S-ADL: Exploring Smartphone-based Activities of Daily Living to Detect Blood Alcohol Concentration in a Controlled Environment

Hansoo Lee
hansoo@kaist.ac.kr
School of Computing, KAIST
Daejeon, South Korea

Auk Kim
kimauk@kangwon.ac.kr
Department of Computer Science and Engineering,
Kangwon National University
Chucheon, South Korea

Sang Won Bae
sbae4@stevens.edu
Charles V. Schaefer, Jr. School of Engineering and Science,
Stevens Institute of Technology
Hoboken, New Jersey, United States

Uichin Lee
uclee@kaist.ac.kr
School of Computing, KAIST
Daejeon, South Korea

ABSTRACT

In public health and safety, precise detection of blood alcohol concentration (BAC) plays a critical role in implementing responsive interventions that can save lives. While previous research has primarily focused on computer-based or neuropsychological tests for BAC identification, the potential use of daily smartphone activities for BAC detection in real-life scenarios remains largely unexplored. Drawing inspiration from Instrumental Activities of Daily Living (I-ADL), our hypothesis suggests that Smartphone-based Activities of Daily Living (S-ADL) can serve as a viable method for identifying BAC. In our proof-of-concept study, we propose, design, and assess the feasibility of using S-ADLs to detect BAC in a scenario-based controlled laboratory experiment involving 40 young adults. In this study, we identify key S-ADL metrics, such as delayed texting in SMS, site searching, and finance management, that significantly contribute to BAC detection (with an AUC–ROC and accuracy of 81%). We further discuss potential real-life applications of the proposed BAC model.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **Empirical studies in HCI**.

KEYWORDS

Alcohol drinking detection, Activities of daily living, Smartphone app usage, Functional assessment, Machine learning

ACM Reference Format:
Hansoo Lee, Auk Kim, Sang Won Bae, and Uichin Lee. 2024. S-ADL: Exploring Smartphone-based Activities of Daily Living to Detect Blood Alcohol Concentration in a Controlled Environment. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3613904.3642832

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642832

1 INTRODUCTION

The recent COVID-19 pandemic has led to changes in the social system (e.g., stay-at-home orders and relaxation of alcohol restrictions) [11], and the stress and depression caused by social isolation have resulted in a significant increase in alcohol consumption among the younger generation [26, 37]. According to previous studies, approximately 50% of young adults aged 18 to 25 have consumed alcohol in the previous month, with approximately 60% of them experiencing a binge drinking episode within the same time frame [2]. Moreover, 49.7% of the younger generation have recently consumed alcohol on a regular basis [3]. These frequent binge drinking behaviors of young adults have led to various unintentional physical health issues (e.g., bodily injuries, diseases) and social problems (e.g., unprotected sex, productivity loss, drunk driving) [1, 89, 91]. However, young adults often struggle to change their frequent binge drinking behaviors compared with other age groups because of factors such as a lack of psychological maturity for impulse control in alcohol use disorder, lack of awareness of their alcohol tolerance, and increased opportunities for alcohol consumption owing to increased social activities accompanied by peer pressure [19, 67]. Therefore, there is a need for a tool designed for young adults that can assist in intervening against alcohol abuse through continuous monitoring of alcohol consumption anytime and anywhere.

Traditional methods measure BAC through self-reporting, transdermal alcohol monitoring, or breathalyzers. Self-reporting methods use formulas (e.g., the Widmark formulation [116]) that require personal information (e.g., sex, weight) and alcohol consumption information (e.g., alcohol content, amount, and time of consumption) to be manually input through a survey or experience sampling method. Nevertheless, these methods rely on the memory of the drinker, which leads to potentially inaccurate results and user burden for repetitive reporting [10]. The common method of transdermal alcohol monitoring (e.g., SCRAM and WrisTAS) involves attaching an ankle bracelet to the skin [104]. However, this measure is delayed by several hours after drinking, making it inappropriate for timely BAC detection [68], and there is a stigma

related to wearing ankle bracelets [12]. Breathalyzers are the most widely used [25]. Recently, Bluetooth-based portable breathalyzers (e.g., BACtrack Mobile Pro [8]) have been developed. Nonetheless, users must always carry the device, and false detections may occur depending on the oral environment and certain diseases (e.g., liver, diabetes, and kidney diseases) [25]. Thus, it is essential to develop a new BAC detection method that can lower user burdens while simultaneously increasing portability to enable immediate self-monitoring of BAC.

At present, 80% of people carry smartphones for 22 hours in their daily lives [4]. People interact with their smartphones for an average of 3 hours and 15 minutes per day [74] and touch their smartphones an average of 2,617 times per day [120], even when they drink alcohol. Therefore, the influence of alcohol consumption can be tracked using smartphones. In the field of HCI, smartphone-enabled functional assessment methods have been developed to automatically measure BAC. Given that after drinking, a functional decline occurs while intoxicated, prior studies on BAC detection have assessed the physical functional decline in terms of motor or psychomotor coordination via smartphones for such detraction [6, 76]. However, the domain and degree of functional decline due to changes in BAC vary among individuals [38]. Although detecting BAC of 0.03% or 0.08%, which is the legal limit for drunk driving in most countries [117, 118], is important, a decline in the motor coordination (e.g., walking, balancing) is not typically evident at these BAC levels [47, 113].

Therefore, in cases where there is a decline in cognitive functions other than motor coordination functions after drinking, it is challenging to detect certain BAC levels (e.g., 0.03% or 0.08%) using the motor function tracking method (e.g., [6]). Therefore, Mariakakis et al. [76] detected BAC by assessing psychomotor control based on a simple choice reaction involving reflexes (e.g., fine motor control and balancing) through smartphone-enabled neuropsychological tests. However, the mild functional decline that arises at BAC of 0.04% is not sensitive to the simple fine motor or psychomotor performance (e.g., stimulus and reaction) [76, 77], varies in domain and level among individuals; thus assessments that are more sensitive to complex cognitive functions than simple cognitive screening tests are required, such as neuropsychological tests [81, 123]. Furthermore, such cognitive screening tests have learning effect issues when measured repeatedly [13, 86].

Activities of daily living (ADL) instruments are fundamental skills required to independently care for oneself [57]. Among ADL instruments, the Instrumental ADL (I-ADL) requires more complex activities and thinking skills related to the ability to live independently in a community (e.g., money transfer and communication with others) [65]. Moreover, before a noticeable cognitive decline occurs in various cognitive domains, there is a decline in I-ADL performance. This makes I-ADL-based functional assessments particularly attuned to detecting mild functional decline compared with conventional neuropsychological tests [81, 123]. Moreover, ADL-based functional assessments have a lower learning effect than neuropsychological tests, making them useful for repetitive BAC measurements [14]. Therefore, ADL-based functional assessments can be more useful for determining varying BAC because people typically exhibit mild or severe functional declines after drinking.

In this study, we aimed to develop Smartphone-based activities of daily living (S-ADL), which require more complex functional skills with a mental workload than the simple choice reaction tasks utilized in prior studies, to automatically detect mild functional changes associated with varying BAC phases (normal: 0%, mild drinking: 0.03%–0.04%, heavy drinking: 0.07%–0.08%) and explore the feasibility of using S-ADL for BAC detection. Therefore, we answered the following research questions: RQ1. How can S-ADL be effectively designed to identify BAC? RQ2. Among the S-ADL-based performance metrics considered for building a machine learning model, which specific metrics demonstrate the most substantial influence on the accuracy and reliability of the BAC model?

We first developed the S-ADL method by adopting an ADL-based functional assessment and expanding the existing smartphone-enabled functional assessment [76]. We designed seven representative S-ADL tasks based on common daily app usage scenarios and developed the metrics for performance assessment related to BAC changes. We then conducted a laboratory study with 40 participants by following protocols similar to those in other alcohol-based studies [38, 62, 76]. In this study, participants performed seven S-ADL tasks and three CNTs (N-BACK, SART, Task Switching) while intoxicated at three BAC phases (0%, 0.03%–0.04%, and 0.07%–0.08%). The CNT was performed alongside S-ADL at each BAC phase to verify the effectiveness of S-ADL for measuring BAC compared with CNT, which has been traditionally used for cognitive state assessment according with BAC in previous research [38, 41, 70, 94].

Finally, we built and compared the performances of machine learning models based on CNT and S-ADL. We also evaluated which S-ADL tasks and metrics exhibited the best performance and investigated whether BAC detection was effective using only the top one or two tasks. Our results showed that both the binary and multi-class models could effectively detect BAC with an approximately AUC–ROC and accuracy of 80%–81%. Moreover, the BAC-based model showed better performance than the traditional CNT-based model, which has been used in previous studies for detecting BAC. In addition, BAC detection with an accuracy of 80% could be achieved within one minute or less by performing only the two best-performing S-ADL tasks (information search and SMS reply).

In addition, we discuss the advantages of S-ADL usage over traditional BAC detection methods (e.g., efficiency, usability, and accessibility) based on user experience according to in-depth interviews with participants, as well as limitations and future studies considering potential bias (e.g., demographic factors, OS difference), noise problems, privacy concerns, potential psychological effects (e.g., false positives/negatives and over-reliance), and other ADLs with other smartphones or smart device sensors for real-life applicability.

Our study is novel in that it develops a performance-based S-ADL instrument for BAC detection that can assess an individual's ADL functional decline, such as a decline in perception, cognition, and motor coordination, by conducting scenario-based common daily use smartphone app tasks. Our design detects BAC in the ranges of 0.03%–0.04% and 0.07%–0.08% as a classification model rather than a regression model for 0.01% intervals because (1) the BAC criterion for binge drinking is 0.08% [88], (2) additionally, the legal threshold for drunk driving is set at 0.03% or 0.08% in most

countries around the world [117, 118], and (3) according to previous research on cognitive state differences due to alcohol consumption and NIAAA [38, 41, 70, 87, 94], the difference in cognitive decline due to acute alcohol consumption is more pronounced in interval ranges of 0.03%–0.04% rather than in intervals of 0.01% or smaller decimal units. Previous smartphone-based alcohol consumption detection research [6, 9, 10] focused on detecting mild and heavy drinking based on BAC phases of 0.03%–0.05% and 0.06%–0.08%.

2 BACKGROUND AND RELATED WORK

2.1 Symptoms of Functional Decline via Acute Alcohol Intake

After alcohol drinking, it takes approximately six minutes to reach the brain through the stomach [16]. Alcohol absorbed by the brain interferes with brain functions, leading to various functional declines (e.g., gross motor skill/planning, attention, amnesia, motor planning, peripheral vision, dysequilibrium, reflexes, and slurred speech), and these effects can persist for several hours until the alcohol is detoxified by the liver [47, 84, 90].

Blood alcohol concentration (BAC) represents the alcohol concentration dissolved in the blood. It is expressed as a percentage either by the mass of alcohol (w) per volume of blood (v) (% w/v) or by the mass of alcohol (w) per mass of blood (w) (% w/w) [31]. The symptoms of cognitive or physical functional decline based on BAC have been reported by the National Institutes on Alcohol Abuse and Alcoholism (NIAAA) [88]. At BAC levels of 0.03%–0.059%, mild declines occur (e.g., mild speech, memory, and fine motor coordination). At BAC levels of 0.06%–0.1%, moderate declines occur, such as effects on reasoning, peripheral vision, and depth perception. At BAC levels of 0.1%–0.15%, moderate declines occur (e.g., speech, memory, attention, motor coordination, and balance). Finally, at BAC levels of 0.16%–0.3%, severe declines are observed, such as effects on gross motor skills, motor planning, reflexes, and memory blackouts. Cognitive decline (e.g., executive function and attention) has been observed within the BAC of 0.04% or 0.08%, which is the legal limit for drunk driving in most countries [15, 117, 118]; however, motor coordination issues are not prominently exhibited [47, 113]. Note that even if the same amount of alcohol is consumed, BAC levels can differ among individuals owing to various factors, such as the type of alcohol, race, age, sex, health status, body mass, and individual tolerance [22].

2.2 Theoretical Backgrounds of BAC Measurement through Functional Assessment

Physical or cognitive functional decline due to BAC changes can be measured using various functional assessment methods. Functional assessments refer to the methods used to measure acute or chronic functional declines caused by various factors (e.g., drinking, stress, dementia, and strokes) such as functioning in activities of daily living (ADL), cognition, and physical mobility [29]. Traditional functional assessment methods can be classified into four types of tests: survey-based cognitive screening tests (e.g., MMSE, MOCA) [51], motor function tests (e.g., TUG) [46, 78], neuropsychological test-based cognitive screening tests (e.g., N-Back, Stroop test) [42], and ADL instruments such as self/informant ADL report questionnaires (e.g., Katz ADL, ADCS-ADL, B-ADL, I-ADL, FIM) [54, 57, 65, 96], performance-based tests (e.g., DAFS) [79], and naturalistic observations (e.g., MET [85]) [29].

Survey-based cognitive screening tests are challenging to use for multiple BAC measurements after drinking because of the learning effects and user burden. Motor function tests (e.g., TUG) primarily focus on assessing basic physical functions (e.g., balance and fall risks) in older individuals [46, 78]. Therefore, these tests are not very sensitive to measuring functional decline below the legal BAC limit 0.08% [15, 117, 118], as the decline in cognitive function is more pronounced than significant motor coordination issues at this BAC level [47, 113].

Additionally, neuropsychological test-based cognitive screening tests also have limitations in measuring BAC. Previous studies have quantitatively measured participants' functional performance using neuropsychological test-based cognitive screening tests (e.g., neuropsychological tests or computerized neuropsychological tests) to understand the functional decline associated with alcohol intake or BAC for each cognitive function domain [38, 41, 70, 94]. Lister et al. [70] found that alcohol at doses of 0, 0.3, and 0.06g/kg had a selective effect on memory, affecting only explicit memory processes and not implicit memory processes. Peterson et al. [94] determined that there were differences in functional performance in planning, verbal fluency, memory, and complex motor control through neuropsychological tests under the conditions of low (0.132ml/kg), moderate (0.66 ml/kg), and high dose (1.32 ml/kg) alcohol intake. Matthew et al. [38] assessed performance in various cognitive functions such as working memory, motor response, strategic optimization, vigilance, psychomotor function, cognitive flexibility, and response inhibition using six neuropsychological tests at BAC levels of 0%, 0.048%, 0.082%, and 0.10%. They demonstrated a decline in cognitive function with an increase in BAC. However, previous research has shown that, even as BAC or alcohol consumption increases, performance in certain cognitive functions (e.g., logical memory, reaction time, flexibility, psychomotor function, strategic optimization) either improves or remains unchanged [38, 41, 70, 94]. Therefore, even at the same BAC or alcohol dose, individuals showed significant variation in performance across all cognitive function domains, and the levels to which cognitive functions are affected vary. In addition, several neuropsychological tests are difficult to administer and require clinician guidance. Furthermore, repeated measurements of these tests pose a learning effect issue [13, 86]. Therefore, objective cognitive impairments may not be observed for each cognitive function domain. To detect mild functional declines below a BAC of 0.08% (the legal limit of DUI [15]).

ADL instruments that require a higher mental workload (e.g., cognitive processes) may be more appropriate for measuring BAC than the three other types of functional assessment methods. The common ADL instrument, also known as the basic ADL (B-ADL) or physical ADL (P-ADL), was designed to assess the treatment and prognosis of acute or chronic problems by observing the fundamental skills required to independently care for oneself. P-ADL consists of six tasks: feeding, continence (regulating bowel and urinary functions), transferring/ambulating, toileting, dressing, and bathing [57]. However, assessing mild cognitive impairment (MCI) based solely

on basic ADL is challenging [54, 65]. This motivated the development of another instrument called the I-ADL instrument, which requires a more complex mental workload to discern various functional declines, including MCI, compared to basic ADL [65]. Initially, I-ADL comprised seven tasks: communication, shopping, preparing food, household chores, transportation, medication intake, and handling finances [65]. To date, 50 tasks have been developed from 37 I-ADL instruments in 25 studies [54]. The strength of the I-ADL instruments lies in their demand for intricate cognitive abilities, allowing them to discern minor functional declines more effectively than P-ADL and neuropsychological assessments *without learning effects* [14, 81, 123]. Furthermore, as the complexity of the I-ADL task increases (e.g., banking tasks), its capability to detect nuanced and minor functional decline improves [123].

The traditional methods for measuring I-ADL mainly include self/informant reporting questionnaires and performance-based tests [54]. Self/informant reporting questionnaires are used to score the extent of ADL performance using 4–5 item questionnaires based on daily self-reporting or informant's observations. However, this method relies heavily on the subjectivity of an individual or observer daily, causing reliability issues. The performance-based test method involves executing scenario-based ADL tasks. Given that evaluators measure a patient's performance [79], this method can be more reliable and quantitative than self/informant reporting questionnaires, potentially making it possible to detect BAC changes. However, traditional I-ADL test methods require evaluation based on observer ratings or self-reports, which entails time and cost limitations. Therefore, measuring BAC immediately after alcohol consumption can be challenging.

Recently, with technological advancements, the potential to use technology-enabled ADL methods has emerged to overcome the limitations of traditional I-ADL test methods (e.g., long duration, high cost, reliability, non-automated performance scoring, and inaccuracy) for detecting BAC changes[29]. A representative method for integrating digital technology with I-ADL for assessment is the ADL-based tests on computer use [7, 56, 58, 93, 102, 103, 114]. Previous studies on computer-use ADL utilized interaction sensing with a mouse and keyboard to determine the functional state by evaluating computer usage performance. The test methods for computer-use ADL include real-life monitoring-based tests and performance-based tests. Real-life monitoring-based tests [58, 101–103] use daily or monthly statistics-based performance metrics such as the number of days in use per month, mean daily use, and time spent on mouse movement. However, applying these metrics for BAC detection within a few hours is challenging. In contrast, the performance-based test methods [7, 56, 93, 114] conducts functional assessments with a single-time measurement of pre-defined scenarios by using web browsing metrics (e.g., websites visited), search typing metrics (e.g., number of words per minute), and keystrokes metrics (e.g., keystroke rate). This type of performance-based test with one-time measurements has the potential to be applied for immediate BAC detection.

Owing to the recent trend of mobile-only lifestyles, most young adults perform I-ADL tasks (e.g., financial management, message texting, calling, searching for information, navigating) through various apps on smartphones rather than on desktop computers. Additionally, while computers are mainly used in offices or homes, thus having location constraints, smartphones can be carried around anytime and anywhere, even when drinking alcohol. Furthermore, the differences in display size and interaction methods between computers and smartphones influence how humans perceive information and make decisions based on Human-Computer Interaction (HCI) theory (i.e., the processes of perception, cognition, and motor functions) differently. This makes it challenging to directly apply the performance metrics used in computer use ADL directly to smartphones. Therefore, developing a smartphone-based ADL design that conducts traditional I-ADL tasks based on smartphone applications commonly used in daily life will make immediate BAC detection after drinking feasible.

2.3 Detecting Alcohol Consumption and BAC with Smartphones

Prior HCI studies have used smartphone context sensing or smartphone enabled functional assessment methods to automatically detect alcohol consumption episodes and BAC levels. Arnold et al. [6] utilized a smartphone accelerometer to detect alcohol consumption (normal, mild, and heavy drinking) through gait analysis (e.g., number of steps and gait velocity). Unlike the detection of alcohol consumption detection, the determination of BAC is difficult when there is no significant movement. Furthermore, according to previous research, many individuals do not exhibit a decline in simple motor coordination because that requires minimal cognitive abilities below BAC of 0.08% [47], which limits its effectiveness in detecting moderate alcohol consumption. Dai et al. [30] placed a smartphone accelerometer sensor inside a vehicle to detect drunk-driving movements.

Several studies leveraged smartphone-based passive sensing, such as those by Phan et al. [95] and Bae et al. [9, 10]. These studies detected normal, sober, and heavy drinking episodes by understanding user contexts such as interaction behavior (e.g., app usage, calls, messaging, key typing), location, and battery status, utilizing various built-in smartphone context data (e.g., GPS, app usage, and system status). However, these approaches focused on identifying the severity of drinking episodes and cannot be used to detect BAC levels immediately after alcohol intake. Thus, while previous studies have used smartphone sensors to monitor users' motor functions or context to determine alcohol consumption, there are limitations in the immediate detection of BAC levels after drinking.

Mariakakis et al. [76] developed a smartphone enabled functional assessment tool for BAC detection by adapting traditional neuropsychological tests to smartphones. This tool utilized touch interactions (e.g., swiping, typing, and tapping) and photoplethysmogram (PPG) and heart rate sensing to gauge various aspects of psychomotor coordination (e.g., fine motor coordination and psychomotor control/speed) for BAC detection. However, the smartphone-enabled neuropsychological test used captures only the human motor processes based on simple human perceptual processes (e.g., reflex actions) [76]. Because tasks designed to evaluate the cognitive processes (i.e., thinking skills) were not included, the ability of the test to identify mild cognitive decline due to moderate alcohol consumption is limited. Additionally, traditional neuropsychological tests were implemented on smartphones instead of using common daily–use smartphone apps, it implemented in the smartphone.

Our study extends the existing smartphone-enabled functional assessment methods for detecting drinking episodes and BAC levels by demonstrating the feasibility of analyzing typical phone usage behaviors (e.g., calling, texting, and map searching) that demand complex cognitive functions as in traditional I-ADL.

3 S-ADL INSTRUMENT DESIGN FOR DETECTING BAC

We propose a new instrument called the Smartphone Activities Daily of Living (S-ADL) for BAC identification (RQ 1). First, we discuss the preliminary S-ADL design with a rationale for BAC detection. Then, we propose scenario-based S-ADL task scripts for performance-based functional assessments. Finally, we suggest performance metrics to measure interaction performance.

3.1 Preliminary S-ADL Design

We defined the S-ADL as follows: "S-ADL is a sequence of interaction behaviors of smartphone apps frequently performed in everyday life" by referring to the existing ADL. To develop the representative S-ADL, we primarily focused on smartphone apps that can be commonly used among existing I-ADL tasks [17, 65, 92].

The major domains of I-ADL tasks can be defined as using phones (e.g., social and communication), shopping, food preparation, housekeeping, laundry, community mobility (e.g., transportation), taking medication, handling finances, and obtaining information [17, 65, 92]. Among these I-ADL tasks, tasks performed through recent smartphone apps include communication ADL through short messaging services (SMS) and phone calls, shopping ADL tasks through shopping apps, mobility ADL tasks through navigation apps, finances ADL tasks through banking apps, and information ADL tasks through information searching apps (e.g., Google). In addition to traditional I-ADLs, ADLs can also be designed based on the unique characteristics of smartphones. ADLs such as screen on/off and typing are performed exclusively on smartphones and are not limited to specific applications. We define these as "generic smartphone usage ADLs." As a result, we proposed five S-ADL categories: communication ADLs, photo taking/management ADLs, finance ADLs, information searching ADLs, and generic smartphone usage ADLs, as shown in Figure 1.

To design the preliminary S-ADL tasks, we first investigated the statistics of the smartphone apps and functions that young adults use most frequently. The most frequently used smartphone apps for individuals aged 18–34 are communication, photos and videos, news/weather information, music and media, games, and navigation [107]. Referring to the most frequently used apps, we defined six specific apps (phone, messaging, camera, banking, weather search, and location search) based on our S-ADL categories, as shown in Figure 1. Furthermore, referring to the most commonly used features in studies utilizing smartphone-based interaction data-driven functional health detection [23, 66], we defined a representative generic smartphone usage ADL including actions such as notifications, screens, typing, and app transition-related actions. We then defined the representative generic usage ADL tasks corresponding to these actions, as illustrated in Figure 1. As a result, 28 S-ADL tasks were derived from the five S-ADL categories as shown in Figure 1.

3.2 Scenario-based S-ADL Task Design for BAC Detection

Our instrument design builds upon the HCI theory and I-ADL research. According to the human information processing models in the HCI theory [21, 115], when interacting with a computer or smartphone, humans perceive information and make decisions through the processes of perception, cognition, and motor coordination sequentially. Human information processing requires the use of attention resources and mental workload. This highlights the fact that we can design S-ADL scenarios with diverse mental workloads (e.g., cognition and motor workloads) to observe the functional declines in terms of information processing.

Previous I-ADL studies [81, 123] have shown that tasks involving thinking skills can better differentiate mild functional decline than simple neuropsychological tests. Overall, the functional decline in information processing for S-ADL tasks can effectively detect BAC; thus, we designed various S-ADL task scripts with different mental workload levels.

As presented in Table 1, we finalized 17 of the 24 S-ADL tasks (Figure 1) and designed the S-ADL task scripts as follows. Tasks requiring a higher mental workload (i.e., various complex cognitive and fine-motor skills required) include Banking, Information search and share (IS), and SMS receives & reply (R&R). Tasks requiring a moderate mental workload (i.e., one or two cognitive functions) include SMS conversation (e.g., association skill), phone number register & call (e.g., working memory, recall), and photo delete (e.g., working memory). Tasks requiring minimum mental workload (simple stimulus & responses or fine motor tasks) include generic usage (e.g., notification response, screen unlock), phone receive & reply (R&R), which are similar to computerized neuropsychological tests [76, 110].

For instance, the money transfer task in a banking app is a prime example of a complex task that is commonly used in daily life without a learning effect. However, the execution process involves a more complex usage process than other S-ADL tasks (e.g., launching the banking app, authentication, selecting the bank, typing the account number and transfer amount, unlocking with a password, and initiating the transfer). This complexity results in increased mental and physical workload during the three steps of the perception, cognition, and motor processes of HCI theory [21, 115], i.e., perceptive loads (e.g., interpreting the app's user interface), cognitive loads (e.g., working memory, decision-making, calculations, information retrieval), and motor loads (e.g., typing passwords or account numbers). Additionally, IS and SMS R&R require high fine motor and cognitive loads. For instance, the SMS R&R task requires cognitive processes because it involves calculating future meeting times and dates from the given information, considering the meeting location to formulate a response, and then typing the response. This is in contrast to a previous study [76] that primarily focused on simple typing tasks involving repeating provided sentences and mainly assessed fine motor coordination. On the other hand, tasks such as phone R&R, notification response, and screen unlocking are similar to those tasks utilized by Mariakakis et al. [76] and primarily rely on reflexive responses to simple visual stimuli (e.g., choice reaction). These tasks do not demand as much in terms of processes of cognitive and motor load compared to complex tasks

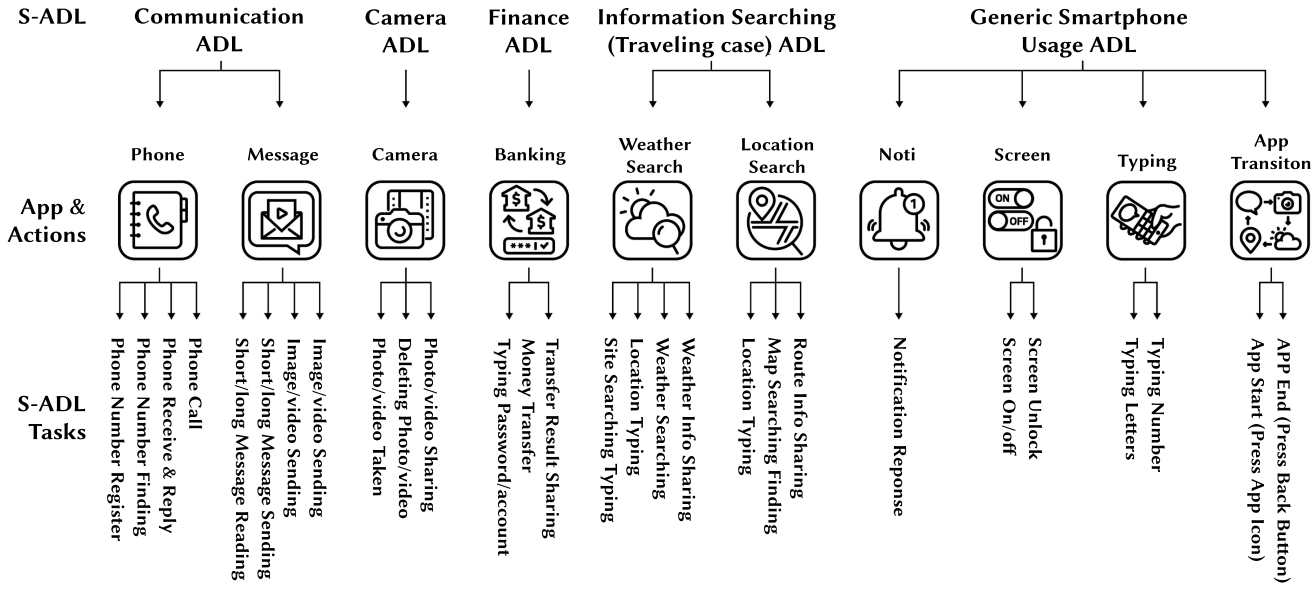


Figure 1: Description of Preliminary S-ADL Overview

Table 1: Scenario-based S-ADL Tasks and Scripts

S-ADL	S-ADL Task	Summary of Scenario-based S-ADL Task Scripts
Communication	Phone number register	Save the sender's phone number and name
	Phone call	Make a phone call and tell your personal information (e.g., email address)
	Phone receive and reply	Leave a different missed message depending on the caller
	SMS conversation	Answering a short question (e.g., fruit color) message
	SMS receive & reply	Send a reply message to make an appointment with time, place, date
Photo Take & Delete	Photo take	Take photos in the order of the cards presented
	Photo delete	Delete a particular photo among photos taken by instruction
Finance Management (Banking)	Money Transfer	Transfer the calculated amount to the provided bank account
	Transfer information share	Share bank transfer information to the sender via message
Information Search and Share (IS)	Weather information search	Enter the weather website and find the weather information on a specific day and location
	Weather information share	Find and share the weather information on a specific day and location
	Location information search	Find a restaurant with a high rating on a map for presented food name
	Location information share	Find and share the restaurant location route found via message
Generic Usage	Screen pattern unlocking	Unlocking the specific screen lock pattern
	Screen on by notification response	Turning on the screen in response to an instruction message notification
	App start after screen unlock	Starting the app after unlocking the screen
	App start by notification response	Starting the app in response to a instruction message notification

(e.g., money transfer), as they are primarily based on automatic reactions to visual cues.

Consequently, we created a set of eight S-ADL task scripts: phone number register & call, phone R&R, SMS short conversation and R&R, photo take and delete, banking, finance management, location & weather IS as shown in the example Figure 2. These S-ADL task scripts were created and revised to fit with smartphone app usage tasks by referring to existing performance-based I-ADL task scenarios [54, 79]. Our data were collected by performing S-ADL tasks, which were later used to derive performance metrics. When conducting a task in each session, we slightly altered the variables in the task instructions to eliminate learning effects in the S-ADL tasks (e.g., person name, fruit name, time, place, date, business card, and food name). The generic smartphone usage ADLs selected in Figure 1 were naturally performed in the process of performing the eight S-ADL tasks. Eight types of S-ADL tasks were performed continuously during data collection. Detailed explanations of the eight S-ADL tasks' scenario scripts are provided in the Supplement material for instruction and response message, as well as the task execution procedures (e.g., app start and end sequence) in Table 11.

We conducted a preliminary user test with six participants. Based on the test results, we excluded several S-ADL tasks and subtasks. In the location search and share task, we needed to find other food types and restaurants each time to eliminate the learning effect. However, it was difficult to maintain the consistency of the difference in difficulty depending on the type of food (e.g., there are too many cafes and too few steak restaurants on the maps). Therefore, we excluded the location search and share task. The SMS conversation task with fruit color answers was also excluded from the S-ADL task because of the variations in user knowledge of fruit colors and the lack of discriminative power.

3.3 Design of S-ADL Performance Metrics

Users perform predefined S-ADL tasks in controlled lab environments, from which we could extract various interpretable metrics that are useful for BAC detection. From the S-ADL tasks and subtasks, we derived a total of 57 performance metrics related to seven task correctness scoring metrics (referenced by traditional ADL performance-based test metrics [54, 79]) in Tables 6 and 7, 21 task completion time metrics including response time (e.g., notification response time) in Table 8, eight numbers of transitions (e.g., number of app transition or screens unlocks trials) in Table 9, and 21 types of SMS or information site searching typing-related metrics such as the error rate (e.g., COER), character level measure (e.g., intercharacter time), entry rates (e.g., CPS), and efficiency measure (e.g., UB) referring to Mackenzie et al. [75] in Table 10. These metrics can be calculated by collecting interaction data from built-in Android smartphone APIs such as Accessibility Service [33], UsageStatsManager [36], NotificationManager [35], and NotificationListenerService [34]. For a more detailed explanation of the S-ADL performance metrics, please refer to Appendix A.

4 CONTROLLED LAB EXPERIMENT

4.1 Participant Recruitment and Selection

We conducted a laboratory study to assess the feasibility of the proposed S-ADL for BAC detection. The laboratory study was used to supplement the limitations of inaccurate alcohol consumption measures in previous studies involving real-life experiments (e.g., participant's inaccurate memory, no reporting of alcohol consumption, no calculation of BAC) [6, 9, 10, 95] and to enable the evaluation of the performance of S-ADL tasks at precise BAC levels. The previous study [76] that identified BAC using smartphones in a laboratory environment only collected data from 14 individuals. However, 14 participants are insufficient to minimize potential bias for the impact on alcohol-induced cognitive abilities since such abilities may vary based on demographic information (e.g., sex, body weight) [40, 64]. Therefore, we chose a larger sample size of 40 participants to ensure sufficient validation of the effectiveness of S-ADLs while considering the impact of various demographic parameters on alcohol-induced cognitive abilities. Our study targeted young adults, specifically in their early 20s to 30s, as these ages exhibit the highest frequency and risk of binge drinking among all age groups [2, 3]. We selected 40 university students aged 20–32 based on the results of a pre-screening survey conducted before the experiment, comprising equal or slightly different distribution numbers with differences in demographic information (e.g., sex, age, and weight), as summarized in Table 2.

Furthermore, in addition to their demographic information (i.e., sex, weight, and age), the pre-screening survey obtained the following pieces of information to prevent potentially risky situations due to drinking-related health and psychological and physical health-related problems:

- Drinking-related health states: An alcohol history was obtained via an AUDIT [100] survey to ensure the safety of the participants. We also collected information on drinking habits, drinking capacity, alcohol-related personality traits, and genetic disorders.
- Psychological and physical health states: To consider participants with normal cognitive status before alcohol intake, we checked whether participants had any mental health issues such as ADHD, dementia, depression, stress, and general health issues through the six different health surveys (CAARS [28], GHQ-12 [44], PSS [27], PHQ-9 [63], EQ-5D-5L [50], and PSQI [20]).

Additionally, to account for differences in learning effects, recruitment was limited to participants with experience in using Android OS-based smartphones, S-ADL task-related apps, and QWERTY keyboards for at least one year. As indicated in Table 2, participants were divided into two groups based on their experience using Android OS (use for over five years or less five years) and current use of different types of smartphone OS to consider potential bias in S-ADL use depending on the OS.

The criteria selected through the pre-screening survey were as follows: (1) To eliminate the learning effect, participants who did not have at least one year or more than 10 times the presented S-ADL-related app usage under the given conditions (Android, QWERTY keyboard) were not allowed to participate in the experiment. (2) To participate in the study, no history of alcohol misuse or addiction could be present (both the participants and their families). Individuals who consumed alcohol within one week before the experiment were not allowed to participate. (3) Participants who were pregnant or had major physical or mental health issues or diseases were excluded from the study. All of these details were documented in the Institutional Ethics Review Board (IRB) submission, and the experiment was conducted with our university's IRB approval.

4.2 Evaluation of the Functional Decline with CNT to Detect BAC

The primary objective of this study is to detect BAC using human functional assessment. Therefore, we conducted a computerized neurocognitive screening test (CNT), which is a conventional performance-based functional assessment test that has been widely employed in previous medical research, to measure functional deterioration associated with BAC. We formulated a BAC detection model using the performance metrics obtained from the CNT and compared its performance with that of our S-ADL-based BAC detection model.

We selected the following popular CNT tasks: N-Back (NB) [61], Task Switching (TS) [55, 80, 83, 111], and Sustained Attention to Response Task (SART) [5, 97, 98] as shown in Figure 3. These tasks

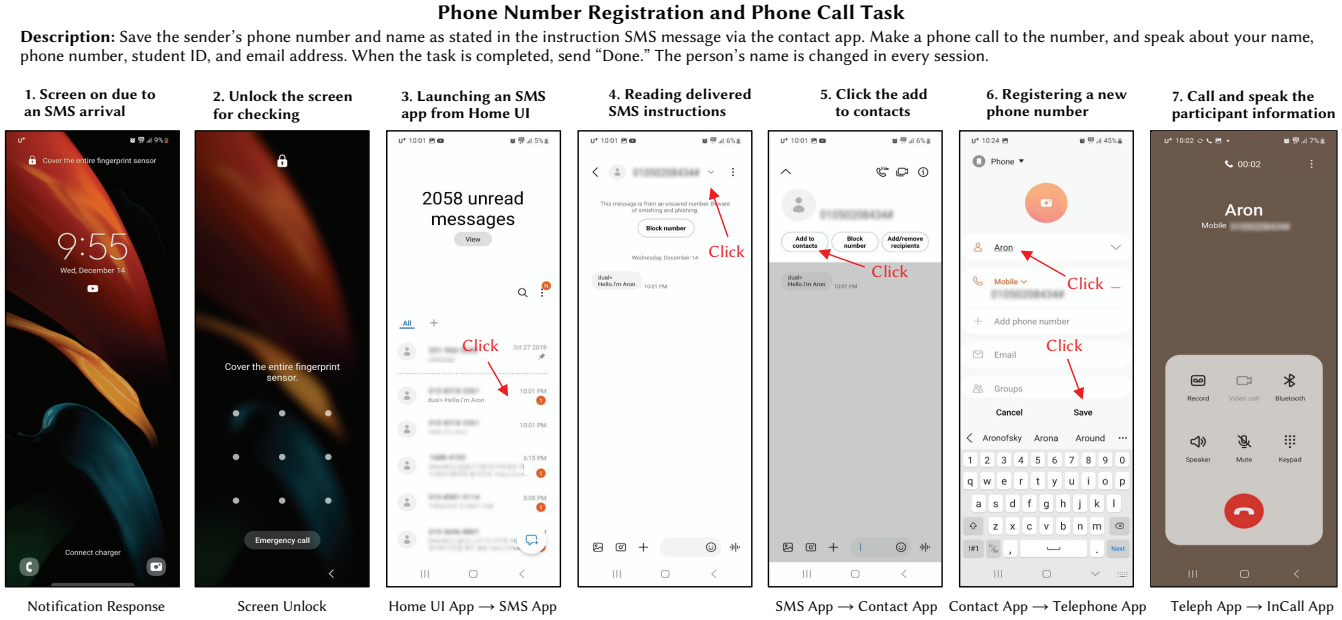


Figure 2: Illustration of S-ADL Task Script Example: Phone Number Registration and Calling Task

Table 2: Demographic factors and smartphone OS use experience of participants

Variable	Category	Numbers	Percent (%)	Range	Mean	SD
Gender	Female	20	50.0			
	Male	20	50.0			
Age (in years)	20-24	26	65.0	20-32	23.63	3.26
	25-32	14	35.0			
Weight (kg)	<51	10	25.0	41-78	61.00	10.95
	51-61	10	25.0			
	61-71	10	25.0			
	71 <	10	25.0			
Smartphone Android OS Use Experience	<5 years	10	25.0			
	>5 years	30	75.0			
Type of OS currently in use	Android	25	62.5			
	iOS	15	37.5			

were designed to evaluate various cognitive capabilities of the executive function (e.g., attention, working memory, processing speed, pattern recognition, cognitive flexibility, and response inhibition) governed by the frontal lobe of the human brain because alcohol consumption results in a temporary decrease in frontal lobe function. The three CNTs were configured for the web-based tests by utilizing and modifying the libraries provided in the popular software toolkit called PsyToolkit [108, 109]. As the performance metrics of the CNT, each individual's mean/median response time (ms) and accuracy (%) for each of the CNTs, as well as the sum scores of the three CNTs, were calculated.

4.3 BAC Phase Design for Safety-aware Experimental Setup

BAC levels were consistently monitored and maintained below the legal threshold for driving of 0.08% (defined as binge drinking by NIAAA [88]) in most states in the United States (US) [15], in strict compliance with the guidelines established by the IRB, as outlined in the NIAAA guidelines [87], and by the specified DUI limit, as documented in [15]. In addition, BAC of 0.03% or 0.04% is also the legal limit for drunk driving in many countries (e.g., most European and Asian countries) [117, 118]. Therefore, the detection of BAC of 0.03% or 0.08% is also very meaningful. Furthermore, following previous studies [38, 76, 82, 94, 119], the experiment was conducted at BAC levels with intervals of BAC 0.03%–0.04% (none drinking: 0%, mild drinking: 0.03%–0.04%, and heavy drinking: 0.07%–0.08%),

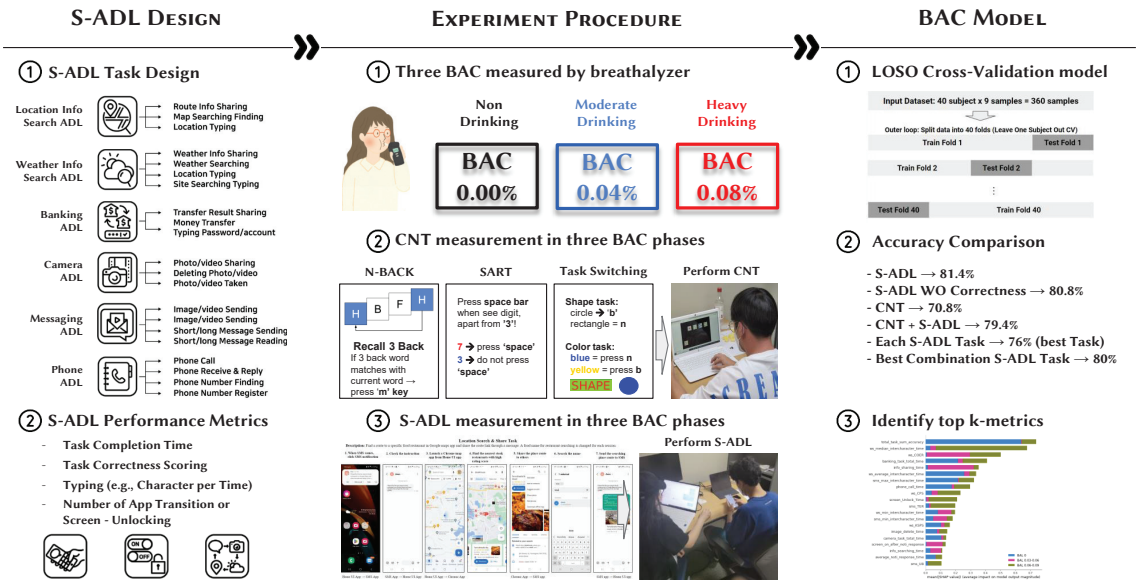


Figure 3: Overview for S-ADL Design, Experiment, and Result

which represents a significant change in the functional decline state.

To prevent alcohol overconsumption, the amount of alcohol that should be consumed by individuals over the three BAC phases was calculated in advance using Widmark's formula [116]. To calculate Widmark's formula, we collected weight and sex information from each participant. To avoid additional alcohol overdose, participants were asked to drink Soju with 20.1% alcohol by volume, a popular Korean distilled alcoholic beverage, once every 30 minutes using a 25 mL plastic cup. BAC was measured using a digital breathalyzer to ensure that the target BAC level was reached. According to Armin et al. [16], it takes approximately 20 minutes for alcohol to reach the liver, which metabolizes approximately eight grams of alcohol per hour.

Owing to the continuous increase or decrease in BAC levels during the progression of the experiment, BAC measurements were taken after completion of each session's CNT or S-ADL tasks to ensure the BAC was maintained at 0.03%–0.04% or 0.07%–0.08%. In addition, if there were no abnormalities in the BAC level, the experiment was conducted continuously. However, if the BAC level was higher than expected, the experiment was paused until the BAC level dropped to the desired range. If the BAC level was lower than expected, an additional 25 mL of alcohol was consumed, followed by a 20-minute wait. After re-measuring the BAC level and achieving the desired BAC level range, the next session was carried out.

Safety criteria were established to ensure the experimental stability. We provided sufficient rest, water, and hangover remedies to participants during the experiment. Fortunately, no significant body reactions were observed in the 40 participants during the experiement. For ethical experiment execution, in the pre-experiment orientation, participants were explicitly informed about precautions regarding alcohol consumption, as well as the option to immediately withdraw from and discontinue the experiment at any time upon the participant's request or the experimenter's judgment. After the experiment, we provided taxi fares to ensure the participant returned home safely and participants were not allowed to use private vehicles.

4.4 Apparatus and Experiment Procedure

We conducted a laboratory study involving 40 participants. An overview of the experimental procedure is shown in Figure 3. In the experiment, nine sessions were performed for the S-ADL and three CNT tasks at three BAC phrases (0%, 0.03%–0.04%, and 0.07%–0.08%) in three sessions. We collected nine samples across three repetitions of each BAC phase to ensure reliability, validity, and reproducibility, according to Design of Experiments (DoE) principles. By repeating the experiment under the same conditions three times, we can estimate the variability of the results and increase the accuracy of the estimate, assuming no systematic error. The reason for limiting the experiment to three repetitions is due to the practical and ethical limitations of lab studies involving alcohol consumption, balancing the need for statistical significance with participant health and ethical considerations.

The CNT tasks were performed using the same laptop model. Participants performed S-ADLs on the same model Samsung Galaxy Android smartphone, and we collected smartphone usage data using a usage data logger made with Android APIs such as [33–36]. BAC was measured using a digital breathalyzer (AL8000 model) to quantitatively determine the degree of alcohol intake. To minimize the learning effect that may occur as the number of sessions increases, both the S-ADL and CNT groups performed at least three

training sessions in advance. In addition, we counterbalanced the orders of the CNT and S-ADL tasks.

5 MACHINE LEARNING MODEL FOR BAC DETECTION

Our goal in RQ2 is to build a machine-learning model for BAC detection using S-ADL-based performance metrics and to identify specific metrics that demonstrate the most substantial influence on the accuracy and reliability of the BAC model. Toward this end, we posed the following three detailed evaluation questions. We first assessed the performance of a BAC detection machine learning (ML) model using S-ADL performance metrics (RQ2.A). We then compared this model with the computerized neuropsychological test (CNT) performance metrics-based models. Secondly, we explored what the key performance metrics are in the best BAC detection model based on S-ADL (RQ2.B). Third, we identified the S-ADL task-based metrics that were the most effective for BAC detection when used individually with separate S-ADL tasks. We then compared the performances of the best-performing S-ADL task-based metrics when used exclusively with the overall S-ADL task-based metrics to examine the feasibility of detecting BAC through a single S-ADL in a short period (RQ2.C). Finally, we explored whether demographic factors and smartphone OS use experience influenced the model by incorporating these features into it and comparing the performance with the S-ADL task-based metrics model (RQ2.D).

5.1 Machine Learning Based Model Building and Evaluation Methods

5.1.1 Binary and Multi-class Model Building. This study examined both binary and multi-class models for three BAC phases (0%, 0.03%–0.04%, and 0.07%–0.08%) and two BAC phases (0%–0.04% and 0.07%–0.08%), as in previous smartphone-based alcohol consumption detection studies [6, 9, 10]. Through this, we aimed to ascertain whether there was a difference in model performance between the two and three BAC phases. The reason for using a classifier model instead of a regression model was presented in Section 4.3, due to the definition of binge drinking by the NIAAA being a BAC of 0.08% [88], and most countries having a legal threshold for driving at 0.03% or 0.08% [15, 15]. Therefore, it is important to detect BAC within this range. In the multi-class model, there is a balanced dataset with three samples per class, whereas, in the binary-class model, there is an imbalanced dataset with six samples for one class and three samples for the other class. Therefore, to address the issue of an imbalanced dataset, we employed the Adaptive Synthetic Sampling (ADASYN) oversampling methods [49] and class weights to evaluate the performance of the model.

5.1.2 Model Selection and Evaluation Methods. We utilized leave-one-subject-out cross-validation (LOSOCV) to minimize bias (i.e., underfitting) by considering the sample data for all participants and enhance the generalizability by considering potential between-subject variation (i.e., the variance in the participant's unique smartphone usage performance capabilities or behavior habits such as typing speed & accuracy, and task completion time under normal conditions) in the training and validation process. LOSOCV involves excluding one subject (n=1) from the entire dataset (n=40), training the model using the remaining subjects (n=39), and then evaluating the model's performance with the excluded subject (n=1). This process was repeated for all 40 participants in the dataset. Compared with other cross-validation methods, LOSOCV is an effective method for enhancing generalizability in situations with limited data samples [39, 48]. This approach considers the differences between individual subjects, which is especially important in cognitive decline-related research where individual differences are high [39, 60]. Therefore, we conducted LOSOCV for validation to reduce bias and improve generalizability by using the data for all participants.

Furthermore, we employed the bagging-based ensemble models such as Random Forest (RF) [18] and boosting-based ensemble models which are Gradient Boosting Machine (GBM) [43], eXtreme Gradient Boosting (XGB) [24], and Light Gradient Boosting Machine (LGBM) [59] from the Sklearn library. These ensemble models are known for their ability to improve the model's performance by preventing overfitting or underfitting by reducing bias or variance, applicability to various datasets, robustness against noise, and an ability to identify feature importance in recent studies [24, 43, 59]. Moreover, these models have been proven in other smartphone data-driven cognitive impairment detection studies [24, 45]. To validate the superiority of the ensemble model's performance, we employed additional classifier models such as Naive Bayes (NB), Decision Tree (DT), and Logistic Regression (LR), which have been used in previous studies todetermine alcohol consumption using smartphone-based context data [6, 9, 10, 95]. To assess the model performance, we primarily relied on the commonly used classifier metrics, such as the area under the ROC curve (AUC–ROC) and accuracy (macro-average).

5.1.3 Model Agnostic Model Explanation. The SHapley Additive exPlanation (SHAP) value [72] was used to calculate the feature importance of the inference models trained in the outer loop. SHAP values were used instead of the built-in feature importance methods in the ensemble model because 1) SHAP values are model agnostic, meaning they can be applied regardless of the model type, 2) they provide consistent interpretations even when the model's architecture and parameters change, 3) they calculate feature importance more fairly and accurately by distributing the marginal contribution compared with ensemble model's built-in feature importance technique, thus overcoming the opacity of complex and hard-to-interpret ensemble models, and 4) they allow for the harmony of local and global model's interpretations, enabling the identification of feature importance not only for the overall model but also through SHAP values of each of the 40 individual test results [72, 99]. We obtained the SHAP values of the 360-sample dataset (40 subjects with nine samples per subject) by using the SHAP value and ranking them in order of importance to determine the S-ADL performance metrics that had the most significant influence on the best model.

Table 3: Model Performance of S-ADL and CNT based Metrics

Combination of metrics	ML Model	Binary Model		Multiple Model		Combination of metrics	ML Model	Binary Model		Multiple Model	
		AUCROC	Accuracy	AUCROC	Accuracy			AUCROC	Accuracy	AUCROC	Accuracy
S-ADL	NB	0.594	0.597	0.608	0.411	**S-ADL (WO correctness scoring)**	NB	0.588	0.692	0.608	0.411
	DT	0.650	0.667	0.646	0.528		DT	0.681	0.711	0.625	0.500
	LR	0.665	0.675	0.725	0.494		LR	0.654	0.711	0.715	0.481
	RF	0.780	0.811	0.793	0.592		RF	0.775	0.800	0.787	0.578
	GBM	0.779	0.806	0.800	**0.642**		GBM	**0.779**	0.797	0.790	0.580
	XGB	0.779	0.797	0.808	0.636		XGB	0.773	**0.808**	0.793	0.589
	LGBM	**0.783**	**0.814**	**0.814**	0.636		LGBM	0.770	0.794	**0.796**	**0.603**
CNT	NB	0.500	0.667	0.500	0.333	**S-ADL & CNT**	NB	0.594	0.597	0.608	0.411
	DT	0.631	0.667	0.563	0.417		DT	0.602	0.633	0.613	0.483
	LR	0.606	0.667	0.742	**0.492**		LR	0.723	0.717	0.752	0.492
	RF	0.644	0.692	**0.746**	0.475		RF	0.754	**0.794**	**0.861**	**0.606**
	GBM	**0.675**	**0.708**	0.696	0.475		GBM	0.715	0.747	0.834	0.594
	XGB	0.663	0.692	0.696	0.483		XGB	**0.763**	0.789	0.841	0.597
	LGBM	0.619	0.642	0.721	**0.492**		LGBM	0.727	0.739	0.820	0.600

5.2 RQ2.A: Performance Comparison of the BAC Detection Models using S-ADL and CNT

5.2.1 Model Performance with S-ADL. As summarized in Table 3, The S-ADL-based binary class model exhibited the best performance, with the AUC–ROC of 78.3% and the accuracy of 81.4% using LGBM. The performance of the S-ADL-based multi-class model exhibited the best performance, with an AUC–ROC of 81.4% using LGBM and an accuracy of 64.2% using XGB, as presented in Table 3. The average performance across the four ensemble models was as follows: for the binary-class models, the AUC–ROC was 78.0% and the accuracy was 80.7%; for the multi-class models, the AUC–ROC was 80.4% and the accuracy was 62.6%. In comparison to the ensemble models, single classifiers such as NB, DT, and LR showed an average performance in the binary-class models with the AUC–ROC of 63.6% and the accuracy of 64.6%, whereas in multi-class models, they exhibited the AUC–ROC of 66.0% and the accuracy of 47.8%. Therefore, ensemble models demonstrated an average improvement of approximately 14%–15% in both AUC–ROC and accuracy compared to single classifiers in both binary-class and multi-class models. Ensemble models perform better than single classifiers because they combine decisions from multiple individual models, thus reducing errors, bias, and variance. This allows ensemble models to perform well even in complex datasets with many features and noise. Accordingly, ensemble models exhibit better performance than single classifiers because the 57 features of the S-ADL-based model constitute a high-dimensional dataset. Additionally, the relatively lower accuracy of all multi-class models compared with all binary-class models, despite the higher AUC-ROC, is likely because there are three classes in the multi-class model, making accurate classification more difficult.

5.2.2 Comparison of the Model Performance with S-ADL and CNT. The best CNT-based models, both binary and multi-class, showed a lower AUC-ROC by approximately 9%–11% and a lower accuracy by approximately 11%–15% than the best S-ADL-based models, indicating that the S-ADL-based models outperformed the CNT-based models as indicated in Table 3. Therefore, we conclude that the S-ADL method performs better than the CNT in detecting BAC-related functional decline. This indicates that similar to previous research findings, I-ADL instruments are more sensitive to functional decline than CNT [81, 123]. In contrast to the S-ADL-based models, in the CNT-based models, the ensemble model did not show a significant difference in performance results compared to the single-classifier models for both the binary and multi-class models. This was attributed to the smaller number of features in the CNT-based models, leading to a reduced effect of the ensemble model.

5.2.3 Comparison of the Model Performance with S-ADL vs Combination of S-ADL and CNT. Furthermore, we evaluated whether combining the performance metrics of the S-ADL and CNT would achieve a better BAC detection performance. As indicated in Table 3, the best performance of the S-ADL and CNT-based binary class model exhibited an AUC-ROC of 76.3% using XGB and an accuracy of 79.4% using RF. The best performance of the S-ADL-based multi-class model showed an AUC–ROC of 86.1% and an accuracy of 60.6% using RF. Similar to the S-ADL-based models, S-ADL and CNT-based models consist of a large number of features (i.e., high-dimensional data). Consequently, the performance results for the ensemble models (RF, GBM, XGB, and LGBM) were generally higher than those of the single classifiers (NB, LR, and DT). The best performance of the binary class model that used the performance metrics of both S-ADL and CNT was lower than that of the S-ADL performance metrics-based model, possibly because of overfitting. The multi-class model showed a slightly higher AUC–ROC than the model that used only S-ADL performance metrics, although the accuracy was lower in this case. Therefore, it appears that there is no compelling need to use both S-ADL and CNT together for BAC detection because the results were not notably better than when using S-ADL alone.

5.3 RQ2.B: BAC Detection Model for Ranking the Importance of Performance Metrics

As shown in Figure 4, the top 20 features were derived from the LGBM-based multi-class model, which showed the highest performance, with an AUC-ROC of 81.4% as summarized in Table 3. As

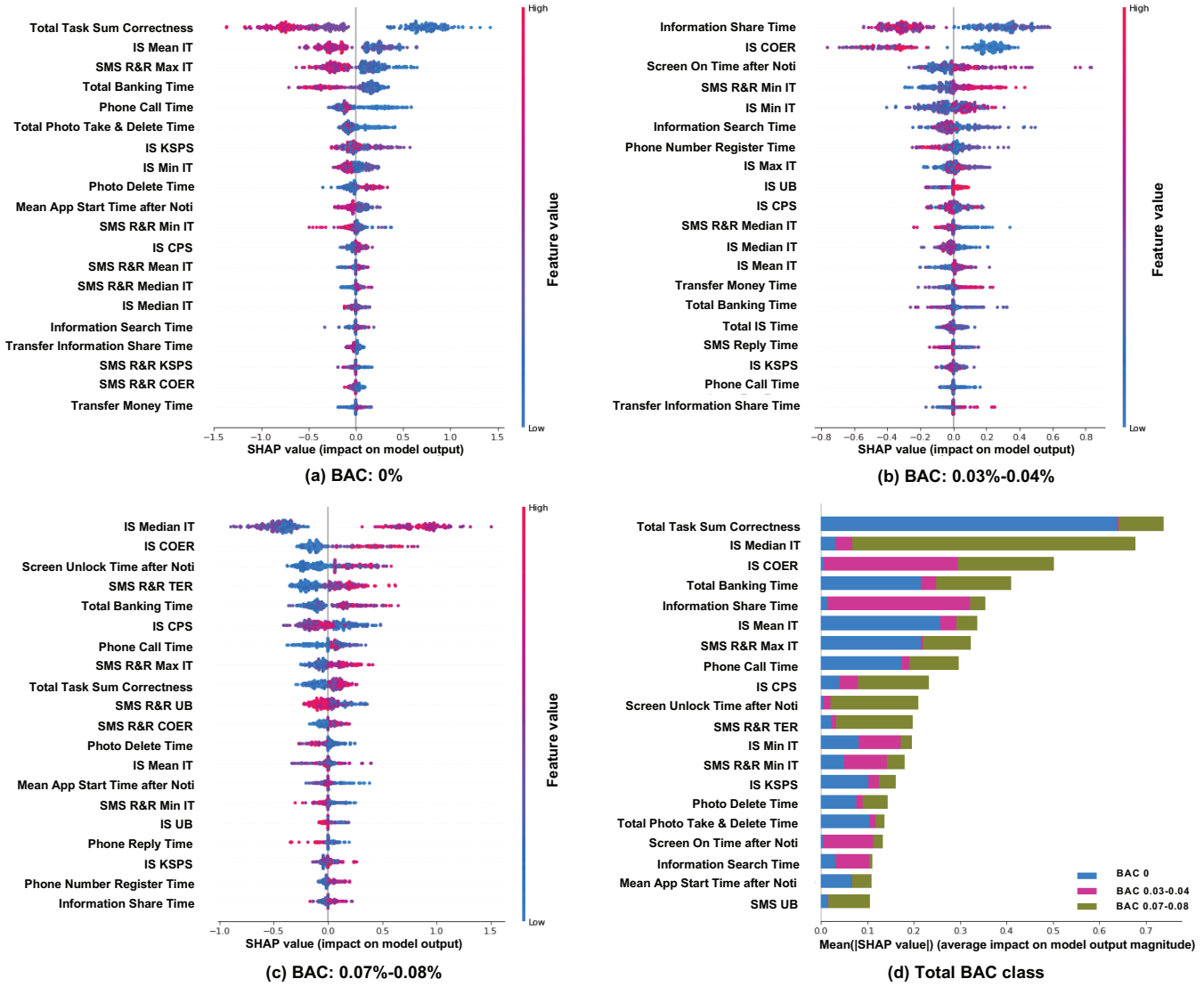
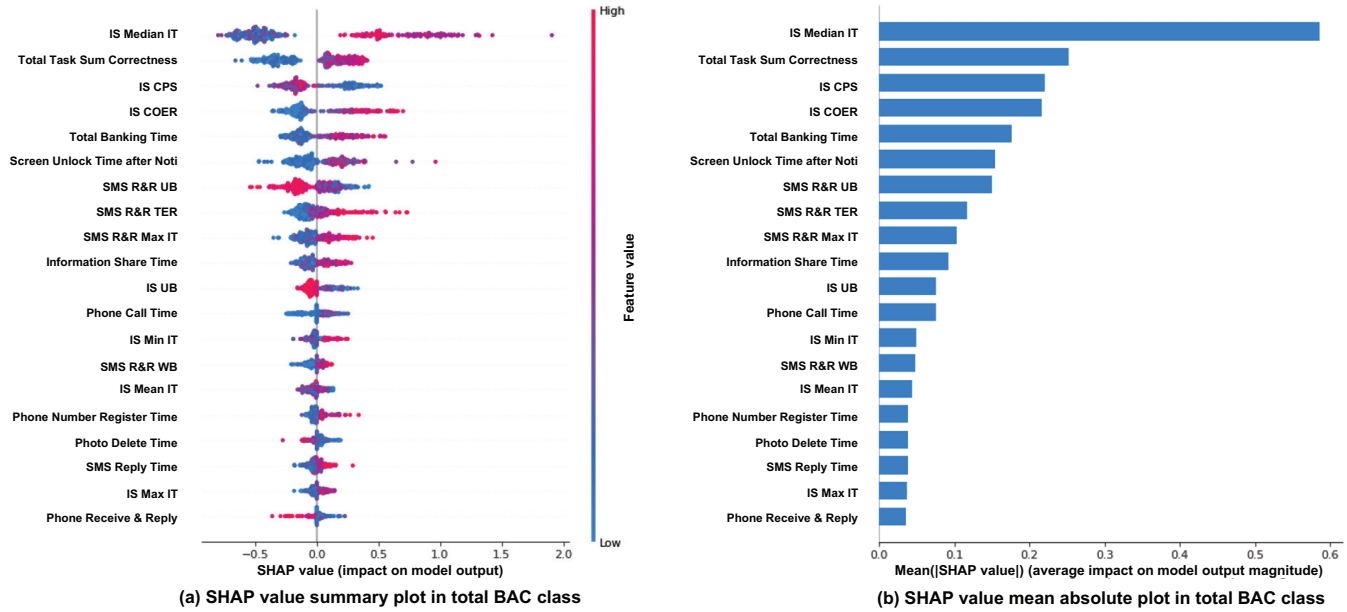


Figure 4: S-ADL performance metrics importance in the best multiclass model: SHAP value summary plots in (a) the non-drinking class (BAC 0%), (b) the mild drinking class (BAC 0.03%–0.04%), (c) the heavy drinking class (BAC 0.07%–0.08%), (d) the total class

shown in Figure 4(d) (absolute mean plot of the SHAP values), the typing and task completion time metrics were identified as part of the top 20 metrics in terms of importance. However, none of the S-ADL task-related task correctness scores or transition metrics were included, except for the *Total Task Sum Correctness* metric. The typing-related metrics, such as *Intercharacter Time (IT), Total/Corrected Error Rate (TER, COER)*, and *Character/Keystrokes per Second (CPS, KSPS)*, accounted for 10 of the top 20 metrics. Nine task completion time metrics for banking, information search, phone call, screen unlocking, and photo deletion were included among the top 20 metrics in that top five order, indicating a high level of influence on the model. When considered on a per-S-ADL task basis, information search & share (IS) task-related metrics were included in the most prevalent eight metrics among the top 20 metrics. Following, the SMS reply task and generic usage task-related metrics included each included three of the top 20 metrics.

In the binary-class model as well, typing and task completion time-related metrics had the highest influence on the best performance (accuracy of 81.4%) of the LGBM-based model, as shown in Figure 5. In the binary-class model, the median intercharacter time (IT) metric for the IS task exhibited an influence that was twice that of the total task sum correctness scoring metric, which showed the highest influence in the multi-class model. It surpassed all of the other metrics by a significant margin. Task completion time-related metrics also comprised eight of the top 20 metrics, with the top five in the order of banking, screen unlocking, information share, phone



(a) SHAP value summary plot in total BAC class

(b) SHAP value mean absolute plot in total BAC class

Figure 5: S-ADL performance metrics importance in the binary class model (BAC 0%–0.04%, BAC 0.07%–0.08%): (a) SHAP value summary plot in total class (b) SHAP value of mean absolute plot in total class

call, and phone number register task completion time exhibiting a high influence. In the S-ADL tasks, IS task-related metrics were the most prevalent with eight metrics, followed by five metrics related to SMS reply, among the top 20 metrics. Ultimately, the key metrics included in the top 20 were similar for the binary and multi-class models.

Based on the key metric results (Figures 4 and 5), this study found that each S-ADL-based task correctness scoring metric, developed by referencing the task correctness scoring method of traditional ADL instruments [53, 79] using automated scoring technology, was not a significantly important metrics in either the binary or multi-class models, even if the total task sum correctness scoring metric was one of the top three significant metrics. Furthermore, task correctness scoring metrics were limited to specific app tasks compared with other metrics. Because task correctness scoring metrics are based on the S-ADL task script in this study, it may be challenging to apply them to other apps with similar purposes. Therefore, we further analyzed the performance of binary and multi-class models using only typing, task completion time, and transition metrics, which can be generalized to other similar apps, and explored the feasibility of S-ADL-based BAC detection through different apps. As summarized in Table 3, we identified by excluding the task correctness scoring metrics-based model, the best binary-class model exhibited a difference of 0.4% in AUC–ROC and 0.6% in accuracy, and the best multi-class model exhibited a difference of 0.8% in AUC–ROC and 3.9% in accuracy compared to the best overall S-ADL metrics based model. This result demonstrates that the S-ADL model without the task correctness scores metrics can perform well for BAC detection. Thus, we have identified the potential for BAC detection using S-ADL performance metrics that can be applied to other similar apps without being limited to specific tasks.

5.4 RQ2.C: Comparison of the Model Performance of Each S-ADL Task

This study compared the actual performance of binary and multi-class BAC detection models using metrics from each S-ADL task with the metric importance results based on the SHAP values, as presented in Figures 4 and 5. The S-ADL tasks exhibited the best performance. Furthermore, it was possible to achieve good performance using only one or two S-ADL tasks instead of all tasks. Selecting only a few tasks can reduce the time required for BAC detection. As depicted in Table 4, we developed a total of seven binary and multi-class BAC detection models, each utilizing S-ADL task-related metrics, including the task completion time, task correctness scoring, typing, and transitions, for each respective S-ADL task (Appendix Section A). The typing metric was included only in the SMS reply and IS tasks. Table 4 clearly indicates that in the binary-class model, the best performance in terms of AUC–ROC and accuracy was obtained for metrics related to the following tasks, in descending order: IS, SMS reply, banking, phone number register & call, photo take & delete, and phone receive & reply tasks. In the multi-class model, the rankings were the same, except for a change in the order of the banking task-related metrics and phone number register & call-related task metrics.

The models with IS task-related metrics showed a performance difference compared to the models with all S-ADL tasks-related metrics, with a difference of approximately 5%–6% in both AUC–ROC and accuracy in the binary model and showed a more substantial difference of 11.1% in AUC-ROC and 7.7% in accuracy, as indicated in Table 4. The two best models based on the combination of metrics related to S-ADL tasks (IS and SMS reply) exhibited a performance difference of approximately 0.6% in AUC-ROC and 1.4% in accuracy in the binary model and 6.3% in AUC-ROC and 6.8% in accuracy in

Table 4: Model Performance of Each S-ADL Task

Combination of metrics	ML Model	Binary Model		Multiple Model		Combination of metrics	ML Model	Binary Model		Multiple Model	
		AUCROC	Accuracy	AUCROC	Accuracy			AUCROC	Accuracy	AUCROC	Accuracy
All S-ADL	NB	0.594	0.597	0.608	0.411	**Photo take & delete ADL**	NB	0.508	0.356	0.515	0.350
	DT	0.650	0.667	0.646	0.528		DT	0.515	0.547	0.540	0.386
	LR	0.665	0.675	0.725	0.494		LR	**0.571**	0.556	**0.608**	0.392
	RF	0.780	0.811	0.793	0.592		RF	0.508	0.606	0.536	0.353
	GBM	0.779	0.806	0.800	**0.642**		GBM	0.469	0.572	0.549	0.347
	XGB	0.779	0.797	0.808	0.636		XGB	0.515	0.578	0.559	0.403
	LGBM	**0.783**	**0.814**	**0.814**	0.636		LGBM	0.529	**0.611**	0.572	**0.406**
Phone number register & call ADL	NB	0.479	0.319	0.506	0.333	**Finance management ADL**	NB	0.460	0.378	0.513	0.347
	DT	0.525	0.550	0.521	0.361		DT	0.527	0.544	0.519	0.358
	LR	**0.590**	0.586	**0.629**	**0.436**		LR	**0.590**	0.628	**0.619**	0.381
	RF	0.521	0.614	0.537	0.361		RF	0.525	0.622	0.566	0.375
	GBM	0.521	0.608	0.544	0.358		GBM	0.542	**0.633**	0.588	**0.417**
	XGB	0.510	0.583	0.499	0.331		XGB	0.538	0.606	0.562	0.394
	LGBM	0.535	**0.617**	0.548	0.353		LGBM	0.525	0.597	0.530	0.369
Phone receive & reply (R&R) ADL	NB	0.500	0.350	0.508	0.347	**Information search & share (IS) ADL**	NB	0.585	0.575	0.546	0.400
	DT	0.492	0.531	0.502	0.336		DT	0.702	0.731	0.656	0.542
	LR	**0.538**	0.553	**0.547**	**0.383**		LR	0.675	0.683	0.696	0.472
	RF	0.527	0.561	0.488	0.300		RF	0.712	0.750	**0.703**	0.554
	GBM	0.517	0.539	0.501	0.339		GBM	0.727	0.753	0.685	0.532
	XGB	0.533	**0.564**	0.473	0.278		XGB	**0.733**	**0.756**	0.694	**0.565**
	LGBM	0.523	0.547	0.499	0.311		LGBM	0.725	0.744	0.679	0.557
SMS receive & reply (R&R) ADL	NB	0.542	0.425	0.544	0.369	**IS and SMS R&R ADL**	NB	0.594	0.592	0.585	0.414
	DT	0.612	0.625	0.552	0.403		DT	0.704	0.739	0.635	0.514
	LR	**0.638**	0.642	**0.658**	**0.461**		LR	0.696	0.706	0.715	0.472
	RF	0.617	0.628	0.604	0.428		RF	0.760	0.792	0.738	**0.574**
	GBM	0.619	0.633	0.633	0.414		GBM	**0.777**	**0.800**	0.720	0.564
	XGB	0.610	0.639	0.593	0.403		XGB	0.769	0.794	**0.751**	0.562
	LGBM	0.590	**0.647**	0.610	0.397		LGBM	0.760	0.792	0.722	0.566

the multi-class model when compared with the models containing all of the S-ADL task-related metrics. These results demonstrate that there was little difference in the BAC detection performance using only the S-ADL tasks of IS and SMS reply, which were the most frequently included in the top 20 metrics derived from the SHAP values, compared to using all of the S-ADL tasks, as shown in Figures 4 and 5. Therefore, considering that the execution time for both S-ADL tasks was less than one minute, this highlights the possibility of achieving rapid BAC detection without performing all of the S-ADL tasks.

As shown in Table 4, except for IS ADL, IS, and SMS R & R ADL, the results of the logistic regression model, a single classifier, were slightly better than those of the ensemble models in terms of accuracy for the binary-class model and both accuracy and AUC–ROC for the multi-class models. This outcome is contrary to the results of all of the S-ADL tasks-related metrics-based models in Table 3. While ensemble models are more suitable for complex data modeling, particularly high-dimensional data, logistic regression can be advantageous in cases where the data are simple and have a clear linear relationship [69]. Therefore, the logistic regression model shows higher performance than the ensemble models because the individual S-ADL task-related metrics-based models, in contrast to all S-ADL tasks-related metrics-based models, use only the performance metrics corresponding to each S-ADL task, thus resulting in models trained on relatively fewer features, or low-dimensional data.

5.5 RQ2.D: Comparison of the Model Performance with S-ADL vs S-ADL with Personal Attributes

We examined the impact of demographic features (age, sex, and weight) and smartphone OS use experience (Android OS usage experience and the type of OS currently in use), as summarized in Table 2, on the S-ADL-based BAC detection model. When building the models, we considered an approach of *fairness through awareness* [112] by incorporating these features into the machine learning model. The results are presented in Table 5. Compared to the best existing S-ADL-based metrics model, the best binary model showed a slight improvement of 1.3% in AUC–ROC and approximately 0.3% in accuracy, whereas the best multiple model exhibited an approximately 0.6% increase in AUC–ROC but a 2.3% decrease in accuracy. Models incorporating only demographic data showed an improvement of around 1% in AUC–ROC in both binary and multiple models, with a slight increase or decrease in accuracy. Models including only smartphone OS usage experience showed a 1.4% decrease in accuracy in the best binary model, whereas the multiple models exhibited a marginal improvement of approximately 0.2%–0.3%. These results suggest that the variance caused by the addition of demographic and smartphone OS usage experience features leads to some performance improvements in certain models; however, the overall impact on the performance of the S-ADL-based BAC detection model is minimal. Regarding feature importance, neither of these two feature types was ranked within the top 20 SHAP

values. Therefore, we conclude that including personal attributes has a minimal impact on the S-ADL-based detection model.

6 DISCUSSION

6.1 A Summary of Major Findings and Contributions

We developed S-ADL tasks and performance metrics for BAC detection and identified the key metrics by building machine learning models. S-ADL tasks are based on scenario-based common daily use smartphone app tasks and can assess an individual's ADL functional decline, such as a decline in perception, cognition, and motor coordination. The S-ADL-based performance metrics could detect BAC after drinking, achieving AUC–ROC and an accuracy of approximately 81%. Furthermore, we validated the superiority of the S-ADL-based performance metrics in detecting BAC compared with traditional performance metrics based on neuropsychological tests that have been widely used to measure functional decline associated with BAC [38, 41, 70, 94]. These findings are consistent with previous findings that ADL functional assessment tools are more sensitive to functional decline than neuropsychological tests [14, 81, 123]. Additionally, in the case of CNT, three were required. However, for S-ADL, because this method involves tasks utilizing commonly used apps and operating systems in daily life, no additional practice was required, even for complex S-ADL tasks (e.g., banking and information searching). Thus, we concluded that S-ADL showed less of a learning effect than CNT, as mentioned in previous studies [13, 14, 86]

Feature importance analyses using SHAP (Figures 4 and 5) revealed that task completion time and typing-related metrics were the key metrics among the five types of metrics. In particular, the banking task completion time and SMS & information searching (IS) typing metrics were the key metrics. Furthermore, the BAC detection model based on IS, SMS receive & reply (R&R), and banking task-related metrics showed better performance than the other S-ADL-task-based models, as indicated in Table 4. This is because IS, SMS R&R, and banking tasks require more perception and cognitive skills (e.g., computational ability and short-term memory) along with fine motor skills (e.g., keystroke typing) than other S-ADL tasks, as indicated in Table 1. The results of previous I-ADL studies also showed that the finance management ADL, which requires complex thinking skills, is more sensitive for detecting functional decline than other I-ADLs [14, 65, 81, 123]. In contrast, photos take & delete and phone receives & reply (R&R) metrics, which require less cognitive and motor loads (i.e., relying predominantly on psychomotor control and speed), exhibited lower performance, as depicted in Table 4. Hence, we found that S-ADL tasks demanding more cognitive and motor processes tended to perform better in binary and multi-class BAC detection models. Moreover, the model based on the two tasks that involved the highest levels of perception, cognition, and motor load (IS and SMS R&R) showed a minimal difference compared with the model based on all of the S-ADL-task-related metrics. This suggests it is possible to detect BAC within less than one minute if users perform only the IS and SMS R&R tasks.

Additionally, generic usage ADL tasks (e.g., screen unlocking, notification responses), photos take & delete, and phone R&R tasks related metrics were not included in the top 20 metrics in the BAC 0.03%–0.04% class of the multi-class model, as shown in Figure 4(b). In contrast, IS, SMS R&R, and banking tasks metrics were included in seven metrics of the top 20 features in the BAC 0.03%–0.04% class model as shown in Figure 4(b). This highlights that the S-ADL-related metrics demand more cognitive and motor processes and have a greater influence on discerning mild functional decline resulting from mild drinking (BAC 0.03%–0.04%). These results are consistent with those of previous studies [76, 77] in which the BAC detection methods based on psychomotor performance and response tasks had difficulties in detecting mild drinking (BAC 0.03%–0.05%). Indeed, a previous study [76] also used a typing task, but it primarily involved simply repeating given sentences without engaging in a significant thinking process. However, the typing task in our study required elaborate cognitive processes, such as thinking about meeting places and times for replies, memorizing responses, considering typing timing, and decision-making. Furthermore, a previous study [76] used only two efficiency metrics (e.g., utilized bandwidth and participant conscientiousness) from the metrics presented by MacKenzie et al. [75]. In contrast, we expanded the scope by incorporating a variety of 12 typing-related performance metrics, as summarized in Table 10, including the error rate (e.g., COER), character level measure (e.g., intercharacter time), entry rates (e.g., CPS), and efficiency measures (e.g., UB and WB) which can be utilized for BAC detection, as shown in Figures 4 and 5. Therefore, we believe that the sensitivity of the S-ADL to cognitive functioning could make it effective for detecting functional declines associated with mild drinking (BAC 0.03%–0.04%) or heavy drinking (BAC 0.07%–0.08%), and S-ADL based models achieved a better detection performance than the models in previous studies [76].

6.2 Privacy Issues and Potential Risks of S-ADL Use

The S-ADL-based assessment tool does not require personal identification of information, as it records extracted features such as the time spent per task in a certain app, the frequency of screen transitions within an app or between apps, typing measures (e.g., character per time, error rate), and/or notification response time extracted by scenario-based app tasks. Hence, this study method has minimal potential privacy risks. Nonetheless, to generalize this test in daily life with similar applications, the technical effort is necessary to ensure privacy protection during the process of data collection and processing as follows. One promising strategy is the use of on-device learning, which can be adapted to create a personalized model to prevent the potential leakage of personal data to an external server. Raw data can be deleted after feature extraction and aggregation, and categorical data (e.g., app names) can be encrypted using a one-way hash function to prevent potential data leakage.

We determined whether there were potential privacy concerns when collecting S-ADL performance metrics data based on actual user surveys and interviews through a questionnaire employing a seven-point Likert scale. The details of the follow-up user study are described in the Supplementary Material (Supplement: Section D). Additionally, we assessed whether privacy protection mechanisms (e.g., on-device learning or a one-way hash function for

Table 5: Model Performance of S-ADL with Personal Attributes: Demographic and OS Experience Features

Combination of metrics	ML Model	Binary Model		Multiple Model		Combination of metrics	ML Model	Binary Model		Multiple Model	
		AUCROC	Accuracy	AUCROC	Accuracy			AUCROC	Accuracy	AUCROC	Accuracy
S-ADL	NB	0.594	0.597	0.608	0.411	**S-ADL (with demographic features + OS use experience features)**	NB	0.592	0.603	0.608	0.411
	DT	0.650	0.667	0.646	0.528		DT	0.631	0.658	0.669	0.558
	LR	0.665	0.675	0.725	0.494		LR	0.669	0.683	0.733	0.492
	RF	0.780	0.811	0.793	0.592		RF	0.777	0.813	0.800	0.597
	GBM	0.770	0.806	0.800	**0.642**		GBM	**0.796**	**0.817**	0.802	**0.619**
	XGB	0.779	0.797	0.808	0.636		XGB	0.788	0.806	0.800	0.617
	LGBM	**0.783**	**0.814**	**0.814**	0.636		LGBM	0.783	0.806	**0.820**	**0.619**
S-ADL (with demographic features)	NB	0.588	0.594	0.608	0.411	**S-ADL (with OS use experience features)**	NB	0.592	0.603	0.608	0.411
	DT	0.698	0.717	0.652	0.536		DT	0.675	0.700	0.646	0.528
	LR	0.677	0.689	0.725	0.483		LR	0.675	0.689	0.726	0.508
	RF	0.764	0.789	0.815	0.616		RF	0.766	0.798	0.782	0.586
	GBM	**0.794**	**0.819**	0.806	0.622		GBM	0.767	0.792	0.798	0.628
	XGB	0.790	0.817	0.798	0.631		XGB	**0.783**	**0.800**	0.798	**0.645**
	LGBM	0.783	0.811	**0.823**	**0.639**		LGBM	0.771	0.794	**0.816**	0.642

data leakage) could mitigate users' privacy concerns. As shown in Figure 12 of Supplement: Section D, positive responses were obtained regarding the collection of performance metrics data, both on-device and to an external database, for detecting BAC while performing scenario S-ADL tasks and other types of S-ADL tasks through commonly used apps in everyday life. Conversely, it was noted that there was more positivity towards data collection performed on-device than in an external database, highlighting the need for privacy protection mechanisms in real-life applications. In addition, even if the data were collected in an external database, the responses indicated that it would not significantly affect the usage of S-ADL methods, as other health diagnostic apps collect even more detailed data. Among the performance metrics data, typing-related metrics received relatively lower positive scores than the other data. This was because the most sensitive information (e.g., bank account passwords, login IDs/passwords, and text message contents) was collected through typing. Although raw data (e.g., typed characters) were not stored, the participants were concerned that some data might have been erroneously stored on the device. This highlights the importance of transparently sharing the information on the collected data and their usage with the users to mitigate privacy concerns.

6.3 User Experiences of S-ADL-based BAC Detection: A Preliminary Examination

The S-ADL approach leverages widely accessible technology, potentially offering a convenient tool for users to monitor BAC levels and make safer decisions, such as avoiding binge drinking. Our approach provides an alternative to traditional BAC identification methods and their smartphone-based applications, such as computerized neuropsychological tests, survey-based formulation applications (e.g., the Widmark formulation), and breathalyzers. As previously stated, a follow-up user study with surveys and interviews was conducted, as described in the supplementary material (Supplement: Section C). For a quantitative evaluation of S-ADL usability, we customized the usefulness, ease of use, ease of learning, and satisfaction (USE) questionnaire [71]. Most participants rated the usefulness, ease of use, ease of learning, and satisfaction positively, with an average score of 6–7 out of 7 in Supplement: Section C (Figures 8–11). Participants mostly responded that they preferred the S-ADL method to traditional methods because it allowed for automatic BAC determination through the smartphone that they normally carried, without the need for separate measurement devices (e.g., breathalyzer) or additional applications (e.g., CNT).

The other user experience dimensions examined were related to users' perceptions of the machine learning algorithms. A significant risk associated with the use of ML models in health-related fields is the potential for over-reliance by users. If individuals trust these systems blindly, they may overlook the inherent limitations and potential errors such as false positives (i.e., the model incorrectly identifies a higher BAC than the actual amount of alcohol consumed or indicates that alcohol consumption when it has not occurred) and false negatives (i.e., the model incorrectly identifies a lower BAC than the actual amount of alcohol consumed or indicates no alcohol consumption when it has occurred) in ML predictions [10, 52]. For example, if a BAC detection app through S-ADL based on ML algorithms inaccurately classifies a user's alcohol level as safe when it is not, the consequences could be dangerous, potentially leading to decisions such as driving when it is unsafe to do so. To understand the user experience regarding over-reliance and concerns about false positives/negatives, we interviewed participants from our experiment about their needs for BAC measurements and their concerns about misclassifications. Most participants expressed more concern about false negatives than false positives, as detailed in Supplement's Section C. This was because most participants wanted to use S-ADL to *raise awareness* about alcohol consumption through quantitative indicators such as BAC, rather than relying on their subjective judgment. They responded that while extreme accuracy was not necessary (e.g., BAC measurement within 0.01% unit), they would appreciate knowing the margin of error for the measured BAC or the range of BAC (e.g., indicating mild or binge drinking phases), possibly through notification alarms or data visualizations.

Therefore, while the application of ML in HCI for functions such as BAC detection is promising, it is crucial to approach the implementation of such systems with careful consideration of the user experience and potential psychological impacts. It is especially important to inform users about the capabilities and limitations of the

ML model to prevent risky decisions due to over-reliance and to enhance trustworthiness. Additionally, the continuous improvement and rigorous testing of these systems are essential to minimize errors and enhance reliability. Understanding and addressing these aspects is crucial before we can conclusively deem such systems to be wholly beneficial. Compared to existing smartphone-based alcohol consumption determination models, the S-ADL method is designed to be more interpretable and transparent through its ML model, allowing users to better understand how the system operates from their perspective. The operation of S-ADL can be explained through the human information processing process in HCI theory [21, 115]. After drinking, when a user interacts with their smartphone using S-ADL, it automatically measures changes in functional decline in human information processing (perception, cognition, and motor coordination) to determine the BAC, which can be categorized as a situational impairment [76, 122]. The S-ADL method allows visualization of the causes of incorrect judgments or errors by presenting task-specific information to the users. Task-specific interpretable features in S-ADL represent a major departure from existing black-box models [6, 9, 10, 95]. The S-ADL allows for the identification of specific tasks being performed, enabling more interpretation from the user's perspective compared to the previous black-box models [6, 9, 10, 95]

In addition, it is essential to educate users about the system's accuracy and margin of error to prevent risky decisions due to over-reliance on the system. For example, information that identifies the results of heavy drinking (BAC of 0.07%) as mild drinking (BAC of 0.04%) can be provided to users to prevent serious consequences (e.g., drunk driving and binge drinking) due to over-reliance. In future research, we can use visualization techniques or alarms to help young adults proactively reflect on their drinking patterns and motivate them to encourage the regulation of their drinking patterns. However, because this study was conducted in a controlled laboratory environment, applying the current system directly to real-life situations poses challenges owing to various real-world factors such as environmental noises (e.g., weather, multi-tasking, interruption by unintended notifications, and other persons), and demographic factors and smartphone OS differences. Therefore, to build a reliable system, it is necessary to conduct further verification that considers real-life contexts, including the surrounding environment, system environment, physical activity, noise (e.g., interruptions), and potential biases (e.g., demographic factors, device variations, and smartphone operating systems). In the following sub-section, we discussed the limitations of our laboratory-based BAC detection method and possible directions for future work.

6.4 Limitations and Future Work

Can S-ADL be generalizable across different demographics data? Although BAC is influenced by various demographic factors (e.g., age, sex, weight, and alcohol tolerance) and reflects the results of different amounts of alcohol consumption, already considers these factors, there is still a potential for bias due to differences in smartphone usage abilities between individuals experiencing functional decline and those in a normal state at the same BAC level. To address this potential bias, our study targeted a healthy younger demographic and included 40 participants, considering age, sex, and weight for training, as shown in Table 2. This approach helped us develop a model that considered differences in demographic factors within the young population to some extent, thereby assessing the impact of these factors on the bias in the S-ADL-based BAC detection model. However, in real-world scenarios, the need for the S-ADL methodology extends beyond healthy young individuals and encompasses a variety of demographic factors, including the elderly, people with disabilities, and individuals struggling with alcohol addiction, all of whom can benefit from increased awareness of the risks of binge drinking. Therefore, future studies should broaden the participant pool to include a more diverse set of demographic factors known to affect mental and physical health due to drinking habits. To minimize potential bias and enhance the generalizability of the findings, these factors may include age, academic background, race, occupation, nationality, health status, level of disability, and degree of alcohol addiction.

Can S-ADL be generalizable across different apps, devices, and platform users? We leveraged widely used commercial applications as S-ADL tasks that people commonly use in everyday life, which is the main departure from the existing approach developed by Mariakakis et al. [76]. Our approach avoids the user burden associated with practicing less familiar tasks designed for BAC detection. However, S-ADL may face challenges in generalizing beyond specific scenario-based tasks under given OS platforms and application settings, which require additional user studies for further optimization. S-ADL is defined on the Android platform; thus, iOS users may be required to familiarize themselves with UI differences. We believe that cross-app and cross-device generalizability is a potential possibility. For instance, in our study, the specific scenario-based tasks tested on iOS users showed the potential for generalizability. This was inferred from the quantitative ML results and user interview responses, where users reported no significant difference in the UI within the same app between the iOS and Android platforms. The key metrics (e.g., task completion time and typing-related metrics) may be collected across all apps with various user interfaces corresponding to specific S-ADL tasks such as communication ADL and finance management ADL. However, the current study, which primarily focused on laboratory-based testing, cannot directly apply its key features (e.g., task completion time and typing-related metrics) to real life. For instance, users who have never used Android may experience differences in the S-ADL tasks conducted through other commonly used apps. In addition, real-world data often contains noise, such as interruptions from others and unexpected notifications. Accordingly, we need to consider minimizing such noise and OS differences when applying S-ADL to real-life scenarios for BAC detection in future work.

How can we reduce the noise when applying S-ADL in the real world? As mentioned in Section 6.3, BAC detection through S-ADL performance in the real world has potential risks of misclassifications, including false negatives/positives due to various contextual factors (e.g., system & surrounding context, weather, physical activity state, etc.) and negative smartphone usage habits (e.g., typing errors), as revealed through user interviews in Supplement: Section C. False negatives, in particular, could lead to serious consequences, such as drunk driving. To mitigate noise from environmental and system-related disturbances during the S-ADL tasks and enhance system reliability, this study aims to understand the environmental

and physical context by considering not only the app usage-based S-ADL utilized in this study but also other types of S-ADLs using various smartphone context sensors (e.g., GPS, Wi-Fi, system status, and physical activity). This approach will be helpful for distinguishing between the performance impacts caused by drinking and those caused by environmental or system status factors, ultimately reducing the potential risk of BAC misclassification in real life. Moreover, future research should consider a wider array of demographic factors and smartphone OS environments and collect data from more participants over a longer period. It is possible that long-term repeated measures would involve distinguishing between the average values of performance metrics during non-drinking and drinking periods for each individual. However, even with these considerations, it is important to acknowledge that unpredictable variables in real life mean that exact BAC identification cannot always be guaranteed. As mentioned in Section 6.3, the results indicate that users are willing to accept a certain degree of error in BAC detection and are more focused on raising awareness and reducing alcohol consumption. Therefore, risky decisions can be reduced through a transparent and interpretable model that informs users about the key metrics of the results and the potential range of errors.

Beyond S-ADL: How can we extend S-ADL to include ADLs that can be captured with smartphones? This study developed S-ADLs, focusing on smartphone app tasks primarily performed in daily life, such as making phone calls, managing finances, and searching for information. BAC detection was then performed using these S-ADLs. However, I-ADLs also include mobility tasks both within and outside the home, such as housekeeping, ambulating, and shopping. Therefore, utilizing these I-ADL tasks for BAC detection is expected to further enhance the feasibility of the model in real-world settings. Data from various smartphones or wearable sensors can be utilized to detect these I-ADL tasks. According to Lee et al. [66], smartphone sensing-based mobile usage and sensor data include interaction sensing, context sensing, and system sensing data. If we use context and system sensing-based data, various I-ADLs can be detected. As in previous smartphone context sensing-based drinking episode detection studies [9, 10, 95], the utilization of various context data (e.g., GPS, Wi-Fi, camera, and NFC) can be employed to assess the functional decline in mobility ADLs such as using transportation and shopping ADLs after drinking.

Beyond S-ADL: How can we leverage other types of sensing, such as home IoT or in-vehicle sensors? When alcohol consumption occurs within the household, it is possible to automatically assess functional decline in household ADLs after alcohol consumption by employing embedded sensors (e.g., infrared and motion sensors), as used in previous research on smart home ADLs or by using accelerometer-based activity recognition with smartphones and wearables [73]. Similarly, in driving situations, smartphones or wearable cameras can be utilized to monitor driving ADLs, which can be applied in conjunction with BAC detection [62]. Therefore, while this study focused on BAC detection using S-ADLs developed by applying interaction-based I-ADLs to smartphones, we expected that by exploring various I-ADLs through a wider range of smart devices and sensors, it will be possible to enhance the BAC detection model by capturing a more multifaceted functional decline. Therefore, understanding these ADLs, as inferred from the app usage behavior-based S-ADL tasks presented in this study, can help reduce noise from environmental and system-related disturbances during the S-ADL tasks, thus contributing to improved performance of the BAC detection model in real life.

7 CONCLUSION

Smartphones are tightly wired into our daily lives, significantly expanding the scope of traditional activities of daily living (ADL). We presented smartphone ADL (S-ADL) tasks and built a classification model for automatic BAC detection in this study. The S-ADLs built upon existing Instrumental ADL research, included five S-ADL tasks and 14 subtasks that people use most frequently. We derived 57 performance metrics from the S-ADLs to detect BAC. We considered two phases BAC (0%–0.04% and 0.07%–0.08%) and three phases BAC (0%, 0.03%–0.04%, and 0.07%–0.08%) for BAC label. We demonstrated the feasibility of the proposed method by comparing the S-ADL BAC detection model with the well-known CNT model and identified the key metrics and S-ADL tasks. A laboratory-based study was conducted to collect an interaction dataset with the precise BAC levels using a counterbalanced study design (e.g., task sequence and gender). The results showed that the S-ADL-based BAC detection model achieved an AUC–ROC of over 80% in the binary and multi-class models and showed better performance than the CNT-based model. The key metrics of the best model were task completion time and typing, which can be applied to similar purpose apps in specific S-ADL tasks. Additionally, S-ADL tasks involving high cognitive and motor loads had better predictive power than the other tasks, demonstrating the ability to detect BAC within a short period by performing one or two of the top-performance S-ADL tasks. Our study offers an initial step toward defining and understanding S-ADL instruments, building upon several decades of research on ADL assessments. To generalize the study results, long-term, large-scale studies in everyday life are required. As human behaviors are predictable and a large number of samples can be collected from individuals considering various demographic factors and smartphone use experiences over time, the S-ADL method may have the potential to reliably track within- and between-person variations in diverse areas of functional declines. Beyond alcohol detection, we solicit further studies on using S-ADL-based functional health monitoring, such as to evaluate health risks to young adults associated with substance use disorders (e.g., alcohol and cannabis) and mental health problems (e.g., depression and stress).

ACKNOWLEDGMENTS

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (2022R1A2C2011536) and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-00064). Corresponding authors: Auk Kim and Uichin Lee.

REFERENCES

[1] Antonia Abbey. 2002. Alcohol-related sexual assault: a common problem among college students. *Journal of Studies on Alcohol, supplement* 14 (2002), 118–128.
[2] Substance Abuse and Mental Health Services Administration. 2023. Risk Factors: Varied Vulnerability to Alcohol-Related Harm. Retrieved September 11, 2023 from https://www.niaaa.nih.gov/health-professionals-communities/core-resource-on-alcohol/risk-factors-varied-vulnerability-alcohol-related-harm

[3] Alcohol and Young Adults Ages 18 to 25. 2023. Risk Factors: Varied Vulnerability to Alcohol-Related Harm. Retrieved September 11, 2023 from https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics/alcohol-facts-and-statistics/alcohol-and-young-adults-ages-18-25#:~:text=According%20to%20the%202021%20National%20Survey%20on%20Drug,alcohol%20in%20the%20past%20month.%201%2C2%20This%20includes%3A#:~:text=According%20to%20the%202021%20National,1%2C2%20This%20includes

[4] Stadd Allison. 2013. 79% of people 18–44 have their smartphones with them 22 hours a day. Retrieved March 12, 2022 from http://www.adweek.com/socialtimes/smartphones/480485

[5] Tracy Packiam Alloway and Ross Geoffrey Alloway. 2012. The impact of engagement with social networking sites (SNSs) on cognitive skills. *Computers in Human Behavior* 28, 5 (2012), 1748–1754.

[6] Zachary Arnold, Danielle Larose, and Emmanuel Agu. 2015. Smartphone inference of alcohol consumption levels from gait. In *2015 International Conference on Healthcare Informatics*. IEEE, 417–426.

[7] Johanna Austin, Krystal Klein, Nora Mattek, and Jeffrey Kaye. 2017. Variability in medication taking is associated with cognitive performance in nondemented older adults. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 6 (2017), 210–213.

[8] BACtrack. 2023. BAC track. Retrieved September 11, 2023 from https://www.bactrack.com/

[9] Sangwon Bae, Tammy Chung, Denzil Ferreira, Anind K Dey, and Brian Suffoletto. 2018. Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions. *Addictive behaviors* 83 (2018), 42–47.

[10] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C Puyana, Ryan Kurtz, Tammy Chung, and Anind K Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 2 (2017), 1–36.

[11] Carolina Barbosa, Alexander J Cowell, and William N Dowd. 2021. Alcohol consumption in response to the COVID-19 pandemic in the United States. *Journal of Addiction Medicine* 15, 4 (2021), 341.

[12] Nancy P Barnett, EB Meade, and Tiffany R Glynn. 2014. Predictors of detection of alcohol use episodes using a transdermal alcohol sensor. *Experimental and clinical psychopharmacology* 22, 1 (2014), 86.

[13] Claudia Bartels, Martin Wegrzyn, Anne Wiedl, Verena Ackermann, and Hannelore Ehrenreich. 2010. Practice effects in healthy adults: a longitudinal study on frequent repetitive cognitive testing. *BMC neuroscience* 11 (2010), 1–12.

[14] A Bavazzano, SU Magnolfi, D Calvani, C Valente, F Boni, A Baldini, and JJ Quesada. 1998. Functional evaluation of Alzheimer patients during clinical trials: a review. *Archives of Gerontology and Geriatrics* 26 (1998), 27–32.

[15] Christy Bieber. 2023. Blood Alcohol Level Chart 2023. Retrieved September 11, 2023 from https://www.forbes.com/advisor/legal/dui/blood-alcohol-level-chart/

[16] Armin Biller, Andreas J Bartsch, György Homola, Lászlo Solymosi, and Martin Bendszus. 2009. The effect of ethanol on human brain metabolites longitudinally characterized by proton MR spectroscopy. *Journal of Cerebral Blood Flow & Metabolism* 29, 5 (2009), 891–902.

[17] Carolina L Bottari, Clément Dassa, Constant M Rainville, and Élisabeth Dutil. 2010. The IADL Profile: Development, content validity, intra-and interrater agreement. *Canadian Journal of Occupational Therapy* 77, 2 (2010), 90–100.

[18] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[19] Dana Bryazka, Marissa B Reitsma, Max G Griswold, Kalkidan Hassen Abate, Cristiana Abbafati, Mohsen Abbasi-Kangevari, Zeinab Abbasi-Kangevari, Amir Abdoli, Mohammad Abdollahi, Abu Yousuf Md Abdullah, et al. 2022. Population-level risks of alcohol consumption by amount, geography, age, sex, and year: a systematic analysis for the Global Burden of Disease Study 2020. *The Lancet* 400, 10347 (2022), 185–235.

[20] Daniel J Buysse, Charles F Reynolds III, Timothy H Monk, Susan R Berman, and David J Kupfer. 1989. The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry research* 28, 2 (1989), 193–213.

[21] Stuart K Card. 2018. *The psychology of human-computer interaction*. Crc Press.

[22] Arthur I Cederbaum. 2012. Alcohol metabolism. *Clinics in liver disease* 16, 4 (2012), 667–685.

[23] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. 2019. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2145–2155.

[24] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[25] Yanjiao Chen, Meng Xue, Jian Zhang, Runmin Ou, Qian Zhang, and Peng Kuang. 2021. DetectDUI: an in-car detection system for drink driving and BACs. *IEEE/ACM Transactions on Networking* 30, 2 (2021), 896–910.

[26] James M Clay and Matthew O Parker. 2020. Alcohol use and misuse during the COVID-19 pandemic: a potential public health crisis? *The Lancet Public Health* 5, 5 (2020), e259.

[27] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[28] C Keith Conners, Drew Erhardt, and Elizabeth P Sparrow. 1999. *Conners' adult ADHD rating scales (CAARS): technical manual*. Multi-Health Systems North Tonawanda, NY.

[29] Diane J Cook, Maureen Schmitter-Edgecombe, Linus Jönsson, and Anne V Morant. 2018. Technology-enabled assessment of functional health. *IEEE reviews in biomedical engineering* 12 (2018), 319–332.

[30] Jiangpeng Dai, Jin Teng, Xiaole Bai, Zhaohui Shen, and Dong Xuan. 2010. Mobile phone based drunk driving detection. In *2010 4th International Conference on Pervasive Computing Technologies for Healthcare*. IEEE, 1–8.

[31] Preeti Dalawari. 2019. Ethanol Level. (2019).

[32] Android Developer. 2022. AccessibilityEvent. Retrieved March 12, 2022 from https://developer.android.com/reference/android/view/accessibility/AccessibilityEvent

[33] Android Developer. 2022. AccessibilityService. Retrieved January 13, 2022 from https://developer.android.com/reference/android/accessibilityservice/AccessibilityService

[34] Android Developer. 2022. NotificationListenerService. Retrieved January 13, 2022 from https://developer.android.com/reference/android/location/LocationListener

[35] Android Developer. 2022. NotificationManager. Retrieved January 13, 2022 from https://developer.android.com/reference/android/app/NotificationManager

[36] Android Developer. 2022. UsageStatsManager. Retrieved January 13, 2022 from https://developer.android.com/reference/android/app/usage/UsageStatsManager#constants_1

[37] Ezgi Dogan-Sander, Elisabeth Kohls, Sabrina Baldofski, and Christine Rummel-Kluge. 2021. More depressive symptoms, alcohol and drug consumption: increase in mental health symptoms among university students after one year of the COVID-19 pandemic. *Frontiers in Psychiatry* 12 (2021), 790974.

[38] Matthew J Dry, Nicholas R Burns, Ted Nettelbeck, Aaron L Farquharson, and Jason M White. 2012. Dose-related effects of alcohol on cognitive functioning. *PloS one* 7, 11 (2012), e50977.

[39] Michael Esterman, Benjamin J Tamber-Rosenau, Yu-Chin Chiu, and Steven Yantis. 2010. Avoiding non-independence in fMRI data analysis: leave one subject out. *Neuroimage* 50, 2 (2010), 572–576.

[40] Rosemary Fama, Anne-Pascale Le Berre, and Edith V Sullivan. 2020. Alcohol's unique effects on cognition in women: A 2020 (re) view to envision future research and treatment. *Alcohol research: current reviews* 40, 2 (2020).

[41] Mark T Fillmore, Erik W Ostling, Catherine A Martin, and Thomas H Kelly. 2009. Acute effects of alcohol on inhibitory control and information processing in high and low sensation-seekers. *Drug and alcohol dependence* 100, 1-2 (2009), 91–99.

[42] Sanne Franzen, Esther van den Berg, Miriam Goudsmit, Caroline K Jurgens, Lotte Van De Wiel, Yuled Kalkisim, Özgül Uysal-Bozkir, Yavuz Ayhan, T Rune Nielsen, and Janne M Papma. 2020. A systematic review of neuropsychological tests for the assessment of dementia in non-western, low-educated or illiterate populations. *Journal of the International Neuropsychological Society* 26, 3 (2020), 331–351.

[43] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[44] David P Goldberg. 1988. User's guide to the General Health Questionnaire. *Windsor* (1988).

[45] Mitchell L Gordon, Leon Gatys, Carlos Guestrin, Jeffrey P Bigham, Andrew Trister, and Kayur Patel. 2019. App usage predicts cognitive ability in older adults. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[46] Barry R Greene and Rose A Kenny. 2011. Assessment of cognitive decline through quantitative analysis of the timed up and go test. *IEEE transactions on biomedical engineering* 59, 4 (2011), 988–995.

[47] M.D. Guillermo J. Salazar and M.D. Melchor J. Antuñano. 2005. Federal Aviation Regulation (CFR) 91.17. (2005).

[48] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

[49] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee, 1322–1328.

[50] Michael Herdman, Claire Gudex, Andrew Lloyd, MF Janssen, Paul Kind, David Parkin, Gouke Bonsel, and Xavier Badia. 2011. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research* 20, 10 (2011), 1727–1736.

[51] S Hoops, S Nazem, AD Siderowf, JE Duda, SX Xie, MB Stern, and D Weintraub. 2009. Validity of the MoCA and MMSE in the detection of MCI and dementia in Parkinson disease. *Neurology* 73, 21 (2009), 1738–1745.

[52] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. 2022. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks* 3 (2022), 58–73.

[53] Katrin Jekel, Marinella Damian, Holger Storf, Lucrezia Hausner, and Lutz Frölich. 2016. Development of a proxy-free objective assessment tool of instrumental activities of daily living in mild cognitive impairment using smart home technologies. *Journal of Alzheimer's Disease* 52, 2 (2016), 509–517.

[54] Katrin Jekel, Marinella Damian, Carina Wattmo, Lucrezia Hausner, Roger Bullock, Peter J Connelly, Bruno Dubois, Maria Eriksdotter, Michael Ewers, Elmar Graessel, et al. 2015. Mild cognitive impairment and deficits in instrumental activities of daily living: a systematic review. *Alzheimer's research & therapy* 7, 1 (2015), 1–20.

[55] Arthur T Jersild. 1927. Mental set and shift. *Archives of psychology* (1927).

[56] Vilas Joshi, Bruce Wallace, A Shaddy, Frank Knoefel, R Goubran, and C Lord. 2016. Metrics to monitor performance of patients with mild cognitive impairment using computer based games. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 521–524.

[57] Sidney Katz, Amasa B Ford, Roland W Moskowitz, Beverly A Jackson, and Marjorie W Jaffe. 1963. Studies of illness in the aged: the index of ADL: a standardized measure of biological and psychosocial function. *jama* 185, 12 (1963), 914–919.

[58] Jeffrey Kaye, Nora Mattek, Hiroko H Dodge, Ian Campbell, Tamara Hayes, Daniel Austin, William Hatt, Katherine Wild, Holly Jimison, and Michael Pavel. 2014. Unobtrusive measurement of daily computer use to detect mild cognitive impairment. *Alzheimer's & dementia* 10, 1 (2014), 10–17.

[59] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[60] Saleha Khatun, Bashir I Morshed, and Gavin M Bidelman. 2019. A single-channel EEG-based approach to detect mild cognitive impairment via speech-evoked brain responses. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27, 5 (2019), 1063–1070.

[61] Wayne K Kirchner. 1958. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology* 55, 4 (1958), 352.

[62] Kevin Koch, Martin Maritsch, Eva Van Weenen, Stefan Feuerriegel, Matthias Pfäffli, Elgar Fleisch, Wolfgang Weinmann, and Felix Wortmann. 2023. Leveraging driver vehicle and environment interaction: Machine learning using driver monitoring cameras to detect drunk driving. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–32.

[63] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine* 16, 9 (2001), 606–613.

[64] Lauren Kuhns, Emese Kroon, Heidi Lesscher, Gabry Mies, and Janna Cousijn. 2022. Age-related differences in the effect of chronic alcohol on cognition and the brain: a systematic review. *Translational Psychiatry* 12, 1 (2022), 345.

[65] M Powell Lawton and Elaine M Brody. 1969. Assessment of older people: self-maintaining and instrumental activities of daily living. *The gerontologist* 9, 3_Part_1 (1969), 179–186.

[66] Hansoo Lee, Joonyoung Park, and Uichin Lee. 2021. A Systematic Survey on Android API Usage for Data-Driven Analytics with Smartphones. *ACM Computing Surveys (CSUR)* (2021).

[67] Matthew R Lee and Kenneth J Sher. 2018. "Maturing out" of binge and problem drinking. *Alcohol research: current reviews* 39, 1 (2018), 31.

[68] Thad R Leffingwell, Nathaniel J Cooney, James G Murphy, Susan Luczak, Gary Rosen, Donald M Dougherty, and Nancy P Barnett. 2013. Continuous objective monitoring of alcohol use: twenty-first century measurement using transdermal sensors. *Alcoholism: Clinical and Experimental Research* 37, 1 (2013), 16–22.

[69] Joshua J Levy and A James O'Malley. 2020. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. *BMC medical research methodology* 20, 1 (2020), 1–15.

[70] Richard G Lister, Clarice Gorenstein, Debra Risher-Flowers, Herbert J Weingartner, and Michael J Eckardt. 1991. Dissociation of the acute effects of alcohol on implicit and explicit memory processes. *Neuropsychologia* 29, 12 (1991), 1205–1212.

[71] Arnold M Lund. 2001. Measuring usability with the use questionnaire12. *Usability interface* 8, 2 (2001), 3–6.

[72] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[73] Maxime Lussier, Monica Lavoie, Sylvain Giroux, Charles Consel, Manon Guay, Joël Macoir, Carol Hudon, Dominique Lorrain, Lise Talbot, Francis Langlois, et al. 2018. Early detection of mild cognitive impairment with in-home monitoring sensor technologies using functional measures: a systematic review. *IEEE journal of biomedical and health informatics* 23, 2 (2018), 838–847.

[74] Jory MacKay. 2019. Screen time stats 2019: Here's how much you use your phone during the workday. Retrieved March 12, 2022 from https://blog.rescuetime.com/screen-time-stats-2018/

[75] I Scott MacKenzie and Kumiko Tanaka-Ishii. 2010. *Text entry systems: Mobility, accessibility, universality*. Elsevier.

[76] Alex Mariakakis, Sayna Parsi, Shwetak N Patel, and Jacob O Wobbrock. 2018. Drunk user interfaces: Determining blood alcohol level through everyday smartphone tasks. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[77] Paul Maruff, Marina G Falleti, Alex Collie, David Darby, and Michael McStephen. 2005. Fatigue-related impairment in the speed, accuracy and variability of psychomotor performance: comparison with blood alcohol levels. *Journal of sleep research* 14, 1 (2005), 21–27.

[78] Samuel Mathias, US Nayak, and Bernard Isaacs. 1986. Balance in elderly patients: the" get-up and go" test. *Archives of physical medicine and rehabilitation* 67, 6 (1986), 387–389.

[79] Graham J McDougall, Heather Becker, Phillip W Vaughan, Taylor W Acee, and Carol L Delville. 2010. The revised direct assessment of functional status for independent older adults. *The Gerontologist* 50, 3 (2010), 363–370.

[80] Nachshon Meiran. 1996. Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, memory, and cognition* 22, 6 (1996), 1423.

[81] Mario F Mendez and Jeffrey L Cummings. 2003. *Dementia: a clinical approach*. Butterworth-Heinemann.

[82] Fredrik Modig. 2013. *Effects of acute alcohol intoxication on human sensory orientation and postural control*. Lund University.

[83] Stephen Monsell. 2003. Task switching. *Trends in cognitive sciences* 7, 3 (2003), 134–140.

[84] Stacy Mosel. 2023. Mental Effects of Alcohol: Effects of Alcohol on the Brain. Retrieved September 11, 2023 from https://americanaddictioncenters.org/alcoholism-treatment/mental-effects

[85] Alderman Nick, Knight Caroline, Henman Collette, et al. 2003. Ecological validity of a simplified version of the multiple errands shopping test. *Journal of the International Neuropsychological Society* 9, 1 (2003), 31–44.

[86] Rafaela Sanches de Oliveira, Beatriz Maria Trezza, Alexandre Leopold Busse, and Wilson Jacob Filho. 2014. Learning effect of computerized cognitive tests in older adults. *Einstein (Sao Paulo)* 12 (2014), 149–153.

[87] National Institute on Alcohol Abuse and Alcoholism. 2015. Alcohol overdose: the dangers of drinking too much. Retrieved March 12, 2022 from http://www.udelas.ac.pa/site/assets/files/4306/alcohol_overdose.pdf

[88] National Institutes on Alcohol Abuse and Alcoholism. 2021. Understanding Binge Drinking. Retrieved March 12, 2022 from https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/binge-drinking

[89] National Institutes on Alcohol Abuse and Alcoholism. 2022. Alcohol-Related Emergencies and Deaths in the United States. Retrieved March 12, 2022 from https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics/alcohol-facts-and-statistics/alcohol-related-emergencies-and-deaths-united-states

[90] National Institutes on Alcohol Abuse and Alcoholism. 2023. Alcohol's Effects on Health. Retrieved March 12, 2022 from https://www.niaaa.nih.gov/alcohols-effects-health

[91] Josephine Palmeri. 2011. Peer pressure and alcohol use amongst college students. *Online Publication of Undergraduate Studies, New York University. Accessed on February* 24 (2011), 2014.

[92] Marzieh Pashmdarfard and Akram Azad. 2020. Assessment tools to evaluate Activities of Daily Living (ADL) and Instrumental Activities of Daily Living (IADL) in older adults: A systematic review. *Medical journal of the Islamic Republic of Iran* 34 (2020), 33.

[93] Misha Pavel, Holly Jimison, Stuart Hagler, and James McKanna. 2017. Using behavior measurement to estimate cognitive function based on computational models. *Cognitive Informatics in Health and Biomedicine: Understanding and Modeling Health Behaviors* (2017), 137–163.

[94] Jordan B Peterson, Jennifer Rothfleisch, Philip D Zelazo, and Robert O Pihl. 1990. Acute alcohol intoxication and cognitive functioning. *Journal of studies on alcohol* 51, 2 (1990), 114–122.

[95] Thanh-Trung Phan, Florian Labhart, Skanda Muralidhar, and Daniel Gatica-Perez. 2020. Understanding heavy drinking at night through smartphone sensing and active human engagement. In *Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 211–222.

[96] KEITH RA. 1987. The functional independence measure; a new tool for rehabilitation. *Adv Clin Rehabil* 1 (1987), 6–18.

[97] Brandon CW Ralph, David R Thomson, Paul Seli, Jonathan SA Carriere, and Daniel Smilek. 2015. Media multitasking and behavioral measures of sustained attention. *Attention, Perception, & Psychophysics* 77, 2 (2015), 390–401.

[98] Ian H Robertson, Tom Manly, Jackie Andrade, Bart T Baddeley, and Jenny Yiend. 1997. Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia* 35, 6 (1997), 747–758.

[99] Raquel Rodríguez-Pérez and Jürgen Bajorath. 2020. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. *Journal of computer-aided molecular design* 34 (2020), 1013–1026.

[100] John B Saunders, Olaf G Aasland, Thomas F Babor, Juan R De La Fuente, and Marcus Grant. 1993. Development of the alcohol use disorders identification

test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption-II. *Addiction* 88, 6 (1993), 791–804.
[101] Adriana Seelye, Stuart Hagler, Nora Mattek, Diane B Howieson, Katherine Wild, Hiroko H Dodge, and Jeffrey A Kaye. 2015. Computer mouse movement patterns: A potential marker of mild cognitive impairment. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1, 4 (2015), 472–480.
[102] Adriana Seelye, Mira Isabelle Leese, Katherine Dorociak, Nicole Bouranis, Nora Mattek, Nicole Sharma, Zachary Beattie, Thomas Riley, Jonathan Lee, Kevin Cosgrove, et al. 2020. Feasibility of in-home sensor monitoring to detect mild cognitive impairment in aging military veterans: prospective observational study. *JMIR Formative Research* 4, 6 (2020), e16371.
[103] Lisa C Silbert, Hiroko H Dodge, David Lahna, Nutta-on Promjunyakul, Daniel Austin, Nora Mattek, Deniz Erten-Lyons, and Jeffrey A Kaye. 2016. Less daily computer use is related to smaller hippocampal volumes in cognitively intact elderly. *Journal of Alzheimer's Disease* 52, 2 (2016), 713–717.
[104] Jeffrey S Simons, Thomas A Wills, Noah N Emery, and Russell M Marks. 2015. Quantifying alcohol consumption: Self-report, transdermal assessment, and prediction of dependence symptoms. *Addictive behaviors* 50 (2015), 205–212.
[105] R William Soukoreff and I Scott MacKenzie. 2003. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 113–120.
[106] R William Soukoreff and I Scott MacKenzie. 2004. Recent developments in text-entry error rate measurement. In *CHI'04 extended abstracts on Human factors in computing systems*. 1425–1428.
[107] Statista. 2020. Most frequently used smartphone apps in South Korea as of September 2020, by category. Retrieved January 13, 2022 from https://www.statista.com/statistics/897227/south-korea-frequently-used-smartphone-apps-by-category/
[108] Gijsbert Stoet. 2010. PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior research methods* 42, 4 (2010), 1096–1104.
[109] Gijsbert Stoet. 2017. PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology* 44, 1 (2017), 24–31.
[110] Nikki H Stricker, Emily S Lundt, Sabrina M Albertson, Mary M Machulda, Shehroo B Pudumjee, Walter K Kremers, Clifford R Jack Jr, David S Knopman, Ronald C Petersen, and Michelle M Mielke. 2020. Diagnostic and prognostic accuracy of the cogstate brief battery and auditory verbal learning test in preclinical Alzheimer's disease and incident mild cognitive impairment: implications for defining subtle objective cognitive impairment. *Journal of Alzheimer's Disease* 76, 1 (2020), 261–274.
[111] André Vandierendonck, Baptist Liefooghe, and Frederick Verbruggen. 2010. Task switching: interplay of reconfiguration and interference control. *Psychological bulletin* 136, 4 (2010), 601.
[112] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.
[113] VirginiaTech. 2024. Alcohol Resources. Retrieved February 11, 2024 from https://hokiewellness.vt.edu/students/alcohol.html
[114] Lisa M Vizer and Andrew Sears. 2015. Classifying text-based computer interactions for health monitoring. *IEEE pervasive computing* 14, 4 (2015), 64–71.
[115] Christopher D Wickens, Justin G Hollands, Simon Banbury, and Raja Parasuraman. 2015. *Engineering psychology and human performance*. Psychology Press.
[116] EMP Widmark. 1932. Die theoretischen Grundlagen und die praktischen Verwendbarkeit der gerichtlichen-medizinischen Alkoholbestimmung. Fortschr Naturwiss Forschung 11, 1–140. Available in English as Baselt RC (1981) Principles and applications of medicolegal alcohol determinations.
[117] Wikipedia. 2023. Driving under the influence — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Driving%20under%20the%20influence&oldid=1174186426. [Online; accessed 14-September-2023].
[118] Wikipedia. 2023. Drunk driving law by country — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Drunk%20driving%20law%20by%20country&oldid=1173501806. [Online; accessed 14-September-2023].
[119] Ann M Williamson, Anne-Marie Feyer, Richard P Mattick, Rena Friswell, and Samantha Finlay-Brown. 2001. Developing measures of fatigue using an alcohol comparison to validate the effects of fatigue on performance. *Accident Analysis & Prevention* 33, 3 (2001), 313–326.
[120] Michael Winnick. 2016. Putting a Finger on Our Phone Obsession. Retrieved March 12, 2022 from https://blog.dscout.com/mobile-touches
[121] Jacob Wobbrock, Brad Myers, and Brandon Rothrock. 2006. Few-key text entry revisited: mnemonic gestures on four keys. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 489–492.
[122] Jacob O. Wobbrock, Krzysztof Z. Gajos, Shaun K. Kane, and Gregg C. Vanderheiden. 2018. Ability-Based Design. *Commun. ACM* 61, 6 (may 2018), 62–71.
[123] YoungSoon Yang, Hyun Duk Yang, Yun-Jung Hong, Jung Eun Kim, Moon-Ho Park, Hae Ri Na, Il-Woo Han, and SangYun Kim. 2012. Activities of daily living and dementia. *Dementia and Neurocognitive Disorders* 11, 2 (2012), 29–37.

A DETAILED EXPLANATION OF S-ADL PERFORMANCE METRICS

A.1 Correctness-based performance metrics

These metrics measure task correctness scoring by checking a user's response messages. This allowed us to determine whether the tasks were performed correctly. If the correct response was recorded for each S-ADL task, a given task received one point; otherwise, it received zero points as shown in Table 6. The total sum was calculated by assigning 1 point if it was completely correct for each sub-task and 0 points if it was partially incorrect. In addition, the correctness summation scores for all six S-ADL tasks were extracted. A more detailed explanation of the selected seven correctness metrics is provided in Table 7.

A.2 Completion time-based performance metrics

These metrics were derived by the interaction sensing of fine-grained smartphone-specific app usage task completion time (e.g., SMS R&R time, transfer money time) and generic smartphone usage response time(e.g., notification response time).

A.3 Transition-based performance metrics

Transition aims to check how many app transitions (including erroneous transitions) have been made when performing a given S-ADL task. For the transition measure calculation, each app start/end frequency was counted by sensing while performing an S-ADL task script. It was calculated by comparing how many more app transitions were performed compared to the number of app transitions required by the S-ADL script (when performed without any mistakes).

A.4 Typing-based performance metrics

We used the typing measures by Mackenzie et al. [75] about site searching (total 15 typing entries: "www.weather.com") and SMS receive & reply typing tasks (total 45 typing entries: "Alright, let's meet there by 4:15 PM on Aug. 14th"). According to Mackenzie et al. [75], there are four categories (i.e., character per time, character level analysis, error rate, and efficiency) and 23 typing metrics (e.g., keystrokes per second and corrected error rates). We selected 12 typing metrics out of the total 23 based on the criteria; if the measures were similar, we selected the recently developed and verified measures from previous studies[75, 105, 106, 121]. A more detailed explanation of the selected 12 typing metrics is provided in Table 10. The weather site search typing of the IS task does not have the "incorrect typing not fixed" in contrast to the SMS R&R typing task because site typing requires fixing all the incorrectly typed letters to be able to access the site. Thus, site address typing tasks can be used only *"COER"* among error rate metrics. Moreover, site address letters are not uppercase letters or special characters, entering a shift or switching key is not required. Therefore, the *"GPS"* does not exist in the IS typing tasks.

In this study, all S-ADL performance metrics were automatically extracted and calculated by Android built-in APIs such as

AccessibilityService [32, 33], UsageStatsManager [36], and NotificationManager and NotificationListerService [34, 35]. The detailed process of extraction metrics is provided in supplement materials.

B DETAILED EXPLANATION OF S-ADL TASK SCRIPTS

B.1 Communication ADL

Communication ADL consists of four S-ADL tasks (phone number register & call, phone receive & reply, SMS conversations, SMS receive & reply). Detailed task script descriptions in Supplement Materials.

B.2 Photo Take & Delete ADL

Photo Take & Delete ADL consisted of two S-ADL tasks (photo taking and deletion). Detailed task script descriptions in Supplement Materials.

B.3 Finance Management ADL

Finance Management ADL consists of two S-ADL subtasks (transferring money and sharing information by sending messages). Detailed task script descriptions in Supplement Materials.

B.4 Information Search and Share ADL

Information Search and Share ADL consists of two S-ADL task groups (location search & share, weather search & share). Detailed task script descriptions in Supplement Materials.

Table 6: Task Correctness Scoring metric criteria for each S-ADL task

S-ADL Task	S-ADL Sub Task	Correctness Criteria	Score (Correct: 1 point)
Phone Number Register & Call	Phone Number Register	Does the registered phone number match the presented phone number?	1 point
	Phone Call	Were the name, contact number, student ID, and e-mail mentioned over the phone the same as they actually are?	1 point
Photo Take & Delete	Take photos of business cards	Were the three business cards photographed in the order presented?	1 point
	Delete one of the business card photos	Was the presented card deleted among the three business cards?	1 point
Phone Receive & Reply	Send absence message	Was the absent message sent properly according to the caller?	1 point
SMS Receive & Reply	Reply location	Was the location entered correctly?	1 point
	Reply time	Was the time zone entered correctly?	1 point
	Reply day of the week	Was the entered day of the week correct?	1 point
	Typed total sentence	Were the spellings of total sentences entered correctly?	1 point
Banking	Typed recipient number	Were bank transfer receipts messages sent to the presented phone number?	1 point
	Typed banking remittances	Were the bank transfer amounts matched to the presented remittance?	1 point
Location Search & Share	Search and type restaurant name	Was the searched & typed restaurant name matched to the presented restaurant name?	1 point
	Search and type the restaurant's rating	Was the searched & typed restaurant's rating matched to the presented restaurant's rating?	1 point
Weather Search & Share	Search and type weather information search location	Was the searched & typed weather location matched to the presented searched location?	1 point
	Search and type temperature & humidity	Was the searched & typed weather information matched to the weather of the presented date and location?	1 point
Sum			**15 point**

Table 7: S-ADL task correctness scoring-related performance metrics

S-ADL Task	Terms of Performance Metrics	Description of Performance Metrics
Phone number Register & Call	Total Phone Number Register & Call Correctness	Whether phone number register & call tasks were performed correctly
Phone Receive & Reply (R&R)	Total Phone R&R Correctness	Whether phone R&R tasks were performed correctly
SMS Receive & Reply (R&R)	Total SMS R&R Correctness	Whether SMS R&R tasks were performed correctly
Photo Take & Delete	Total Photo Take & Delete Correctness	Whether photo take & delete tasks were performed correctly
Banking	Total Banking Correctness	Whether Banking tasks were performed correctly
Information Search & Share (IS)	Total IS Correctness	Whether IS tasks were performed correctly
Total S-ADL Tasks	Total Task Sum Correctness	Whether Total S-ADL tasks were performed correctly

Table 8: S-ADL task completion time-related performance metrics

S-ADL Task	Terms of Performance Metrics	Description of Performance Metrics
Phone number Register & Call	Phone Number Register Time	Time taken register phone number and name
	Phone Call Time	Time taken to personal information (e.g., name, phone number, email address)
	Total Phone Number Register & Call Time	Total time taken to conduct phone number register & call time task
Phone Receive & Reply (R&R)	Phone Reply Time	Time taken to leave an absent message after calling
	Total Phone R & R Time	Total time taken to conduct phone R&R task
SMS Receive & Reply (R&R)	SMS Reply Time	Time taken to type SMS replying
	Total SMS R&R Time	Total time taken to conduct SMS R&R task
Photo Take & Delete	Photo Take Time	Time taken to take number cards with a camera app
	Photo Delete Time	Time taken to delete a specific image among number cards in a gallery app
	Total Photo Take & Delete Time	Total time taken to conduct photo take & delete task
Banking	Transfer Money Time	Time taken to authenticate the bank app and type account and money amount
	Transfer Information Share Time	Time taken to share remittance information as the message sending
	Total Banking Time	Total time taken to conduct banking task
Information Search & Share (IS)	Information Search Time	Time taken to type weather site and search weather information
	Information Share Time	Time taken to share weather information
	Total IS Time	Total time taken to conduct IS task
Generic Usage	Mean App Start Time after Noti	The average of time taken to start the messaging app after noti in All S-ADL tasks
	Median App Start Time after Noti	The median of time taken to start the messaging app after noti in All S-ADL tasks
	Screen On Time after Noti	Time taken to Turn off the screen from screen off after noti
	Message App Start Time after Screen Unlock	Time taken to start message app after screen unlock
	Screen Unlock Time after Noti	Time taken to unlock screen unlock pattern

Table 9: S-ADL transition-related performance metrics

S-ADL Task	Terms of Performance Metrics	Description of Metrics
Phone number Register & Call	Phone Number & Call App Transition	Number of apps converted to perform total phone number register & call task
Phone Receive & Reply (R&R)	Phone R&R App Transition	Number of apps converted to perform total phone R&R task
SMS Receive & Reply (R&R)	SMS R&R App Transition	Number of apps converted to perform total SMS R&R task
Photo Take & Delete	Photo Take & Delete App Transition	Number of apps converted to perform total Photo Take & Delete task
Banking	Banking App Transition	Number of apps covered to perform total Banking task
Information Search & Share (IS)	IS App Transition	Number of apps converted to perform total IS task
Generic Usage	Number of Screen Unlock	Number of attempts to unlock screen unlocking pattern
Total S-ADL Tasks	Total Task Sum Transition	Number of apps converted to perform total S-ADL tasks except number of screen unlock

Table 10: S-ADL typing-related performance metrics. R&R=receive & reply, , IT=Intercharacter time, IS=Information Search & Share.

S-ADL Task	Terms of Performance Metrics	Description of Performance Metrics
SMS Receive & Reply (R&R)	SMS R&R Mean IT	The mean of average inter-keystroke interval time in SMS replying typing
	SMS R&R Median IT	The median of average inter-keystroke interval time in SMS replying typing
	SMS R&R Min IT	The mix of average inter-keystroke interval time in SMS replying typing
	SMS R&R Max IT	The max of average inter-keystroke interval time in SMS replying typing
	SMS R&R Characters per Second (CPS)	CPS (i.e., (|T|-1)/S) in SMS replying typing
	SMS R&R Keystrokes per Second (KSPS)	KSPS (i.e., (|IS|-1)/S) in SMS replying typing
	SMS R&R Gestures per Second (GPS)	Gestures (i.e., atomic action) per Second (i.e., (|ISϕ|-1)/S) in SMS replying typing
	SMS R&R Total Error Rate (TER)	TER (i.e., IF+INF/C+INF+IF) in SMS replying typing
	SMS R&R Corrected Error Rate (COER)	COER (i.e., IF/C+INF+IF) in SMS replying typing
	SMS R&R Uncorrected Error Rate (UER)	UER (i.e., INF/C+INF+IF) in SMS replying typing
	SMS R&R Utilized Bandwidth (UB)	UB (i.e., C/C+INF+IF+F) is the proportion of transmitted keystrokes that contribute to the correct aspects of the transcribed string in SMS replying typing
	SMS R&R Wasted Bandwidth (WB)	WB (i.e., INF+IF+F/C+INF+IF+F) in SMS replying typing
Information Search & Share (IS)	IS Mean IT	The mean of average inter-keystroke interval time in weather searching typing
	IS Median IT	The median of average inter-keystroke interval time in weather searching typing
	IS Min IT	The min of average inter-keystroke interval time in weather searching typing
	IS Max IT	The max of average inter-keystroke interval time in weather searching typing
	IS CPS	CPS (i.e., (|T|-1)/S) in weather searching typing
	IS KSPS	KSPS (i.e., (|IS|-1)/S) in weather searching typing
	IS COER	COER (i.e., IF/C+INF+IF) in weather searching typing
	IS UB	UB (i.e., C/C+INF+IF+F) is the proportion of transmitted keystrokes that contribute to the correct aspects of the transcribed string in weather searching typing
	IS WB	WB (i.e., INF+IF+F/C+INF+IF+F) in weather searching typing

Table 11: Overview of S-ADL task with task sequence and extracted S-ADL subtask. '*' These tasks were excluded after a preliminary study (see the details in Section 3.2

S-ADL Task	Task Sequence: App Start & End	S-ADL subtask Extraction
Phone Number Register & Call	Screen off (start) → SMS notification → Screen on → Screen pattern unlock → Home UI app start → Home UI app end → SMS app start → SMS app end → Contact app start → Contact app end → SMS app start → Calling start → Calling end → SMS app end	Notification response (screen on) Screen unlock Starting app from Home UI Instruction SMS reading Phone number register (typing) Phone call App end (click back button)
Phone Receive & Reply	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS app end → Calling notification → Calling end → SMS app start → SMS app end	Notification response (app start) Starting app from Home UI Instruction SMS reading Phone reply App end (click back button)
SMS Conversation*	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS receive → SMS send → SMS app end	Notification response (app start) Starting app from Home UI SMS short conversation App end (click back button)
SMS Receive & Reply	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS app end	Notification response (app start) Starting app from Home UI SMS reply task & App end
Photo Take & Delete	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS app end → Home UI app start → Home UI app end → Camera app start → Camera app end → Gallery app start → Gallery app end → Camera app start → Camera app end → SMS app start → SMS app end	Notification response (app start) Starting app from Home UI Instruction sms reading Starting app from Home UI Photo take Photo delete Photo transfer* App end (click back button)
Banking Transfer & Share	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS app end → Banking app start → Banking app end → SMS app start → SMS app end	Notification response (app start) Starting app from Home UI Banking app start Banking password/transfer Banking sharing task & App end
Location Search & Share*	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS app end → Home UI app start → Home UI app end → Google map app start → Google map app end → SMS app start→ SMS app end	Notification response (app start) Starting app from Home UI Instruction sms reading Starting app from Home UI Navigation Route search & share App end (click back button)
Weather Search & Share	Home UI app start → SMS notification → Home UI app end → SMS app start → SMS app end → Home UI app start → Home UI app end → Chrome site start → Chrome site end → SMS app start → SMS app end	Notification response (app start) Starting app from Home UI Instruction SMS reading Chrome app start Weather search & share App end (click back button)