Jointly Predicting Job Performance, Personality, Cognitive Ability, Affect, and Well-Being

Pablo Robles-Granda, Suwen Lin, Xian Wu, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Aaron Striegel, and Nitesh V. Chawla[§], University of Notre Dame, USA
Sidney D'Mello, Julie Gregg, University of Colorado Boulder, USA
Kari Nies, Gloria Mark, Ted Grover, University of California, Irvine, USA
Andrew T. Campbell, Shayan Mirjafari, Dartmouth College, USA
Koustuv Saha, Munmun De Choudhury, Georgia Institute of Technology, USA
Anind D. Dey, University of Washington, USA

Abstract—**Assessment of individuals' job performance, personalized health and psychometric measures are domains where data-driven ubiquitous computing will have a profound impact in the near future. Existing work in these domains focus on techniques that use data extracted from questionnaires, sensors (wearable, computer, etc.), or other traits to assess well-being and cognitive attributes of individuals. However, these techniques can neither predict individuals' well-being and psychological traits in a global manner nor consider the challenges associated with processing the often incomplete and noisy data available. In this paper, we create a benchmark for the predictive analysis of individuals from a perspective that integrates physical and physiological behavior, psychological states and traits, and job performance. We develop a novel data mining framework that can extract meaningful predictors from noisy and incomplete data derived from wearable, mobile and social media sensors to predict nineteen constructs based on twelve standardized and well-validated tests. The framework can be used to build a predictive model of outcomes of interest. We validate the framework using data from 757 knowledge workers in organizations across the United States with varied work roles. Our framework and resulting model provides the first benchmark that combines these various instrument-derived variables in a single framework to understand people's behavior. The results show that our framework is reliable and capable of predicting our chosen variables better than the baselines when prediction includes the noisy and incomplete data.**

Index Terms—**Personality, wellness, job performance, psychometric, prediction**

I. Introduction

WEARABLE devices are opening new avenues to improve our understanding of our health and well-being by tracking individual in-situ patterns of activity and affect. These patterns can be extended into precise and objective measures of health and well-being, which can not only benefit an individual's health goals, but also aid organizational efforts to promote well-being [1]–[3]. Detailed and continuous data collection can yield valuable insights into health and well-being, e.g., stress, sleep, work performance, and physical activity [4]–[7].

In recent years, literature that assesses well-being and its effect on productivity has expanded. Well-being, personality, and cognitive ability can affect job performance [8]–[15]. Personalized well-being assessment [1], [16] is also receiving more attention due to the ubiquity of wearable devices.

Despite advances in computational methods that use wearable data to assess well-being, e.g., [1], [16]–[18], substantial challenges still exist. Previous work has been limited by data characteristics such as small samples, homogeneity (specific demographics and work roles), or controlled environments (within specific scenarios and locations). Additionally, predicting well-being, where productivity or performance is considered [19], becomes difficult due to scarcity of situational (contextual) information [20], [21], privacy, or other concerns [8]–[15]. Assessing both well-being and workplace performance requires machine learning strategies that overcome these issues. In our work, we identify three main challenges: 1) unobtrusive data collection using large samples of heterogeneous individuals in a wide range of work settings and geographic locations; 2) need for multi-modal sources for holistic representations of physical and behavioral patterns; and 3) missing and noisy values that stem from the nature of the data and sensors and idiosyncrasies of individuals (e.g., gaps in measurements, variable compliance in wearing the devices, failure of sensors) [22], [23].

In this paper, we report our approach in addressing these challenges and building a machine learning pipeline for the Tesserae Project [24]. In Tesserae, we implemented a system comprised of a broad suite of sensing modalities for automated modeling of individual physical, psychological, and job performance differences. Figure 1 shows a diagram of the specialized sensors that were used to gather data in an unobtrusive manner. The goal of these sensors was to collect information that was representative of an individual's behavior, health, and mental states. These sensors also collected data related to job and daily activities (offline and online interactions, phone and computer usage) both at home and at work. We enrolled a diverse cohort of 757 knowledge workers performing varied roles at several organizations across the United States.

Research Questions. We consider the following questions: RQ1) How do we integrate data from different sensors and modalities? RQ2) How do we develop machine learning meth-

[§] Corresponding Author: nchawla@nd.edu ods that can deal with the challenges of varying levels of missingness and noise, and inter-individual, intra-individual, and inter-sensor variance? RQ3) Are individual job performance, well-being, and personality predictable from these sensors?

Critically, real-world wearable data can also be affected by (ir)regularities, temporal variations, and differences in an individual in addition to missing values, noise and other issues mentioned before. Thus, generalizability issues arise in models trained on samples from this data because the training sample distribution may be different than the test sample distribution used in the final applications. While sensor data has been used to predict and assess human behavior and well-being [25]–[28], these predictions are done on highly curated and often homogeneous data. Such models may be overly-optimistic about what can be achieved in real-world, messier scenarios.

We address these data issues and create a machine learning solution to our joint prediction problem. Specifically, we model nineteen survey-based variables [29]–[40]. The 927 predictors (features) were obtained from sensors assigned to each participant in Tesserae: a wearable (Garmin Vivosmart 3), a phone agent (an app for iPhone and Android), four Bluetooth beacons (office, home, and two portable), and social media (Facebook). As part of our work, we identified the most relevant subset of these features for each of the nineteen variables.

While sensor data has been used to create models of human behavior, including: daily activities, mobility, [25]–[27], well-being [28], and academic performance [17], to our knowledge there is no study that harnesses multi-modal data sources into a unified comprehensive framework that addresses data messiness and algorithmic challenges to create a generalizable machine learning pipeline predicting person-level behavior, physical and psychological well-being, and job performance.

To create a generalizable joint model of the nineteen constructs, we implement several strategies. First, we consider various imputation approaches as well as co-dependencies among the variables for both feature selection and prediction (RQ1). Unlike other approaches that deal with missing data for longitudinal scenarios (e.g., [22], [23]), our approach does not require complex likelihood-based techniques. Second, we use a fusion technique to synthesize multi-modal sensor-derived features (RQ1). Third, we consider an ensemble learning technique to incorporate various machine learning models (RQ2). Fourth, in addition to our data policy and machine learning design, we use higher order networks (HON) [41] to obtain descriptions of each individual (RQ3). Finally, to ensure generality we perform 5-fold cross-validations at all stages of the model creation: feature selection, dimensionality reduction, hyper-parameter tuning, training, prediction, and tests.

In summary, our contributions are as follows: 1) We create a framework whereby noisy, heterogeneous, multi-modal data can be fused without the need for highly specialized curation; 2) We provide a benchmark that leverages data fused from the modalities to produce more integrated predictions of human behavior than existing techniques; and 3) We verified experimentally the predictive capability of our approach using data from our longitudinal study. The results show that our model perform favorably with respect to theory-driven baselines. We verify the results using various reliability tests.

II. Background

In order to acquire a comprehensive view of an individual's physical attributes, psychological properties, personality, and job performance, we used the battery of psychometric surveys to construct our ground truth variables listed in Table I. Surveys were administered at the beginning of the study, and shorter versions periodically over the first 60 days of this year-long study. We present the analysis for the initial battery of surveys (results for the shorter versions are presented in [42]).

A. Job Performance

We considered five variables that assess job performance from three perspectives: task performance, organizational citizenship behavior, and counterproductive work behavior [21], [43]–[46]. The surveys are designed to capture behaviors associated with achieving organizational goals [47]–[49].

1) Task Performance: We measure task performance using two variables: In-Role Behavior (IRB) [29] and Individual Task Proficiency (ITP) [30] using instruments validated with significant samples [29], [30]. The former measures an individual's perception of her job performance based on completion of tasks associated with that individual's position. The latter measures the individual's perception of how frequently she completed her core job tasks, completed these tasks well, and verified that these tasks were completed well.

2) Organization Citizenship Behavior: Organizational citizenship behavior was assessed using the Organizational Citizenship Behavior Checklist (OCB-C; [31]). OCBs are optional actions that are not rewarded by a worker's organization. OCB-C is a validated instrument as described in [31].

3) Counterproductive Work Behavior: (CWBs) are actions taken by employees that intentionally harm either the organization or individuals within the organization [50]. To measure CWB, we use the Interpersonal and Organizational Deviance (IOD) scale [32]. The 19-item instrument is broken into two major categories of items: (1) Interpersonal Deviance (7 items) and (2) Organizational Deviance (12 items). Each item has a seven-point frequency score: 1 (never) to 7 (daily). In our predictions, we consider each major category as a separate variable. IOD was validated as described in [32].

B. Psychological Constructs

We focus on four psychological constructs: cognitive ability, personality, affect, and anxiety.

1) Cognitive Ability: We use the Shipley Institute of Living Scales 2 (Abstraction and Vocabulary sub-tests) [33] to measure fluid and crystallized intelligence respectively [51], [52]. See [53] for a study on the relation of cognitive ability and job performance. Shipley 2 has high reliability and internal consistency [33].

2) Personality: Personality was measured in the initial ground truth battery via the Big Five Inventory-2 (BFI-2; [34]). The BRI-2 traits are: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Open-Mindedness. Each of the Big Five Personality Traits has varying levels of association to job performance [11], [12], [14], [15], [54]–[56]. BFI-2 was validated with four datasets [34].

3) Affect: We use the Positive and Negative Affect Schedule-Expanded Form (PANAS-X; [35]). Affect variation is a key indicator of a person's mental health and is critical for job performance and other job behaviors [57]–[60]. PANAS-X is shown to be a reliable instrument [61].

4) Anxiety: We use the State-Trait Anxiety Inventory (STAI; [36]). Anxiety is another key indicator of a person's mental health. STAI was validated by [36].

C. Health and Physical Variables

1) Alcohol Consumption: We use the Alcohol Use Disorders Identification Test (AUDIT), which was developed by the World Health Organization (WHO; [37]). The effects of alcohol consumption on job performance and other areas of people's lives are well documented [62]–[67]. AUDIT validity has been widely confirmed, e.g., [68]–[70].

2) Physical Activity: We use the International Physical Activity Questionnaire (IPAQ; [39]). Physical activity affects not only physical health but also mental well-being and job performance [71]–[74]. The test-retest reliability for the IPAQ questionnaires is reported in [39].

3) Sleep: We use the Pittsburgh Sleep Quality Index (PSQI; [40]). Sleep is critical to incorporate because poor sleep directly impacts job performance [75], cognitive ability, and mental health [76], [77]. PSQI has both good internal reliability and good test-retest reliability [40].

4) Tobacco Use: We used IARPA's modified version of the Global Adult Tobacco Survey (GATS) from the World Health Organization (WHO, [38]), which focuses only on the individual's consumption by considering three items: whether the participant is a current smoker, if they use tobacco daily, and the quantity used in the past week. We predict the last item only. Tobacco use is associated with stress, negative emotionality, lower agreeableness, [67], [78]–[82], and work performance [83]. GATS was reviewed and approved by the GATS Questionnaire Review Committee of the WHO.

III. Related Work

We discuss related work for each of the categories of dependent variables in our study.

Psychological Variables. Empirical evidence shows that cognitive ability is related to personality traits. For example, personality (introversion) and abstraction are shown to be associated with intellectual curiosity, which along with other traits can partially explain crystallized intelligence [86].

Personality has been predicted by an individual's behavior, ranging from, e.g., a person's interactions with computers [91]–[93] to social pressures [102]. The field of automated personality modeling has focused on predicting the Big 5 personality traits from sensing and system adaptability. For a recent review of the most popular computational approaches for automated personality detection (including datasets, applications, and machine learning methods), see [95] and most recently [126]–[128] that improved the results on the Facebook, Essays, and Kaggle and Essays datasets, respectively. Some of these techniques combine various types of complex features using Deep Residual Networks and sophisticated techniques requiring large amounts of data.

TABLE I: List of Dependent Variables

Type	Subtype	Variable
Job Performance	Task	IRB [29]
		ITP [30]
	Org. Cit. Behavior	OCB [31]
	Deviance [32]	Interpersonal
		Organizational
Psychological	Cognitive [33]	Vocabulary
		Abstraction
	Personality [34]	Extraversion
		Agreeableness
		Conscientiousness
		Neuroticism
		Openness
	Affect [35]	Positive
		Negative
	Trait Anxiety [36]	Anxiety
Health	Consumption	Alcohol [37]
		Tobacco [38]
	Activity	Physical [39]
		Sleep [40]

TABLE II: Literature per Variable Group and Predictor Types

Variable Group	Sensor/Social	Attributes & Traits
Job Performance	[17] [18] [42]	[9]–[15], [84]–[87]
Cognitive Ability	[88], [89]	[90]
Personality	[18], [91]–[101]	[86], [102]
Affect	[103]–[109]	[106]
Anxiety	[104], [110], [111]	[36], [112], [113]
Alcohol	[114], [115]	[116]
Tobacco	[117]	[118]
Physical Activity	[119], [120]	[121]
Sleep	[122]–[125]	[125]

Affect detection has been an active area of research in the field of affective computing [94], [103]–[105], [129]. Positive and negative affect, as well as perception and satisfaction with health and life, have been predicted using wearable sensors [106]. Affective modeling also involves sentiment analysis, including predicting sentiment intensity [108] using a stacked ensemble that contains neural networks as part of the architecture. In the case of sentiment analysis, however, the number of instances used for training in [108] is in the order of thousands, while in our case we collected posts from social media for 392 participants. The interested reader is referred to the survey in [107], [129], which uses common tasks in affective computing and sentiment analysis (emotion recognition, polarity detection, multimodal fusion) to present a taxonomy of platforms: knowledge based, statistical methods, and hybrid. Notice that sentiment analysis is often done in the context of specific scenarios, e.g., with respect to social media posts and news. In contrast, our paper is devoted more to sensor-based modeling for understanding individual traits. Emotion categorization involves psychological models, e.g., [130]–[134] and sensing algorithms [91], [107], [135], [136] at the intersection of psychology linguistics, computer science, engineering, and other fields. Most emotion categorization models differ on the number and the list of emotions. However, emotions can be roughly grouped into three types: positive, neutral, and negative, as discussed in [109]. In our paper, we model generalized affect (positive and negative) [35] as a fairly stable trait that describes the experience of emotions because affect is associated with well-being and productivity.

Beyond positive and negative affective states we consider anxiety. Previous work has focused on predicting anxiety from sensors, e.g., [104], [110]. Methods that predict stress and anxiety based on ECG monitoring have been proposed [111].

Physical and health variables have been predicted by a variety of methods. Alcohol consumption is found to be related to both a trait (self-control) and cognitive ability (working memory capacity) [116] and has been predicted by EEG signals and trans-dermal devices [114], [115]. Physical activity (IPAQ) is commonly predicted by using mobile sensing, e.g., [119], or specialized devices such as accelerometers [120]. Physical activity patterns appear to be stable, as patterns earlier in life can predict some activity patterns later in life [121].

Sleep quality has been estimated using wearables by [122], [123]. More invasive techniques can use chest sensors and polysomnography (to measure body acceleration and position) [124]. Specialized work to estimate sleep for patients with schizophrenia also exists (see a comparative analysis in [125]).

Tobacco consumption has been monitored using air sensors as in [117]. Also, smoker group membership (never, established, former, non-daily, and daily) has been predicted using family history, depression, consumption of other substances, and demographics [118]. To our knowledge, tobacco consumption has not been monitored using wearable sensors.

Job performance is usually measured through either subjective rating scales [48], [137] or objective performance outcomes, such as sales amounts and production numbers. [48]. Using wearable sensor data to estimate job performance has been explored by [17], who demonstrate that wearables can be used to detect when a person is focused on their work via physiological features. Another approach estimates job performance based on personality and individual traits [11], [13]–[15] with varying degrees of success. This includes conscientiousness [12], [15], [84], and extraversion [13]. Links have been found between personality, cognitive ability, and other traits with job performance [86]. However, as mentioned, estimation of job performance more commonly relies on various types of questionnaires including self-reports, and supervisory and peer evaluations [11], [48], [137]. Job performance varies depending on demographic information (e.g., age, gender) [85] and individual traits (personality, emotional intelligence) [9], [10]. Cognitive ability, personality, affect, and anxiety are all not only related to each other as discussed, but can both affect and be affected by job performance, and physical and health variables. The mutual effects of personality traits and job activities are well documented [87].

Mobile and wearable sensor data are powerful sources of information about human behavior which could help us identify patterns of activities, human mobility, [25]–[27], well-being [28], and job and academic performance [17]. Machine learning models have been demonstrated to achieve high accuracy on very specific tasks on very small samples, e.g., estimating work load category using wearables on a cohort of twenty academic participants [17]. Subjective perceptions of job performance are, however, harder to predict using wearable data even in small samples and specific work locations and environments [18], as opposed to various work locations and occupations as in the present case. To our knowledge, ours is the first model that jointly predicts health, job-performance, and psychometric variables of individuals in a global manner.

TABLE III: Participants per Cohort Used for Modeling

Cohort	# Participants
1. Multinational Consultancy Company	217
2. Multinational Technology Company	138
3. Small Software Company	21
4. Various Smaller Companies	147
5. Local University	31

The present work is a broad and personalized analysis using instruments from the longitudinal Tesserae Project [24]. An initial analysis based solely on job performance was presented in [42], which reported a model to differentiate low from high job performance but did not estimate the specific job performance score directly. [42] reported predictions of the daily battery. In contrast, our analysis is done on the single initial battery of twelve standardized tests not only of job performance, but of all other measures as well.

IV. Data Collection and Description

From Fall 2017 to Summer 2018 we recruited 757 individuals working in knowledge fields in the US as part of a large-scale longitudinal research study. We collected data from these participants for a period of one year starting from January 2018. Individuals' participation in the study was voluntary and those who participated received a monetary incentive to stay in the study and comply with the data-collection protocols. This monetary compensation varied according to levels of compliance and was allocated throughout the year of study. The monetary compensation for participants was also specific for one of the companies, per the rules of the company.

Our project was conducted in accordance with the Institutional Review Board of the University of Notre Dame (under protocol number 17-05-3870) and similar authorities of all the institutions involved. All participants provided written informed consent prior to taking part in the study. No Personal Identifiable Information (PII) was shared.

To handle the heterogeneity of our dataset, a subset of participants was selected from each cohort in Table III for external validation and was not considered during development of our models in order to prevent bias and data leakage. The remaining 554 participants came from various organizations in the USA and can be grouped into five cohorts, as shown in Table III. Another source of heterogeneity, particularly at the job-performance level, comes from the participants' roles. A total of 254 participants self-reported holding a supervisory whereas 297 reported a non-supervisory role; 3 participants declined to mention their role within their companies.

The data collection protocols could be classified into two stages: 1) an initial set of surveys used to collect the initial battery of ground truth variables and social media data; and 2) daily data-gathering of data streams from various sensors (daily varying predictors). We analyzed the initial ground truth battery using the daily sensor data streams and social media.

A. Data Sources — Sensing Streams

In order to model individuals' behaviors and physical attributes, we selected multiple modalities that unobtrusively collect physiological, psychological, behavioral, and physical states of individuals; their offline and online interactions; their phone, social media activity, and workplace routines; and health and well-being both at work and at home. Specifically, we used a wearable to capture an individual's physical and physiological state. In order to capture the context of an individual's actions we used a phone agent (app) and proximity beacons that allowed us to identify the individuals' relative locations (home/work) during the day. Finally, we capture higher level information using social media, together with the wearable and phone agent data, provided insights about a person's psychological states. All data was de-identified to protect the participants privacy. In addition to raw features, we considered features derived from the sensors.

Wearable: Garmin Vivosmart 3. This wristband is a commercial smart wristband (a wearable device) that is widely used as a fitness, activity, and well-being monitoring device. The device collects physical/physiological data, (e.g., heartrate, step count, number of floors climbed, calories burned, physical activity such as running, and walking), sleep quality data (e.g., sleep staging, duration), and psychological data (e.g., stress—which is based on physical signals such as heart rate). The wearable was paired via Bluetooth with Connect, a Garmin App that participants installed on their phones. The wearable was also paired with an app we developed for our study (see PhoneAgent below). Both apps collected data from the wearable which was transferred to our collecting servers and into databases that anonymized the data. We computed daily summaries from each of the signals collected.

App: PhoneAgent. We created an app (the PhoneAgent) for both iOS and Android devices. The app ran in the background and periodically collected data, saving it temporarily as JSON files that were later transmitted to servers when the phone was connected to Wi-Fi. The data collected by our app included location, physical activity (walking, bicycling, driving, etc.), phone usage (e.g., screen lock/unlock) and ambient light levels. Our app also connected to the wearable and the beacons (described below) via Bluetooth. The PhoneAgent streamed data directly from the wearable, which allowed for more fine-grained and real-time data collection compared to Garmin's Connect app (e.g., beat-to-beat interval in the PhoneAgent compared to average heart rate every minute from Connect). Our app collected the following wearable generated time-series: heart rate (HR), steps, floors climbed, calories burned, and stress levels. From the beacons (see below), our app collected information about the proximity of an individual (through a key-chain beacon and backpack beacon) to either of the fixed beacons (home, office). This provided details of interactions using the strength of the signal as described next.

Beacons: Gimbals. Beacons are low energy devices that transmit and receive Bluetooth signals to and from other devices. We used four Gimbal beacons per participant in our study. Two beacons were the static Gimbal Series 21, with one placed at the participant's home and the other placed at the participant's workplace. The other two Series 10 beacons are small, coin-size, mobile beacons that participants carried, one in their key-chain or wallet, and one in their backpacks or purses. Beacon signals were detected by the phone through our PhoneAgent app which uses the Gimbal API library to detect proximity. When a PhoneAgent enabled smartphone approaches a beacon, the phone will detect a Bluetooth signal and will record the signal strength which is inversely proportional to the distance between the phone and the beacon. The beacons provided information about the location of an individual relative to their home, work, or to other participants. This allowed us to derive features that describe the mobility of individuals and other daily location-based routines while hiding actual physical locations. These features were stored by the PhoneAgent into a local server and on Gimbal servers.

Social Media: During the recruiting process, we requested read access to the participant's accounts on Facebook and LinkedIn. As with all of the other data sources, we anonymized their data but, in the case of social media, we also modified the data so as to avoid storing raw information that may affect privacy. After data collection from social media, we applied feature extraction techniques and stored only the anonymized features. For the present analysis, we considered 5,075 raw features computed from participants Facebook data; most of these were n-grams (words/phases) of posts. However, a feature selection step was applied to select only the relevant features, as detailed in Section V-C. These raw features corresponded to a variety of categories—1) psycholinguistic attributes [138] (that captured language usage across keywords related to affect, cognitive attributes, perception, interpersonal focus, temporal references, biological concerns, and social and personal concerns), 2) open vocabulary n-grams (the 5,000 most frequent uni-, bi-, and tri-grams used by the participants), 3) sentiment in posts, and 4) social capital (e.g., by measuring check-ins to places, posting/sharing updates, uploading media, changing relationship status, and hanging out with friends).

B. Predictors

We used a total of 927 candidate features (filtered out later with dimensionality reduction and feature selection techniques, as detailed below) based on the sensor data from the PhoneAgent, Garmin wearable, Gimbal beacons, and social media. We extracted additional information from two wearable generated time-series per participant: heart rate and stress measurements. We used these time series as separate components to extract features that facilitate discriminatory prediction based on signatures extracted using a higher order network (HON) approach with one HON per time series. We also used the heart rate to build an additional component for the ensemble using a special representation for the time series. Table IV details the number of features used per data source. For features collected as time series we computed the daily mean, median, mode, minimum, and maximum.

TABLE IV: Low-level Sensor-Derived Features.

Source	Sub-Modality	#
Wearable	Higher Order Network—Heart Rate	5*
	Higher Order Network—Stress	5*
	Heart Rate	28
	Other Physical	26
Phone App	Physical Activity	19
	Context	8
	User State	47
	Phone Usage	56
	Regularity	580
Beacon	Work Activities	16
	Other	7
	Home Activities	5
Social Media		200*

* =post PCA

Examples of features collected from Garmin (through the Connect API) include sleep staging (duration in light, deep, and REM sleep) and bed time, daily step counts, daily floors climbed, physical activity (duration of light, medium, heavy activity), calories burned, and stress level (in range 0–100).

Examples of features collected by the PhoneAgent include phone usage (number of screen locks and unlocks, duration of locks and unlocks, etc.) and daily aggregations of physical activity such as mobility features (e.g., places visited, distance traveled, duration of sedentary state, driving and biking time). The PhoneAgent also collected fine-grained data from the wearable, e.g., heart rate, sleep, stress and steps. We computed time series features at a daily level (which we call *epoch-0*) but also in epochs within the day: early morning (12AM – 9AM), day (9AM – 6PM) and evening (6PM – 12AM). We used the *epochs* to identify differences of behavior for the times that are associated with sleep, work, and nightly activities.

Examples of features collected through the beacons include various measurements of closeness of the static and mobile beacons. These features in their raw form do not provide direct insights about the participants' activities, but in combination with the type of beacon and the duration of the interactions we can capture information such as the time spent at work (total duration a participant spends at work from the first to the last sighting of the work-beacon), the time spent at desk (percentage of the time a participant spends at their desk), and the number of breaks taken away from the desk that exceed 5, 15 and 30 minutes (captured by gaps in beacon sightings).

We experimented with various time resolutions to derive the summary statistics, as the distributions may have non-linear relations that may not fully capture the individual's behavior. We report predictions for individuals with at least two weeks of data. We constructed HON representations of people's behaviors through the heart rate and stress time series as we describe in Section V.

The predictors are highly heterogeneous due to the multi-modal nature of our dataset. This made it prudent to apply ensemble-learning strategies. Another source of heterogeneity and noise in the features was due to the compliance of the participants, the quality of the data transfer, and missing data.

C. Missing Data and Data Difficulties

In addition to the heterogeneity of the data sources, the main challenge of building predictive models with our data set was caused by missing values. The data sources most affected by *feature missingness*, i.e., missing values of specific predictors, were the wearable and the PhoneAgent. In particular, missingness in the latter was critical as the PhoneAgent was used to collect data from the wearable and the beacons.

PhoneAgent. Missing data from the PhoneAgent was mostly due to technical issues. In particular, participants had a variety of phone models with different operating systems (and versions) and capabilities. This variability in devices made it difficult to provide user support. Also, some adjustments were needed because both the Garmin platform and the beacons did not record data properly in some iOS versions.

Wearable. Missingness was due to loss, failure (e.g., did not hold charge, did not charge at all, data did not sync, unusual report of floors climbed, inability to connect to the phone, inadequate sleep tracking on public transport.), or damage to the device (e.g., strap, screen) or charger. One participant reported an allergic reaction to the nickel in the buckle.

Social Media. This stream presented two challenges: not all individuals had Facebook accounts and the level of engagement of individuals in their online profiles varied greatly.

Beacons. Some participants placed home beacons at the work place and vice versa. This made extracting meaningful features related to location challenging. Additionally, noisy data also presents some challenges. Bluetooth signal strength can be affected by certain objects and their properties (e.g., number of walls and the materials used in construction). Beacons (and the wearable) can also be affected by noise because sleep during the day may not be reliably detected.

Finally, in addition to feature missingness, a major challenge is full-modality missingness, i.e., participants with information missing for the entire modality, as in the case of social media, where no data was available for several participants. In such cases we used group imputation methods as detailed next.

V. Joint Prediction Model

We developed an ensemble learning method for joint prediction of the physical, psychological, and job-performance variables. As we detailed in Section IV-A, the data sources included social media, Garmin wearable, phone agent, and beacon data. Additionally, we computed heart-rate variability and used it as a separate stream. Furthermore, we constructed a HON based on heart rate and wearable/sensor stress measures as detailed below. Within each model, a set of candidate models are trained per ground truth variable as outlined in our model's schema in Figure 1. The components of the ensemble are supervised techniques (regression and classification) as detailed next. In order to deal with the complexity of the data as well as the missingness, we considered the following: a) the ensemble components that identify both linear and non-linear partitions and regressions, b) the pre- and post-processing that ensure generality and avoid outliers, c) the feature selection that eliminates redundant dimensions and selects relevant features, d) a higher-order representation of temporal data that extracts non-Markovian patterns (long-term temporal dependencies), e) imputations, both at the feature and modality level, f) a fusion strategy for the various modalities, and g) an algorithm to coordinate the model selection framework.

A. *Design of Components*

Our design goal was to automate the discovery of variable relations and data separability for linear, multicollinear, and nonlinear relations. In each case we consider low and high dimensional cases. Thus, we considered the following regression methods as candidates for the components: linear regression (low-dimensional cases), linear regression with L_2-norm (multicollinearities cases), linear regression with built-in cross-validation with L_2-norm (high dimensional multicollinear relations), lasso model with least angle regression (high-dimensional linear cases), Bayesian ridge regression (high dimensional cases), support vector regressor (SVR) with either linear, radial basis function, or polynomial kernel (for linear and non-linear high-dimensional relations). Finally, decision trees (CART), and random forest regression were used for non-linear relations. The selection of the optimal technique and corresponding features was done using cross-validation, as detailed below, which allows us to pick the best performer per ground truth variable. The best performers were then used for training and prediction. Likewise, we used classification counter-parts for linear and non-linear separability and high vs. low-dimensional problems. Specifically, we considered: k nearest-neighbors, linear support vector machine, support vector machine with radial-basis function, decision trees, and random forest. At a lower level, various intermediate steps were performed: transformation, mapping, dimensionality reduction, fusion of sub-datasets, and feature selection.

B. *Pre- and Post-Processing*

1) Cross-Validation: To insure generality of our models we used 5-fold cross-validation for both model design and for final predictions. We considered both static and dynamic partitioning of the data. In order to maintain a homogeneous experimental setting, we considered a fixed partitioning so that the models created across the experiments in the various experimental stages could be comparable. Thus, this static partition was applied across all the variables. However, we also considered dynamic partitioning when evaluating new models, feature selection algorithms, and techniques before the final comparison of performance, which was done on the statically partitioned sets. Imputations were also done using the static 5-fold partitioning by imputing data in a per-fold approach to avoid overfitting due to data-leakage at all levels.

2) Outliers: We performed an outlier analysis where out of range errors for sensing streams were analyzed for extraction issues (the most common case), resulting in integrity checks and script validation from the raw sensing streams. We verified the measurements of sleep time in wrong ranges (fixed via computations of sleep and awake times), negative commute times (due to beacons wrongly placed), and various other aspects. Error sources included typos in the enrollment process, re-assigning of devices from dropped participants to newly enrolled participants, and several edge conditions with respect to the enrollment/ingestion process.

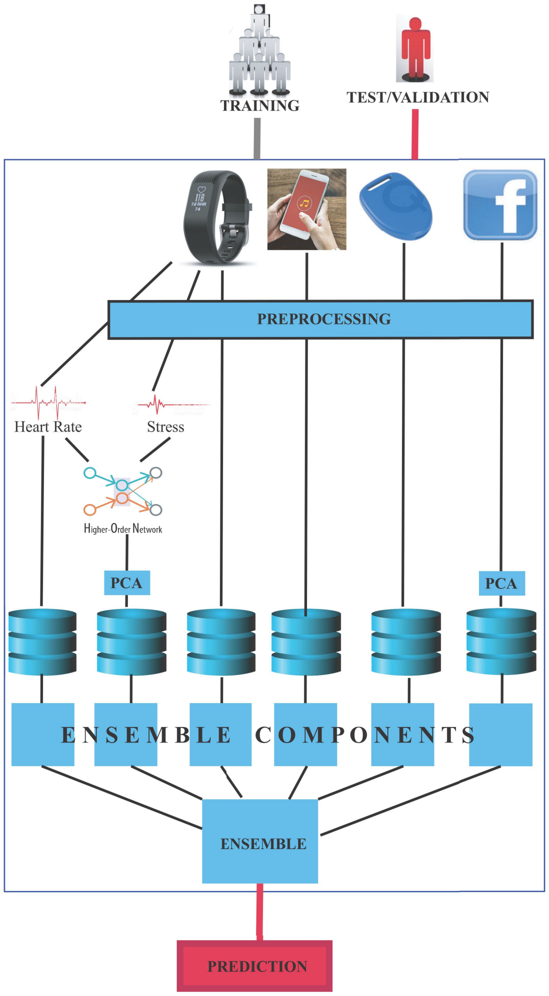

Fig. 1. Model Diagram: We use a set of weak modality-based learners to produce a strong prediction for each variable. Database icons represent repositories of data that is fully anonymized and ready to be processed.

3) Data Range Transformation: We applied systematic verification to ensure the predicted values are not outside of the prescribed ground truth ranges. Code corrections were applied to properly bound prediction results.

C. *Feature Selection*

A sequential exploration of various combinations of features to identify a set of predictive features per construct was conducted. The result was a curated subset of features. First, we introduced the social media feature selection process. It is worth noting that raw social media data is neither shared nor processed for privacy purposes. We only used reformatted features to remove personally identifiable information (PII).

The relevant social media features were selected using principal component analysis (PCA) to identify the top 200 latent features for predicting the ground-truth variables—we aimed to capture complex behaviors latent in the data which are not directly observable in the raw signals. As detailed in Section IV, we used psycholinguistic features, n-grams, sentiment, social engagement that lead to 5075 features per individual and applied PCA to reduce dimensionality to 100 features that account for multicollinearities.

The features from other sources were treated under the same selection policy to define the set of models (components). This involved five stages of selection, in addition to the feature pre-selection and social media selection. First, features were selected based on correlations per fold during cross-validation (the features selected are the ones that overlapped across folds, but for each fold only the training folds are used to compute correlations, to avoid leakage across folds). Second, features were selected by the individual candidate models. Third, a selection was done on the overall final training by the best model. Fourth, a subset of latent features was mapped using PCA for specific feature sets. Lastly, we ranked the models on predictive performance and chose the best model.

D. Higher Order Networks (HON) of Temporal Data

Most real sequential data does not fulfill the Markov property [41]. HON are powerful tools that allow us to overcome this challenge by representing high order dependencies. Non-Markovian patterns provide unique information about the problem under study. For this reason, we use a HON algorithm to provide a multi-scale representation of sequential data on a per-feature basis (e.g., heart rate and sensor-measured stress). When extracting features in sequential data, conventional methods (e.g., Markov model) might lead to information loss on the state transition with the assumption that the next status only depends on the current status. To address this limitation, we utilized a HON method to make a sufficient representation by exploring higher order dependencies in sequential data. Building the HON model consists of the following steps.

We applied discretization to the time series as shown in Figure 2. The discretization step works as a pattern recognition technique that identifies regularities in the time series that are grouped to remove high frequency components. Since the network representation of the time series (e.g., heart rate) is not directly available, the discretization of the raw data can be used to construct a network. We divided time into equal-size (half hour) time slots. x_i is the state in i-th time slot, i.e., the mean value for the heart rate during the corresponding slot.

Given the discretized heart rate data, the output is the conditional probabilities of each individual

$$P(x_t|x_{t-n},\ldots,x_{t-1}) = \frac{\mathrm{I}(x_{t-n},\ldots,x_{t-1},x_t)}{\mathrm{I}(x_{t-n},\ldots,x_{t-1})}$$

where n denotes the network order, $\mathrm{I}(x)$ indicates the number of occurrences of x.

HON applies a low-pass filter to ternary relations among the selected patterns derived from the discretization step in Figure 2. For instance, consider the heart rate time-series illustrated in Figure 3. An algorithm that only identifies first order relations could describe the probabilities of going from a heart rate of 90 bpm to 100 or to 120. HONs, which can identify more than first order relations, can describe different probabilities for heart rate transitions that go from 80 to 90 to 100 (or to 120) compared to heart rate transitions that go from only 100 to 90 to 100 (or to 120 bpm). PCA was used to reduce feature dimensions to a target $n\_component = 5$ from 727 original features (transition probabilities). HON captures transition probabilities across individuals, while heart-rate and heart-rate variability are within subject features. Thus, we capture both general and individual heart patterns.

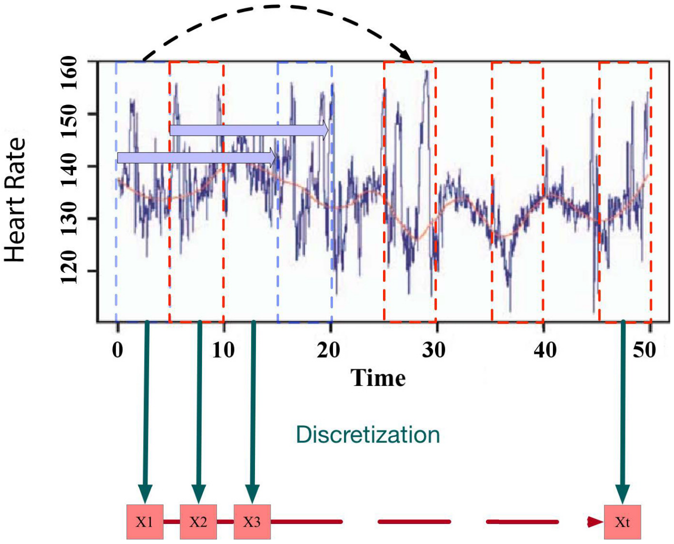

Fig. 2. Discretization, Stabilization, Regularization. Blue continuous lines = the time series; red continuous lines = mean value of the series; the dashed curve arrow = example of a sequential long-term dependency that could be lost without the use of HONs; blue thick arrows = set of higher order dependencies considered by HON.

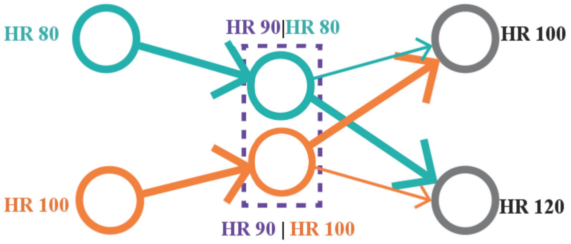

Fig. 3. HON—Heart rate case example. The node HR90 is broken down with HONs by including information about the path. Thus, HR90 originated from HR80 will have different probabilities to link to HR110 (or HR120) than HR90 originated from HR100. The arrows have different width to represent the relative difference in probability.

We investigated different small orders (1-5) of HONs and chose order 2. The number of transition probabilities exponentially increases as the order of the network increases, which lead to a sparsity problem. In particular, most transition probabilities of each individual might be zero as the order increases. As the number of elements in a possible transition increase, the transition may not be associated to a participant.

E. Imputation

Two approaches were used: (1) a theoretically-driven approach that attempted to fuse data across multiple sensing streams using the knowledge of subject-matter experts (e.g., sleep can be fused between the wearable, and smartphone [139]) and (2) a data-driven approach that can vary across the various features (impute via the mean, impute via zeros, etc.). For the joint prediction of the physical, psychological, and job-performance variables, we also performed sensor-wide imputation. For this purpose, we considered the data from one stream and performed clustering on it. This allowed us to impute missing data in one stream from data in another based on the relationships between sensor streams. Other techniques applied include mean and median value imputation. We also performed data imputation using individual rolling means, i.e., individual mean value up to the specific moment. If there was no record at all, we used the global mean.

The level of sparsity was a critical challenge for the phone agent data at the raw data level. However, this was overcome by carefully selecting regularity-based features. Regularity features can capture rhythms and routines within a participant, namely the patterns within hourly phone usage, physical activity and mobility across the participant's time series. Additionally, we had to deal with sparsity for heart rate variability (HRV) when the size of the window used to compute the HRV was not adequate over the 5-minute windows [140]. Some sparsity was also due to data quality issues. Since HRV windows are calculated using Beat-to-Beat-Interval (BBI), many windows did not have a minimal number of BBI readings. This was due to inconsistencies in the data updates from the wearable. HON selection also had sparsity constraints, as higher order networks provided no further information than lower order ones. Namely, we combined the features from each stream/data source and then we applied our regression models for prediction purposes.

F. Fusion

For the joint prediction of the physical, psychological, and job-performance variables, we also used a feature fusion method to combine the various modalities. The features from each stream/data source are combined and fed into our regression and classification candidates for the automated selection of the model at the model selection step. Thus, we capture moments from the distribution of features that provide a summary of each of the modalities. For numerical features, we use summary statistics: mean, median, standard deviation, minimum, and maximum of the distributions. For time series data, we use the features extracted from HON and from other summary statistics. For the PhoneAgent we considered regularity based high-level representations, as well as the imputed values that help model building at the component-of-the-ensemble level. The specific prediction models as well as the relevant features were selected by the cross-validation process. For the final ensemble, we considered a model selection.

G. Model Selection

Using the elements described so far, we built the components of the ensemble learning model by combining the HON features (heart and stress), heart rate, social media, beacons, phone agent, and wearable. We did so with the following steps:

1) *Feature pre-selection.* We use both the sequential exploration of various combinations of features to identify a set of predictive features per construct and the social media anonymization of features described before.
2) *Relevance-based feature selection.* Each specific technique uses an a priori relevance (measured by correlation) on the training set (linear or non-linear correlation).
3) *Model selection.* Automated machine learning methods are applied to decide the best set of features along with the best classifier/regressor per construct.
4) *Proxy ground truth.* We considered the predicted values for AUDIT and OCB, due to higher predictability, in order to perform prediction of other values. We then use these predictions and loop back to the previous step.

Dimensionality reduction through principal component analysis was applied on HON construction (both stress and heart rate) and on social media data. The candidate-components were described in Section V-A, where a component is selected as a member of the final predictor ensemble. The main training and predictions are shown in Algorithm 1.

Algorithm 1 Joint Model

```
Input: Multimodal Data $D$
Output: Predictions
Divide $D$ in training and validation sets $T, V$
Use $T$ to apply feature selection (top 20 features per modality with highest correlation to the nineteen constructs) to select candidate features
for each ground truth variable do
    for each parameter-set do
        for fold $= 1$ to 5 of $T$ do
            for each candidate-component do
                Predict on current fold using candidate-component trained on the remaining folds
            end for
        end for
        SelectedComponent ← candidate with highest score across the folds
        Add SelectedComponent to Ensemble
        Add the fold-wise SelectedComponent predictions $F$
    end for
end for
model ← train the ensemble on $T$
Set predictions $P$ ← Predict($V$,model)
return $F, P$
```

The data and code are in the process of being released through the supervision of IARPA. In order to protect human subjects, only non-identifiable information will be released. We will make the data available through the Open Science Foundation and a corresponding Data Use Agreement (DUA). Information can be found at https://tesserae.nd.edu.

VI. Experiments

We evaluate our approach using four sets of experiments. First, we investigate the performance of our model when compared to a baseline constructed with estimators derived from the ground truth values as detailed below. Second, we verify the bivariate and discriminant criterion validity. Third, we verify the model reliability under 5-fold cross-validation to ensure generalizability. Finally, an external team validated our model on a sub-cohort of participants whose data was withheld from our team during model development.

A. Data Setup

We consider the twelve standardized tests administered in our initial ground-truth battery that contained all 19 dependent variables used for prediction. The independent variables came from various data sources (a wearable, a phone app, four beacons, and social media data) as described in Section IV-A.

1) Data Selection and Feature Set: We perform preprocessing of all the streams previous to fusion of the features as described in Section V-B.

2) Metrics: We use the Kendall's τ correlation coefficient which is a non-parametric measure of correlation based on rank statistics and, thus, assumes no specific structure of the data. Specifically, to compute τ scores, we apply the General Monotone Model (GeMM) [141].

B. Results

We test our framework using 4 sets of evaluations:

1) Validation vs. Theory-Driven Baseline: In this set of experiments, we verify that the performance of our model is comparable to theory-driven (survey-based) predictions. To create a baseline theoretical model, we use the distributions of each variable and take the expected value of the training folds to estimate the values for the test fold. Table V shows the symmetric mean absolute percentage error (SMAPE) for each of the variables and the two models. This table demonstrates that using only sensor-based estimates (our framework) leads to estimations with smaller errors compared to a baseline that is based on surveys. The sensor-based predictions are entirely based on wearable sensor and social media data; no survey or demographic data that could otherwise facilitate identification of patterns based on personal traits were used. Our method shows improved performance over the baseline.

TABLE V: Performance SMAPE (%)—sensors vs. baseline

Variable	Sensor-Based	Baseline
IRB	3.8	7.9
ITP	4.6	9.4
OCB	6.8	14.2
Interpersonal Deviance	18.7	32.9
Organizational Deviance	14.8	28.5
Abstraction	6.4	13.4
Vocabulary	4.2	8.8
Extraversion	8.3	17.5
Agreeableness	5.7	11.6
Conscientiousness	6.8	14.2
Neuroticism	12.6	26.0
Openness	6.4	13.2
Positive Affect	6.6	13.5
Negative Affect	11.4	22.2
Anxiety	10.1	19.9
Alcohol	30.6	70.4
Tobacco	92.2	195.6
Physical Activity	30.8	68.9
Sleep	13.4	27.3

2) Job Performance—Improvement Assessment Over Participant-Oriented Baseline: We evaluate the interplay of psychological and job-performance variables. Thus, we compare the τ score accounted for by sensor-derived estimates, beyond what is accounted for by two known predictors of job performance, personality ([11], [12], [14], [15], [30], [54]) and cognitive ability [53], [142]. This baseline is then compared with the estimation of job performance using our sensor-based framework. To assess the relevance of our estimations, we considered the τ-score of each job performance variable when predicted with the baseline and with the sensor-derived estimates. Table VIII shows the expected τ-score for both the theoretical baseline and our framework. This table shows our sensor-based model has a higher expected performance for all the variables. However, in order to fully validate this result, we also performed a comparative analysis of the full set of predictions using a 5-fold cross-validation approach. Then, we built the distribution of differences of estimations based on our model minus the estimations based on theory. The mode of the distributions of all the job performance variables lie in the region $\Delta_\tau > 0$ where Δ_τ is the difference of τ scores. Thus, our model performs better than the theory-based estimation in the majority of cases.

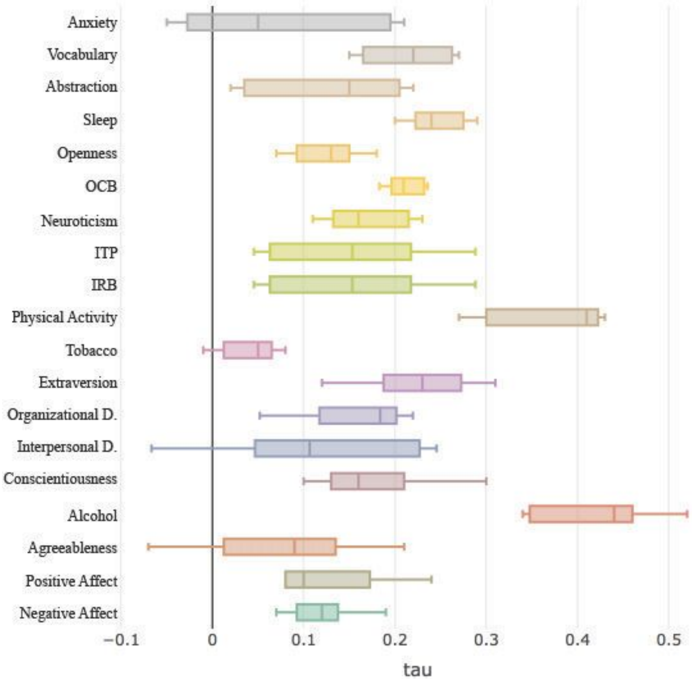

Fig. 4. Model Reliability—cross validation: Kendall's τ confidence interval of the socio-psycho-physiological variable predictions. Distribution 5-fold cv.

3) Discriminant Validity: We examined correlations between constructs and their predictions. The discriminant validity is shown in Table VII. The objective was to verify whether our model sufficiently discriminated the constructs which is signaled by low inter-correlations among constructs and the predictions of other constructs (main diagonal = -). As shown in Table VII, the model has good discriminant validity with correlations in the range $[-0.21, 0.2]$.

4) Model Reliability: Figure 4 shows the distribution of τ scores for each of the variables estimated. We build this distribution by running a 5-fold cross-validation and thus, all the values obtained are created from independently built models. Our assessment goes beyond single-dimensional summaries of performance to include measures of variation, confidence,

TABLE VI: Model Reliability—External Validation

Variable	Min	Max	Mean
Vocabulary	0.00	0.26	0.10
Abstraction	0.00	0.27	0.11
Extraversion	0.02	0.36	0.19
Agreeableness	-0.09	0.22	0.07
Conscientiousness	0.05	0.31	0.15
Neuroticism	-0.19	0.18	0.00
Openness	-0.11	0.305	0.14
Positive Affect	-0.07	0.32	0.16
Negative Affect	-0.16	0.18	0.01
Anxiety	0.00	0.31	0.14
Alcohol	0.22	0.49	0.37
Tobacco	-0.13	0.00	-0.04
Physical Activity	0.20	0.52	0.37
Sleep	0.03	0.35	0.20

and distributional information. As we can see in the figure, the variables for which our framework performs the best are physical variables, followed by job performance and psychological constructs. Tobacco use and job performance are the most challenging constructs to predict.

5) External Validation—Totally Unknown Cohort: We provided an external evaluation team with a pipeline and data to corroborate our results on a sub-cohort of participants whose data was totally unavailable to us during model development. This validation was administered by the Testing and Evaluation Team of the Project Sponsor. The independent evaluation was performed on variables other than job performance and can be seen in Table VI. These results are consistent with what we report in Figure 4. The evaluation shows τ scores in similar ranges as those obtained in our experiments Figure 4. It also shows that variables such as agreeableness, neuroticism, openness, affect (negative affect in particular), and tobacco consumption are the hardest variables to predict, and some variables have less stability than others. However, despite this challenge, some of these variables (e.g., openness, positive affect), have a mean τ performance greater than 0.14. A few variables show better performance in our 5-fold validation (Figure 4) and a few show better performance in the external evaluation. When assessing these results the most conservative values should be considered as reference.

VII. Discussion

The experiments suggest our model is stable with non-trivial predictive performance that is better than construct-based alternative baselines. The performance of our technique is usually better for physical variables of well-being such as alcohol consumption, sleep, and physical activity. The performance is also competitive for psychological and job performance variables. We verified the significance of our sensor-based predictions compared to a participant-oriented baseline to predict job performance variables. The linear-mixed model predictions based on our estimates produced better τ-scores than predictions based on survey estimates. Thus, our framework has better bivariate criterion validity.

The discriminant validity analysis show that our framework sufficiently identifies the various constructs with small absolute values [143] for the correlations [-0.21,0.2]. The reliability analysis shows that the model is reliable as a reflection of the prediction performance. Roughly speaking, physical variables are the most reliably sensed and estimated, followed by psychological and job performance constructs. The most challenging variables in terms of reliability are anxiety, tobacco consumption, interpersonal deviance, and agreeableness. The lower limit for the performance ranges can be explained from the difficulty of modeling social constructs in general and human performance in particular [19]. These reliability results were verified externally.

Our work provides a realistic assessment of the performance of prediction algorithms. This was done by respecting the nature of the data. We did not curate nor select data objects for optimal performance. Instead, we worked with the full original dataset which included individuals with both full and partial sets of features and modalities. While we experimented with neural networks techniques as candidate components (not included in this article to simplify exposition), the models with highest weight in the final ensemble were always parsimonious models. This is consistent with previous observations [144] that simple models perform as well as more complex models, in an empirical realization of the Occam Razor in the social sciences. The first week or two add noise (low/irregular compliance) yet our model's performance is stable despite the missingness. Our model provides a simpler alternative for dealing with missingness without the use of complex likelihood-based or similar techniques (e.g., [22], [23]).

Our work focuses on individual traits assessed through psychological constructs. Specific context about emotionality may be an area for future work. Sentiment analysis [107], [145], [146] could be applied for feature extraction. Another area of possible work is adding features detailing context, e.g., by combining text features with commonsense knowledge.

VIII. Conclusions

Assessing workplace performance, psychological, and physical characteristics of individuals usually relies on existing full traditional questionnaires, on subjective evaluations. Furthermore, current predictive techniques are effective in limited situations including small subsets of variables, subsets of highly curated data, or with a focus on a few variables without a global overview of an individual. In this paper, we presented the first modeling framework and benchmark that leverages sensor data from multimodal sources to jointly predict psychological, physical and physiological, and job-performance constructs. We used traditional social and psychological questionnaires to create the ground truth variables. We used objective mobile and personal sensing data from social media, phones, wearable and beacons as predictors and offer new insights into behavioral patterns that distinguish our various constructs. We presented results from our study of 757 information workers collected over a period ranging from 15 days to 60 days. We created a global ensemble learning algorithm that takes advantage of various data mining techniques and feature extraction approaches to achieve our joint prediction problem on messy data. Our results indicated that our modeling framework allows for a prediction performance above baselines. Despite the wealth of sources and features we used, predicting job performance and psychological constructs is a harder task than predicting physical well-being (alcohol consumption, sleep, etc.). While predicting job performance is a difficult task, most physical variables were predicted well. The predictions in our 5-fold validations and in the externally validated sample were comparable.

Our contribution is three-fold. First, we identified strategies for integrating highly heterogeneous data without curation, and thus, maintained the data integrity. Second, we analyzed the different challenges presented by non-curated data with a systematic feature mining approach. Third, we created a benchmark for predictive tasks by leveraging the identified challenges of the real noisy or incomplete multi-modal high-dimensional data to create a comprehensive prediction and assessment of well-being: physical, psychological, and workplace well-being characteristics of individuals. Development of effective affect-computing systems must include the century-long research on emotion created by psychology. Thus, we contributed in this area as well. Our work's realistic assessment of machine learning applied to performance prediction could also provide benefits for mitigating bias [147]. Our work can be used towards the creation of more objective measures of job performance, and as a realistic and sound baseline for analysis.

ACKNOWLEDGEMENT

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017-17042800007. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] A. Abbas, M. Ali, M. U. S. Khan, and S. U. Khan, "Personalized healthcare cloud services for disease risk assessment and wellness management using social media," *Pervasive and Mobile Computing*, vol. 28, pp. 81–99, 2016.

[2] R. Z. Goetzel, R. M. Henke, M. Tabrizi, K. R. Pelletier, R. Loeppke, D. W. Ballard, J. Grossmeier, D. R. Anderson, D. Yach, R. K. Kelly, T. McCalister, S. Serxner, C. Selecky, L. G. Shallenberger, J. F. Fries, C. Baase, F. Isaac, K. A. Crighton, P. Wald, E. Exum, D. Shurney, and R. D. Metz, "Do workplace health promotion (wellness) programs work?" *Journal of Occupational and Environmental Medicine*, vol. 56, no. 9, pp. 927–934, 2014.

[3] S. G. Aldana, R. M. Merrill, K. Price, A. Hardy, and R. Hager, "Financial impact of a comprehensive multisite workplace health promotion program," *Preventive Medicine*, vol. 40, no. 2, pp. 131–137, 2005.

[4] X. Yang, C. Ge, B. Hu, T. Chi, and L. Wang, "Relationship between quality of life and occupational stress among teachers," *Public Health*, vol. 123, no. 11, pp. 750–755, 2009.

[5] A. Sano, P. Johns, and M. Czerwinski, "Designing opportune stress intervention delivery timing using multi-modal data," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 346–353.

[6] J. M. Smyth, M. J. Sliwinski, M. J. Zawadzki, S. B. Scott, D. E. Conroy, S. T. Lanza, D. Marcusson-Clavertz, J. Kim, R. S. Stawski, C. M. Stoney, O. M. Buxton, C. N. Sciamanna, P. M. Green, and D. M. Almeida, "Everyday stress response targets in the science of behavior change," *Behaviour Research and Therapy*, vol. 101, pp. 20–29, 2018.

[7] M. Quante, R. Wang, J. Weng, E. R. Kaplan, M. Rueschman, E. M. Taveras, S. L. Rifas-Shiman, M. W. Gillman, and S. Redline, "Seasonal and weather variation of sleep and physical activity in 12–14-year-old children," *Behavioral Sleep Medicine*, vol. 17, no. 4, pp. 398–410, 2019.

[8] T. A. Wright and R. Cropanzano, "Psychological well-being and job satisfaction as predictors of job performance," *Journal of Occupational Health Psychology*, vol. 5, no. 1, p. 84, 2000.

[9] D. S. Chiaburu, I.-S. Oh, C. M. Berry, N. Li, and R. G. Gardner, "The five-factor model of personality traits and organizational citizenship behaviors: A meta-analysis," *Journal of Applied Psychology*, vol. 96, no. 6, p. 1140, 2011.

[10] D. Kamdar and L. Van Dyne, "The joint effects of personality and workplace social exchange relationships in predicting task performance and citizenship performance," *Journal of Applied Psychology*, vol. 92, no. 5, p. 1286, 2007.

[11] M. R. Barrick and M. K. Mount, "The Big Five personality dimensions and job performance: a meta-analysis," *Personnel Psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[12] M. R. Barrick, M. K. Mount, and T. A. Judge, "Personality and performance at the beginning of the new millennium: What do we know and where do we go next?" *International Journal of Selection and Assessment*, vol. 9, no. 1-2, pp. 9–30, 2001.

[13] H. J. Eysenck, Ed., *A Model for Intelligence*. Springer Berlin Heidelberg, 1982.

[14] J. F. Salgado, "The five factor model of personality and job performance in the European Community," *Journal of Applied Psychology*, vol. 82, no. 1, p. 30, 1997.

[15] R. P. Tett, D. N. Jackson, and M. Rothstein, "Personality measures as predictors of job performance: a meta-analytic review," *Personnel Psychology*, vol. 44, no. 4, pp. 703–742, 1991.

[16] J. Andreu-Perez, D. Leff, H. M. Ip, and G.-Z. Yang, "From wearable sensors to smart implants – toward pervasive and personalized healthcare," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2750–2762, 2015.

[17] F. Schaule, J. O. Johanssen, B. Bruegge, and V. Loftness, "Employing consumer wearables to detect office workers' cognitive load for interruption management," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, p. 32, 2018.

[18] D. O. Olguín, P. A. Gloor, and A. S. Pentland, "Capturing individual and group behavior with wearable sensors," in *Proceedings of the 2009 AAAI Spring Symposium on Human Behavior Modeling, SSS*, vol. 9, 2009, pp. 68–74.

[19] M. J. Salganik, I. Lundberg, A. T. Kindel, C. E. Ahearn, K. Al-Ghoneim, A. Almaatouq, D. M. Altschul, J. E. Brand, N. B. Carnegie, R. J. Compton, D. Datta, T. Davidson, A. Filippova, C. Gilroy, B. J. Goode, E. Jahani, R. Kashyap, A. Kirchner, S. McKay, A. C. Morgan, A. Pentland, K. Polimis, L. Raes, D. E. Rigobon, C. V. Roberts, D. M. Stanescu, Y. Suhara, A. Usmani, E. H. Wang, M. Adem, A. Alhajri, B. AlShebli, R. Amin, R. B. Amos, L. P. Argyle, L. Baer-Bositis, M. Büchi, B.-R. Chung, W. Eggert, G. Faletto, Z. Fan, J. Freese, T. Gadgil, J. Gagné, Y. Gao, A. Halpern-Manners, S. P. Hashim, S. Hausen, G. He, K. Higuera, B. Hogan, I. M. Horwitz, L. M. Hummel, N. Jain, K. Jin, D. Jurgens, P. Kaminski, A. Karapetyan, E. H. Kim, B. Leizman, N. Liu, M. Möser, A. E. Mack, M. Mahajan, N. Mandell, H. Marahrens, D. Mercado-Garcia, V. Mocz, K. Mueller-Gastell, A. Musse, Q. Niu, W. Nowak, H. Omidvar, A. Or, K. Ouyang, K. M. Pinto, E. Porter, K. E. Porter, C. Qian, T. Rauf, A. Sargsyan, T. Schaffner, L. Schnabel, B. Schonfeld, B. Sender, J. D. Tang, E. Tsurkov, A. van Loon, O. Varol, X. Wang, Z. Wang, J. Wang, F. Wang, S. Weissman, K. Whitaker, M. K. Wolters, W. L. Woon, J. Wu, C. Wu, K. Yang, J. Yin, B. Zhao, C. Zhu, J. Brooks-Gunn, B. E. Engelhardt, M. Hardt, D. Knox, K. Levy, A. Narayanan, B. M. Stewart, D. J. Watts, and S. McLanahan, "Measuring the predictability of life outcomes with a scientific mass collaboration," *Proceedings of the National Academy of Sciences*, vol. 117, no. 15, pp. 8398–8403, 2020.

[20] T. A. Judge and C. P. Zapata, "The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the Big Five personality traits in predicting job performance," *Academy of Management Journal*, vol. 58, no. 4, pp. 1149–1179, 2015.

[21] C. Viswesvaran and D. S. Ones, "Perspectives on models of job performance," *International Journal of Selection and Assessment*, vol. 8, no. 4, pp. 216–226, 2000.

[22] C. H. Mallinckrodt, T. M. Sanger, S. Dubé, D. J. DeBrota, G. Molenberghs, R. J. Carroll, W. Z. Potter, and G. D. Tollefson, "Assessing and interpreting treatment effects in longitudinal clinical trials with missing data," *Biological Psychiatry*, vol. 53, no. 8, pp. 754–760, 2003.

[23] G. Molenberghs, H. Thijs, I. Jansen, C. Beunckens, M. Kenward, C. Mallinckrodt, and R. Carroll, "Analyzing incomplete longitudinal clinical trial data," *Biostatistics*, vol. 5, no. 3, pp. 445–64, 2004.

[24] S. M. Mattingly, J. M. Gregg, P. Audia, A. E. Bayraktaroglu, A. T. Campbell, N. V. Chawla, V. Das Swain, M. De Choudhury, S. K. D'Mello, A. K. Dey, G. Gao, K. Jagannath, K. Jiang, S. Lin, Q. Liu, G. Mark, G. J. Martinez, K. Masaba, S. Mirjafari, E. Moskal, R. Mulukutla, K. Nies, M. D. Reddy, P. Robles-Granda, K. Saha, A. Sirigiri, and A. Striegel, "The Tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, p. CS11.

[25] A. Madan, S. T. Moturu, D. Lazer, and A. S. Pentland, "Social sensing: obesity, unhealthy eating and exercise in face-to-face networks," in *Wireless Health 2010*. ACM, 2010, pp. 104–110.

[26] N. J. Yuan, F. Zhang, D. Lian, K. Zheng, S. Yu, and X. Xie, "We know how you live: Exploring the spectrum of urban lifestyles," in *Proceedings of the First ACM Conference on Online Social Networks*. ACM, 2013, pp. 3–14.

[27] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, 2013.

[28] B. Mariani, M. C. Jiménez, F. J. G. Vingerhoets, and K. Aminian, "On-shoe wearable sensors for gait and turning assessment of patients with Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 155–158, 2013.

[29] L. J. Williams and S. E. Anderson, "Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors," *Journal of Management*, vol. 17, no. 3, pp. 601–617, 1991.

[30] M. A. Griffin, A. Neal, and S. K. Parker, "A new model of work role performance: Positive behavior in uncertain and interdependent contexts," *Academy of Management Journal*, vol. 50, no. 2, pp. 327–347, 2007.

[31] S. Fox, P. E. Spector, A. Goh, K. Bruursema, and S. R. Kessler, "The deviant citizen: Measuring potential positive relations between counterproductive work behaviour and organizational citizenship behaviour," *Journal of Occupational and Organizational Psychology*, vol. 85, no. 1, pp. 199–220, 2012.

[32] R. J. Bennett and S. L. Robinson, "Development of a measure of workplace deviance," *Journal of Applied Psychology*, vol. 85, no. 3, p. 349, 2000.

[33] W. C. Shipley, C. P. Gruber, T. A. Martin, and A. M. Klein, *Shipley-2 Manual*, Western Psychological Services, Los Angeles, CA, 2009.

[34] C. J. Soto and O. P. John, "The next Big Five Inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power," *Journal of Personality and Social Psychology*, vol. 113, no. 1, p. 117, 2017.

[35] D. Watson and L. Clark, *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form*. University of Iowa, 1994.

[36] C. D. Spielberger, G. A. Jacobs, S. Russell, and R. S. Crane, "Assessment of anger: The state-trait anger scale," *Advances in Personality Assessment*, vol. 2, pp. 161–189, 1983.

[37] J. B. Saunders, O. G. Aasland, T. F. Babor, J. R. De la Fuente, and M. Grant, "Development of the alcohol use disorders identification test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II," *Addiction*, vol. 88, no. 6, pp. 791–804, 1993.

[38] K. M. Palipudi, J. Morton, J. Hsia, L. Andes, S. Asma, B. Talley, R. D. Caixeta, H. Fouad, R. N. Khoury, N. Ramanandraibe, J. Rarick, D. N. Sinha, S. Pujari, and E. Tursan d'Espaignet, "Methodology of the global adult tobacco survey—2008–2010," *Global Health Promotion*, vol. 23, no. 2_suppl, pp. 3–23, 2016.

[39] C. L. Craig, A. L. Marshall, M. Sjöström, A. E. Bauman, M. L. Booth, B. E. Ainsworth, M. Pratt, U. Ekelund, A. Yngve, J. F. Sallis, and P. Oja, "International physical activity questionnaire: 12-country reliability and validity," *Medicine & Science in Sports & Exercise*, vol. 35, no. 8, pp. 1381–1395, 2003.

[40] D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research," *Psychiatry Research*, vol. 28, no. 2, pp. 193–213, 1989.

[41] J. Xu, T. L. Wickramarathne, and N. V. Chawla, "Representing higher-order dependencies in networks," *Science Advances*, vol. 2, no. 5, p. 10, 2016.

[42] S. Mirjafari, K. Masaba, T. Grover, W. Wang, P. Audia, A. T. Campbell, N. V. Chawla, V. D. Swain, M. D. Choudhury, A. K. Dey, S. K. D'Mello, G. Gao, J. M. Gregg, K. Jagannath, K. Jiang, S. Lin, Q. Liu, G. Mark, G. J. Martinez, S. M. Mattingly, E. Moskal, R. Mulukutla, S. Nepal, K. Nies, M. D. Reddy, P. Robles-Granda, K. Saha, A. Sirigiri, and A. Striegel, "Differentiating higher and lower job performers in the workplace using mobile sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, p. 37, 2019.

[43] M. Rotundo and P. R. Sackett, "The relative importance of task, citizenship, and counterproductive performance to global ratings of job performance: A policy-capturing approach," *Journal of Applied Psychology*, vol. 87, no. 1, p. 66, 2002.

[44] W. C. Borman and S. Motowidlo, "Expanding the criterion domain to include elements of contextual performance," *Personnel Selection in Organizations*, pp. 71–98, 1993.

[45] J. M. Cortina and J. N. Luchman, "Personnel selection and employee performance," *Handbook of Psychology, Second Edition*, vol. 12, pp. 143–183, 2012.

[46] R. S. Dalal, H. Lam, H. M. Weiss, E. R. Welch, and C. L. Hulin, "A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, and dynamic relationships with affect and overall job performance," *Academy of Management Journal*, vol. 52, no. 5, pp. 1051–1066, 2009.

[47] S. J. Motowidlo and H. J. Kell, "Job performance," *Handbook of Psychology, Second Edition*, vol. 12, pp. 82–103, 2012.

[48] J. P. Campbell, "Modeling the performance prediction problem in industrial and organizational psychology," *Handbook of Industrial and Organizational Psychology*, pp. 687–732, 1990.

[49] J. P. Campbell and B. M. Wiernik, "The modeling and assessment of work performance," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 2, no. 1, pp. 47–74, 2015.

[50] P. R. Sackett, "The structure of counterproductive work behaviors: Dimensionality and relationships with facets of job performance," *International Journal of Selection and Assessment*, vol. 10, no. 1-2, pp. 5–11, 2002.

[51] R. B. Cattell, *Intelligence: Its Structure, Growth and Action*. Elsevier, 1987, vol. 35.

[52] W. J. Schneider and K. S. McGrew, "The Cattell-Horn-Carroll model of intelligence," *Contemporary Intellectual Assessment: Theories, Tests, and Issues*, pp. 99–144, 2012.

[53] F. L. Schmidt and J. Hunter, "General mental ability in the world of work: occupational attainment and job performance," *Journal of Personality and Social Psychology*, vol. 86, no. 1, pp. 162–173, 2004.

[54] G. Anderson and C. Viswesvaran, "An update of the validity of personality scales in personnel selection: A meta-analysis of studies published after 1992," in *13th Annual Conference of the Society of Industrial and Organizational Psychology, Dallas*, 1998, p. 10.

[55] D. Bartram, "The great eight competencies: a criterion-centric approach to validation," *Journal of Applied Psychology*, vol. 90, no. 6, pp. 1185–1203, 2005.

[56] G. J. Feist, "A meta-analysis of personality in scientific and artistic creativity," *Personality and Social Psychology Review*, vol. 2, no. 4, pp. 290–309, 1998.

[57] P. Y. Chen and P. E. Spector, "Relationships of work stressors with aggression, withdrawal, theft and substance use: An exploratory study," *Journal of Occupational and Organizational Psychology*, vol. 65, no. 3, pp. 177–184, 1992.

[58] L. M. Penney and P. E. Spector, "Job stress, incivility, and counterproductive work behavior (CWB): The moderating role of negative affectivity," *Journal of Organizational Behavior*, vol. 26, no. 7, pp. 777–796, 2005.

[59] S. Fox, P. E. Spector, and D. Miles, "Counterproductive work behavior (CWB) in response to job stressors and organizational justice: Some mediator and moderator tests for autonomy and emotions," *Journal of Vocational Behavior*, vol. 59, no. 3, pp. 291–309, 2001.

[60] S. Fox and P. E. Spector, "A model of work frustration–aggression," *Journal of Organizational Behavior*, vol. 20, no. 6, pp. 915–931, 1999.

[61] A. Mackinnon, A. F. Jorm, H. Christensen, A. E. Korten, P. A. Jacomb, and B. Rodgers, "A short form of the positive and negative affect schedule: Evaluation of factorial validity and invariance across demo-

graphic variables in a community sample," *Personality and Individual Differences*, vol. 27, no. 3, pp. 405–416, 1999.

[62] J. Mullahy and J. Sindelar, "Effects of alcohol on labor market success: Income, earnings, labor supply, and occupation," *Alcohol Health & Research World*, vol. 16, no. 2, pp. 134–139, 1992.

[63] D. M. Podolsky and D. Richards, "Investigating the role of substance abuse in occupational injuries," *Alcohol Health & Research World*, vol. 9, no. 4, pp. 42–5, 1985.

[64] P. R. Giancola and H. B. Moss, "Executive cognitive functioning in alcohol use disorders," in *Recent Developments in Alcoholism*. Springer, 1998, pp. 227–251.

[65] M. Galanter, Ed., *Recent Developments in Alcoholism - The Consequences of Alcoholism-Medical, Neuropsychiatric, Economic, Cross-Cultural*. Springer US, 1998, vol. 14.

[66] A. M. Hussong, R. E. Hicks, S. A. Levy, and P. J. Curran, "Specifying the relations between affect and heavy alcohol use among young adults," *Journal of Abnormal Psychology*, vol. 110, no. 3, pp. 449–461, 2001.

[67] S. V. Paunonen, "Big Five factors of personality and replicated predictions of behavior," *Journal of Personality and Social Psychology*, vol. 84, no. 2, pp. 411–424, 2003.

[68] M. J. Bohn, T. F. Babor, and H. R. Kranzler, "The alcohol use disorders identification test (AUDIT): Validation of a screening instrument for use in medical settings," *Journal of Studies on Alcohol*, vol. 56, no. 4, pp. 423–432, 1995.

[69] M. Piccinelli, E. Tessari, M. Bortolomasi, O. Piasere, M. Semenzin, N. Garzotto, and M. Tansella, "Efficacy of the alcohol use disorders identification test as a screening tool for hazardous alcohol intake and related disorders in primary care: a validity study," *BMJ*, vol. 314, no. 7078, pp. 420–424, 1997.

[70] R. J. Volk, J. R. Steinbauer, S. B. Cantor, and C. E. HOLZER III, "The alcohol use disorders identification test (AUDIT) as a screen for at-risk drinking in primary care patients of different racial/ethnic backgrounds," *Addiction*, vol. 92, no. 2, pp. 197–206, 1997.

[71] N. P. Pronk, B. Martinson, R. C. Kessler, A. L. Beck, G. E. Simon, and P. Wang, "The association between work performance and physical activity, cardiorespiratory fitness, and obesity," *Journal of Occupational and Environmental Medicine*, vol. 46, no. 1, pp. 19–25, 2004.

[72] J. J. Ratey and J. E. Loehr, "The positive impact of physical activity on cognition during adulthood: A review of underlying mechanisms, evidence and recommendations," *Reviews in the Neurosciences*, vol. 22, no. 2, pp. 171–185, 2011.

[73] S. Toker and M. Biron, "Job burnout and depression: Unraveling their temporal relationship and considering the role of physical activity," *Journal of Applied Psychology*, vol. 97, no. 3, pp. 699–710, 2012.

[74] P. Salmon, "Effects of physical exercise on anxiety, depression, and sensitivity to stress: A unifying theory," *Clinical Psychology Review*, vol. 21, no. 1, pp. 33–61, 2001.

[75] M. R. Rosekind, K. B. Gregory, M. M. Mallis, S. L. Brandt, B. Seal, and D. Lerner, "The cost of poor sleep: Workplace productivity loss and associated costs," *Journal of Occupational and Environmental Medicine*, vol. 52, no. 1, pp. 91–98, 2010.

[76] L. Barber, M. J. Grawitch, and D. C. Munz, "Are better sleepers more engaged workers? a self-regulatory approach to sleep hygiene and work engagement," *Stress and Health*, vol. 29, no. 4, pp. 307–316, 2013.

[77] C. M. Barnes, "Working in our sleep: Sleep and self-regulation in organizations," *Organizational Psychology Review*, vol. 2, no. 3, pp. 234–257, 2012.

[78] A. C. Parrott, "Does cigarette smoking cause stress?" *American Psychologist*, vol. 54, no. 10, pp. 817–820, 1999.

[79] R. R. McCrae, P. T. Costa Jr, and R. Bossé, "Anxiety, extraversion and smoking," *British Journal of Social and Clinical Psychology*, vol. 17, no. 3, pp. 269–273, 1978.

[80] N. Cherry and K. Kiernan, "Personality scores and smoking behaviour. a longitudinal study," *Journal of Epidemiology & Community Health*, vol. 30, no. 2, pp. 123–131, 1976.

[81] J. D. Kassel, L. R. Stroud, and C. A. Paronis, "Smoking, stress, and negative affect: Correlation, causation, and context across stages of smoking," *Psychological Bulletin*, vol. 129, no. 2, pp. 270–304, 2003.

[82] F. F. Ikard, D. E. Green, and D. Horn, "A scale to differentiate between types of smoking as related to the management of affect," *International Journal of the Addictions*, vol. 4, no. 4, pp. 649–659, 1969.

[83] M. T. Halpern, R. Shikiar, A. M. Rentz, and Z. M. Khan, "Impact of smoking status on workplace absenteeism and productivity," *Tobacco Control*, vol. 10, no. 3, pp. 233–238, 2001.

[84] D. M. Higgins, J. B. Peterson, R. O. Pihl, and A. G. Lee, "Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance," *Journal of Personality and Social Psychology*, vol. 93, no. 2, pp. 298–319, 2007.

[85] T. W. Ng and D. C. Feldman, "The relationship of age to ten dimensions of job performance," *Journal of Applied Psychology*, vol. 93, no. 2, pp. 392–423, 2008.

[86] F. L. Schmidt, "A general theoretical integrative model of individual differences in interests, abilities, personality traits, and academic and occupational achievement: A commentary on four recent articles," *Perspectives on Psychological Science*, vol. 9, no. 2, pp. 211–218, 2014.

[87] M. L. Kohn and C. Schooler, "Job conditions and personality: A longitudinal assessment of their reciprocal effects," *American Journal of Sociology*, vol. 87, no. 6, pp. 1257–1286, 1982.

[88] J. Tamez-Pena, J. Orozco, P. Sosa, A. Valdes, and F. Nezhadmoghadam, "Ensemble of SVM, Random-Forest and the BSWiMS method to predict and describe structural associations with fluid intelligence scores from T1-weighed MRI," in *Challenge in Adolescent Brain Cognitive Development Neurocognitive Prediction*. Springer, 2019, pp. 47–56.

[89] T. P. Alloway and R. G. Alloway, "The impact of engagement with social networking sites (SNSs) on cognitive skills," *Computers in Human Behavior*, vol. 28, no. 5, pp. 1748–1754, 2012.

[90] J. Moutafi, A. Furnham, and L. Paltiel, "Can personality factors predict intelligence?" *Personality and Individual Differences*, vol. 38, no. 5, pp. 1021–1033, 2005.

[91] S. Bai, T. Zhu, and L. Cheng, "Big Five personality prediction based on user behaviors at social network sites," *Preprint arXiv:1204.4809*, 2012.

[92] M. Skowron, M. Tkalčič, B. Ferwerda, and M. Schedl, "Fusing social media cues: Personality prediction from Twitter and Instagram," in *Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 107–108.

[93] J. Shen, O. Brdiczka, and J. Liu, "Understanding email writers: Personality prediction from email messages," in *International Conference on User Modeling, Adaptation, and Personalization*, 2013, pp. 318–330.

[94] M. K. Abadi, J. A. M. Correa, J. Wache, H. Yang, I. Patras, and N. Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, vol. 1, 2015, pp. 1–8.

[95] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artificial Intelligence Review*, vol. 53, pp. 2313–2339, 2020.

[96] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[97] S. Poria, A. Gelbukh, B. Agarwal, E. Cambria, and N. Howard, "Common sense knowledge based personality recognition from text," in *Mexican International Conference on Artificial Intelligence*. Springer, 2013, pp. 484–496.

[98] F. Celli, "Unsupervised personality recognition for social network sites," in *The Sixth International Conference on Digital Society*, 2012, pp. 59–62.

[99] Y. Saez, C. Navarro, A. Mochon, and P. Isasi, "A system for personality and happiness detection." *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 2, no. 5, pp. 7–15, 2014.

[100] J. A. Golbeck, "Predicting personality from social media text," *AIS Transactions on Replication Research*, vol. 2, no. 1, pp. 1–10, 2016.

[101] M. P. Kalghatgi, M. Ramannavar, and N. S. Sidnal, "A neural network approach to personality prediction based on the Big Five model," *International Journal of Innovative Research in Advanced Engineering*, vol. 2, no. 8, pp. 56–63, 2015.

[102] P. E. Tetlock and J. I. Kim, "Accountability and judgment processes in a personality prediction task," *Journal of Personality and Social Psychology*, vol. 52, no. 4, pp. 700–709, 1987.

[103] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[104] A. Ghandeharioun, S. Fedor, L. Sangermano, D. Ionescu, J. Alpert, C. Dale, D. Sontag, and R. Picard, "Objective assessment of depressive symptoms with machine learning and wearable sensors data," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 325–332.

[105] C. Liu, K. Conn, N. Sarkar, and W. Stone, “Online affect detection and robot behavior adaptation for intervention of children with autism,” *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 883–896, 2008.

[106] S. Tuarob, C. S. Tucker, S. Kumara, C. L. Giles, A. L. Pincus, D. E. Conroy, and N. Ram, “How are you feeling?: A personalized methodology for predicting mental states from temporally observable physical and behavioral information,” *Journal of Biomedical Informatics*, vol. 68, pp. 1–19, 2017.

[107] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, “Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis,” *Proceedings of the 2020 ACM International Conference on Information and Knowledge Management*, pp. 105–114, 2020.

[108] M. S. Akhtar, A. Ekbal, and E. Cambria, “How intense are you? predicting intensities of emotions and sentiments using stacked ensemble,” *IEEE Computational Intelligence Magazine*, vol. 15, no. 1, pp. 64–75, 2020.

[109] Z. Wang, S.-B. Ho, and E. Cambria, “A review of emotion sensing: Categorization models and algorithms,” *Multimedia Tools and Applications*, pp. 1–30, 2020.

[110] Y. Zheng, T. C. Wong, B. H. Leung, and C. C. Poon, “Unobtrusive and multimodal wearable sensing to quantify anxiety,” *IEEE Sensors Journal*, vol. 16, no. 10, pp. 3689–3696, 2016.

[111] H. Liu, W. Wen, J. Zhang, G. Liu, and Z. Yang, “Autonomic nervous pattern of motion interference in real-time anxiety detection,” *IEEE Access*, vol. 6, pp. 69 763–69 768, 2018.

[112] R. S. Dalal, D. P. Bhave, and J. Fiset, “Within-person variability in job performance: A theoretical review and research agenda,” *Journal of Management*, vol. 40, no. 5, pp. 1396–1436, 2014.

[113] C. D. Spielberger and E. C. Reheiser, “Assessment of emotions: Anxiety, anger, depression, and curiosity,” *Applied Psychology: Health and Well-Being*, vol. 1, no. 3, pp. 271–302, 2009.

[114] W. Mumtaz, P. L. Vuong, L. Xia, A. S. Malik, and R. B. A. Rashid, “Automatic diagnosis of alcohol use disorder using EEG features,” *Knowledge-Based Systems*, vol. 105, pp. 48–59, 2016.

[115] P. R. Marques and A. S. McKnight, “Field and laboratory alcohol detection with 2 types of transdermal devices,” *Alcoholism: Clinical and Experimental Research*, vol. 33, no. 4, pp. 703–711, 2009.

[116] K. P. Lindgren, S. A. Baldwin, J. J. Ramirez, C. C. Olin, K. P. Peterson, R. W. Wiers, B. A. Teachman, J. Norris, D. Kaysen, and C. Neighbors, “Self-control, implicit alcohol associations, and the (lack of) prediction of consumption in an alcohol taste test with college student heavy episodic drinkers,” *PLOS ONE*, vol. 14, no. 1, pp. 1–23, 01 2019.

[117] M. F. Hovell, J. Bellettiere, S. Liles, B. Nguyen, V. Berardi, C. Johnson, G. E. Matt, J. Malone, M. C. Boman-Davis, P. J. Quintana, S. Obayashi, D. Chatfield, R. Robinson, E. J. Blumberg, W. M. Ongkeko, N. E. Klepeis, and S. C. Hughes, “Randomised controlled trial of real-time feedback and brief coaching to reduce indoor smoking,” *Tobacco Control*, vol. 29, no. 2, pp. 183–190, 2020.

[118] S. L. Ridner, “Predicting smoking status in a college-age population,” *Public Health Nursing*, vol. 22, no. 6, pp. 494–505, 2005.

[119] T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. Klasnja, K. Koscher, A. LaMarca, J. A. Landay, L. LeGrand, J. Lester, A. Rahimi, A. Rea, and D. Wyatt, “The mobile sensing platform: An embedded activity recognition system,” *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 32–41, 2008.

[120] G. Plasqui and K. R. Westerterp, “Physical activity assessment with accelerometers: An evaluation against doubly labeled water,” *Obesity*, vol. 15, no. 10, pp. 2371–2379, 2007.

[121] D. Van Dyck, G. Cardon, B. Deforche, and I. De Bourdeaudhuij, “The contribution of former work-related activity levels to predict physical activity and sedentary time during early retirement: Moderating role of educational level and physical functioning,” *PLOS ONE*, vol. 10, no. 3, pp. 1–14, 03 2015.

[122] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, “Toss’n’turn: Smartphone as sleep and sleep quality detector,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 477–486.

[123] A. Sano, Z. Y. Amy, A. W. McHill, A. J. Phillips, S. Taylor, N. Jaques, E. B. Klerman, and R. W. Picard, “Prediction of happy-sad mood from daily behaviors and previous sleep history,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, pp. 6796–6799.

[124] J. Razjouyan, H. Lee, S. Parthasarathy, J. Mohler, A. Sharafkhaneh, and B. Najafi, “Improving sleep quality assessment using wearable sensors by including information from postural/sleep position changes and body acceleration: A comparison of chest-worn sensors, wrist actigraphy, and polysomnography,” *Journal of Clinical Sleep Medicine*, vol. 13, no. 11, pp. 1301–1310, 2017.

[125] P. Staples, J. Torous, I. Barnett, K. Carlson, L. Sandoval, M. Keshavan, and J.-P. Onnela, “A comparison of passive and active estimates of sleep in a cohort with schizophrenia,” *NPJ Schizophrenia*, vol. 3, no. 1, pp. 1–6, 2017.

[126] V. Lynn, N. Balasubramanian, and H. A. Schwartz, “Hierarchical modeling for user personality prediction: The role of message-level attention,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 5306–5316.

[127] A. Kazemeini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, “Personality trait detection using bagged SVM over BERT word embedding ensembles,” in *Proceedings of the ACL Workshop on Widening NLP*. Association for Computational Linguistics, 07 2020, p. 4.

[128] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, “Bottom-up and top-down: Predicting personality with psycholinguistic and language model features,” in *Proceedings of the International Conference of Data Mining*. IEEE, 2020, p. 6.

[129] E. Cambria, “Affective computing and sentiment analysis,” *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.

[130] P. Shaver, J. Schwartz, D. Kirson, and C. O’Connor, “Emotion knowledge: Further exploration of a prototype approach,” *Journal of Personality and Social Psychology*, vol. 52, no. 6, p. 1061, 1987.

[131] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge university press, 1990.

[132] P. Ekman, “An argument for basic emotions,” *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[133] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, “The hourglass model revisited,” *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[134] M. Gjurković, M. Karan, I. Vukojević, M. Bošnjak, and J. Šnajder, “PANDORA talks: Personality and demographics on Reddit,” *Preprint arXiv:2004.04460*, 2020.

[135] R. Xia and Z. Ding, “Emotion-cause pair extraction: A new task to emotion analysis in texts,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1003–1012.

[136] Y. Ma, H. Peng, and E. Cambria, “Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM,” in *Proceedings of the 2018 AAAI Conference on Artificial Intelligence*, 2018, pp. 5876–5883.

[137] S. Sonnentag, J. Volmer, and A. Spychala, “Job performance,” *The Sage Handbook of Organizational Behavior*, vol. 1, pp. 427–447, 2008.

[138] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” *Journal of Language and Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[139] G. J. Martinez, S. M. Mattingly, J. Young, L. Faust, A. K. Dey, A. T. Campbell, M. De Choudhury, S. Mirjafari, S. K. Nepal, P. Robles-Granda, K. Saha, and A. D. Striegel, “Improved sleep detection through the fusion of phone agent and wearable data streams,” in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 2020, pp. 1–6.

[140] M. Malik, J. T. Bigger, A. J. Camm, R. Kleiger, A. Malliani, A. Moss, and P. Schwartz, “Heart rate variability: Standards of measurement, physiological interpretation and clinical use,” *European Heart Journal*, vol. 17, pp. 354–381, 1996.

[141] M. Dougherty and R. Thomas, “Robust decision making in a nonlinear world,” *Psychological Review*, vol. 119, no. 2, pp. 321–344, 2012.

[142] F. L. Schmidt and J. E. Hunter, “The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings,” *Psychological Bulletin*, vol. 124, no. 2, pp. 262–274, 1998.

[143] R. Furr and V. Bacharach, “Validity: Estimating and evaluating convergent and discriminant validity evidence,” *Psychometrics: An Introduction*, pp. 191–235, 2006.

[144] J. Jung, C. Concannon, R. Shroff, S. Goel, and D. G. Goldstein, “Simple rules for complex decisions,” *SSRN 2919024*, p. 9, 2017.

[145] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining,” in *The International Conference on Language Resources and Evaluation*. European Language Resources Association, 2010, pp. 2200–2204.

[146] E. Cambria and A. Hussain, *SenticComputing:A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Springer, 2015.

[147] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 469–481.