User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy

Weiwei Jiang
weiwei.jiang@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Zhanna Sarsenbayeva
zhanna.sarsenbayeva@unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Niels van Berkel
nielsvanberkel@cs.aau.dk
Aalborg University
Aalborg, Denmark

Chaofan Wang
chaofanw@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Difeng Yu
difeng.yu@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Jing Wei
jing.wei@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Jorge Goncalves
jorge.goncalves@unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Vassilis Kostakos
vassilis.kostakos@unimelb.edu.au
The University of Melbourne
Melbourne, Australia

ABSTRACT

We investigate the use of a miniaturized Near-Infrared Spectroscopy (NIRS) device in an assisted decision-making task. We consider the real-world scenario of determining whether food contains gluten, and we investigate how end-users interact with our NIRS detection device to ultimately make this judgment. In particular, we explore the effects of different nutrition labels and representations of confidence on participants' perception and trust. Our results show that participants tend to be conservative in their judgment and are willing to trust the device in the absence of understandable label information. We further identify strategies to increase user trust in the system. Our work contributes to the growing body of knowledge on how NIRS can be mass-appropriated for everyday sensing tasks, and how to enhance the trustworthiness of assisted decision-making systems.

CCS CONCEPTS

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

KEYWORDS

NIRS, gluten, mobile sensing, decision-making, trust.

ACM Reference Format:
Weiwei Jiang, Zhanna Sarsenbayeva, Niels van Berkel, Chaofan Wang, Difeng Yu, Jing Wei, Jorge Goncalves, and Vassilis Kostakos. 2021. User Trust in Assisted Decision-Making Using Miniaturized Near-Infrared Spectroscopy. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445710

1 INTRODUCTION

The increasing availability of mobile and ubiquitous computing services has given rise to a wide range of applications that can assist people with their everyday decision-making [3, 7]. Previous work on assisted decision-making has primarily focused on assisting experts, particularly in healthcare, in clinical diagnosis [19] and treatment [54]. In this paper we consider an important everyday scenario with non-expert users that has not been widely considered before – the use of mobile sensing to help identify food allergens, and specifically gluten.

We wish to explore design considerations in assisted decision-making systems concerning object identification scenarios, which remain under-explored from a Human-Computer Interaction perspective. This includes, *inter alia*, the usability of new technologies to assist in decision-making, as well as the presentation of uncertainty quantification [6] with artificial intelligence (AI) based approaches. Existing work shows that various factors can affect decision-making and end-users' trust across different stages of the interaction process [42], such as end-users' tacit knowledge [58] and user interface design [13]. While it is critical to consider these factors when designing an assisted decision-making system, few studies have done so in daily-life decision-making tasks using novel technologies.

Here we focus on ingredient detection, which can be a crucial everyday decision-making task. For instance, gluten labeling regulations vary considerably between countries. In the United States, neither gluten-free labeling nor gluten-contained labeling is mandatory (it is mandatory to label wheat that is one of the main sources of gluten) [73]; conversely in countries like Australia, it is mandatory to label gluten as an allergen on the package [80]. Hence, it is possible that consumers might not be able to acquire sufficient information from package labels, while also a lot of fresh product tend to be sold without packaging.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445710

In this paper, we present a user study on assisted decision-making using a miniaturized near-infrared spectroscopy (NIRS) scanner – an emerging, but not yet widely known mobile technology. Recent studies in the HCI community demonstrate that miniaturized NIRS can be used in a variety of settings such as identifying pills [44] and the probing of ingredients in foods and beverages [37, 45, 52]. In our study, we assess three factors that may affect people's use of this new technology: *decision accuracy*, *time usage*, and *perceived trust* in the technology. We ask participants to complete the task using a fully functional NIRS setup which is highly realistic, and representative of NIRS-capable services that may become prevalent in the future. Our fully functional NIRS system can be used to detect ingredients, and it this study we focus solely on gluten detection as a task that is familiar to people while the technology is not. It is expected that in the future such services will facilitate allergen detection, and more broadly ingredient detection.

We present a 3x3 mixed within-between subject user study with 36 participants in a gluten detection task with 18 different tortilla wraps. We consider three different package labeling conditions as the between-subjects factor to avoid learning effects: 1) Gluten information is available and understandable for the end-user (with English labels); 2) Gluten information is available but not understandable for the end-user (with foreign language labels); 3) Gluten information is not available. We also consider different visualization methods to display the confidence of the device's suggestions as needed in practice [6] as a within-subjects factor. This enables us to better understand the effects of visualization on end-users' time usage for decision-making.

To summarize, this paper makes three key contributions:

- We provide a rich account of how a novel technology, miniaturized NIRS, can assist *in situ* decision-making in a daily-life scenario – identifying gluten in a consumer product.
- We identify multiple factors that can affect user's efforts, in particular time usage, as well as trust towards this new technology.
- We highlight the implications of our findings regarding users' trust towards new technologies in everyday assisted decision-making scenarios. We also propose strategies that can enhance users' trust when designing an assisted decision-making system for daily usage.

2 RELATED WORK

We first summarize existing work focusing on how technology can be used to assist humans in decision-making in various professional scenarios such as health care and business, as well as recent work on interaction design for decision-making tools. As an extension of professional scenarios, in this paper, we focus on daily-life decision-making tasks for object identification tasks. In particular, we summarize the use of miniaturized NIRS, which has broad applications not only in industry (*e.g.*, food monitoring), but also for consumers, for example through allergen detection (*e.g.*, gluten, peanuts).

2.1 Assisted Decision-Making Methods

A wide range of studies have explored the topic of assisted decision-making. Of particular interest has been the use of machine learning-based systems, which have been considered in different disciplines [62]. For instance, Bashir *et al.* present a medical decision support application (IntelliHealth) to assist in medical decision-making tasks, such as predicting diseases including heart diseases, breast cancer, diabetes, and liver disease [5]. Later, De Fauw *et al.* demonstrate a deep learning method to make referral recommendations for sight-threatening retinal diseases, achieving comparable or even better performance as compared to human experts [19]. Other studies also show that the machine learning methods can significantly assist medical decision-making in a number of different tasks (see [3]).

The effectiveness of different assisted decision-making systems has also been shown in other fields. For example, in business studies, Bilel *et al.* highlight that information systems can greatly improve decision-making and strategies in companies [7]. Similarly, Bohanec *et al.* present an intelligent system that allows users to create a sales prediction model to assist decision-making for the business environment [8]. Highlighting an application targeted on daily life among a wide group of end-users, Wences *et al.* demonstrate a mobile application that provides real-time transportation information aimed to assist passengers in arranging their journeys [76].

Nevertheless, most studies focus on expert scenarios including medical decisions and business. In this paper, we extend the works on assisted decision-making to *in situ* scenarios in daily life to further understand end-users' considerations on decision-making using novel technologies.

2.2 Interaction Design for Decision-Making

In addition to technological developments focused on decision-making, a substantial number of studies have focused on the interaction between a decision-making support tool and the end-user. For instance, Rau *et al.* present an interactive social robot to provide recommendations on several decision-making tasks (such as selecting items to carry). The authors find that highly autonomic robots have more influence on human decision-making than robots with a low level of autonomy [64]. Furthermore, for multi-people decision-making tasks, Tong *et al.* show a multi-surface environment (Pickit) tool for supporting collaborative decision-making activities such as learning activities [74].

Moreover, existing works show that users may encounter "algorithm aversion" – where users stop trusting machines after seeing mistakes [20], or "automation bias" – where users over-rely on the machine's decision [17]. To alleviate this issue, a trust calibration process may be required to help users develop a mental model of the machine's error boundaries. For example, Zhang *et al.* show that displaying confidence scores can help calibrate people's trust in an AI-assisted decision-making model, which can help human experts apply their knowledge to improve final decision outcomes [82]. Furthermore, Okamura *et al.* present an adaptive trust calibration to avoid users to "overtrust" the machine [59, 60]. Specifically, authors proposed a "trust calibration cues" framework to detect inappropriate calibration status by monitoring users' reliance behavior and cognitive clues to reinitiate trust calibration. However, trust calibration may not be sufficient to improve performance of assisted decision-making. It is important that the users have enough domain knowledge with comparable performance to complement machine's errors (*e.g.*, machines may make mistakes in some specific cases while human experts may not) [82]. Therefore, it is not practical to adopt trust calibration to our study, as we focus on general users instead of knowledgeable experts.

Nevertheless, with increasingly prevalent tools and studies for assisting decision-making tasks, there are still some challenges remain in interaction design, such as information visualization. Existing work suggests that information visualization may significantly affect users' *trust*, their willingness to rely on and use the digital information, as well as their decision accuracy [61, 81]. In particular, information regarding uncertainty is considered as one major factor in health care related decision-making scenarios [30], which may also impact user trust [6, 13, 42, 53]. For example, Antifakos *et al.* found that displaying the confidence level of presented information has a positive effect on user trust [2]. However, the increase in user trust does not necessarily lead to an improvement of the final outcome. Rukzio *et al.* demonstrated that users were more error prone and took a longer time to trust the results when the confidence was being displayed [67]. Moreover, Kong *et al.* found that a misalignment between the information visualization and the title text influences credibility of digital information [47].

Aligned with the findings above, prior literature also suggests to take into account the visual design of interface elements [15, 43, 65]. For instance, Fernandes *et al.* show that a quantile droplets visualization method can significantly improve the decision-making quality on a bus-transit task [26]. Moreover, Zhou *et al.* shows that presenting uncertainty to users can increase users' trust when their cognitive load is low [83]. In this study, we explore the effect of visualization design – in particular on time usage for the users – in relation to assisted decision-making using miniaturized NIRS.

2.3 NIRS and Its Applications

Our study focuses on the use of NIRS to assist in decision-making in daily life. NIRS technology enables users to perform a high-quality ingredient assessment and identification of everyday objects. NIRS emits near-infrared light into a sample and measures the absorbance of the light at various wavelengths, thus allowing for ingredient identifications [9]. Therefore, NIRS is an inherently useful technology across many different fields including pharmaceuticals [22], health care [68], food industry [21] and other fields. In this paper, we consider the applications on allergen detection and food monitoring using NIRS.

2.3.1 Allergen Detection. Since food allergens may cause significant health risks, it is useful to be able to identify their presence. Existing studies demonstrate NIRS as an efficient method for such a task. For example, Mishra *et al.* successfully detected peanut traces in wheat flours using NIRS [55]. Furthermore, Rady *et al.* reported 100% accuracy on detecting food allergens in 50 different powdered food materials using NIRS [63].

Most relevant to our work, Bruun and colleagues used NIRS to detect changes in gluten levels in complex gluten protein structures [11, 12]. The authors used two methods to alternate gluten levels in samples: 1) increasing the water content, and 2) heat treatment. They found that the structural changes in gluten proteins were well captured with the NIR spectra [11, 12].

However, existing studies were limited to laboratory settings with either powdered or hydrated materials that cannot be directly applied to everyday scenarios. In this study, we detect presence of gluten in tortilla wraps, which is an everyday consumer product that can be analyzed *in situ*.

2.3.2 Food Monitoring. Besides allergen detection in food, NIRS has also been used in food monitoring in general [32]. For example, NIRS has previously been used to monitor grain quality (*e.g.*, protein and moisture content) [51], meat [31], fruits [14, 23, 28], and dairy products [41].

Moreover, NIRS has been used not only for controlling food quality, but also for determining food content [32]. For instance, Isaksson *et al.* used NIRS to determine fat, moisture, and protein contents in fish [35], and ground beef [34]. The authors used multiple linear regression as the calibration method and reached prediction errors of $0.73-1.50\%$ for fat, $0.75-1.33\%$ for moisture, and $0.23-0.32\%$ for protein detection; however, their sample size was relatively small [34].

Furthermore, Kawano *et al.* used NIRS to detect sugar levels in peaches [40] and mandarins [39], with prediction errors of $0.50\,^{\circ}Brix$ and $0.32\,^{\circ}Brix$, respectively. Not only the content but also freshness of fruits has been examined using NIRS. For instance, Clark *et al.* used NIRS to detect if 'Braeburn' apples were undergoing internal browning [14]. The authors demonstrate that NIRS can successfully be used to sort apples; hence, reduce food waste in retail [14]. Prior research has also shown that NIRS can be used to determine the components of different grains [32]. Maertens *et al.* measured the protein levels in crop yield and achieved 0.57% standard cross-validation error in protein prediction [51].

In addition to the use cases reporting professional NIRS methods, recent works in the HCI community show a great potential of introducing miniaturized NIRS in everyday decision-making tasks, including the aforementioned pill identification task [44] that can also be applied to food monitoring tasks, alcohol concentration estimation [52] and beverage identification tasks [36, 37]. These examples highlight the potential of NIRS being widely used by non-experts in the near future [44, 46].

While the majority of existing literature involves NIRS used by either domain experts or as part of an automated control system, we focus on exploring the use of NIRS by non-experts in everyday settings, and particularly on understanding design considerations for assisted decision-making in object identification tasks.

3 METHOD

We design a study to assess users' trust towards NIRS and investigate how they interact with the system in the context of everyday decision-making. Our experiments focus on the use of NIRS to detect gluten in tortilla wraps.

3.1 Experimental Setup

The experimental setup is shown in Figure 1. We designed and fabricated a 3D printed case for a miniaturized NIRS scanner (Texas Instruments DLP NIRscan Nano) with a sample platform for holding wraps. The wrap can be placed on top of the scanner to retrieve a NIRS spectrum of its composition and identify gluten presence. The scanner is connected via Bluetooth to a smartphone (OnePlus 6) and controlled by our custom-built Android app. The Android app retrieves the NIRS spectrum from the scanner, and sends it (via WiFi) to our Django server running on a laptop. The server processes the spectrum and used machine learning to estimate whether

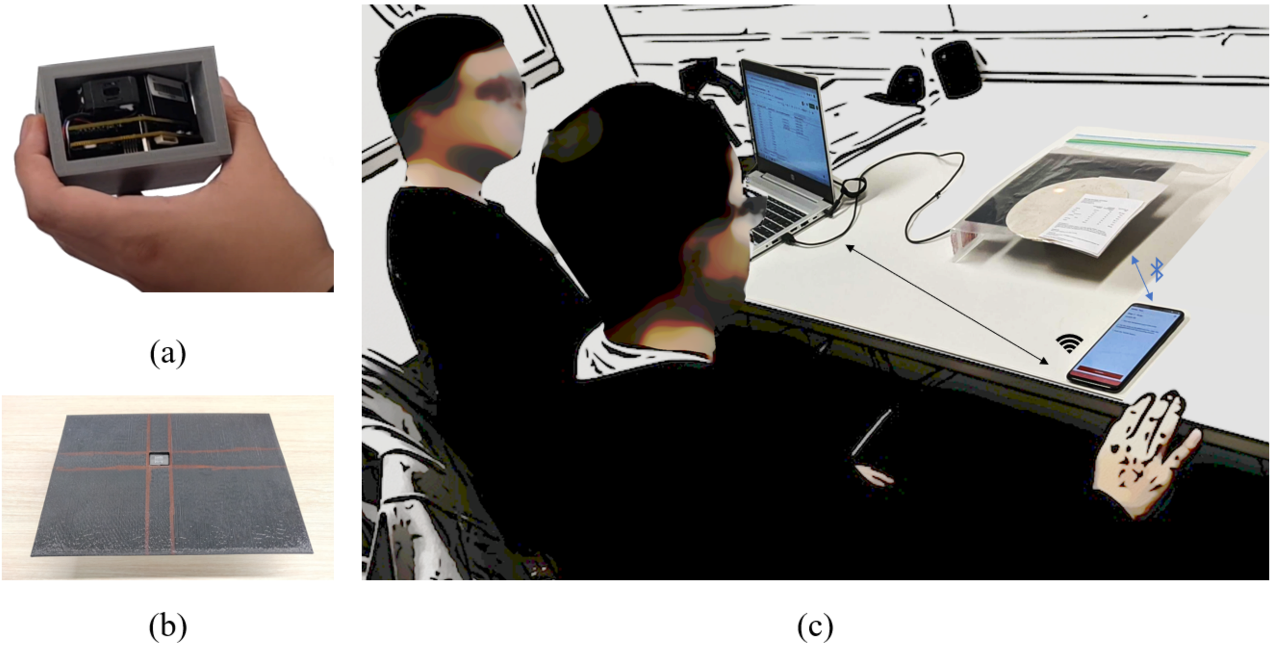


Figure 1: Experimental setup for the study. The miniaturized NIRS scanner is encapsulated in a 3D printed case (a), with a scanning platform used to place samples (b). The user study is illustrated in (c).

the scanned sample contains gluten or not. The result is returned with an estimation probability (*i.e.*, an estimation of the scanner's confidence) to the Android app, which then visualizes the results.

3.1.1 Tortilla wrap samples. We chose 18 different types of tortilla wraps (9 gluten-free, 9 containing gluten) that are commonly available in local grocery stores. Our experiment had three conditions regarding the packaging labels:

(1) *English label*: Wraps are in their original commercial packages. This corresponds to the scenario that the gluten (allergen) information is available and understandable. It is worth noting that not all countries require allergen information to be shown on the package. In our study, all the gluten-free wraps are marked as "Gluten Free" with big fonts on the package's front side, as illustrated in Figure 2 (a), with a not gluten-free package in Figure 2 (d). We use the English label condition as the baseline for comparing with other label conditions.
(2) *Russian label*: Wraps are in packages containing nutrition information in a language foreign to our participants. We chose Russian since it is not based on a Latin alphabet and, therefore, cannot be easily recognized or guessed by English-speakers. Furthermore, Russian language is not widely spoken in our community. This condition corresponds to the scenario when buying a wrap in a foreign country and not understanding the printed language. For this study, we created the Russian labels by translating the nutrition information and reconstructing the front packages from English to Russian. The nutrition information was translated by a native Russian speaker for accuracy, while the front packages were translated using the Google AR translate app to replicate their original appearance. For validation, we asked the native Russian speaker to check the texts on the Google translated packages. The Russian-labeled packages are illustrated in Figure 2 (b) and (e).
(3) *No label*: Wraps are in a transparent package without any label. This corresponds to the scenario when consuming a wrap without a commercial label, *e.g.*, at a food stand. The no-label packages are illustrated in Figure 2 (c) and (f).

For each packaging condition, there are 18 types of wraps (9 gluten-free and 9 containing gluten), resulting in $18\times3=54$ different wrap samples in total.

3.1.2 Gluten classifier. We scanned all 54 wrap samples by placing the package on the NIRS scanner. The collected spectra were used to develop machine learning models that can detect gluten. For scanning, we used the scanner settings recommended in prior literature [36, 37, 44]. Specifically, we adopted the Hadamard method with a wavelength range from 900 nm to 1700 nm, 7.03 nm generated light pattern width, 228 digital resolution, 0.635 ms exposure time, and 6 repeated scans for averaging. For each package, we scanned three positions on the pack with three scans per position. The positions are distributed on the transparent parts in the front side of the package in Figure 2. As a result, for each type of wrap, we collected 3 *scans per position* × 3 *positions per package* × 3 *types of package per wrap* = 27 NIRS spectra, resulting in $27\times18=486$ spectra in total for model training.

3.1.3 Probability Visualizations. In practice, it is often required to show the classifier's estimation probability [6]. To investigate how different visualizations of NIRS uncertainty affect user's decision-making, we implemented three different probability visualization techniques, as based on commonly used visualization techniques in

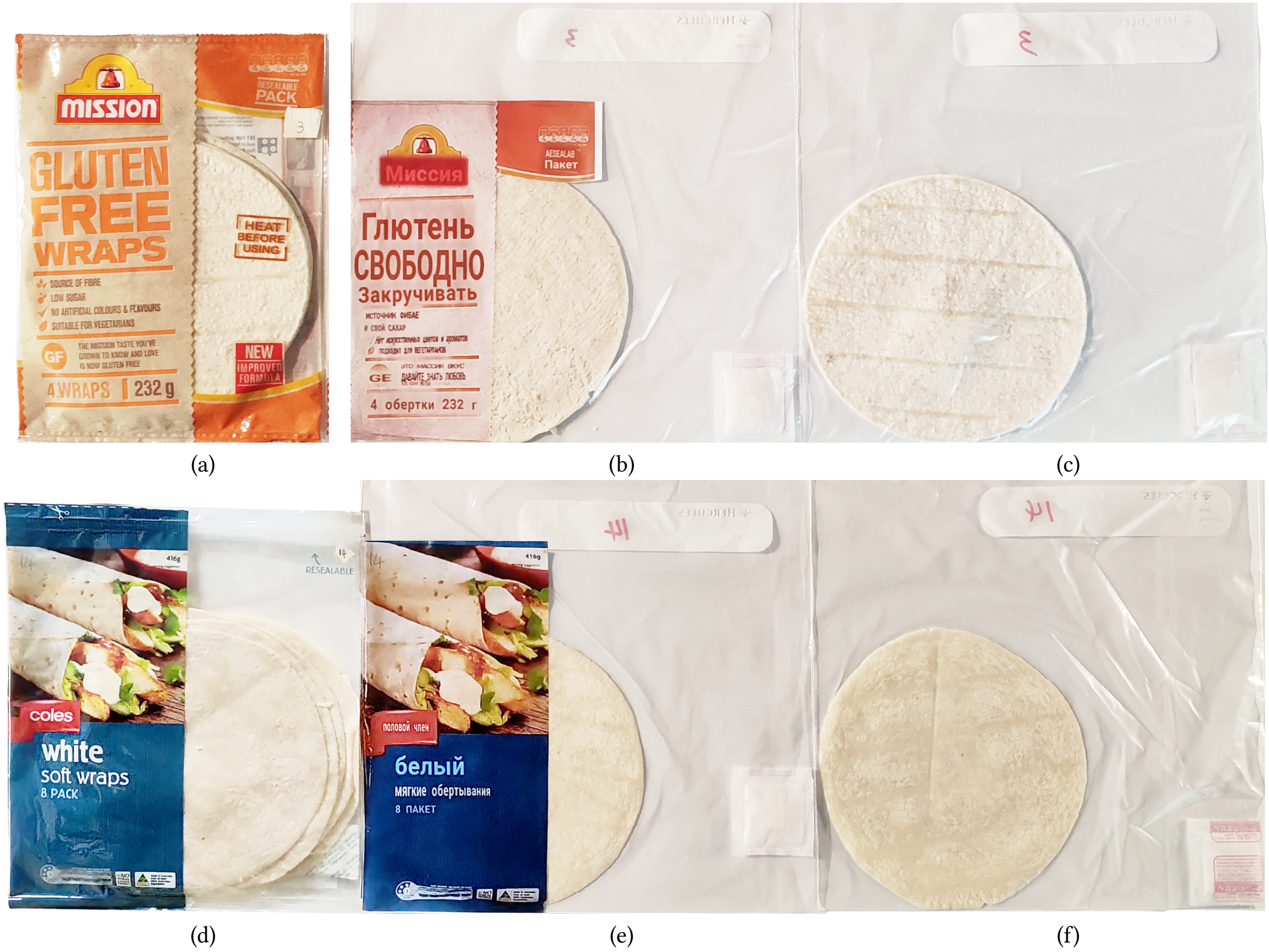


Figure 2: Illustrations of wraps in different labels. Figures (a), (b) and (c) show gluten-free wraps in the original package with the commercial label in English, a made-up package with the label translated into Russian, and a transparent package with no label, respectively. Figures (d), (e) and (f) show conventional wraps and their respective packaging.

decision support applications across different fields [1, 6, 77], detailed below:

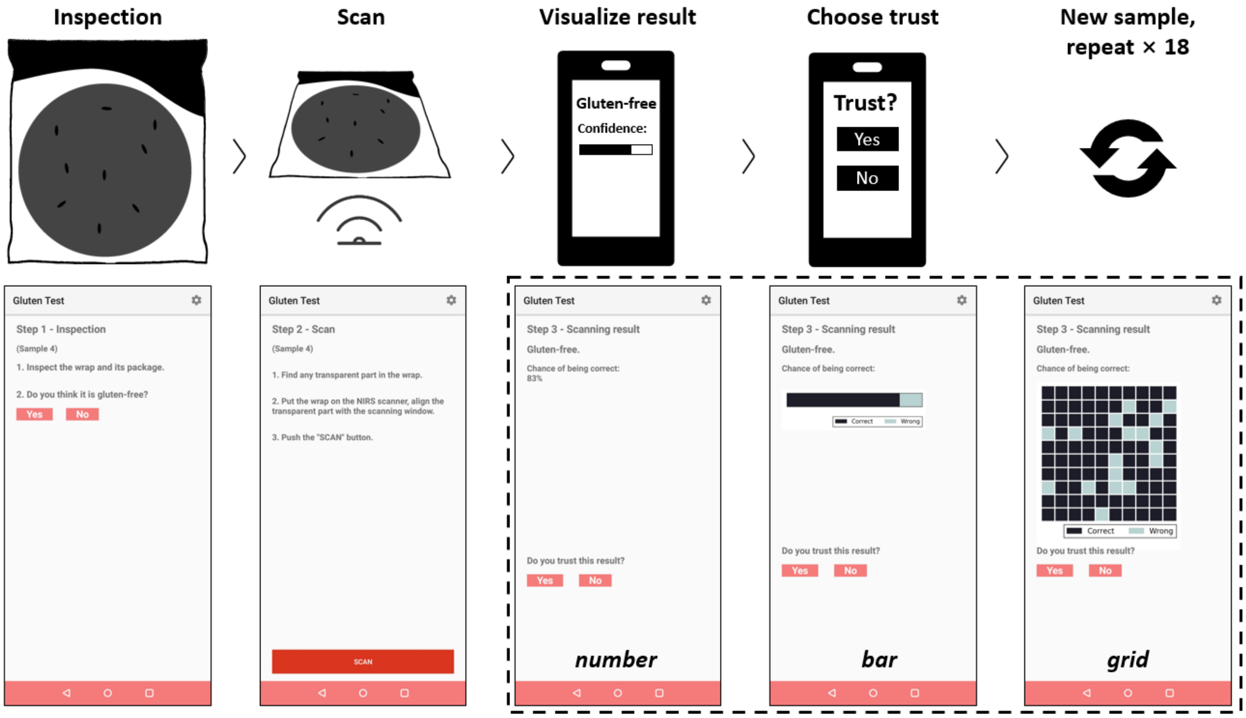
(1) *Number*: Shows a percentage as a rounded integer (0-100), which is simple and precise [1, 50, 77].
(2) *Bar*: Shows a progress-bar (0-100), which also features part-to-whole in graph designs [1, 77].
(3) *Grid*: Shows a 10×10 grid with dark-and-shallow cells, also known as the "icon array" or "frequency framing" method in a two-dimension space [1, 66, 77]. Here we randomize the cells (icons), as existing studies suggest that such an arrangement may better convey the idea of chance [1, 49].

To avoid the possibility of color affecting user preference and bias, we only use dark gray and light gray in the bar and grid visualizations [29]. Ticks and numbers are removed from the bar and grid visualizations to avoid any effect from the number itself. Illustrations of the visualizations are shown in the bottom of Figure 3.

3.1.4 Pilot study. In a pilot study we recruited 3 participants using our university's notice board (visible to all students and staff), and rewarded them with a $20 gift card. The purpose of the pilot study was to perform cross-validation against the collected ground-truth data, and identify which class of machine learning model should be used for the main user study. Although we could apply cross-validation on our own sample data to identify a suitable classifier, we decided to avoid doing so as it could lead to overfitting. In our pilot study, each participant was assigned to one label condition (English label, Russian label, and no label) with 18 wraps in each condition, after a training session. In total 18×3=54 NIRS spectra were collected for cross-validation and model selection. Participant performance during the pilot study was not included in the data analysis of the paper.

3.2 User Study

We recruited 36 participants (18F/18M) for our main user study using the university's notice board with a questionnaire for basic background information. Participants were randomly selected while



Gluten Test

Step 1 - Inspection

(Sample 4)

1. Inspect the wrap and its package.

2. Do you think it is gluten-free?

Yes No

Gluten Test

Step 2 - Scan

(Sample 4)

1. Find any transparent part in the wrap.

2. Put the wrap on the NIRS scanner, align the transparent part with the scanning window.

3. Push the "SCAN" button.

SCAN

Gluten Test

Step 3 - Scanning result

Gluten-free.

Chance of being correct:
83%

Do you trust this result?

Yes No

Gluten Test

Step 3 - Scanning result

Gluten-free.

Chance of being correct:

Do you trust this result?

Yes No

Gluten Test

Step 3 - Scanning result

Gluten-free.

Chance of being correct:

Do you trust this result?

Yes No

Figure 3: The protocol for the experiment on gluten detection task. Top: Illustration of steps for the experiment. Bottom: Screenshots for the mobile app, including three different visualization methods (a number, a bar or a grid).

simultaneously accounting for a distribution of gender, background, and age. In general, participants' age ranged between 18 and 38 (mean = 26.17, SD = 4.46) and they came from various academic background (*e.g.*, Finance, Engineering, Business, History, Food Science, Public Health).

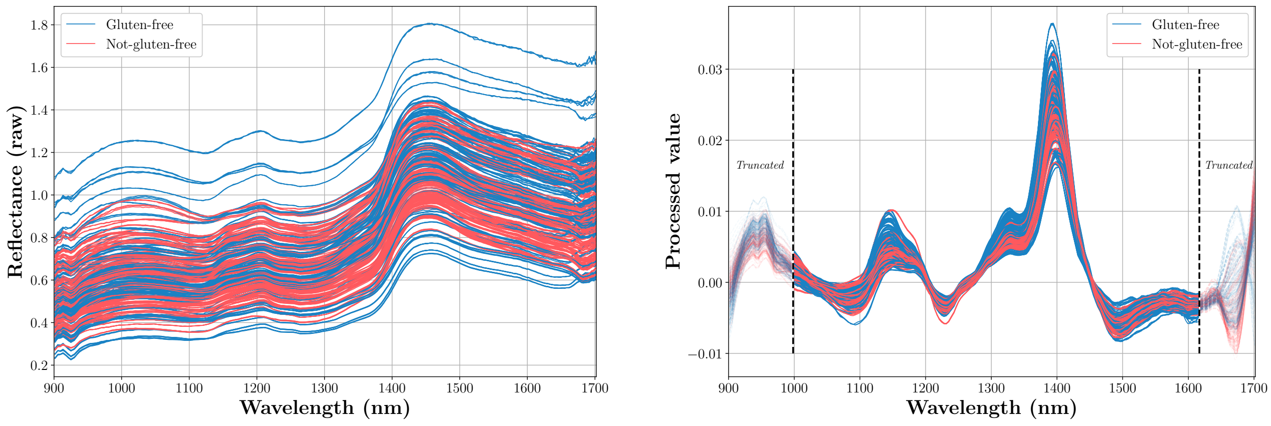
Upon arrival to our usability lab we briefed each participant on the purpose of the study, provided them with a written plain language statement, and collected their written consent to participate in the experiment. Once the participant agreed to take part in the study, we then conducted a training session with them to introduce the scanning technique using our NIRS scanner (the training session is identical to the pilot study). Once the participants completed the training session and were acquainted with the device and user interface, we randomly assigned them to one of the three main conditions (English label, Russian label, or no label), allocating 12 participants to each condition. Gluten-intolerant participants were spread equally across the three label conditions. We also confirmed that participants assigned to the Russian language condition did not have any knowledge of Russian. We then asked participants to perform the following procedure with the 18 wraps.

(1) Participant inspects a wrap for the presence of gluten and decides whether the wrap is gluten-free or not, and records their judgment in the mobile app (referred to as *inspection result* in this paper).
(2) Participant places the wrap on the NIRS holder and scans the wrap using the mobile app developed to interact with the miniaturized NIRS scanner.
(3) The mobile app then displays a scanning result (gluten-free or not) together with the scanning accuracy, shown as a number, a bar, or a grid, counterbalanced with random permutations.
(4) Participant decides to either trust or not the scanning results as their final judgment in the mobile application.
(5) Participant takes a new sample for scanning and repeats the above steps.

The protocol is illustrated in Figure 3. It should be noted that we chose binary ratings for trust. Although a fine-grained measure such as Likert scale can give more insights to compare human-device confidence levels [38], a binary rating aligns better with real-life scenarios, where end-users must make a binary decision *in situ*. Also, a complex decision can be decomposed to multiple binary decisions.

All participant inputs are recorded at each step and timestamped. Since the protocol follows a pipeline flow (*i.e.*, participants start scanning the next sample immediately after finishing the previous one), we consider the inspection time as the time between the end of the previous sample and the end of the first step. The order in which the 18 wraps were presented to each participant was counterbalanced. For this experiment, we employed two independent variables in a 3×3 mixed study design (between-by-within):

(1) 3 label conditions (between-subjects): English, Russian, and no label.
(2) 3 visualization conditions (within-subjects): number, bar, and grid.



(a) Training data (raw)

(b) Training data (processed)

Figure 4: NIRS spectra for gluten detection model training. Pre-collected raw NIRS spectra are shown in (a), processed NIRS spectra are shown in (b).

At the end of the study, we held individual exit interview sessions with each participant. Upon finishing the study, we rewarded each participant with a $20 gift card.

4 RESULTS

We first describe the accuracy of the gluten detection model, followed by an analysis of the participants' interaction with the NIRS in the aforementioned scenarios.

4.1 Gluten Detection Accuracy

To build the gluten classifier as used in the user study, we first trained multiple models for gluten detection using the collected training data as described in Section 3.1.1. Subsequently, a grid search method was applied for model selection. We utilized the model with the highest test score for the user study.

For the model preparation, we adopted pre-processing and training methods that have been shown to work well for miniaturized NIRS [37, 44]. The raw training data is shown in Figure 4 (a). Our first step in pre-processing is the application of a Savitzky-Golay filter to the raw spectra (window length = 21, polynomial order = 3) for smoothing. Next, we take the first-order gradient of the smoothed spectra. Finally, the over-sensitive two ends of the spectra are truncated, resulting in 175 wavelengths (features) ranging between 1000 nm and 1615 nm. The NIRS spectra after preprocessing are shown in Figure 4 (b).

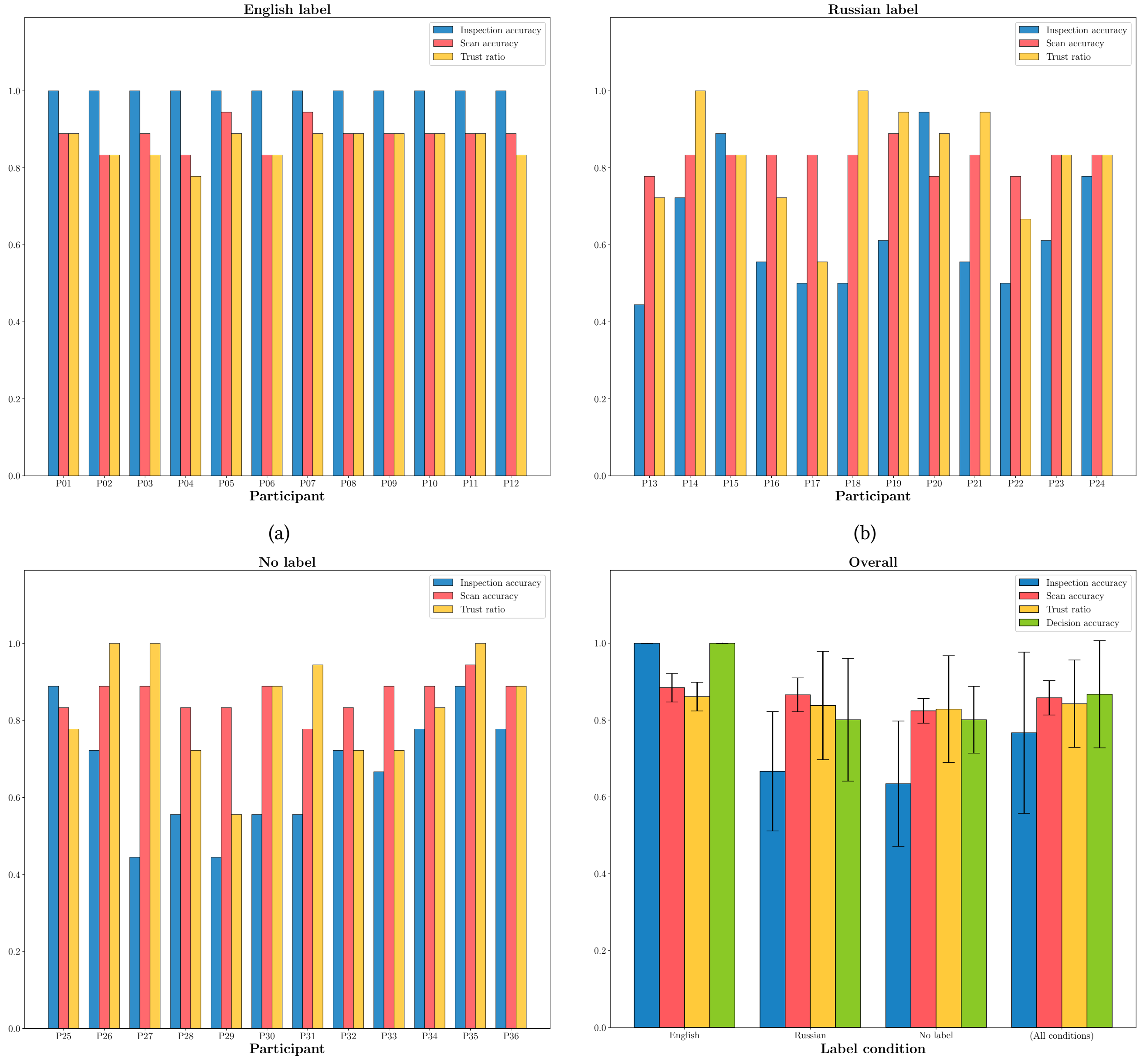
Utilizing these pre-processed spectra, we adopted a Random Forest (RF) classification model and a Support Vector Machine (SVM) classification that were shown to be both accurate and robust in literature [37, 44]. To tune the model for our study, we conducted a grid search for both the RF and SVM models, with numbers for estimators (for the RF model) and regulation parameter C (for the SVM model) ranging between 1 and 100. Both models used the data collected by our scanning for training, and the data from the pilot study for validation. The RF models achieved 100% accuracy in training, however, the test accuracy was relatively low (< 80%). For the SVM models, although the training accuracy did not achieve 100%, the test accuracy was more stable than the RF models with higher accuracy. Therefore, we selected the best SVM model (C=16) with the highest test score as the first criterion and the highest training score as the second criterion (training accuracy = 0.901, test accuracy = 0.815).

During the user study we collected a total of 18 *scans per participant* × 36 *participants* = 648 *scans*. Overall, the mean accuracy of our classifier was 0.858 (SD = 0.045) during the user study. Within the 91 false cases, 80 cases were false-positives (false-gluten-detected) while 11 were false-negatives (false-gluten-free). Considering the fact that the consequence of eating a gluten-free wrap is not as severe as the opposite, a higher false-positive rate is preferable to a higher false-negative rate. In other words, the benign ratio for NIRS is 0.983.

4.2 Participant accuracy and trust

We summarize participants' performance in Figure 5. Across all participants, the mean inspection accuracy (how accurately participants can detect gluten on their own) is 0.767 (SD = 0.210), the mean scan accuracy (how accurate is our system in detecting gluten) is as reported already 0.858 (SD = 0.045), and the mean trust ratio (how often people trust the scanner result) is 0.842 (SD = 0.114). For each condition (*i.e.*, different packaging labels), the observations are as follows:

- **English label:** All 12 participants that were assigned to the English label condition achieved 100% inspection accuracy. The mean scanning accuracy is 0.884 (SD = 0.037), and the mean trust ratio is 0.861 (SD = 0.037). These results confirm that all participants could read and interpret the English labels correctly.
- **Russian label:** For the 12 participants that were assigned to the Russian label condition, the mean inspection accuracy is 0.634 (SD = 0.163), the mean scanning accuracy is 0.824 (SD = 0.032) and the mean trust ratio is 0.829 (SD = 0.139). Two participants achieved much higher accuracy than conditions' average (P15 with 0.889 accuracy, P20 with 0.944 accuracy).



(a)

(b)

(c)

(d)

Figure 5: Inspection accuracy, scan accuracy, and trust ratio in main study. (a) English label condition. (b) Russian label condition. (c) No label condition. (d) Means and standard deviations for each condition and all conditions.

- **No label:** For the 12 participants that were assigned to the no label condition, the mean inspection accuracy is 0.667 (SD = 0.155), the mean scanning accuracy is 0.866 (SD = 0.044), and the mean trust ratio is 0.838 (SD = 0.141). Also, two participants achieved much higher accuracy than average (P25 and P35, both with 0.889 accuracy).

For all the 151 inspection errors, 94 were false-positive and 57 false-negative, showing that participants tended to over-report wraps as gluten contained. In particular, participants in the Russian label and no label conditions reported higher inspection accuracy for gluten-contained wraps (0.72 in Russian label condition, and 0.75 in no label condition) than what would be expected by random guess (0.50). This may be caused by participants recognizing some features (such as wheat grains) on several gluten-contained wraps. Nevertheless, we accounted for this issue by randomizing their presentation order. For the scanning results, the accuracy dropped for the Russian label and no label conditions. This is likely caused by the sensitivity of NIRS

Table 1: Variance and post-hoc tests for the recorded and perceived inspection time between pairs of label conditions.

Time usage group	ANOVA	Post-hoc tests (Tukey's test)		
		English vs. Russian	*English vs. no label*	*Russian vs. no label*
Inspection (recorded)	F = 11.01 (p < 0.05)	p < 0.05	p = 0.08	p < 0.05
Scan (recorded)	F = 34.77 (p < 0.05)	p < 0.05	p < 0.05	p < 0.05
Inspection (perceived)	F = 0.61 (p = 0.54)	-	-	-
Scan (perceived)	F = 5.89 (p < 0.05)	p < 0.05	p < 0.05	p = 0.90

and high dimensional data (175 features), resulting in challenges to tune a stable classifier. Yet this did not affect the final decision accuracy for Russian and no label conditions (as shown below in Section 4.2.1). Furthermore, participants had higher trust ratio of the gluten-detected scans (0.618 trust ratio) compared to gluten-free scans (0.392 trust ratio).

4.2.1 Decision accuracy. Next, we calculate participants' *decision* accuracy. We define '*decision*' as a binary variable indicating the final judgment by participants (gluten-detected vs. gluten-free), which was obtained after considering their own initial opinion and the scanner's estimation. The results are shown in Figure 5 (d). Overall, the mean decision accuracy was 1.0 (SD = 0.0) for the English label condition, 0.801 (SD = 0.087) for the Russian label condition, and 0.801 (SD = 0.160) for the no label condition. With paired t-tests, we find that the mean decision accuracy is significantly higher than the mean inspection accuracy in both Russian label condition (p < 0.05, Cohen's d=1.84) and no label condition (p < 0.05, Cohen's d=1.25). Furthermore, among the 86 decision errors in all conditions, 63 were false-positive with 23 false-negatives, which follows a similar bias as the inspection accuracy (*i.e.*, participants over-report gluten-detected results).

4.2.2 Participant Trust. Finally, we define '*consensus*' as a binary variable to indicate whether participants' initial inspection produced the same gluten-detection result as the subsequent NIRS scan, detailed as the following four cases

(1) *With consensus*: Participant inspects as gluten-free, scanner estimates as gluten-free.
(2) *With consensus*: Participant inspects as not gluten-free, scanner estimates as not gluten-free.
(3) *Without consensus*: Participant inspects as gluten-free, scanner estimates as not gluten-free.
(4) *Without consensus*: Participant inspects as not gluten-free, scanner estimates as gluten-free.

Overall, participants trusted the scanner results 84% of the time. We also investigate whether trust is affected by consensus, *i.e.*, is there more trust towards the scanner when there is consensus (*null hypothesis*), or is trust unrelated to the consensus (*alternative hypothesis*)? We test for this using a McNemar's test (*consensus vs. trust*, paired data points) for each label condition.

In the English label condition there is no significant difference of distributions (*i.e.*, proportions of True and False) between consensus and trust (effective size ϕ = 0.12, p = 0.07), while there was a significant difference of distributions (proportions) in the Russian label condition (effective size ϕ = 0.39, p < 0.05) and the no label condition (effective size ϕ = 0.41, p < 0.05). These results suggest that, on one hand, in the English label condition participants' decision to trust (or not) the NIRS scanner was affected by the existence of consensus (0.974 trust ratio with consensus vs. 0.0 trust ratio without consensus). This was expected, since the participants in this condition effectively knew the ground-truth, and only trusted the scanner if it agreed with their own opinion. On the other hand, in the Russian label and no label conditions participants' decision to trust (or not) the scanner was not significantly affected by consensus (*i.e.*, participants tended to trust the scanner regardless of consensus). This observation also validates the results above that the mean decision accuracy is higher than the inspection accuracy in the Russian label condition (0.634 vs. 0.801 accuracy) and the no label condition (0.667 vs. 0.801 accuracy).

4.3 Participant completion time

Next, we investigate the completion time for inspection and scanning. We emphasize that scan time includes the time taken to place the wrap on the scanner and the time spent waiting for the results to be displayed on the smartphone. Here, we consider completion time as the effort required to achieve decision support. The results are shown in Figure 6 and Table 1.

In general, we observe that the recorded completion time varies across the label conditions, for both inspection and scanning (p < 0.05 for both). For inspection, participants spent less time on wraps with English label (mean = 16.36 sec, SD = 7.92) and no label (mean = 18.41 sec, SD = 10.18), and more time on wraps with a Russian label (mean = 20.79 sec, SD = 10.32). The results show that participants could quickly make a decision when there was sufficient information or an absence of information, while the processing time increased with information that was not understood by our participants (*i.e.*, information in a foreign language).

For scanning, participants spent the least time on the no label condition (mean = 12.75 sec, SD = 2.50), slightly longer on the English label condition (mean = 14.63 sec, SD = 2.98), and longest on the Russian label condition (mean = 15.51 sec, SD = 4.47). The increasing time usages on scanning for the English label and the Russian label conditions might be caused by placing and positioning of the wraps on the scanner in such a way that the NIRS scanner can scan the wraps without obscure. In contrast, a wrap without any label was covered in a fully transparent package that does not require additional efforts on positioning (as shown in Figure 2).

Next, we consider the length of decision time under consensus or no consensus, *i.e.*, when participants' inspection is the same as or different to the scanner's estimation, as shown in Figure 7 (b). We observe that the decision time is significantly different ($F_{1,35}$ = 41.78, p < 0.05), and specifically when there is no consensus the decision time is significantly longer (mean = 7.80, SD = 4.58) than with consensus

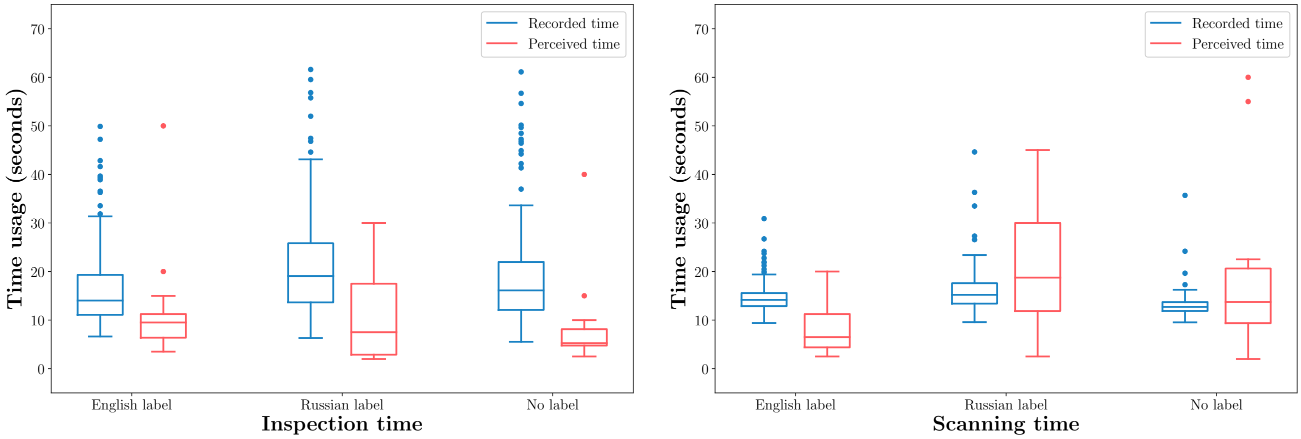


Figure 6: Recorded and perceived completion time for inspection (left) and scanning (right).

(mean = 4.97, SD = 3.30). This result also holds for the different visualization methods ($p < 0.05$ for *with vs. without consensus* pairs in all visualization methods). Between visualization methods, we do not find significant differences under the without consensus condition. However, participants spent shorter time with the number or the bar visualization method as compared to the grid visualization method.

Finally, we analyze the recorded decision time for each visualization method in Figure 7 (a). For each visualization method (number, bar, grid) the mean time values in seconds for decision-making are 5.32 (SD = 3.56), 5.53 (SD = 3.87), and 6.49 (SD = 4.22) ($F_{2,70} = 4.56$, $p < 0.05$) respectively. In particular, we can observe that the decision time with the number visualization is shorter as compared to the grid visualization. For the other two pairs we do not find significant differences.

4.4 Participant Perceptions

4.4.1 Perceptions of time. During the exit interview, all participants were asked to estimate the time spent on both the inspection and scanning tasks. For the inspection time, we do not find a significant difference between the three label condition groups. For the scanning time, participants in the English label group perceived a shorter time spent (mean = 8.79 sec, SD = 6.79) than in the Russian label condition (mean = 20.83 sec, SD = 13.50) and the no label condition (mean = 20.375 sec, SD = 17.82). This might be caused by the fact that participants in the English label condition did not anticipate to obtain any additional information from the scanner, while participants in the other two conditions were relying on the scanning results. We contrast the recorded completion time and the participants' perceived completion time (Table 2). Participants in the English label condition tended to underestimate the scan time, while participants in the other two label conditions tended to underestimate the inspection time.

4.4.2 Preference for Visualizations. In the exit interview, we asked participants to provide their preference rank on the three visualization methods (number, bar, and grid). As shown in Figure 7 (b), of the 36 participants, 23 participants found the number visualization (a percentage) the most preferred method, 8 selected the bar visualization, and 5 preferred the grid visualization. The least preferred method of visualization was the grid (30 participants). We then adopted the Borda count [25] method to assign points to all the participants' rankings (3 points to the most preferred, 2 to the second preferred, and 1 to the least preferred), followed by a repeated ANOVA test with multiple paired T-test with Bonferroni correction as post-hoc test. The result shows the points for the three visualization groups are significantly different ($F_{2,70} = 27.22$, $p < 0.05$). The number and the bar conditions show no significant difference ($p = 0.55$), while both the number method and the bar method have a significantly higher score than the grid method ($p < 0.05$ for both pairs), indicating that the grid method is less preferred than the other two methods.

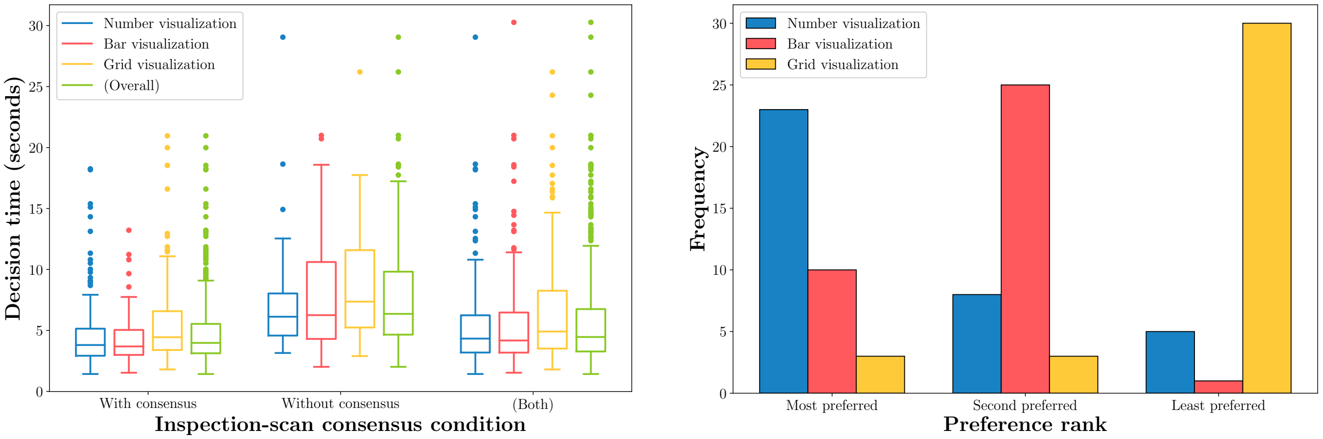
Table 2: Paired T-test between mean recorded and perceived completion time for inspection and scanning across different label conditions.

Label condition	Inspection time (*recorded vs. perceived*)	Scan time (*recorded vs. perceived*)
English label	d = -0.41 (p = 0.38)	d = -1.43 (p < 0.05)
Russian label	d = -1.11 (p = 0.05)	d = 0.58 (p = 0.19)
No label	d = -1.10 (p < 0.05)	d = 0.61 (p = 0.17)

4.4.3 Threshold for Trust. Furthermore, we investigate the effect of *a priori* information on participants' trust decisions. Participants were asked to provide a probability value as the threshold to trust in the exit-interview (*i.e.*, participants would lean to trust the scanner's result if the probability is above this threshold). In total, 33 participants out of 36 reported a threshold (three participants were not able to provide a probability). The mean trust thresholds and standard deviations are 0.81 (SD = 0.08) for the English label condition, 0.72 (SD = 0.07) for the Russian label condition, and 0.72 (SD = 0.06) for the no label condition, respectively. An ANOVA test shows that there is a significant difference between the label conditions for the trust threshold ($F = 5.32$, $p < 0.05$). Specifically, participants in the English label condition reported higher threshold probabilities (mean = 0.81, SD = 0.08) than the other two label conditions ($p < 0.05$ for both pairs

Table 3: Variance test for decision-making time with different visualization methods.

Consensus condition	Repeated ANOVA	Paired T-test with Bonferroni correction		
		Number vs. bar	*Number vs. grid*	*Bar vs. grid*
With consensus	F = 6.17 (p < 0.05)	p = 1.00	p < 0.05	p < 0.05
Without consensus	F = 0.65 (p = 0.53)	-	-	-
Both	F = 4.56 (p < 0.05)	p = 1.00	p < 0.05	p = 0.08



(a) decision-making time in the final step.

(b) Visualization preferences.

Figure 7: Visualization preference and decision-making time.

with Tukey's tests), while participants reported similar threshold probabilities for the Russian label and no label conditions (p = 0.90). This indicates a significant impact of *a priori* information on the participants' trust decisions for the scanner's estimation results.

4.5 Qualitative Results

Finally, we carried out a thematic analysis based on the exit-interviews. Three authors were involved in the thematic analysis, which was conducted in two stages. In the first stage, three authors read all the responses and independently performed initial coding on the data. In the second stage, the same authors reviewed the initial codes and agreed on the final codes. We merged similar codes to obtain our list of final codes. Finally, we reviewed the final codes and agreed on the themes.

In general, the majority of participants commented that the device is easy to use and reliable. However, there are several factors that impacted participants' trust – as categorized in three themes below.

4.5.1 Theme 1 – Prior Knowledge. Some participants mentioned that their prior experience affected their trust in the device, in particular, they stated they would trust more on their knowledge, rather than the machine unless it matches their initial decision. A couple of participants stated that they were able to identify wraps themselves: *"I do eat a lot of wraps, so I know how to identify if the wraps were gluten-free or not. I trusted the machine if it matched my guess. "* (P34, no label condition). Also, a few of our participants claimed that they could distinguish gluten-free wraps based on their characteristics. For example, gluten-free wraps may have specific colors and textures: *"I looked at the texture, I can recognize the (gluten-free) brand from its whiteness and thickness. If the scanner agrees with me then I choose to trust.* (P20, Russian label condition), *"I've got experience in cooking so I could have guessed from the colors and the texture and the type of the flour. Whether the scanner agreed with my guess affects my trust."* (P35, no label condition). This indicates that participants may trust more their own prior knowledge if they have experience with gluten-free wraps, even though they are gluten tolerant but cook or consume gluten-free food.

4.5.2 Theme 2 – Consensus. Furthermore, several participants, regardless of their prior knowledge or experience on gluten detection, mentioned that the consensus between their initial decision and the device's estimation had a greater effect on their trust towards the device: *"[Whether] my guess is confirmed by the scanner affects my trust [of the results suggested by the scanner]"* (P26, no label condition). More interestingly, some participants trusted the device only if the device produced the same result as their initial decisions: *"I wasn't very sure if it is trustworthy, but later there were two more cases showing the same as my guesses so I got more trust"* (P36, no label condition). In addition, a couple of participants highlighted that a high confidence level of the scanner's estimation was required in order for them to trust the device if the reading differed from their initial decision: *"If the result is the same as my guess then I don't care about the confidence, if not, then [I would trust the device if the probability is] above 80%"* (P13, Russian label condition).

4.5.3 *Theme 3 – Risk Aversion.*

As people may have health concerns regarding gluten including those who are not gluten-intolerant (*e.g.*, they have gluten-intolerant friends or relatives, or simply choose a gluten-free lifestyle [57]), some participants tended to be risk-averse. In particular, a couple of participants were aware of the risks of false-positives (*i.e.*, classifying a gluten-free wrap as gluten-contained) and false-negatives (*i.e.*, classifying a gluten-contained wrap as gluten-free), as the consequence of false-positives being benign while false-negatives could cause acute symptoms for gluten-intolerant people: *"I'd be more conscious if I was gluten-intolerant"* (P05, English label condition), *"It won't hurt as long as always detects gluten [false positive is fine]"* (P08, English label condition). Another participant stated that gluten-intolerant people should always trust the label: *"If I were gluten-intolerant, then I would trust the label"* (P35, No label condition).

5 DISCUSSION

5.1 Decision Accuracy

Unsurprisingly, participants assigned to the English label condition had a 100% accuracy in identifying gluten-free items. This application of NIRS is aimed at situations in which end-users are unable to infer the required information from, *e.g.*, an available ingredient label that cannot be read or trusted. Furthermore, although the scanner did not have perfect accuracy, the errors were mostly false-positives (reporting a gluten-free sample as gluten-detected) rate, which is benign when compared to false-negatives. The same pattern was also observed in participants' behavior, both for inspection and their final decision: their behavior was conservative and biased towards avoiding a false-negative decision (*i.e.*, mis-reporting a gluten-contained wrap as gluten-free) that may have an adverse consequence. To summarize, our participants were aware that false-positive results are less dangerous for the health of gluten-intolerant people as compared to false-negative results. This led them to better identifying gluten-contained wraps which resulted in a high false-positive self-detection rate discussed in Section 4.2. These insights are in line with our quantitative and qualitative results presented in Sections 4.2 and 4.5 respectively.

This bias might be explained by participants' tacit knowledge [10], which implies that participants were conscious of the potential harmful consequence caused by the false-negative result. Specifically, previous work suggests that tacit knowledge may not be easily recognized or acknowledged but can potentially affect the outcome of decision-making. Our findings also suggest an important consideration on the effect of tacit knowledge when developing ubiquitous systems, *e.g.*, users may be risk-averse when making a final decision, as also highlighted by Ocegueda *et al.* [58]. Potentially, it might be possible that users tend to abandon the use of an assisted decision-making system if they encounter recommendations that are wrong and harmful.

5.2 Decision Time

In this study we had an opportunity to study how participants use our technology to make decisions on an object identification task. The results showed variations in the decision time of participants, as measured by the amount of time they took for each step in the experiment. For the inspection task, participants spent the longest time on inspecting Russian labels. In contrast, participants in the English label condition and the no label conditions made decisions in shorter time. This shows that additional information that cannot be understood may require additional time for decision-making, as compared to the conditions with either comprehensible information or no information at all. This finding implies a distraction effect of the Russian label condition, as participants struggled to find necessary information in a foreign language. However, it also suggests that NIRS User Interfaces should avoid providing superfluous information. Our UI design was minimal, and unlike typical NIRS UIs targeted at experts, we did not include information on wavelengths, amplitude, spectra, or technical settings that are required for operation [71].

For the scanning task, participants could successfully use the scanner without *a priori* knowledge after completing a training session. Specifically, besides the time waiting for the data transmission and signal processing, participants spent most of their time on placing and positioning the transparent part of the package in the wrap on the scanner, with less than 3-second difference on average. This observation indicates an acceptable usability level of our experimental setup. Nevertheless, some improvements are required to further reduce the time spent on positioning the sample. One possibility would be to let the users place and position the scanner on the wraps (rather than the wraps on the scanner). However, additional training and signal processing methods might be necessary as different background noise can be induced in such a scenario as highlighted by previous work [45].

Moreover, we observed that participants tended to underestimate the time they spent on the inspection task in general. Specifically, we found a significant difference in the no label condition for the inspection task, and in the English label for the scanning tasks. This is similar to the findings highlighted by Van Berkel *et al.* in smartphone usage scenarios [75]. This result suggests that participants' perception might be influenced by the *a priori* information. In particular, participants in the no label condition could not find sufficient clues to determine gluten presence, and they perceived themselves to spend less time on the inspection task. In contrast, participants in the English label condition were already satisfied with the ground truth, they did not expect additional information from the scanner, hence perceived waiting time on the scanner appeared to be shorter. A similar effect on the perceived waiting time by information availability and expectations has been previously reported by Thompson *et al.* in a study regarding waiting time perceptions in an emergency department [72].

Finally, we also observed that the choice of visualization affects the time taken to make a decision. Overall, participants took more time to decide when using the grid visualization as compared to the number or the bar visualizations. This result is aligned with the preference of the participants, as the majority of the participants preferred the number or the bar visualizations. This observation shows that the grid visualization method may pose higher cognitive load compared to the other two methods, since it has more unrelated information (higher information entropy). Furthermore, we found that participants took longer to make the decision when there was no consensus between their inspection result and the scanner's result. This finding aligns with previous work by Rukzio *et al.* , where they demonstrated that users were more error prone and took longer time to make a trust decision when the confidence was being displayed [67].

5.3 Emergent Trust Issues

Our intention in this study is to elicit trust responses from participants in different experimental conditions. Hence, our study used different labeling and visualization conditions as a means to give rise to a wide range of trust issues during decision-making. We observe that when there is no clear evidence to validate participants' inspection results, participants tend to follow the scanner's estimation results. In contrast, as expected, participants tend to trust the scanner when it provided a scan result that aligned with their initial perceived ground truth obtained during the inspection. Interestingly, however, as an extreme case, one participant in the Russian label condition also claimed that they only trusted themselves because they had no idea about the technology itself (*i.e.*, they only trusted the scanner when it agreed with their inspection results). This might be driven by the personality characteristics of the participant; however, personality traits and their effect on decision-making was beyond our study's scope. Nevertheless, such a factor should be considered in future studies to further understand the personality's effect on assisted decision-making with new technology.

In addition, we observed that participants in the English label condition had a higher threshold for trust than those in the other two conditions. This implies that higher reliability is required to affect people's decision-making when there exists some other reliable *a priori* information. Also, prior work has shown that users trust and accept system information more easily if they can relate or link it to their prior knowledge [16]. Our findings are in line with the literature [16, 18] and demonstrate that users are more likely to accept the results of the scanner if they align with their own initial observation (*i.e.*, participants spent longer time on the trust decision when they had no consensus with the scanner).

Furthermore, users' perceived risk has previously been shown to influence their trust in information, *e.g.*, the users tend to reevaluate their trust if the decision comes at a potentially high cost [70]. In line with this finding, we demonstrate that participants tended to over-report wraps as gluten-detected, and their trust in the scanner's results increased if the scanner reported gluten-detected.

To this end, along with the aforementioned findings, we summarize three main considerations when designing an assisted decision-making system for everyday object identification scenarios:

(1) *Unnecessary information should be avoided.* In our study, participants spent longer time on non-understandable information (Russian label) and unrelated information (grid visualization), without improving their decision accuracy. Hence, showing excessive information or settings may induce extra time for decision-making without substantial benefits. This factor may be increasingly important in an interconnected-world in which travelers are required to make *in situ* decisions such as in stores [33] or restaurants [48].

(2) *Risks should be highlighted.* For instance, in our study, participants tended to over-report gluten-detected as the consequence of making an error (false-positive) was less risky than making a false-negative decision. However, there might be situations when users may not know the risks of making a wrong decision. Hence, risks should be outlined when showing a recommendation to users for decision-making. This factor may be amplified in high-stakes decision-making scenarios, such as healthcare [6] and commercial decisions [7].

(3) *Highly confident recommendations are required to influence a user's decision.* In particular, literature shows that uncertainty information (*e.g.*, a probability showing the confidence of the recommendation) is needed to improve users' trust [6]. Through our study, we further demonstrate a minimal of 75% confidence on average for the system's recommendation was required for the end-users to trust the scanner. In particular, this threshold may be higher for the applications where users have competitive performance with the machine, such as clinical diagnosis [19], law enforcement [79], and agriculture automation [56].

5.4 Limitations

We acknowledge a number of limitations in our study. First, the use case scenario of our experiment was limited to tortilla wraps. As clarified in Section 3, we chose tortilla wraps as the used sample due to their wide availability in daily life, but lower popularity when compared to other types of bread. However, the use cases for miniaturized NIRS can be extended to other scenarios as we clarified in the related work section, such as peanuts, milk, and other food or beverages [37, 44, 51, 78]. We also note that the miniaturized NIRS scanner has physical limitations (*e.g.*, wavelength, light intensity, etc). Furthermore, we restricted the position of the scanner to perform upside-down scanning, which was necessary to prevent user-induced errors that might occur due to various factors (*e.g.*, device motion, sample motion, sample angle) [45]. We only adopted three fundamental visualization techniques according to their dimensions: zero-, one-, and two-dimension for number, bar and grid (icon-array) respectively. We did not include other factors that might have possibly influenced user decision-making (*e.g.*, color [29], hierarchy [24], codification [27]). Future studies can be conducted to further study the effects of those factors on task load, decision outcomes, etc. Also, other trust theories can be bridged for simulating other realistic contexts for in situ decision-making tasks, such as trust calibration for participants with specific domain knowledge [82] and users' personality [84].

In addition, our study was conducted in a laboratory setting for simulating real-life contexts. We acknowledge this as a limitation of our study; nevertheless, by doing so we strived to maximize the internal validity of the study through rigorous control and conditions. However, we agree that to increase the ecological validity of our results, a future field study is required to further our understanding of the design factors that may affect user trust in an *in situ* decision-making task. Also, in our study scenarios, we consider all labels are reliable regardless of their understandability. However, in practice, it is possible that the labels are not trustworthy, which may be another important factor that is worth investigating in a future study. Furthermore, 14% of our participants (5 out of 36) in the study were gluten intolerant. This representation of gluten intolerant participants aligns with worldwide statistics (5.7%~16.3%, with ~1.4% celiac disease [69] and 4.3%~14.9% wheat sensitivity [4]). Moreover, since NIRS is a generic device for material sensing, our methodology can be generalized to other allergens beyond gluten – as discussed in Section 5.3. Nevertheless, future study with a stronger focus on gluten intolerant participants may yield more insights specific to those living with this condition.

Finally, we have limited options in our decision-making task, *i.e.*, only binary options that might be one of the most common cases in real-life. However, for more complex tasks, it is possible to break down the decision process into multiple binary decisions. Hence, some of our findings might also be applied in more complex decision-making tasks.

6 CONCLUSION

In this work, we investigate how end-users interact with an emerging miniaturized NIRS scanner in an assisted decision-making task of identifying gluten in tortilla wraps. We conducted a user study with 36 participants to identify the effects on users' decision-making process when using NIRS technology. Our findings reveal that different factors, including information availability and visualizations, affect users' decision accuracy, time usage, and trust towards the technology. Based on our findings, we provide design considerations that could benefit the development of assisted decision-making systems using NIRS or other novel technologies for object identification tasks in everyday settings.

ACKNOWLEDGMENTS

This work is partially funded by University of Melbourne Graduate Research Scholarships 10.13039/501100000987, ARC Discovery Project DP190102627 and NHMRC grant 1170937. Zhanna Sarsenbayeva is supported by a Doreen Thomas Postdoctoral Fellowship.

REFERENCES

[1] Jessica S Ancker, Yalini Senathirajah, Rita Kukafka, and Justin B Starren. 2006. Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association* 13, 6 (2006), 608–618.

[2] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards Improving Trust in Context-Aware Systems by Displaying System Confidence. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services* (Salzburg, Austria) *(MobileHCI '05)*. Association for Computing Machinery, New York, NY, USA, 9–14. https://doi.org/10.1145/1085777.1085780

[3] Abdullah Awaysheh, Jeffrey Wilcke, François Elvinger, Loren Rees, Weiguo Fan, and Kurt L Zimmerman. 2019. Review of medical decision support and machine-learning methods. *Veterinary pathology* 56, 4 (2019), 512–525.

[4] Imran Aziz. 2018. The global phenomenon of self-reported wheat sensitivity.

[5] Saba Bashir, Usman Qamar, and Farhan Hassan Khan. 2016. IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of biomedical informatics* 59 (2016), 185–200.

[6] Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence* 1, 1 (2019), 20–23.

[7] R Bilel and AH Alaa. 2017. Impact of the information system on decision-making within the company. In *2017 9th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. IEEE, IEEE, New York, NY, USA, 1–8.

[8] Marko Bohanec, Mirjana Kljajić Borštnar, and Marko Robnik-Šikonja. 2017. Explaining machine learning models in sales predictions. *Expert Systems with Applications* 71 (2017), 416–428.

[9] L. Bokobza. 1998. Near Infrared Spectroscopy. *Journal of Near Infrared Spectroscopy* 6, 1 (1998), 3–17. https://doi.org/10.1255/jnirs.116 arXiv:https://doi.org/10.1255/jnirs.116

[10] Erich N Brockmann and William P Anthony. 2002. Tacit knowledge and strategic decision making. *Group & Organization Management* 27, 4 (2002), 436–455.

[11] Susanne Wrang Bruun, Ib Søndergaard, and Susanne Jacobsen. 2007. Analysis of protein structures and interactions in complex food by near-infrared spectroscopy. 1. Gluten powder. *Journal of agricultural and food chemistry* 55, 18 (2007), 7234–7243.

[12] Susanne Wrang Bruun, Ib Søndergaard, and Susanne Jacobsen. 2007. Analysis of protein structures and interactions in complex food by near-infrared spectroscopy. 2. Hydrated gluten. *Journal of agricultural and food chemistry* 55, 18 (2007), 7244–7251.

[13] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14.

[14] CJ Clark, VA McGlone, and RB Jordan. 2003. Detection of Brownheart in 'Braeburn' apple by transmission NIR spectroscopy. *Postharvest Biology and Technology* 28, 1 (2003), 87–96.

[15] Cynthia L. Corritore, Beverly Kracher, and Susan Wiedenbeck. 2003. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58, 6 (2003), 737 – 758. https://doi.org/10.1016/S1071-5819(03)00041-7 Trust and Technology.

[16] E. Costante, J. den Hartog, and M. Petkovic. 2011. On-line trust perception: What really matters. In *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST)*. IEEE, New York, NY, USA, 52–59.

[17] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*. AIAA, Reston, VA, USA, 6313.

[18] Aritra Dasgupta, Susannah Burrows, Kyungsik Han, and Philip J. Rasch. 2017. Empirical Analysis of the Subjective Impressions and Objective Measures of Domain Scientists' Visual Analytic Judgments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1193–1204. https://doi.org/10.1145/3025453.3025882

[19] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine* 24, 9 (2018), 1342–1350.

[20] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[21] Claudia A Teixeira Dos Santos, Miguel Lopo, Ricardo NMJ Páscoa, and João A Lopes. 2013. A review on the applications of portable near-infrared spectrometers in the agro-food industry. *Applied spectroscopy* 67, 11 (2013), 1215–1233.

[22] A Durand, O Devos, C Ruckebusch, and JP Huvenne. 2007. Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton–viscose textiles. *Analytica Chimica Acta* 595, 1-2 (2007), 72–79.

[23] H John Elgar, Nagin Lallu, and Christopher B Watkins. 1999. Harvest Date and Crop Load Effects on a Carbon Dioxide–related Storage Injury of Braeburn' Apple. *HortScience* 34, 2 (1999), 305–309.

[24] N. Elmqvist and J. Fekete. 2010. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE Transactions on Visualization and Computer Graphics* 16, 3 (2010), 439–454.

[25] Peter Emerson. 2013. The original Borda count and partial voting. *Social Choice and Welfare* 40, 2 (2013), 353–358.

[26] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173718

[27] Carla M. D. S. Freitas, Paulo R. G. Luzzardi, Ricardo A. Cava, Marco Winckler, Marcelo S. Pimenta, and Luciana P. Nedel. 2002. On Evaluating Information Visualization Techniques. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Trento, Italy) *(AVI '02)*. Association for Computing Machinery, New York, NY, USA, 373–374. https://doi.org/10.1145/1556262.1556326

[28] Antihus Hernández Gómez, Yong He, and Annia Garcia Pereira. 2006. Non-destructive measurement of acidity, soluble solids and firmness of Satsuma mandarin using Vis/NIR-spectroscopy techniques. *Journal of food engineering* 77, 2 (2006), 313–319.

[29] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. 2017. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 521–530.

[30] Paul KJ Han, Austin Babrow, Marij A Hillen, Pål Gulbrandsen, Ellen M Smets, and Eirik H Ofstad. 2019. Uncertainty in health care: Towards a more systematic program of research. *Patient education and counseling* 102, 10 (2019), 1756–1766.

[31] Kjell Ivar Hildrum, Bjørg Narum Nilsen, Frank Westad, and Nils Magnus Wahlgren. 2004. In-line analysis of ground beef using a diode array near infrared instrument on a conveyor belt. *Journal of Near Infrared Spectroscopy* 12, 6 (2004), 367–376.

[32] Haibo Huang, Haiyan Yu, Huirong Xu, and Yibin Ying. 2008. Near infrared spectroscopy for on/in-line monitoring of quality in foods and beverages: A review. *Journal of food engineering* 87, 3 (2008), 303–313.

[33] J. Jeffrey Inman, Russell S. Winer, and Rosellina Ferraro. 2009. The Interplay among Category Characteristics, Customer Characteristics, and Customer Activities on in-Store Decision Making. *Journal of Marketing* 73, 5 (2009), 19–29. https://doi.org/10.1509/jmkg.73.5.19 arXiv:https://doi.org/10.1509/jmkg.73.5.19

[34] T Isaksson, BN Nilsen, G Tøgersen, RP Hammond, and KI Hildrum. 1996. On-line, proximate analysis of ground beef directly at a meat grinder outlet. *Meat Science* 43, 3-4 (1996), 245–253.

[35] Tomas Isaksson, Geir Tøgersen, Arve Iversen, and Kjell Ivar Hildrum. 1995. Non-destructive determination of fat, moisture and protein in salmon fillets by use of near-infrared diffuse spectroscopy. *Journal of the Science of Food and Agriculture* 69, 1 (1995), 95–100.

[36] Weiwei Jiang, Gabriele Marini, Niels van Berkel, Zhanna Sarsenbayeva, Chu Luo, Xin He, Tilman Dingler, Yoshihiro Kawahara, and Vassilis Kostakos. 2018. A Mobile Scanner for Probing Liquid Samples in Everyday Settings. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (Singapore, Singapore) *(UbiComp '18)*. Association for Computing Machinery, New York, NY, USA, 1172–1177. https://doi.org/10.1145/3267305.3274764

[37] Weiwei Jiang, Gabriele Marini, Niels van Berkel, Zhanna Sarsenbayeva, Zheyu Tan, Chu Luo, Xin He, Tilman Dingler, Jorge Goncalves, Yoshihiro Kawahara, and Vassilis Kostakos. 2019. Probing Sucrose Contents in Everyday Drinks Using Miniaturized Near-Infrared Spectroscopy Scanners. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 136 (Dec. 2019), 25 pages. https://doi.org/10.1145/3369834

[38] Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *Current Journal of Applied Science and Technology* 7 (2015), 396–403. https://doi.org/10.9734/BJAST/2015/14975

[39] Sumio Kawano, Takayuki Fujiwara, and Mutsuo Iwamoto. 1993. Nondestructive determination of sugar content in satsuma mandarin using near infrared (NIR) transmittance. *Journal of the Japanese Society for Horticultural Science* 62, 2 (1993), 465–470.

[40] Sumio Kawano, Hisayoshi Watanabe, and Mutsuo Iwamoto. 1992. Determination of sugar content in intact peaches by near infrared spectroscopy with fiber optics in interactance mode. *Journal of the Japanese Society for Horticultural Science* 61, 2 (1992), 445–451.

[41] Masataka Kawasaki, Shuso Kawamura, Hiroki Nakatsuji, and Motoyasu Natsuga. 2005. Online real-time monitoring of milk quality during milking by near-infrared spectroscopy. In *2005 ASAE Annual Meeting*. American Society of Agricultural and Biological Engineers, ASAE, Washington, DC, USA, 1.

[42] Kari Kelton, Kenneth R. Fleischmann, and William A. Wallace. 2008. Trust in digital information. *Journal of the American Society for Information Science and Technology* 59, 3 (2008), 363–374. https://doi.org/10.1002/asi.20722 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20722

[43] Jinwoo Kim and Jae Yun Moon. 1998. Designing towards emotional usability in customer interfaces—trustworthiness of cyber-banking system interfaces. *Interacting with Computers* 10, 1 (1998), 1 – 29. https://doi.org/10.1016/S0953-5438(97)00037-4 HCI and Information Retrieval.

[44] Simon Klakegg, Jorge Goncalves, Chu Luo, Aku Visuri, Alexey Popov, Niels van Berkel, Zhanna Sarsenbayeva, Vassilis Kostakos, Simo Hosio, Scott Savage, et al. 2018. Assisted Medication Management in Elderly Care Using Miniaturised Near-Infrared Spectroscopy. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 69.

[45] Simon Klakegg, Jorge Goncalves, Niels van Berkel, Chu Luo, Simo Hosio, and Vassilis Kostakos. 2017. Towards Commoditised Near Infrared Spectroscopy. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. ACM, New York, NY, USA, 515–527. https://doi.org/10.1145/3064663.3064738

[46] Simon Klakegg, Chu Luo, Jorge Goncalves, Simo Hosio, and Vassilis Kostakos. 2016. Instrumenting Smartphones with Portable NIRS. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct* (Heidelberg, Germany) *(UbiComp '16)*. ACM, New York, NY, USA, 618–623. https://doi.org/10.1145/2968219.2971590

[47] Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. 2019. Trust and Recall of Information across Varying Degrees of Title-Visualization Misalignment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article 346, 13 pages. https://doi.org/10.1145/3290605.3300576

[48] JoAnne Labrecque and Line Ricard. 2001. Children's influence on family decision-making: a restaurant study. *Journal of Business Research* 54, 2 (2001), 173 – 176. https://doi.org/10.1016/S0148-2963(99)00088-0 Retail Consumer Decision Processes.

[49] Leslie A Lenert and Daniel J Cher. 1999. Use of meta-analytic results to facilitate shared decision making. *Journal of the American Medical Informatics Association* 6, 5 (1999), 412–419.

[50] Isaac M Lipkus. 2007. Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. *Medical decision making* 27, 5 (2007), 696–713.

[51] Koen Maertens, P Reyns, and Josse De Baerdemaeker. 2004. On-line measurement of grain quality with NIR technology. *Transactions of the ASAE* 47, 4 (2004), 1135.

[52] Hidenori Matsui, Takahiro Hashizume, and Koji Yatani. 2018. Al-Light: An Alcohol-Sensing Smart Ice Cube. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 126 (Sept. 2018), 20 pages. https://doi.org/10.1145/3264936

[53] Eva Mayr, Nicole Hynek, Saminu Salisu, and Florian Windhager. 2019. Trust in information visualization. In *TrustVis workshop. The Eurographics Association, Porto*. The Eurographics Association, Geneve, Switzerland, 1 pages. https://doi.org/10.2312/trvis.20191187

[54] Georg Meyer, Gediminas Adomavicius, Paul E Johnson, Mohamed Elidrisi, William A Rush, JoAnn M Sperl-Hillen, and Patrick J O'Connor. 2014. A machine learning approach to improving dynamic decision making. *Information Systems Research* 25, 2 (2014), 239–263.

[55] Puneet Mishra, Ana Herrero-Langreo, Pilar Barreiro, Jean Michel Roger, Belén Diezma, Nathalie Gorretta, and Lourdes Lleó. 2015. Detection and quantification of peanut traces in wheat flour by near infrared hyperspectral imaging spectroscopy using principal-component analysis. *Journal of Near Infrared Spectroscopy* 23, 1 (2015), 15–22.

[56] H. Navarro-Hellín, J. Martínez del Rincon, R. Domingo-Miguel, F. Soto-Valles, and R. Torres-Sánchez. 2016. A decision support system for managing irrigation in agriculture. *Computers and Electronics in Agriculture* 124 (2016), 121 – 131. https://doi.org/10.1016/j.compag.2016.04.003

[57] Benjamin Niland and Brooks D Cash. 2018. Health benefits and adverse effects of a gluten-free diet in non–celiac disease patients. *Gastroenterology & hepatology* 14, 2 (2018), 82.

[58] Violeta Ocegueda-Miramontes, Mauricio A Sanchez, and Leocundo Aguilar. 2019. Towards Intelligent Systems for Ubiquitous Computing: Tacit Knowledge-Inspired Ubicomp. In *Applied Decision-Making*. Springer, AG, Switzerland, 65–94.

[59] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.

[60] Kazuo Okamura and Seiji Yamada. 2020. Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation. *IEEE Access* 8, 1 (2020), 17 pages. https://doi.org/10.1109/ACCESS.2020.3042556

[61] Alvitta Ottley, Evan M Peck, Lane T Harrison, Daniel Afergan, Caroline Ziemkiewicz, Holly A Taylor, Paul KJ Han, and Remco Chang. 2015. Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 529–538.

[62] Foster Provost and Tom Fawcett. 2013. Data science and its relationship to big data and data-driven decision making. *Big data* 1, 1 (2013), 51–59.

[63] Ahmed Rady, Joel Fischer, Stuart Reeves, Brian Logan, and Nicholas James Watson. 2020. The effect of light intensity, sensor height, and spectral pre-processing methods when using NIR spectroscopy to identify different allergen-containing powdered foods. *Sensors* 20, 1 (2020), 230.

[64] Pei-Luen Patrick Rau, Ye Li, and Jun Liu. 2013. Effects of a social robot's autonomy and group orientation on human decision-making. *Advances in Human-Computer Interaction* 2013 (2013), 14 pages. https://doi.org/10.1155/2013/263721

[65] Jens Riegelsberger, M. Angela Sasse, and John D. McCarthy. 2005. The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies* 62, 3 (2005), 381 – 422. https://doi.org/10.1016/j.ijhcs.2005.01.001

[66] Azzurra Ruggeri, Laurianne Vagharchakian, and Fei Xu. 2018. Icon arrays help younger children's proportional reasoning. *British Journal of Developmental Psychology* 36, 2 (2018), 313–333.

[67] Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 2006. Visualization of Uncertainty in Context Aware Mobile Applications. In *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services* (Helsinki, Finland) *(MobileHCI '06)*. Association for Computing Machinery, New York, NY, USA, 247–250. https://doi.org/10.1145/1152215.1152267

[68] Nicoletta Sinelli, Sara Limbo, Luisa Torri, Valentina Di Egidio, and Ernestina Casiraghi. 2010. Evaluation of freshness decay of minced beef stored in high-oxygen modified atmosphere packaged at different temperatures using NIR and MIR spectroscopy. *Meat science* 86, 3 (2010), 748–752.

[69] Prashant Singh, Ananya Arora, Tor A Strand, Daniel A Leffler, Carlo Catassi, Peter H Green, Ciaran P Kelly, Vineet Ahuja, and Govind K Makharia. 2018. Global prevalence of celiac disease: systematic review and meta-analysis. *Clinical Gastroenterology and Hepatology* 16, 6 (2018), 823–836.

[70] David Sprague and Melanie Tory. 2012. Exploring how and why people use visualizations in casual contexts: Modeling user goals and regulated motivations. *Information Visualization* 11, 2 (2012), 106–123. https://doi.org/10.1177/1473871611433710 arXiv:https://doi.org/10.1177/1473871611433710

[71] KS Technologies. 2016. NIRScanNano Android. https://github.com/kstechnologies/NIRScanNano_Android Accessed: 2020-09-10.

[72] David A Thompson, Paul R Yarnold, Diana R Williams, and Stephen L Adams. 1996. Effects of actual waiting time, perceived waiting time, information delivery, and expressive quality on patient satisfaction in the emergency department. *Annals of emergency medicine* 28, 6 (1996), 657–665.

[73] Tricia Thompson, Rhonda R Kane, and Mary H Hager. 2006. Food allergen labeling and consumer protection act of 2004 in effect. *Journal of the Academy of Nutrition and Dietetics* 106, 11 (2006), 1742–1744.

[74] Lili Tong, Audrey Serna, Sébastien George, and Aurélien Tabard. 2017. Supporting Decision-making Activities in Multi-Surface Learning Environments. In *Proceedings of the 9th International Conference on Computer Supported Education*. HAL, France, 13 pages.

[75] Niels van Berkel, Jorge Goncalves, Lauri Lovén, Denzil Ferreira, Simo Hosio, and Vassilis Kostakos. 2019. Effect of experience sampling schedules on response rate and recall accuracy of objective self-reports. *International Journal of Human-Computer Studies* 125 (2019), 118–128.

[76] Pedro Wences, Alicia Martinez, Hugo Estrada, and Miguel Gonzalez. 2017. Decision-Making Intelligent System for Passenger of Urban Transports. In *International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer, Springer, AG, Switzerland, 128–139.

[77] Claus O Wilke. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media, Sebastopol, CA, USA.

[78] Jerome Workman, Ken E Creasy, Steve Doherty, Leonard Bond, Mel Koch, Alan Ullman, and David J Veltkamp. 2001. Process analytical chemistry. *Analytical Chemistry* 73, 12 (2001), 2705–2718.

[79] Yifei Xue and D. E. Brown. 2003. A decision model for spatial site selection by criminals: a foundation for law enforcement decision support. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 33, 1 (2003), 78–85. https://doi.org/10.1109/TSMCC.2003.809867

[80] Food Standards Australia New Zealand. 2019. *Food Standard Codes - Allergen Labelling*. Food Standards Australia New Zealand. https://www.foodstandards.gov.au/consumer/foodallergies/pages/allergen-labelling.aspx

[81] Hang Zhang and Laurence T Maloney. 2012. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience* 6 (2012), 1.

[82] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[83] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. 2017. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP Conference on Human-Computer Interaction*. Springer, Springer, AG, Switzerland, 23–39.

[84] Jianlong Zhou, Simon Luo, and Fang Chen. 2020. Effects of personality traits on user trust in human–machine collaborations. *Journal on Multimodal User Interfaces* 14 (2020), 387–400.