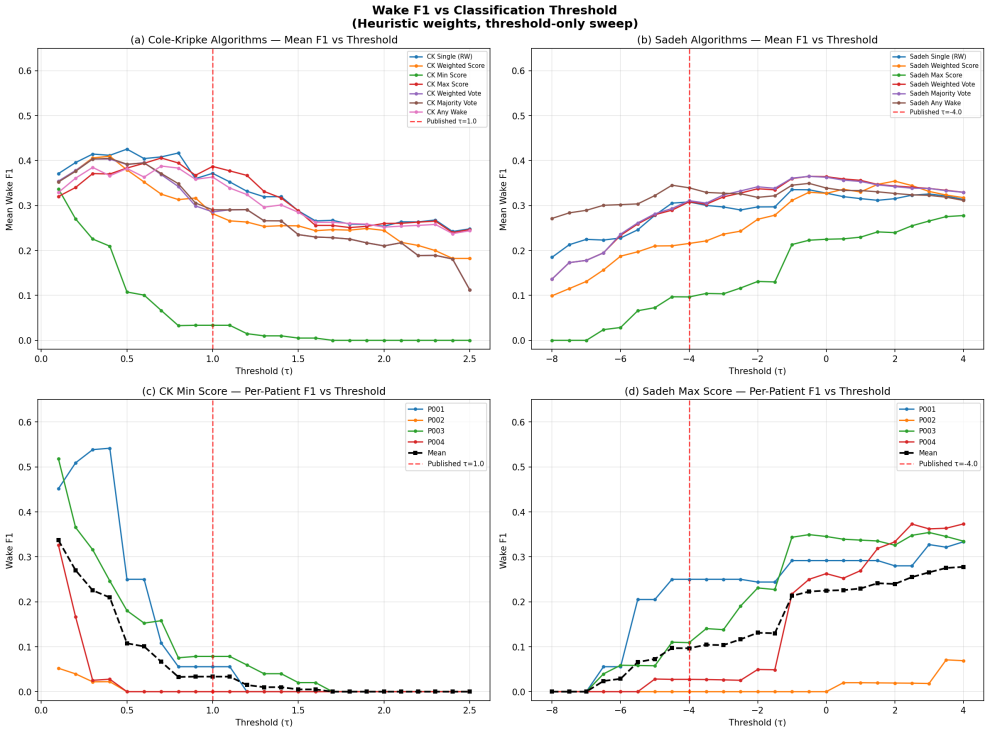

Fig. 3 Visualizing F_1 score in relation to different thresholds based on LOOCV optimization with fixed heuristic wrist and ankle weights (1.0 and 0.5 respectively). The red dashed line in each panel marks the published single-sensor threshold ($D_\tau = 1.0$ for Cole-Kripke, $PS_\tau = -4.0$ for Sadeh). **(a)** All Cole-Kripke fusion variants, showing mean F_1 across all four participants. **(b)** All Sadeh fusion variants, showing the analogous landscape.**(c)** and **(d)** CK Min Score and Sadeh Max Score broken out by individual patient, illustrating that the threshold shift is consistent across participants. In both (c) and (d), P002 (orange) is consistently the weakest performer across all thresholds, corroborating the anomalous activity profile identified in Table 4.

Weight Optimization unearthed marginal contributions in contrast to the threshold shifting. Only four of the algorithmic variants have fully tunable weight parameters (CK/Sadeh Weighted Vote and CK/Sadeh Weighted Score). When exploring these algorithms more closely, outlined in Table 7, the contributions of weight optimizations are minimal at best, and in the case of CK Weighted Score, actively negative. The

slight performance in the weighted variants of Sadeh, F_1 increases of 0.058 and 0.018 may indicate that weight optimization is more appropriate for the statistical approach of the Sadeh algorithm, whereas our better performing Cole-Kripke algorithm has less optimization room with weighting factors.

Algorithm	**Heuristic F_1**	**τ-only LOOCV F_1**	**Full LOOCV F_1**	**Δ_τ**	**Δ_w**
CK Weighted Score	0.363	0.411	0.377	0.048	−0.035
CK Weighted Vote	0.362	0.418	0.418	0.056	0.000
Sadeh Weighted Score	0.281	0.384	0.442	0.103	0.058
Sadeh Weighted Vote	0.388	0.359	0.377	−0.029	0.018

Decomposition of LOOCV optimization gains into threshold and weight contributions for the four algorithms with tunable sensor weights. *τ-only LOOCV F_1*: is the resulting F_1 with heuristic weights held constant. *Full LOOCV F_1*: LOOCV result when both threshold and weights are optimized jointly. Δ_τ = (τ-only F_1) − (Heuristic F_1); Δ_w = (Full F_1) − (τ-only F_1). Figure 4 visualizes the flatness of the weight landscape across all four algorithms.

Table 7 Decomposition of LOOCV F_1 gains into threshold (Δ_τ) and weight (Δ_w) contributions ($N = 3$, P002 excluded).

Reviewing the search space of the weights once the optimal threshold was set reveals the relatively flat landscape across the weight search space (Figure 4). While there are slight trends visible in CK Weighted Vote with wrist-biased weights and a slight trend towards ankle-biased weights across the Sadeh variants, the overall contribution is small. The maximum observed variation across any algorithm's F_1 due to weight is 0.058, which is modest compared to the threshold impact of 0.103.

With a larger cohort, minor optimizations with weight parameters may be validated and set, but our limited sample size does not support this action. Furthermore, the core optimizations are found in adjusting the thresholds rather than searching the weight space.

5 Discussion

Our multi-sensor exploration revealed multiple key assertions: multi-sensor approaches attack the brittleness of single sensor algorithms through redundant data streams, fusion algorithms consistently outperform worst-case single sensor algorithms and closely match the best case, fusion architecture is more important than individual parameters, threshold shifts are more impactful than sensor weights, wake-biased strategies outperform consensus seeking approaches, and Cole-Kripke consistently outperforms Sadeh on all fusion levels.

The most significant contribution of multi-sensor algorithms is avoiding the significant downside risk presented in Cole-Kripke, with a relative 31% performance drop running the exact same algorithm, on the exact same patient, with the only variance being wrist selection. In cohorts such as our own, where wrist dominance (which is the assumed causal lever between the variance, but not empirically proven), is not provided, there is simply no way for researchers to guarantee the appropriate

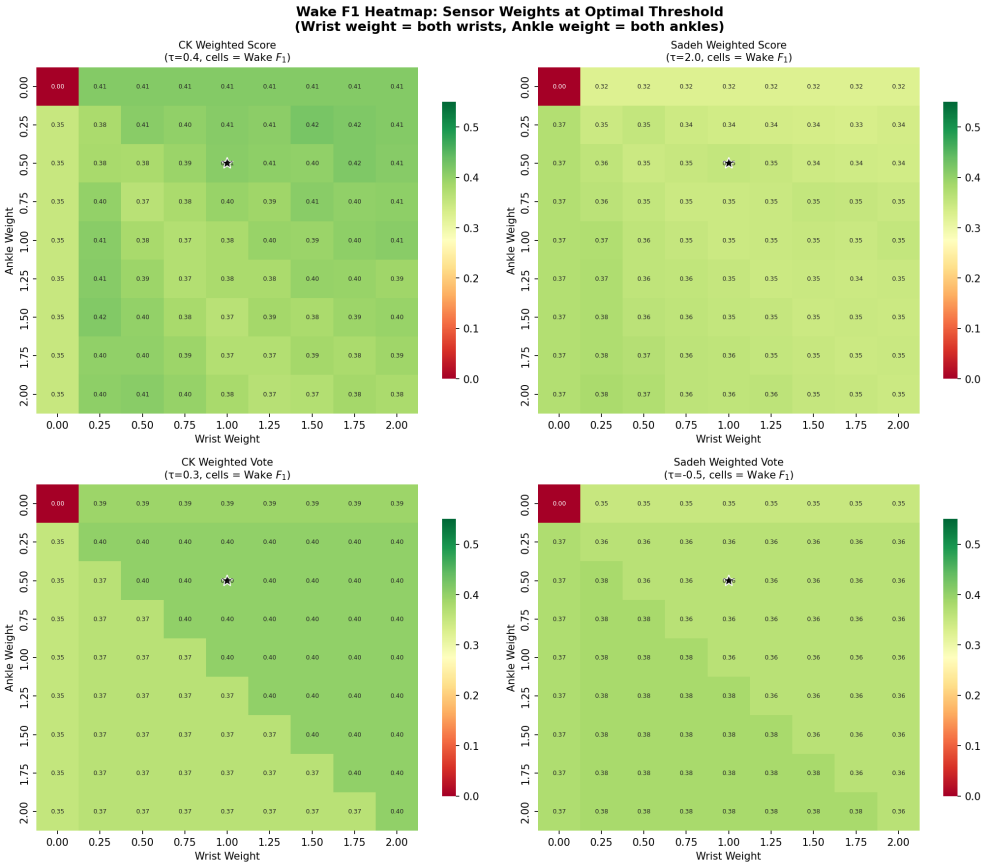

Fig. 4 Wake F_1 across the sensor weight space for the four algorithms with tunable wrist and ankle weights, evaluated at each algorithm's optimal threshold (shown in parentheses). Each cell reports the mean F_1 across P001, P003, and P004; the black star marks the heuristic weight configuration ($w_{\text{wrist}} = 1.0$, $w_{\text{ankle}} = 0.5$). All four landscapes are notably flat: the total F_1 variation across the weight grid (excluding the degenerate $(0, 0)$ corner) is at most $\sim$0.06 for any algorithm.

limb selection. In the Sadeh family, the case for multi-sensor approaches is even stronger, with five of the seven multi-sensor algorithms performing better than both single-sensor approaches.

Further analyzing the data, when looking at the top three performing algorithms out of the fourteen multi-sensor variants and the single-sensor baselines, CK Single (RW), CK Max Score, and CK Any Wake, all perform relatively similarly with fair Cohen's Kappa, fair to moderate MCC, and an F_1 variance of 0.052 between the highest and lowest performer of the three. Essentially, trading the 31% F_1 downside risk for a modest change in F_1 while retaining similar classification quality is a tradeoff that warrants consideration for future studies.

Evaluating the fusion architecture, another key point becomes evident— the choice of underlying architecture is far more important than any individual weight or

threshold optimization. Across both Sadeh and Cole-Kripke, maximally wake-favored algorithms performed within the top three algorithms in their family. Maximally sleep-favored algorithms performed at the bottom. Further threshold optimization slightly improved the top performers, while bringing the bottom performers closer to average performance. Picking a poor fusion strategy inherently results in optimization essentially "fighting against" the underlying architectural decision. Higher-level consensus architectures, decision, and score level for Cole-Kripke and the decision level for Sadeh consistently score higher than the data level fusion. Data level fusion does not seem to leverage the redundant signals in a way that most amplifies their usefulness, as the higher level variants do.

Furthermore, as seen in Figure 4, individual weight parameterization for cohorts is likely unnecessary, as running multi-sensor algorithmic variants with adjusted thresholds captures a majority of the performance improvement. The most significant adjustment to the base algorithms, necessary to support multi-sensor approaches, is to adjust the thresholds to account for the signal attenuation that occurs when using multiple sensors. The scores are inherently compressed towards sleep, given the frequent inclusion of at least one inert limb. With further exploration, and a larger cohort of $15 - 20$ individuals, threshold values can be set, allowing for immediate use of multi-sensor variants without the need for further weight optimization.

Intuitively, the superior performance of wake-biased fusion strategies matches expected sleep behavior. Physiologically, when one is awake but trying to sleep, it is unlikely that they are going to move all limbs at once. CK Max Score/Any Wake, and Sadeh Min Score/Any Wake surface individual limb movement that occurs when naively combining the underlying data. With further threshold research, these algorithms likely will continue to provide similar accuracy to single sensor variants while improving the underlying wake sensitivity.

As a whole, it is worth refocusing on the broader finding that Cole-Kripke outperforms Sadeh across all fusion levels. Cole-Kripke's temporal weighting (shifting coefficients for neighboring epochs) are better exploited in multi-sensor architectures than the statistical features in Sadeh. This holds both for the base heuristic and non-optimized algorithms as well as the post-optimization algorithms.

Multi-sensor algorithms provide an algorithmic robustness missed in single-sensor approaches. With threshold optimization, they maintain comparable performance to optimal single-sensor variants while eliminating the significant downside of wrong limb choice. The underlying algorithmic approach is the most important decision in multi-sensor approaches, and no amount of optimization can overcome poor architectural decisions. Threshold optimization returns a more significant F_1 optimization than individual weights in weight-oriented algorithms. Wake biased algorithms, and particularly, algorithms in the Cole-Kripke family, are the most promising within the multisensor approach.

5.1 Limitations

It is important to recognize that the broad generalizability of our study was significantly impacted by the small population size, especially because our limited dataset included an outlier. Furthermore, polysomnography studies are commonly

prescribed for individuals who have underlying conditions. We initially were hoping to analyze the underlying medical conditions of individuals and select healthier populations to improve our ability to generalize these results to the broader public. With only four participants, such an analysis was not possible, as we needed to include all members in the study regardless of preexisting condition.

The significant difference between CK (RW) and (LW) also raises the point that hand dominance is another significant criterion that likely should have been included in the study. Without this data point, we were not able to empirically test whether the inter-wrist performance gap was causally linked to handedness.

Additionally, in-lab settings do not perfectly mirror the naturalistic environment in which individuals are used to sleeping. Particularly with the long-term extension of applying this to harsh conditions in military environments, more robust testing in degraded environments will be important to validate the multi-sensor approach.

Finally, not all PSG scorers were blinded to the presence of actigraphy sensors. While it is unlikely that knowledge of additional sensors impacted the standard AASM PSG scoring, it is worth recognizing that an ideal study would remove this variable in its entirety.

6 Future Work

The most immediate and necessary future work is to test these algorithms with larger population sizes. While this pilot study identified the general magnitude and directional threshold changes for optimal algorithmic performance, larger populations would allow specific numerical thresholds to be determined. With this, gathering a large enough population size that more discriminative criteria can be applied to select healthier participants would allow for broader claims about multi-sensor actigraphy and its application to the general public to be made.

This study also highlighted a unique gap— the single-sensor Cole-Kripke algorithm has not been tested across different limbs. Future studies should aim to record hand dominance and assess if our observed discrepancy between the left and right wrist are causally linked to handedness. More robust findings supporting the difference between the two would further promote the necessity of multi-sensor approaches.

Furthermore, once these algorithms have been further validated on larger cohorts, beginning to test them in more authentic environments (individuals sleeping in their rooms, or soldiers sleeping in the field) is critically necessary to make claims about these algorithms' efficacy within daily life.

In tandem with efforts to test these algorithms in multiple environments, there needs to be additional testing with lower-cost, individually built sensors as described by Intille et al. [10]. Finding a way to gather and potentially stream data from these cheap actigraphy devices would be critical for user adoption, as the friction of needing to plug and unplug four sensors on a daily basis would quickly degrade the quality of the user experience.

Finally, we mentioned promising machine learning algorithms [6, 42], which could be directly extended to use our multi-sensor dataset. We are still actively consenting patients at the time of this publication, and as such, will approach having a substantial

enough dataset that we can test the state-of-the-art ML models alongside our traditional algorithmic approaches.

The interdisciplinary nature of Sleep Detection Algorithms provides multiple vectors for future study; sleep researchers, electrical engineers, and computer scientists alike have multiple ways to contribute to the future use and adoption of multi-sensor algorithmic approaches.

7 Conclusion

Accurate total sleep detection remains a priority for both consumers and clinical practitioners alike. Fundamentally, sleep-sensing algorithms are challenged by the underlying class imbalance problem, in that algorithms that optimize for sleep often overlook and significantly misclassify the minority class of wakefulness. A naive classifier that labels every epoch as sleep achieves $80 - 90\%$ accuracy simply by exploiting the class distribution. Our approach with multi-sensor actigraphy not only attacks the CI problem directly through the implementation of algorithms more biased towards wake-detection, but also addresses the inherent brittleness of single-sensor setups, which can be subject to high variance between limbs.

Our initial hypothesis— that multi-sensor fusion would improve wake classification compared to single-sensor algorithms— was partially supported. While Multi-sensor variants did not consistently outperform single-sensor variants, they performed similarly while mitigating the significant downside risk of inter-wrist variance. We did not initially expect robustness to be the primary contribution, though this arguably has greater practical significance.

Threshold recalibration is necessary to account for the multi-sensor approach, but further optimization with individual algorithmic parameters is less impactful. Higher-level consensus algorithms, particularly those that are biased towards wake-detection, outperform algorithms that fuse data streams at lower levels.

With larger population validation, efforts to reduce the cost of the experimental setup, and further validation in robust environments, multi-sensor actigraphy presents a cheap, disposable, operationally safe sleep monitoring capability that meets both military and consumer demands.

Acknowledgments

Special thanks to Dr. Richard Millman, Dr. Sohaib Ansari, and Ms. Diane Incerpi at the Brown University Health Sleep Lab. Your contributions over the past year and a half to facilitate both the administrative process and physical administration of our study have been critical; we could not have done this without you. Thank you to Myan Nguyen and Kevin Yang of the Brown University Computer Science department for your efforts in exploring and implementing multi-sensor machine learning approaches. Finally, thank you to Dr. Gray, Dr. Webb, and Dr. Huang for your continued advice and guidance throughout these two years of research.

References

[1] Rajaraman, S., Gribok, A.V., Wesensten, N.J., Balkin, T.J., Reifman, J.: An improved methodology for individualized performance prediction of sleep-deprived individuals with the two-process model. Sleep **32**(10), 1377–1392 (2009) https://doi.org/10.1093/sleep/32.10.1377

[2] Paquet, J., Kawinska, A., Carrier, J.: Wake detection capacity of actigraphy during sleep. Sleep **30**(10), 1362–1369 (2007) https://doi.org/10.1093/sleep/30.10.1362

[3] Khan, A.A., Chaudhari, O., Chandra, R.: A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications **244**, 122778 (2024) https://doi.org/10.1016/j.eswa.2023.122778

[4] Marino, M., Li, Y., Rueschman, M.N., Winkelman, J.W., Ellenbogen, J.M., Solet, J.M., Dulin, H., Berkman, L.F., Buxton, O.M.: Measuring sleep: Accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. Sleep **36**(11), 1747–1755 (2013) https://doi.org/10.5665/sleep.3142

[5] Fuster-Garcia, E., Bresó, A., Martínez-Miranda, J., Rosell-Ferrer, J., Matheson, C., García-Gómez, J.M.: Fusing actigraphy signals for outpatient monitoring. Information Fusion **23**, 69–80 (2015) https://doi.org/10.1016/j.inffus.2014.08.003

[6] Logacjov, A., Bach, K., Mork, P.J.: Long-term self-supervised learning for accelerometer-based sleep–wake recognition. Engineering Applications of Artificial Intelligence **141**, 109758 (2025) https://doi.org/10.1016/j.engappai.2024.109758

[7] Empatica Store: E4 wristband (n.d.). https://store.empatica.com/products/e4-wristbands?srsltid=AfmBOorxrUuUWaWBIXgzY-QmZPruvEe3o7mH6ev6d5Ki3uAQZW8w00ST&variant=17719039950953

[8] Sathyanarayana, A., Joty, S., Fernandez-Luque, L., Ofli, F., Srivastava, J., Elmagarmid, A., Arora, T., Taheri, S.: Sleep quality prediction from wearable data using deep learning. JMIR mHealth and uHealth **4**(4) (2016) https://doi.org/10.2196/mhealth.6562

[9] Pereira, D.d.L.O.D.: Sleep at the wheel: Wearable sensors for detection of drowsy driving. Dissertation, University of Porto (2021)

[10] Intille, S.S., Albinali, F., Mota, S., Kuris, B., Botana, P., Haskell, W.L.: Design of a wearable physical activity monitoring system using mobile phones and accelerometers. In: 2011 Annual International Conference of the IEEE

Engineering in Medicine and Biology Society, pp. 3636–3639 (2011). https://doi.org/10.1109/IEMBS.2011.6090611

[11] Patel, A.K., Reddy, V., Shumway, K.R., *et al.*: Physiology, Sleep Stages, Updated 2024 jan 26 edn. StatPearls Publishing, Treasure Island (FL) (2024). Accessed: April 10, 2025. https://www.ncbi.nlm.nih.gov/books/NBK526132/

[12] Elmenhorst, E.-M., Elmenhorst, D., Luks, N., Maass, H., Vejvoda, M., Samel, A.: Partial sleep deprivation: Impact on the architecture and quality of sleep. Sleep Medicine **9**(8), 840–850 (2008) https://doi.org/10.1016/j.sleep.2007.07.021

[13] Killgore, W.D.: Effects of sleep deprivation on cognition. Progress in Brain Research **185**, 105–129 (2010) https://doi.org/10.1016/B978-0-444-53702-7.00007-5

[14] Kuosmanen, E., Visuri, A., Kheirinejad, S., Berkel, N., Koskimäki, H., Ferreira, D., Hosio, S.: How does sleep tracking influence your life? experiences from a longitudinal field study with a wearable ring. Proceedings of the ACM on Human-Computer Interaction **6**(MHCI), 185–118519 (2022) https://doi.org/10.1145/3546720

[15] Watson, N.F., Badr, M.S., Belenky, G., Bliwise, D.L., Buxton, O.M., Buysse, D., Dinges, D.F., Gangwisch, J., Grandner, M.A., Kushida, C., Malhotra, R.K., Martin, J.L., Patel, S.R., Quan, S.F., Tasali, E.: Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society. Sleep **38**(6), 843–844 (2015) https://doi.org/10.5665/sleep.4716

[16] Shrivastava, D., Jung, S., Saadat, M., Sirohi, R., Crewson, K.: How to interpret the results of a sleep study. J Community Hosp Intern Med Perspect **4**(5), 24983 (2014) https://doi.org/10.3402/jchimp.v4.24983

[17] Augner, C.: Associations of subjective sleep quality with depression score, anxiety, physical symptoms and sleep onset latency in students. Central European Journal of Public Health **19**(2), 115–117 (2011). PMID: not listed

[18] Cole, R.J., Kripke, D.F., Gruen, W., Mullaney, D.J., Gillin, J.C.: Automatic sleep/wake identification from wrist activity. Sleep **15**(5), 461–469 (1992) https://doi.org/10.1093/sleep/15.5.461

[19] Sadeh, A., Sharkey, M., Carskadon, M.A.: Activity-based sleep-wake identification: An empirical test of methodological issues. Sleep **17**(3), 201–207 (1994) https://doi.org/10.1093/sleep/17.3.201

[20] PsychDB: Polysomnography (PSG). Accessed: April 15, 2025 (2024). https://www.psychdb.com/neurology/polysomnography#indications

[21] American Academy of Sleep Medicine: Sleep Study. https://sleepeducation.org/patients/sleep-study/. Reviewed by Rafael Sepulveda, MD and Seema Khosla, MD. Accessed: 2026-02-19 (2020)

[22] Latshang, T., Mueller, D., Lo Cascio, C., Stöwhas, A., Stadelmann, K., Tesler, N., Achermann, P., Huber, R., Kohler, M., Bloch, K.: Actigraphy of wrist and ankle for measuring sleep duration in altitude travelers. High Altitude Medicine & Biology **17** (2016) https://doi.org/10.1089/ham.2016.0006

[23] Kamdar, B.B., Kadden, D.J., Vangala, S., Elashoff, D.A., Ong, M.K., Martin, J.L., Needham, D.M.: Feasibility of continuous actigraphy in patients in a medical intensive care unit. American Journal of Critical Care **26**(4), 329–335 (2017) https://doi.org/10.4037/ajcc2017660 . Available in PMC 2018 January 01

[24] Hoque, E., Dickerson, R.F., Stankovic, J.A.: Monitoring body positions and movements during sleep using wisps. In: Wireless Health 2010. WH '10, pp. 44–53. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1921081.1921088 . https://doi.org/10.1145/1921081.1921088

[25] Adams, J.L., Dinesh, K., Xiong, M., Tarolli, C.G., Sharma, S., Sheth, N., Aranyosi, A.J., Zhu, W., Goldenthal, S., Biglan, K.M., Dorsey, E.R., Sharma, G.: Multiple wearable sensors in parkinson and huntington disease individuals: A pilot study in clinic and at home. Digital Biomarkers **1**(1), 52–63 (2017) https://doi.org/10.1159/000479018

[26] Yuan, H., Plekhanova, T., Walmsley, R., Reynolds, A.C., Maddison, K.J., Bucan, M., Gehrman, P., Rowlands, A., Ray, D.W., Bennett, D., *et al.*: Self-supervised learning of accelerometer data provides new insights for sleep and its association with mortality. NPJ digital medicine **7**(1), 86 (2024)

[27] Brønd, J., Andersen, L., Arvidsson, D.: Generating actigraph counts from raw acceleration recorded by an alternative monitor. Medicine and science in sports and exercise **49** (2017) https://doi.org/10.1249/MSS.0000000000001344

[28] Berry, R.B., Quan, S.F., Abreu, A.R., *et al.*: The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications, Version 3 edn. American Academy of Sleep Medicine, Darien, IL (2023)

[29] Neishabouri, A., Nguyen, J., Samuelsson, J., Guthrie, T., Biggs, M., Wyatt, J., Cross, D., Karas, M., Migueles, J.H., Khan, S., Guo, C.C.: Quantification of acceleration as activity counts in actigraph wearable. Scientific Reports **12**(1), 11958 (2022) https://doi.org/10.1038/s41598-022-16003-x

[30] Gao, C., Li, P., Morris, C.J., Zheng, X., Ulsa, M.C., Gao, L., Scheer, F.A.J.L., Hu, K.: Actigraphy-based sleep detection: Validation with polysomnography and comparison of performance for nighttime and daytime sleep during simulated shift work. Nature and Science of Sleep **14**, 1801–1816 (2022) https://doi.org/10.

2147/NSS.S373107

[31] Cheung, J., Leary, E.B., Lu, H., Zeitzer, J.M., Mignot, E.: PSG validation of minute-to-minute scoring for sleep and wake periods in a consumer wearable device. PLOS ONE **15**(9), 0238464 (2020) https://doi.org/10.1371/journal.pone.0238464

[32] ActiGraph, LLC: ActiLife 7 User's Manual. ActiGraph, LLC (Ametris), (2025). ActiGraph, LLC (Ametris). Version G, released April 15, 2025. https://6407355.fs1.hubspotusercontent-na1.net/hubfs/6407355/User%20Manuals/ActiLife%207%20Users%20Manual.pdf

[33] Petkov, D.: actigraph.sleepr: Python port of sleep detection algorithms from ActiGraph's R package. https://github.com/dipetkov/actigraph.sleepr. Accessed: 2025-04-22 (2021)

[34] Quante, M., Kaplan, E.R., Cailler, M., Rueschman, M., Wang, R., Weng, J., Taveras, E.M., Redline, S.: Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. Nature and Science of Sleep **10**, 13–20 (2018) https://doi.org/10.2147/NSS.S151085

[35] Gravina, R., Alinia, P., Ghasemzadeh, H., Fortino, G.: Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. Information Fusion **35**, 68–80 (2017) https://doi.org/10.1016/j.inffus.2016.09.005

[36] Kittler, J.: Combining classifiers: A theoretical framework. Pattern Analysis and Applications **1**, 18–27 (1998) https://doi.org/10.1007/BF01238023

[37] Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recognition **38**(12), 2270–2285 (2005) https://doi.org/10.1016/j.patcog.2005.01.012

[38] Martin, J., Hakim, A.: Wrist actigraphy. Chest **139**(6), 1514–1527 (2011) https://doi.org/10.1378/chest.10-1872

[39] Cawley, G.C., Talbot, N.L.C.: Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. Journal of Machine Learning Research **8**, 841–861 (2007)

[40] Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[41] Chicco, D., Jurman, G.: The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining **16**(1), 4 (2023) https://doi.org/10.1186/s13040-023-00322-4

[42] Logacjov, A., Herland, S., Ustad, A., Bach, K.: SelfPAB: Large-scale pre-training

on accelerometer data for human activity recognition. Applied Intelligence (2024) https://doi.org/10.1007/s10489-024-05322-3