Multimodal Machine Learning Framework for Soldier Fatigue Prediction

Louis Kim, Ryan Dougherty Connor Diehl Kelly Hale Andrea Webb	Hope Mango	Seth Elkin-Frankston Victoria G. Bode
The Charles Stark Draper Laboratory	Johns Hopkins University Applied Physics Laboratory	U.S. Army CCDC Soldier Center
Cambridge, MA	Laurel, MD	Natick, MA
lkim; rdougherty; cdiehl; khale; awebb@draper.com	hope.mango@jhuapl.edu	seth.elkin-frankston.civ@army.mil, victoria.g.bode.civ@army.mil

ABSTRACT

With an ever changing, fast paced, global initiative, the U.S. must adapt to the increasing operational tempo of modern warfare. The need to rapidly adapt to changing conditions has contributed to the growing number of Soldiers experiencing fatigue-related injuries. These injuries cost the military billions of dollars annually, while impacting training opportunities and reducing the number of combat-ready Soldiers that support critical missions. Existing methods of injury prevention rely on Soldiers' self-reporting their fatigue state which can be less reliable compared to objective physiological, cognitive, and physical indicators of fatigue. Additionally, lack of methods for early warning and prediction of emergent fatigue conditions hinders the military's ability to proactively intervene and reduce the likelihood of injury. Motivated by these challenges, a machine learning framework was developed to predict fatigue with individual multimodal data collected as part of the US Army supported, MASTR-E (Measuring and Advancing Soldier Tactical Readiness and Effectiveness) program.

Through the MASTR-E program, this study leveraged a subset of data collected from over 200 Soldiers during a 72-hour training exercise. Collected data included baseline physical and cognitive status, demographics, standard surveys tracking information across health, physical, social-emotional, and cognitive domains, biomarker data from blood and saliva samples, and physiological and kinematic data from wearable sensors. Using the available data, a multimodal machine learning framework was developed to process data and perform supervised learning to predict individual fatigue up to 60 minutes forward in time. Fatigue states are labeled based on acute to chronic workload ratio (ACWR) along with training impulse, which enables accurate prediction of excessive fatigue conditions that can lead to injuries. The framework enables tradeoff analysis of different data types and prediction horizons for enhanced decision support. Demonstration of the framework achieved 70% balanced accuracy when predicting 30 minutes forward, with explainable AI embedded to provide data driven actions to mitigate injuries and better assess soldier performance.

ABOUT THE AUTHORS

Louis Kim is Principal Scientist in the Machine Intelligence Group at Draper. He leads machine learning algorithms development and data analytics tasks for Department of Defense, intelligence community, as well as commercial customers, delivering high-performance prediction and classification solutions leveraging various machine learning and deep learning techniques, with applications ranging from human performance, social and behavioral computing, finance, multi-modal sensor fusion, neuroscience, and image analytics. He holds a bachelor's degree in Systems Engineering from the University of Illinois at Urbana-Champaign and a Master's in Operations Research from Massachusetts Institute of Technology.

Ryan Dougherty is Senior Member of Technical Staff in the Machine Intelligence Group at Draper. He develops analysis platforms for the Department of Defense, intelligence community, and commercial customers, leveraging machine learning, data analytics, and human centered design. His application focus areas include human performance predictions & recommendations, multi-modal sensor fusion, multi-omics analysis, and social & behavioral computing. He holds a bachelor's degree in Electrical Engineering and Computer Science from Tufts University, and a Master's in Computer Science from Tufts University.

Connor Diehl is a Senior Member of Technical Staff in the 3D and Mobile Applications Group at Draper. Working to provide situational awareness and readiness solutions to Department of Defense customers, he develops tools for human-performance prediction, situational awareness and real-time simulation. He holds a bachelor's degree in Engineering Robotics from Olin College of Engineering.

Dr. Andrea Webb is Chief Scientist of Human Centered Solutions and a Program Manager at Draper. Her work focuses primarily on human signals, specifically the quantification of human state. She is an expert in credibility assessment and additionally leads a number of programs focused on mental health assessment and cognitive and physical performance. She serves as a Program Manager for Health and Human Performance and for Digital Engineering and Information Management Systems. In addition to her technical and programmatic roles, Dr. Webb also serves as Draper's Human Protections Administrator, ensuring that all work conducted with humans and/or their data is compliant with local and federal rules and regulations, and Draper's Research Integrity Officer. Dr. Webb completed her undergraduate work at Boise State University and received her MS and PhD from the University of Utah.

Dr. Kelly Hale is Principal Engineer and Group Leader of the UX/Human Performance Group at Draper. Dr. Hale has 20+ years of experience in user-centered design, development and evaluation of systems on sponsored by efforts by the Department of Defense, Department of Homeland Security, NASA, DARPA, and IARPA. She has led multiple efforts developing physiologically-based measures of cognitive and affective states (e.g., psychological stress, workload, engagement, boredom, fear) and performance (e.g., visual search, target recognition). Dr. Hale holds bachelor's degree in Kinesiology and received her MS and PhD in Industrial Engineering from the University of Central Florida.

Hope Mango is Modeling and Simulation Analyst for the Air Combat and Strike Mission Analysis Group at Johns Hopkins University Applied Physics Laboratory. She leads data analytics tasks and mission level analysis for the Department of Defense and other defense industry leaders, delivering in-depth technical solutions for engineering advancements and acquisition of critical defense technology. Prior to joining the lab, Hope served in the United States Army as an Air Defense Officer, primarily focused on ballistic missile defense as an operator of the Patriot weapon system. She holds a bachelor's degree in Systems Engineering from the United States Military Academy and a Master's in Engineering Management from Dartmouth College.

Dr. Seth Elkin-Frankston holds the position of Cognitive Scientist at the U.S. Army DEVCOM Soldier Center. Prior to this role, he spent 5 years in industry, successfully leading numerous Government-funded research programs across diverse disciplines. His primary focus was on neuroscience, cognitive science, and behavioral research. Currently, Dr. Elkin-Frankston's research explores advancements in human assessment science and technology. He aims to analyze, model, predict, and enhance future performance outcomes. Dr. Elkin-Frankston earned his doctorate in Neurobiology from Boston University School of Medicine, specializing in the Department of Anatomy and Neurobiology. During his graduate studies, he investigated how different brain areas collaborate to process visual information, with the goal of developing techniques for lasting improvements in visual performance.

Victoria G. Bode is Research Physiologist at the U.S. Army DEVCOM Soldier Center. Ms. Bode is the Principal Investigator for the 72 Hour Field Study within the Measuring and Advancing Soldier Tactical Readiness and Effectiveness (MASTR-E) program. As Principal Investigator, she helped develop the infiltration and exfiltration ruck march tasks, which generated data for training the fatigue prediction model for this paper. She holds a bachelor's degree in Biology from the University of New Haven and a Master's in Kinesiology with a concentration in Exercise Science from the University of New Hampshire.

Multimodal Machine Learning Framework for Soldier Fatigue Prediction

Louis Kim,
Ryan Dougherty
Connor Diehl
Kelly Hale
Andrea Webb
The Charles Stark Draper Laboratory
Cambridge, MA
lkim; rdougherty; cdiehl; khale; awebb@draper.com

Hope Mango
Johns Hopkins University Applied Physics Laboratory
Laurel, MD
hope.mango@jhuapl.edu

Seth Elkin-Frankston
Victoria Bode
U.S. Army CCDC Soldier Center
Natick, MA
seth.elkin-frankston.civ@army.mil,
victoria.g.bode.civ@army.mil

INTRODUCTION

Overuse injuries, linked to physical fatigue, are the primary contributors to lost training days and attrition among military forces worldwide (Forrest et al., 2022). The U.S. Army alone, faces a significant number of overuse injuries, resulting in the loss of near 12% of all active-duty days annually (Schwartz et al., 2018). For example, overuse injuries cost the Army over $3.7 billion a year, with $400 million of that cost accounted for in medical care and $3.3 billion in indirect costs due to limited duty days, with an average cost of $5,929 per overuse injury (Schwartz et al., 2018). The ability to accurately estimate and predict physical fatigue in a manner that enables leaders to make informed decisions and provide timely intervention should limit injury and reduce the cost associated with medical care and training time lost due to injury.

Due to the nature of work conducted, the Army has a unique need for fatigue prediction and prevention. There are many devices and sensors on the market today that collect valuable data regarding an individual’s physiological response to fatigue, however the specialized needs of military training and operations limit the number of devices or tools available that can be tailored to military requirement. First, soldiers already must account for moving or carrying high quantities of equipment during training scenarios, and the benefit of every extra ounce they carry must outweigh the cost of additional weight. Fatigue monitoring and prediction tools should not add significant weight to what a soldier is already carrying and must not get in the way of training or operational requirements. Additionally, fatigue monitoring and prediction must not be invasive. Field training and combat environments are not sterile environments. Puncturing skin or drawing blood in those environments is an unnecessary risk and consumes valuable time during an exercise or mission. Military decisions must be made effectively and efficiently, therefore, information regarding fatigue needs to be presented in a concise, timely manner that allows leaders to make informed decisions about exercise and mission requirements, all while mitigating risk of injury and accident caused by fatigue. Since military exercises and operations already assume elevated levels of risk, fatigue reporting and prediction must be accurate and provided early enough for a leader to plan and execute training or operations in a way that produces the most combat effective soldiers. Accurate prediction of fatigue will help leaders better understand their soldiers and in turn allow them to plan training and rest cycles accordingly.

Existing Research and Commercial Solutions

The focus of this effort is to support the needs of the military by identifying accurate, transparent, and interpretable analytics. These analytics facilitate decision-making based on predictive capabilities using current wearable technology. Focusing on these elements, an analysis was conducted on existing individual wearable devices currently on the market to determine whether their capabilities could meet the needs of reporting and predicting fatigue in a military training or combat environment.
We identified several lightweight sensors that can measure different physiological responses that can help gauge fatigue. The first thing to note is that approximately 5% of consumer wearable technologies have been fully and formally validated regarding their accuracy (Hulin et al., 2013). While heartrate monitoring is accurate in many

sensors, many of the other measurements reported, displayed, or calculated are inaccurate. For example, a study of eleven different wearable devices concluded that all the wearable devices had an inaccurate assessment of overall energy expenditure (Pasadyn et al., 2019). All four of the market prominent fitness trackers, Fitbit, Garmin, Whoop, and Polar, offer some measures of physical exertion that can help quantify physical fatigue. However, they do not provide transparency into how the measures are calculated, how their methods were validated, and what accuracies were obtained through method validations. Regarding the accuracies of those measures, all have a disclaimer stating that the provided measures are intended to be an estimation and may not be accurate. In addition to the lack of accuracy and transparency, they either do not provide interpretability into how certain measures occurred or provide a generic interpretability that may not hold true for a certain individual. Furthermore, while all the devices have means of giving real-time measurements of physiological metrics, currently no devices on the market predict a future level of physical fatigue.

Research studies conducted to integrate the use of wearable technology to predict fatigue using various machine learning techniques have substantial limitations regarding their data and fatigue prediction capabilities. Pinto-Bernal et al. (2021) had some success estimating fatigue within controlled exercise conditions, yet relied on blood lactate measures captured via blood draw with Inertial Measurement Unit (IMU) data. The study was limited to classifying immediate fatigue states on controlled treadmill walks and runs after manual inducement of physical fatigue through performing physical exercises such as high knees, jumping jacks, squats, and short runs. Kathirgamanathan et al. (2022) also utilized IMUs to classify fatigue in runners to show whether feature selection needed to be individualized or if it could be generalized when it comes to estimating fatigue and concluded that generic feature selection can perform well in estimating fatigue in runners. Although high accuracies were obtained for both individualized and generalized models, the models were limited to classifying immediate fatigue states. Luo et al. (2020) assessed cognitive and physical fatigue using wearable sensors where they had a group of subjects wear sensors for a week while conducting their “normal” activities in addition to completing a daily fatigue questionnaire. While this study was conducted outside of a laboratory environment and assessed normal activity and behavior in classifying self-reported fatigue, presented models were limited to classifying the immediate self-reported fatigue at daily time scale. Furthermore, the models used self-reported fatigue as labels, which can be inherently inaccurate due to individual biases and subjectivity in the self-reported fatigue. Kim et al. (2022) collected GPS sensor data during soccer training and match sessions and post-session self-reported Rate of Perceived Exertion (RPE) to develop deep neural network models that can estimate the post-session RPE using the sensor data. While the study leveraged activation mapping techniques to highlight features that impacted the prediction to provide some interpretability of the results, the models had the same limitations of classifying the immediate self-reported fatigue after each event. Russell et al. (2021) conducted a study utilizing a single IMU sensor in a field environment to predict fatigue with a deep learning model demonstrating that a single wearable sensor could be used in conjunction with a neural network model to determine fatigue in an outside, unstructured environment. This study is supportive of fatigue prediction with a wearable sensor outside of a controlled laboratory environment, however, the study only included one participant and predicted a near-immediate time horizon of few minutes.

Table 1. Indication of whether five key criteria are met in the relevant past studies.

	Data collected in real environment	Sufficient data (>30 subjects) used	Wearable sensor data used	Interpretability	Prediction capability
Pinto-Bernal et al. (2021)	No	No	Yes	No	No
Kathirgamanathan et al. (2022)	Yes	No	Yes	Yes	No
Luo et al. (2020)	Yes	No	Yes	Yes	No
Kim et al. (2022)	Yes	No	Yes	Yes	No
Russell (2021)	Yes	No	Yes	No	Yes

Overall, while there’s potential for further development of predictive capabilities, current studies are limited in terms of the amount of data leveraged in developing their models and offering predictions that are supported with interpretations. As outlined in Table 1, five key criteria (data collected in real environment, sufficient data used, wearable sensor data used, interpretability, and prediction capability) are desired to develop accurate and dependable models, which must be addressed to provide decision support for high impact operations. This work aims to address gaps of previous research, where no study directly addressed all five key criteria.

Our Proposed Fatigue Prediction Framework Addressing Aforementioned Limitations

The data collected and the framework developed in this work aim to deal with the limitations of prior studies and existing products to effectively predict physical fatigue. The data collection leveraged for this study was conducted as part of a field study on two separate occasions, with over 100 total participants and 200 metrics for each individual participant. While data processing and filtering substantially reduced the final set of data used for the model development in this effort, the extensive data collection enabled generation of sufficient quality-controlled data to develop robust prediction models.

The framework developed provides three main contributions to address the limitations of the current commercial solutions and research. First, the framework enables development of physical fatigue prediction models that could predict an individual's physical fatigue classification up to 60 minutes in the future. Such prediction capability will provide leaders ability to intervene prior to injury occurring. Second, the framework leverages the heart rate signals collected from wearables to generate more objective fatigue states instead of relying on self-reports, which enables continuous estimation and prediction of excessive physical fatigue conditions and removes the possibility of biases and flawed self-reports from yielding inaccurate fatigue labels. Lastly, the framework provides both the population- and individual-level interpretations of the fatigue predictions. The interpretations allow identification of critical data sources for effective prediction and help evaluate whether the development of deployable prediction models is feasible using data from wearables, which is relatively easier to collect. Additionally, the individual-level interpretations can help devise personalized mitigation and training plans to limit the injuries from excessive physical fatigue.

METHOD

Data Collection

Data for this study was taken from a subset of data collected as part of the US Army led, Measuring and Advancing Soldier Tactical Readiness and Effectiveness (MASTR-E) program 72 Hour Field Study (72-HFS). All data collection activities were conducted in accordance with the approved protocol requirements. The baseline data collection sessions occurred at Fort Liberty, NC, where Soldiers completed a battery of physical and cognitive measures and questionnaires. Data was used to provide a measure of baseline health, social-emotional, physical and cognitive metrics prior to participation in the 72-hour field study. Approximately, one month following the baseline data collection, the field study occurred at Fort Devens, MA, where Soldiers participated in a 72-hour field exercise designed to assess both individual and squad-level mission and combat performance. The field study consisted of a series of pre-mission, mission, and post-mission events. The pre- and post-missions included biomechanical assessments and captured saliva and blood samples to evaluate changes in physiological and general health status after the mission. The mission consisted of 3 days of operational exercises including individual and team shooting scenarios, tactical stress marksmanship assessment, reconnaissance, and infiltration and exfiltration ruck marches, where equipment fit and wearable sensor evaluation, anthropometry, and surveys were conducted before and after each activity to ensure that the sensors were correctly equipped and functional to capture changes in social, cognitive, and affect states. In addition to the standard tactical gear, the Polar® Team Pro device was worn on the chest directly over the skin using the designated chest strap to measure physiology, and Opal IMU sensors from APDM Wearable Technologies were mounted to the helmet, rifle, torso, and ankles to track body, head, and weapon movements. Table 2 summarizes the subset of collected data used in this work.

Table 2. Summary of the subset of collected data used, organized by event, device / collection method, frequency, and collected information.

Event	Device / Collection Method	Frequency	Collected Information
Baseline data collection	Questionnaires	Once	Age Weight Height Years in military Tobacco usage Alcohol consumption Caffeinated beverages consumption Other drink consumption (soda, etc)
	Physical tests	Once	ACFT (Army Combat Fitness Test) total score

			Total sprint time Sprint drag carry total
Pre- and Post-missions	Blood and saliva samples	Before and after mission	CREQ UROQ LACQ CARQ HISQ COR-Q EST-Q TEST-Q DHEA-Q
	Surveys	Before and after activity	Visual Analog Scale of Fatigue (VAS Fatigue) Rate of Perceived Exertion (RPE)
Wearables during mission event	Polar Pro	10 Hz	Heart rate Interbeat interval Speed Distance Acceleration Running cadence
		2 Hz	Heartbeat interval
		1 Hz	Latitude Longitude
	Opal IMU	256 Hz	Torso lean ML (mediolateral) Torso lean AP (anterior-posterior)
		2 Hz	Step length Step width Stride speed Foot yaw
		2 Hz	Toe off to heal strike time delta

For the purposes of this study, data captured during the infiltration and exfiltration ruck marches was used as input to the fatigue prediction model. The ruck activity was selected due to activity duration (approximately one hour for each ruck) and was an example of continuous movement where fatigue could likely occur for some participants. The sample set included a total of 113 participants that have completed both ruck marches. Of that 113, 33 participants demonstrated fatigue during the ruck (in at least one of the infiltration or exfiltration ruck) defined as having at least one ACWR score of 1.3 (explained in following sections). This subset of 33 participants was used to develop the fatigue prediction models in order to have a balanced dataset for the models to better learn and predict the patterns of fatigue.

Fatigue Prediction Framework

The main purpose of the proposed fatigue prediction framework (Figure 1) is twofold: 1) enable efficient and effective development of fatigue prediction models leveraging any combination of extensive multimodal data; and 2) improve upon existing fatigue classification and limitations of predictive analytics by providing capability to generate objective fatigue labels, predict future fatigue states, and interpret prediction outcomes. Motivated by the problem of Soldier overuse injuries in the Army, the framework is set up to address the key question: "At any given time during training or mission, can fatigue be predicted ahead of time to allow for intervention to mitigate potential injuries?"

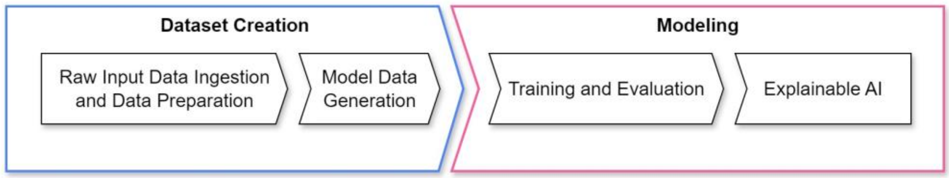

Figure 1. High-level overview of the fatigue prediction framework.

The framework is primarily structured into the dataset creation and modeling stages. The dataset creation stage consists of raw input data ingestion and preparation and model data generation processes to transform multimodal data from various inputs sources into noise-filtered, normalized, structured parings of features and future fatigue states (fatigue labels), which will allow efficient development and evaluation of prediction models in the subsequent modeling stage. During the modeling stage, prediction models are trained and evaluated for varying prediction horizons, and once the final models are identified, explainable artificial intelligence (AI) methods are applied to generate interpretation of the prediction models. Each step will be described in detail in the following sections.

Dataset Creation: Raw Input Data Ingestion and Data Preparation

The pipeline was designed to ingest all the collected data sources from Table 2 and allow the user to select applicable subsets of the data to meet their modeling needs in a later step in the process. Raw data collected in various formats (.h5, .csv, .txt, etc.) was ingested using a combination of open source and custom written MATLAB and Python scripts. Once ingested, non-sensor data collected from the baseline, pre- and post-events were stored in a tabular format, organized by participant. Sensor data collected during the mission events were truncated to match the start and end times of the ruck march activities.

Dataset Creation: Model Data Generation

In the model data generation step, analyst selected datasets outlining features of interest and prediction horizon (how far ahead to predict) were used to drive the following processes: missing data handling, feature extraction, label generation, and data selection and split. This generated observations that could be used for training and testing of machine learning models. Figure 2 shows a general overview of how an observation is generated.

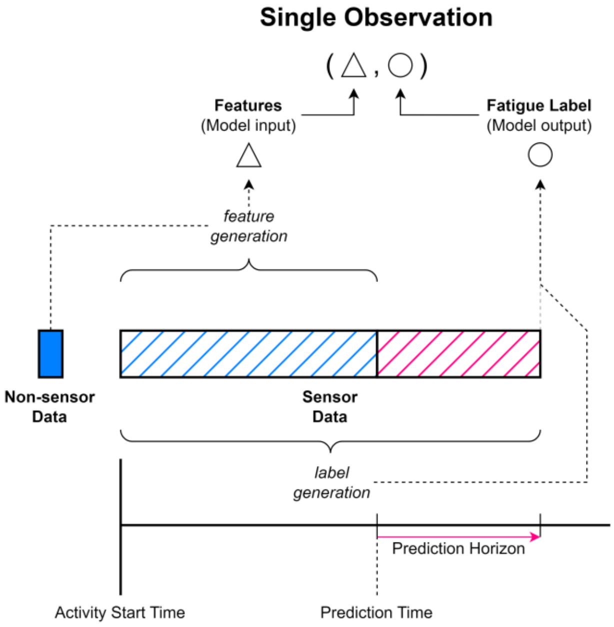

Figure 2. A general overview of how a single observation is generated. From a single participant's data, several observations are generated as the prediction time varies.

Missing data handling Since the sensor data was collected in the wild, frequent data drops were observed which could be due to various reasons such as sensor malfunctions or poor contact between the skin and the chest-worn Polar device as the participant got sweaty during the mission activities. To handle missing data, a combination of interpolation and data removal methods were used. In case of short periods of data drops (less than 70% of prediction horizon length), interpolation methods were used to infer missing data gaps using surrounding known data. Since interpolating long periods of gaps can create data quality issues subsequently resulting in poor model reliability, long gaps (equal to or greater than 70% of prediction horizon length) of data and subsequent data following the long gaps were removed from the analysis. In choosing between interpolation and removal, the two main factors that were considered included quantity of missing data and preservation of fatigue observations. Based on the definition of fatigue used in this study, 33 out of 113 participants experienced fatigue and were thus at risk of possible injury. Preserving those limited observations were essential in developing effective prediction models. Equation 1 is a scoring criteria used determine the ideal data handling rules for datasets generated for model development. f denotes the percent of fatigued observations in the resulting dataset after missing data handling, and d denotes the percent of interpolated and missing data. Ideal dataset would have perfectly balanced ratio of fatigued and non-fatigued observation ($f = 50$) and have no data removed or interpolated ($d = 100$), yielding a perfect score of 1. The dataset scoring criteria for varying interpolation gap limits as a function of the prediction horizon length were evaluated, and the 70% gap limit was chosen as it yielded the best score. Depending on the prediction horizon, between 0 to 19% of the data was dropped, and 4 to 6% of the data was interpolated.

$$s(f,d) = \frac{100 - |50 - f|}{100} \times 0.67 + \frac{100 - d}{100} \times 0.33 \qquad (1)$$

Feature extraction In order to sync time series data collected in various frequencies as stated in Table 1, mean values over a 10 second window were calculated. To capture trends over a longer time horizon, the following list of time series features were calculated using the available data up to the time at which prediction was made:

- Mean
- Median
- Standard deviation
- Root mean squared
- Mean change
- Mean value of central approximation of second derivative
- Kurtosis
- Skewness
- Autocorrelation
- Minimum
- Maximum
- Range
- Interquartile Range
- Polynomial fit (slope and intercept)

Additionally, changes in mean and median values between the four quarters of available sensor data were calculated to capture trends across larger blocks of time. For example, if a 30-minute prediction was to be made 10 minutes into the ruck march, each 10-minute worth of sensor data is used to calculate above features, and mean and median differences between the quarters of the 10-minute window are calculated as features.

Label generation Although standard survey measures of fatigue, VAS Fatigue and RPE, were collected before and after each activity, these were not used to indicate fatigue due to two main limitations: 1) These measures are self-reported and subjective in nature often containing individual biases, which can yield inaccurate assessment. 2) Since the measures are only collected before and after the activity, it is difficult to infer a participant's fatigue level during the activity. To address the limitations posed by generating fatigue labels from self-reported survey responses, Training Impulse (TRIMP) and Acute to Chronic Workload Ratio (ACWR) were used based on the data from the wearable sensors to assess fatigue. The TRIMP method was originally developed by Banister (1991) to quantify the training load of an athlete using heart rate reserve and the duration of the exercise. There are several variations of the TRIMP metric, one of which is the exponential TRIMP metric defined in equation 2, where D is the duration in minutes at a particular heart rate, HR_r is the current heart rate divided by the maximum heart rate, the maximum heart rate is defined by 220 minus the participant's age, and y is the HR_r multiplied by 1.92 for men and 1.68 for women. Exponential TRIMP was used here because it accounts for the constant used to estimate the lactate to heart rate ratios in men and women (the constants used in y variable). It has been shown that lactate increases with heart rate (Ohkuwa et al., 2009), and that lactate reflects muscle metabolism which conveys fatigue related information to the brain when muscle fatigue occurs (Ishii & Nishida, 2013).

$$TRIMP^{exp} = \sum (D \times HR_r \times 0.64e^{y}) \qquad (2)$$

Because TRIMP does not consider an individual's level of fitness it may be more reflective of intensity or exertion. However, by incorporating TRIMP into the ACWR, a better understanding of an individual's fitness and fatigue can be achieved, which has been associated with injury prediction and prevention (Gabbett, 2016). ACWR is defined as the ratio between acute and chronic workload, shown in equation 3. Acute workload (equation 4) is the TRIMP calculated over a singular interval, where t is the time at which the acute workload is calculated, and chronic workload (equation 5) is defined as a rolling average of all previous acute workloads starting from the beginning of the ruck march activity, where n is the number of intervals. Figure 3 shows how acute and chronic workloads are defined with respect to individual TRIMP observations.

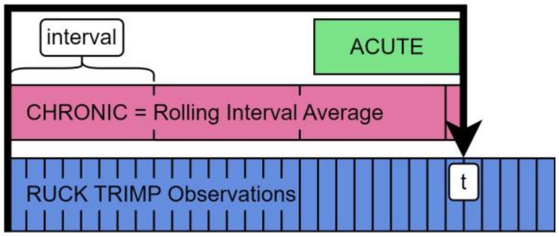

Figure 3. Acute and chronic workload calculations at time t. Acute workload is the sum of the last interval TRIMP values and chronic workload (pink) is the average of all previous intervals since the start of the activity.

$$ACWR = \frac{Acute\ Workload}{Chronic\ Workload} \qquad (3)$$

$$Acute\ Workload = \sum_{D=t-1}^{D=t} (D \times HR_r \times 0.64e^y) \tag{4}$$

$$Chronic\ Workload = \frac{\sum_{w=1}^{w=n} Acute\ Workload}{n} \tag{5}$$

Gabbett (2016) and Seshadri (2019) provided a guidance on interpreting ACWR with respect to injury risk based on data collected from three different sports (cricket, Australian football and rugby league), stating that ACWRs within the range of 0.8-1.3 is optimal “sweet spot” for training and reducing the risk of injury while ACWRs greater than 1.5 represents over-fatigued “danger zone” where injury risk can be up to 4 times greater. Hulin et al. (2013) conducted a study regarding ACWR and rate of injury with a group of elite cricket players and found that when ACWR was below 1, the likelihood of injury for the players over the next 7 days was only 4%; however, when the ACWR was greater than or equal to 1.5, the risk of injury was 2 to 4 times greater. The most common ratio for ACWR is the 7:28 day ratio, however, a benefit of ACWR is that it can scale to fit different sports and training schedules (Griffin et al, 2020). In this effort, ACWR was calculated with 20-minute acute workload intervals over the duration of the ruck march, and an ACWR value threshold of 1.3 was used to generate binary “fatigued” and “non-fatigued” labels as the target variable to predict.

Data selection and split Once the method was applied to the collected data to generate fatigue labels, 33 participants who have experienced fatigue during the ruck march activities were selected so that models can better learn to predict fatigue patterns, as having highly imbalanced dataset with mostly non-fatigue participants and observations can inhibit proper learning of fatigue indications. From the selected participant, an 80/20 strategized split method was performed to generate train and test sets for model training and evaluation.

Modeling

Once the model data was generated, a series of model training and evaluation steps were performed to develop the best performing prediction model for four different prediction horizons predicting 5, 15, 30, and 60 minutes forward in time. A random forest classifier model was developed using the scikit-learn Python module due to its versatility in handling both continuous and categorical variables and demonstrated success in various binary classification and prediction problems (Pedregosa et al., 2011). Random search method using 5-fold cross validation over the training set of data was performed to tune hyperparameters of the random forest classifier to optimize for balanced accuracy (arithmetic mean of recall and specificity). While accuracy, balanced accuracy, recall, precision, specificity, and the area under the Receiver Operating Characteristics curve (AUC ROC) were calculated to evaluate the overall performance of prediction models, the balanced accuracy metric was used as the main criteria for selecting the best performing model due to the label imbalance in the dataset. In addition to the development of prediction models forecasting the varying prediction horizons, a sensitivity analysis was conducted to investigate prediction performance sensitivity to different input data sources.

To provide explanation and reasoning of the developed prediction models, model-agnostic permutation importance (Fisher, 2019), partial dependency plots, and Shapley values (Shapley, 1951) were implemented and applied to the models. The permutation importance method permutes values of a particular feature, holding all other features constant, recomputing predictions, and then moving on to the next feature. Features where the prediction differs significantly from the original data indicate that the feature has more importance in driving the ultimate prediction. While permutation importance describes what features in a model are important, partial dependency plots show how the predicted probability for the model change as a result of a small perturbation in a feature’s value by plotting the predicted probability response relative to the feature values. From an explainability perspective, this provides an understanding of a particular relationship between the feature and prediction. Lastly, Shapley values intuitively take a game theoretic approach to take a black box prediction and allocate a portion of the predicted probability that is estimated by the ML algorithm to each of the features. Whereas the permutation importance and partial dependency plots estimated values for each feature, it did so holding all other values constant; with Shapley values, value allocation to the overall prediction is dynamic with respect to every other variable. As feature values change from Soldier A to Soldier B, Shapley values recalculate to identify how each feature contributes to the predicted probability. In this way, a granular view of an observation can be made. By utilizing these methods, key indicators of fatigue prediction and how they impact the prediction can be understood, and person to person variation can be revealed.

RESULTS

Table 3 summarizes the statistics on the number of fatigued and not fatigued observations in train and test datasets for varying horizons after the dataset creation stage. The observations are extracted using the 33 participants' data from both the infiltration and exfiltration marches, and 54 total features were extracted for each observation. Splitting data by participant ensured that the observations in the test set are from approximately 7 distinct participants not included in the train set. Depending on the prediction horizon, between 4.4% to 9.0% of the observations are fatigued observations, and the 80/20 stratified train/test split ensured that the fatigued to not fatigued label ratio was preserved. While using all participants' data can yield a greater number of total observations, including the rest of the participants that have not experienced fatigue (80 out of the 113) leads to greater imbalance between the fatigued and not fatigued observations, which can make the fatigue prediction more challenging. In this work, the data selection method of choosing the participants that have experienced fatigue limited the label ratio imbalance to support the development of effective fatigue prediction models.

Table 3. Statistics on the number of observations for varying prediction horizons.

Prediction Horizon	Number of Observations		Observation Split			
			Train		Test	
	Fatigued (%)	Not Fatigued (%)	Fatigued (%)	Not Fatigued (%)	Fatigued (%)	Not Fatigued (%)
5 minutes	61 (4.4%)	1329 (95.6%)	54 (4.8%)	1065 (95.2%)	7 (2.6%)	264 (97.4%)
15 minutes	78 (5.5%)	1343 (94.5%)	64 (5.6%)	1087 (94.4%)	14 (5.2%)	256 (94.8%)
30 minutes	78 (6.1%)	1211 (93.9%)	64 (6.3%)	959 (93.7%)	14 (5.3%)	252 (94.7%)
60 minutes	93 (9.0%)	938 (91.0%)	76 (9.2%)	749 (90.8%)	17 (8.3%)	189 (91.8%)

Table 4 summarizes the prediction performance results of the final prediction models evaluated on the test data for the four prediction horizons. A combination of randomized search and grid search methods using 5-fold cross validation was performed using the train data to optimize the random forest classifier hyperparameters in terms of maximizing the balanced accuracy metric. Once the best performing model was identified for each prediction horizon, the final model was evaluated against the test data to capture the prediction performance results. The final model training and evaluation step was repeated 50 times for different stratified splits. Average from the 50 runs and corresponding 95% confidence interval are shown in Table 4.

Table 4. Prediction performance results.

Prediction Horizon	Accuracy*	Balanced Accuracy*	Recall*	Precision*	Specificity*	AUC ROC*
5 minutes	0.87 (± 0.01)	0.84 (± 0.03)	0.81 (± 0.05)	0.25 (± 0.03)	0.87 (± 0.01)	0.93 (± 0.02)
15 minutes	0.77 (± 0.02)	0.74 (± 0.02)	0.71 (± 0.05)	0.16 (± 0.01)	0.78 (± 0.02)	0.85 (± 0.02)
30 minutes	0.70 (± 0.02)	0.70 (± 0.03)	0.70 (± 0.06)	0.13 (± 0.01)	0.70 (± 0.02)	0.78 (± 0.02)
60 minutes	0.76 (± 0.02)	0.67 (± 0.03)	0.55 (± 0.06)	0.21 (± 0.02)	0.78 (± 0.02)	0.72 (± 0.03)

*All measures +/- 95% Confidence Interval

Balanced accuracy of 0.84 is achieved for predicting fatigue 5 minutes forward in time. As the prediction horizon is increased to 15, 30, and 60 minutes, balanced accuracy drops to 0.74, 0.70, and 0.67, respectively. This decreasing trend in performance makes intuitive sense as predicting longer time horizon becomes more challenging. The models show lower precisions for all prediction horizons compared to other metrics. Limited fatigued observations available in the data and the large imbalance between the fatigued and not fatigued observations limit the models' ability to precisely learn future fatigue patterns without falsely identifying not fatigued observations, inevitability yielding the low precision metrics. More fatigued observations and balanced data can help improve precision. In practice, the ability to predict fatigue states as far as 30 minutes into the future could provide leaders time to take preventive actions to limit potential injuries that can occur from excessive physical fatigue. For the 30-minute prediction horizon, the model can correctly predict 70% of all fatigue and non-fatigue observations, demonstrating the predictive power of the developed model.

For the sensitivity analysis and explainability results, the 30-minute prediction horizon model was used as it offers a good balance between early prediction capability and effective prediction performance. Performance sensitivity to varying input data sources are captured in Table 5. In terms of balanced accuracy, using the Polar wearable sensor alone achieves comparable prediction accuracy compared to using all data sources. While using the IMU wearables alone achieves 0.67 balanced accuracy, it performs approximately 3% worse compared to using Polar alone, and addition of the IMU wearables to Polar does not yield performance improvement. Polar tracks general movement information (speed, acceleration, and running cadence) in addition to physiological signals, which can make the additional kinematics information from the IMUs less informative here.

Table 5. 30-minute prediction performance results for varying input data sources. The best score from each metric is shown in bold.

Data Source(s)	Accuracy*	Balanced Accuracy*	Recall*	Precision*	Specificity*	AUC ROC*
Wearable (Polar)	**0.73 (± 0.02)**	**0.69 (± 0.03)**	0.64 (± 0.06)	**0.14 (± 0.02)**	**0.74 (± 0.02)**	**0.79 (± 0.02)**
Wearables (IMUs)	0.68 (± 0.02)	0.67 (± 0.02)	0.64 (± 0.06)	0.12 (± 0.01)	0.69 (± 0.03)	0.76 (± 0.02)
Wearables (Polar, IMUs)	0.60 (± 0.02)	**0.69 (± 0.02)**	**0.80 (± 0.05)**	0.12 (± 0.01)	0.59 (± 0.02)	0.76 (± 0.02)

*All measures +/- 95% Confidence Interval

The permutation importance highlighted the overall importance of features for predicting fatigue, and the top 20 most important features all come from the wearable sensor data that are related to acceleration, distance, speed, torso lean, step width, and heart rate. This result is consistent with the sensitivity analysis results in that wearable sensor data is critical for effective fatigue prediction.

To understand individual differences in the predictors and how they impact the prediction, the Shapley values were calculated. Figure 5 shows the Shapley values for two subjects at the time they were predicted to fatigued. Higher stride speed (indication of faster walking movement) is both positively contributing to their fatigue predictions. On the other hand, for Subject A, alcohol use and increased max speed are positively affecting the subject's fatigue prediction while those were not significant factors contributing to the Subject B's fatigue prediction.

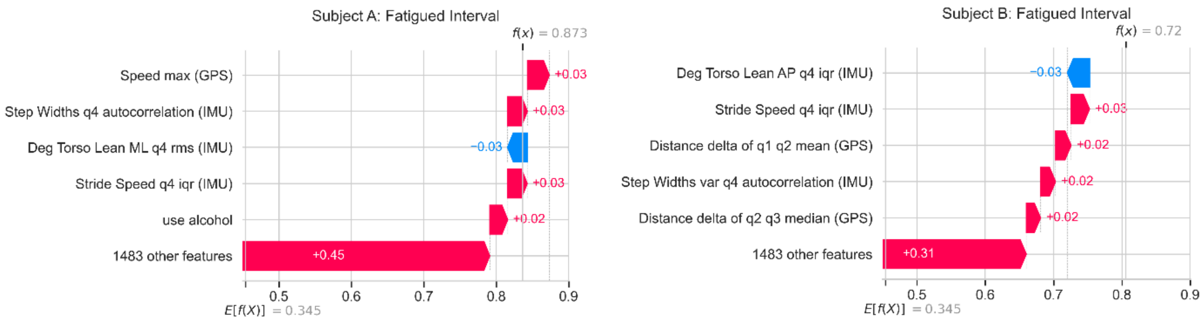

Figure 8. Shapley plots showing top 5 features for two different subjects at a fatigued interval

CONCLUSION AND FUTURE WORK

Motivated by the Army's need for accurate and transparent fatigue prediction capability to support data-driven mitigation strategies that prevent fatigue related injuries, combined with a lack of available research and commercial solutions that meet the need, we presented the multimodal fatigue prediction framework enabling development of accurate and interpretable fatigue prediction models. We demonstrated using the developed framework that the multimodal data collected from many different sources including baseline, biosamples, survey, demographics, and wearable sensor data during activities can be leveraged to develop effective fatigue prediction models predicting fatigue for varying prediction horizons. As for the prediction horizon and predictive power of the developed models, it is worth emphasizing that predicting one's fatigue states 30 minutes into the future using as little as 20 minutes of past sensor data with over 70% balanced accuracy is a noteworthy achievement that has not been accomplished in the past, to the best of our knowledge.

While we demonstrated the ability to leverage the multimodal data to make effective prediction, we acknowledge that data such as subjective survey questionnaires and biosamples are less desired in applied training scenarios given their limitations in objectivity and ease of capture respectively, and prediction models requiring such data to make accurate predictions while training or operating in the real environment are challenging to apply in practice. The flexibility of the proposed framework addresses this, as it allows efficient development of prediction models requiring any set of data sources. As shown in Table 5, performance sensitivity to different type and combinations of data sources showed that models only requiring wearable sensor data still maintain good prediction power.

In terms of the explainability and interpretability of the developed prediction models, methods implemented in the framework enabled both population- and individual-level explanations. For the population-wise explainability, all 20 globally significant features were from the wearable sensor data while predicting 30 minutes forward in time, which highlights the importance of sensor data to make accurate prediction of fatigue. This finding adds some flexibility when transitioning the development in this work to the field, as the non-sensor data such as demographics, biosamples, and baseline data, which can be costly to obtain, may not be necessary to achieve effective prediction performance. As long as the Soldiers are wearing non-invasive and commercially available wearables tracking movement and physiological signals, it is possible to effectively predict future fatigue states to enable informed decision making. For individual-level explanations, the Shapley values provided a number of key features that positively or negatively impacted the provided prediction for a given observation window for an individual. Such explainability outputs accompanying the prediction help users and analysts to gain better understanding of why one is predicted to be fatigued or not, which subsequently will support individually tailored intervention actions and will ease the Army to trust and adapt the models more compared to many existing blackbox models that do not offer the explainability.

The presented framework and results represent a significant research advancement in effectively measuring and predicting Soldier fatigue, and offers a path toward implementing a data-driven decision support system to limit the injuries and resources losses currently experienced by Army. While the presented framework was demonstrated using the data from dozens of soldiers who experienced fatigue, it is easily scalable as the proposed method of fatigue label generation can automatically generate fatigue labels as long as heart rate signals are collected from wearables. The presented methods of feature extraction can also be easily replicated and scaled to large-scale data set using standard modules available in Python.

For future work, continued communication and collaboration with the Army will be essential to demonstrate how the recent advancements in technology and research can be incorporated into day-to-day training scenarios and/or operations to minimize injuries and enhance Soldier performance. While the results demonstrated effective prediction of fatigue, there is still potential for false positive predictions, which could lead to resource costs from taking unnecessary preventative actions. To promote the adaptation of the fatigue prediction capability, a cost-benefit analysis incorporating impacts of both successful and false positive/negative predictions will need to accompany corresponding candidate models to quantify the potential impact of the deployed models. Applying the framework to other physical movement tasks beyond a ruck could assess the generalizability of the fatigue prediction model. Additional examination into prediction windows with subject matter experts could further enhance the appropriate time window that provides Instructors/Leaders with sufficient insights for optimizing training and operations while minimizing injury risk.

To enhance model performance and to establish robust research advancement, collecting high quality human subjects data is critical in human training and performance. As the data collected in operational environments are susceptible to noise and a high-degree of data loss, frequent data quality checks, and sensor tests and evaluations need to become mandatory scheduled events throughout data collection events. Smart sensors with embedded data quality checks and enhancers, and offline data augmentation and interpolation methods are some areas that can be explored further to improve data quality for model performance enhancements. In terms of machine learning model development, techniques such as gradient boosting and ensemble learning can be explored to achieve performance improvements.

ACKNOWLEDGEMENTS

This study was funded by the Development Command Soldier Center (DEVCOM SC) under the Measuring and Advancing Soldier Tactical Readiness and Effectiveness (MASTR-E) program.

REFERENCES

Banister, E. W. (1991). Modeling Elite Athlete Performance. *Physiological Testing of the High-Performance Athlete*.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research: JMLR*, 20, 177.

Forrest, L. J., Schuh-Renner, A., Hauschild V. D., Barnes, S. R., Grier, T. L., Jones, B. H., Steelman, R. A., McCabe, A., Dada, E. O., Canham-Chervak, M. (2022). Estimating the cost of injuries among U.S. Army soldiers. *U. S. Army Public Health Center*. https://apps.dtic.mil/sti/pdfs/AD1166831.pdf

Gabbett, T. J. (2016). The training-injury prevention paradox: should athletes be training smarter and harder?. *British journal of sports medicine, 50*(5), 273-280. https://doi.org/10.1136/bjsports-2015-095788

Griffin, A., Kenny, I. C., Comyns, T. M., & Lyons, M. (2020). The Association Between the Acute:Chronic Workload Ratio and Injury and its Application in Team Sports: A Systematic Review. *Sports medicine (Auckland, N.Z.), 50*(3), 561-580. https://doi.org/10.1007/s40279-019-01218-2

Hulin, B. T., Gabbett, T. J., Blanch, P., Chapman, P., Bailey, D., & Orchard, J. W. (2013). Spikes in Acute Workload are Associated with Increased Injury Risk in Elite Cricket Fast Bowlers. *British Journal of Sports Medicine, 48*(8), 708-712. https://doi.org/10.1136/bjsports-2013-092524

Kathirgamanathan, B., Buckley, C., Caulfield, B., & Cunningham, P. (2022). Feature subset selection for detecting fatigue in runners using time series sensor data. *Pattern Recognition and Artificial Intelligence*, 541-552. https://doi.org/10.1007/978-3-031-09037-0_44

Kim, J., Kim, H., Lee, J., Yoon, J. & Ko, S. (2022). A deep learning approach for fatigue prediction in sports using GPS data and Rate of Perceived Exertion. *IEEE Assess, 10,* 103056-103064. https://doi.org/10.1109/ACCESS.2022.3205112

Luo, H., Lee, P.-A., Clay, I., Jaggi, M., & De Luca, V. (2020). Assessment of fatigue using wearable sensors: A pilot study. *Digital Biomarkers, 4*(Suppl. 1), 59-72. https://doi.org/10.1159/000512166

Ohkuwa, T., Tsukamoto, K., Yamai, K., Itoh, H., Yamazaki, Y., & Tsuda, T. (2009). The relationship between exercise intensity and lactate concentration on the skin surface. *International journal of biomedical science: IJBS, 5*(1), 23-27. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3614747/

Ishii, H., & Nishida, Y. (2013). Effect of lactate accumulation during exercise-induced muscle fatigue on the sensorimotor cortex. *Journal of physical therapy science, 25*(12), 1637-1642. https://doi.org/10.1589/jpts.25.1637

Pasadyn, S. R., Soudan, M., Houghtaling, P., Phelan, D., Gillinov, N., Bittel, B. & Desai, M. Y. (2019). Accuracy of commercially available heart rate monitors in athletes: a prospective study. *Cardiovascular diagnosis and therapy, 9*(4), 379-385. https://doi.org/10.21037/cdt.2019.06.05

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830. https://dl.acm.org/doi/10.5555/1953048.2078195

Pinto-Bernal, M. J., Cifuentes, C. A., Perdomo, O., Rincón-Roncancio, M., & Múnera, M. (2021). A data-driven approach to physical fatigue management using wearable sensors to classify four diagnostic fatigue states. *Sensors (Basel, Switzerland)*, *21*(19), 6401. https://doi.org/10.3390/s21196401

Russell, B., McDaid, A., Toscano, W., & Hume, P. (2021). Predicting fatigue in long duration mountain events with a single sensor and deep learning model. *Sensors, 21*(16). https://doi.org/10.3390/s21165442

Seshadri, D. R., Li, R. T., Voos, J. E., Rowbottom, J. R., Alfes, C. M., Zorman, C. A., & Drummond, C. K. (2019). Wearable sensors for monitoring the internal and external workload of the athlete. *NPJ digital medicine*, *2*, 71. https://doi.org/10.1038/s41746-019-0149-2

Schwartz, O., Malka, I., Olsen, C. H., Dudkiewicz, I., & Bader, T. (2018) Overuse injuries in IDF's combat training units: Rates, types, and mechanisms of injury. *Military medicine, 183*(3-4), e196-e200. https:/dio.org/10.1093/milmed/usx055

Shapley, L. S. (1951). Notes on the N-person game – II: The value of an N-person game. *Rand Corp*. https://www.rand.org/pubs/research_memoranda/RM0670.html

"What's My Daily Readiness Score." Fitbit. https://help.fitbit.com/articles/en_US/Help_article/2470.htm#:~:text=Wear%20your%20device%20consistently%20for,a%20more%20accurate%20personal%20baseline (accessed Jul. 3, 2023).

"What is Whoop Strain?" Whoop. https://www.whoop.com/thelocker/how-does-whoop-strain-work-101/. (accessed Jul. 3, 2023).

"Body Battery™ Energy Monitoring." Garmin. https://www.garmin.com/en-US/garmin-technology/health-science/body-battery/ (accessed Jul. 3, 2023).

"Training Load Pro." Polar. https://support.polar.com/en/training-load-pro (accessed Jul. 3, 2023).