BiliScreen: Smartphone-Based Scleral Jaundice Monitoring for Liver and Pancreatic Disorders

ALEX MARIAKAKIS, MEGAN A. BANKS, LAUREN PHILLIPI, LEI YU, JAMES TAYLOR, and SHWETAK N. PATEL, University of Washington

Pancreatic cancer has one of the worst survival rates amongst all forms of cancer because its symptoms manifest later into the progression of the disease. One of those symptoms is jaundice, the yellow discoloration of the skin and sclera due to the buildup of bilirubin in the blood. Jaundice is only recognizable to the naked eye in severe stages, but a ubiquitous test using computer vision and machine learning can detect milder forms of jaundice. We propose BiliScreen, a smartphone app that captures pictures of the eye and produces an estimate of a person's bilirubin level, even at levels normally undetectable by the human eye. We test two low-cost accessories that reduce the effects of external lighting: (1) a 3D-printed box that controls the eyes' exposure to light and (2) paper glasses with colored squares for calibration. In a 70-person clinical study, we found that BiliScreen with the box achieves a Pearson correlation coefficient of 0.89 and a mean error of -0.09 ± 2.76 mg/dl in predicting a person's bilirubin level. As a screening tool, BiliScreen identifies cases of concern with a sensitivity of 89.7% and a specificity of 96.8% with the box accessory.

CCS Concepts: • **Human-centered computing → Smartphones**; • **Applied computing → Consumer health**;

Additional Key Words and Phrases: Health sensing; smartphones; jaundice; bilirubin; image processing

ACM Reference format:
Alex Mariakakis, Megan A. Banks, Lauren Phillipi, Lei Yu, James Taylor, and Shwetak N. Patel. 2017. BiliScreen: Smartphone-Based Scleral Jaundice Monitoring for Liver and Pancreatic Disorders. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 20 (May 2017), 26 pages.
https://doi.org/10.1145/3090085

1 INTRODUCTION

Among all forms of cancer, Pancreatic cancer has one of the worst survival rates [2]. Many attribute this statistic to the fact that the symptoms associated with pancreatic cancer often go unnoticed until the cancer is in a later stage; 80-85% of patients present themselves with tumors so advanced that they cannot be removed completely through surgery [5, 34]. One of the earliest symptoms to appear is jaundice, a yellow discoloration of the skin and eyes. In the case of pancreatic cancer, jaundice occurs because a cancerous growth obstructs the common bile duct, causing a buildup of bilirubin in the blood [11]. Being able to detect the very first signs of jaundice when levels of bilirubin are minimally elevated could enable an entirely new screening program for at-risk individuals. Jaundice also manifests as a symptom for a variety of other conditions, such as hepatitis and Gilbert's syndrome, but we are primarily motivated by the link between jaundice and pancreatic cancer for the purpose of this paper.

This work is supported by the National Science Foundation Graduate Research Fellowship Program and the Coulter Foundation. Author's addresses: A. Mariakakis and M. A. Banks and S. N. Patel, Computer Science and Engineering Department, University of Washington; L. Phillipi and L. Yu, University of Washington Medical Center; J. Taylor, Department of Pediatrics, University of Washington.

2474-9567/2017/5-ART20 $15.00
https://doi.org/10.1145/3090085

The clinical gold standard for measuring bilirubin is through a blood draw called a total serum bilirubin (TSB). TSBs are invasive, require access to a healthcare professional, and are inconvenient if done routinely for screening. Medical device manufacturers have investigated non-contact alternatives to a TSB for bilirubin. One such device is the transcutaneous bilirubinometer (TcB). A TcB shines a wavelength of light that is specifically reflected by bilirubin onto the skin and measures the intensity that is reflected back to the device. The computations underlying TcBs are designed for newborns; their results simply do not translate correctly for adults. Part of the reason for this is that normal concentrations of bilirubin are much lower in adults compared to newborns (<1.3 mg/dl vs. <15.0 mg/dl [4]). As it so happens, the sclerae are more sensitive than the skin to changes in bilirubin because their elastin has a high affinity for bilirubin [22]. This presents an opportunity for early, non-invasive screening that has been previously unexplored. Our contribution to this space is BiliScreen, a system that estimates the extent of jaundice in a person's eyes through pictures taken from the smartphone and produces an estimate of their bilirubin level.

To be effective, BiliScreen should be sensitive enough to measure the range of bilirubin levels exhibited by adults. Ruiz et al. [31] found that jaundice is not apparent to the trained naked eye until roughly 3.0 mg/dl; however, bilirubin levels greater than 1.3 mg/dl warrant clinical concern. There exists a detection gap between 1.3 and 3.0 mg/dl that is missed by clinicians unless a TSB is requested, which is rarely done without due cause. We hypothesize that diagnoses can be made much earlier and lead to better outcomes with a system that is precise enough to distinguish between bilirubin levels within and outside of those bounds.

Oftentimes, the trend of a person's bilirubin level over time is far more informative than just a single point measurement. If a person's bilirubin exceeds normal levels for one measurement but then returns to normal levels, it could be attributed to normal variation. If, however, a person's bilirubin shows an upward trend after it exceeds normal levels, it is more likely that a pathologic issue is worsening their condition, such as a cancerous obstruction around the common bile duct. Trends are not only important for diagnosis, but also for determining the effectiveness of treatment. One course of action for those affected by pancreatic cancer is the insertion of a stent in the common bile duct. The stent opens the duct so that compounds like bilirubin can be broken down again; a person's bilirubin level should decrease thereafter. If their bilirubin continues to rise, then there are either issues with the stent or the treatment is ineffective. Trends in bilirubin levels are difficult to capture because repeated blood draws can be uncomfortable and inconvenient for many people, especially those in an outpatient setting. BiliScreen takes advantage of the ubiquity of smartphones, dramatically reducing the effort required to perform these measurements.

BiliScreen uses the smartphone's built-in camera to collect pictures of a person's eyes. The sclera, or white part of the eyes, are extracted from the image using computer vision. Features describing the color of the sclera are then produced and analyzed by a regression model to return a bilirubin estimate. Since different lighting conditions can change the colors of the same scene, we evaluate two accessories that account for the ambient lighting conditions. The first accessory is a head-worn box (Fig. 1, top-left), similar to a head-mounted VR display, that simultaneously blocks out ambient lighting and provides controlled internal lighting through the camera's flash. The second accessory is a pair of paper glasses printed with colored squares that facilitate calibration (Fig. 1, bottom-left). The latter accessory is reminiscent of a previous project called BiliCam [10] by some of the co-authors of this work, which uses a color-calibration card to account for ambient lighting conditions in pictures of newborns that are processed to detect neonatal jaundice. Beyond their intent of assessing bilirubin levels by detecting jaundice through the smartphone camera, the two projects are quite different. BiliCam is intended for newborns, who exhibit a far wider range of normal bilirubin levels than adults. Because the sclera does not have a predefined shape, BiliScreen also requires an additional step of segmentation. Although BiliScreen has tighter precision requirements, it benefits from the fact that the typical sclera is race-agnostic; the same cannot be said for skin, which varies across different ethnicities.

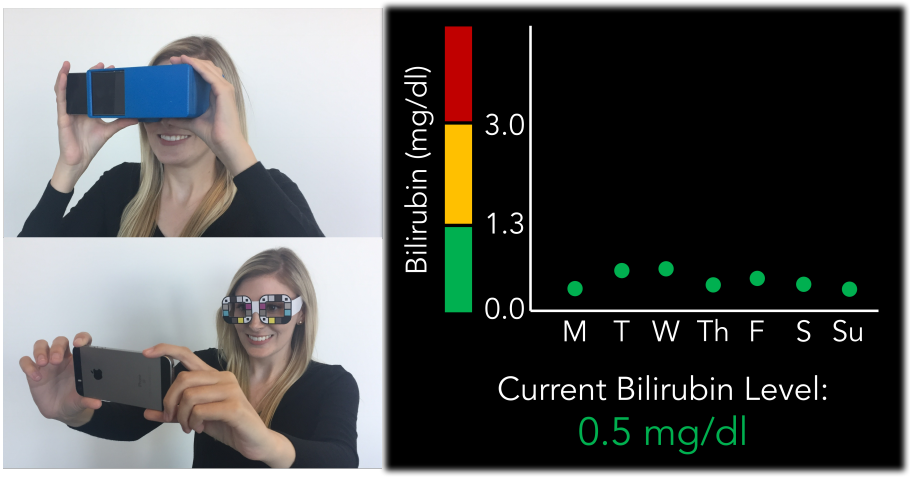

Fig. 1. BiliScreen is a system that measures a person's bilirubin level using the smartphone's camera. We examine two methods for color normalization: **(top-left)** a box similar to a head-mounted VR display that controls the amount of light that reaches the eyes, and **(bottom-left)** paper glasses that provide colored squares for calibration.

We evaluated BiliScreen in a 70-person preliminary study including individuals with normal, borderline, and elevated bilirubin levels. We found that BiliScreen with the box accessory, which leads to better results than the glasses, estimates an individual's bilirubin level with a Pearson correlation coefficient of 0.89 and a mean error of -0.09 ± 2.76 mg/dl when compared to a TSB. BiliScreen with the glasses accessory leads to a Pearson correlation coefficient of 0.78 and a mean error of 0.15 ± 3.55 mg/dl.

Our contribution comes in four parts:

(1) An implementation of the BiliScreen system for convenient bilirubin testing with two different methods for color calibration,
(2) A novel sclera segmentation algorithm that is robust for individuals with jaundice,
(3) Models that relate the color of the sclera to a measure of bilirubin in the blood, and
(4) An evaluation of BiliScreen on 70 participants.

2 RELATED WORK

The BiliScreen algorithm has two fundamental components: automatic segmentation of the sclera and models that map sclera color to bilirubin level. We summarize the literature related to both components below. We then provide a brief overview of other smartphone apps dedicated to diagnosing conditions through the eye.

2.1 Sclera Imaging

To our knowledge, BiliScreen is the first application that automatically segments the sclera for medical purposes. There is, however, a body of literature that has proposed various methods of segmenting the sclera for biometric verification and gaze estimation. For biometrics, individuals are recognized through the uniqueness of the blood vessel patterns in their sclera. For gaze estimation, researchers have relied on the fact that the exposed area of sclera changes as a person makes significant changes in gaze.

The most common method for sclera segmentation relies strictly on color information, noting that the sclera is normally white. Zhou et al. [38] use dynamic thresholds in the RGB and HSV color spaces to create binary

masks that correspond to non-skin- and sclera-colored pixels, respectively. After taking the intersection of those masks, the iris and pupil are removed by using a visible glint within the iris as a seed for an iterative method that moves radially until it reaches the iris-sclera border. Marcon et al. [23] train a linear discriminant analysis classifier on pixel color values to distinguish between sclera and non-sclera pixels. Morphological operations and watershed flooding are applied to form fuller candidate regions for the sclera, after which a classifier trained on shape information is used to select the regions that most resemble the sclera. Das et al. [8] propose a method that involves fuzzy k-means clustering on the pixel color and location to form three clusters: the skin, iris, and sclera. These strictly color-based methods rely on the assumption that the sclera is bright and white, which is not the case for people with jaundice. As the sclera becomes more yellow, its color can be confused with the color of lighter skin tones, making it difficult to train a global classifier. Even if the person's skin tone is known beforehand, there is the chance that its color is too similar to the person's sclera for it to be removed without spatial information.

In a different paper, Das et al. [9] demonstrate a method of sclera segmentation that uses active contour-based segmentation. In active contour-based segmentation, a snake (i.e., deformable spline) is initialized roughly around an object of interest. An energy function is defined based on the presence of lines, edges, and corners in the image, and the position of the snake is iteratively adjusted until that energy function is minimized. For sclera segmentation, Das et al. initialize snakes to the left and right of the automatically detected pupil. This technique is suitable for BiliScreen in that does not depend on the color of the sclera, but the initialization of the snakes can be difficult when the geometry of the eye is not completely constrained. The location of the sclera relative to the pupil depends on both the geometry of the eye and the user's gaze direction. For instance, depending on the narrowness of the eye and how far the user looks up, the sclera may or may not appear directly under the iris. If the initial snakes are too far out from the sclera, they may stop short at glare spots or wrinkles near the eyelids as they constrict. More onus could be placed on the person whose picture is being taken to adjust themselves until their pose satisfies specific constraints, but such a procedure could lead to frustration. Instead of relying on the location of the pupil, eye detection algorithms [20, 35] could be used to standardize a region of interest around the eye; however, such techniques fail when nearby facial features are obstructed, as is the case with the BiliScreen accessories.

One more approach that has been explored for sclera segmentation is the use of dedicated hardware. Crihalmeanu and Ross [7] utilize near-infrared (NIR) lighting to make sclera segmentation straightforward. They observe that the skin has higher NIR reflectance than the sclera since the skin has less water, which makes the separation between the sclera from pale skin more apparent in NIR than in RGB. The use of dedicated hardware in BiliScreen beyond the our box or glasses accessory is undesirable for cost and accessibility purposes.

Overall, these issues motivate the need for a more automated solution. The sclera segmentation approach we propose for BiliScreen uses two iterations of the GrabCut method [30]. The first iteration learns the color characteristics of the skin and removes the skin to isolate the eye. The second iteration isolates the sclerae by assuming that they are the brightest regions within the eyes (not necessarily white).

2.2 Jaundice Assessment

The standard for measuring bilirubin in the blood is through a blood draw called a total serum bilirubin (TSB). The more convenient alternative used in neonatal clinics is a transcutaneous bilirubinometer (TcB). Beyond these two methods, there are several researchers who have investigated bilirubin measurement via the digital photography of areas susceptible to jaundice: the skin and eyes.

Leartveravat [18] proposes a completely manual system for assessing jaundice in a newborn's skin. Photographs of the skin with a color calibration card are captured using a digital camera. Once the photo is uploaded to image editing software (e.g., Adobe Photoshop), the image is color-calibrated and converted to the CYMK color space. A technician then manually selects a pixel representative of the newborn's skin, subtracts its yellow component

from its magenta component, and inputs that value into a linear regression to get a bilirubin estimate. The BiliCam system by de Greef et al. [10] also analyzes pictures of a newborn's skin with a color calibration card to estimate their bilirubin level. It differs from the work of Leartveravat in that BiliCam entails more complicated models that account for skin tone.

Leung et al. [19] compare the performance that a system could achieve by analyzing both the skin and the sclera for newborns. Similar to de Greef et al. and Leartveravat, the authors manually selected regions corresponding to the skin, sclera, and a color calibration card for their analyses. With a fairly modest linear regression model, the authors achieve far better Pearson and Spearman correlations using the sclera (0.75 and 0.72) than using the skin (0.56 and 0.54).

To the best of our knowledge, BiliScreen is the first non-invasive system to quantify an adult's bilirubin level. BiliScreen analyzes the sclera because, as Leung et al. confirmed, the sclera is more sensitive to changes in bilirubin than the skin. This is important because higher precision is needed for adults. Bilirubin levels in healthy newborns may peak as high as 15.0 mg/dl [4], whereas bilirubin levels for healthy adults are normally less than 1.3 mg/dl. BiliScreen is also completely automated, from the segmentation of the glasses and sclera to the feature extraction and machine learning. Finally, BiliScreen benefits from the fact that healthy sclera colors are independent of ethnicity, so less training data should be needed in the long-term.

2.3 Ocular Diagnostic Applications

Pamplona et al. have developed several inexpensive attachments for smartphones to diagnose conditions of the eye. Much like an eye chart, the hardware presents stimuli to the user. Rather than conversing with a clinician, the user interacts with their smartphone depending on what they see; this is an iterative procedure that goes on until a result is reached. In NETRA [27], refractive errors are identified by asking the user to align patterns projected through a microlens display and pinhole. In CATRA [28], cataracts are localized by scanning the eye with a beam of collimated light and asking the user for feedback about the spread of the beam. EyeMITRA [17], being a wearable camera, varies slightly from the other two projects. It is meant for mobile retinal imaging, so it does not perform diagnosis on its own. The user is placed within the loop of the system by being asked to focus on focal points shown in the other eye, which in turn focuses the camera on the opposite side. Others have developed hardware attachments for ocular diagnostics as well. D-Eye[1] is a smartphone adapter for performing fundoscopy. Bastawrous et al. [3] and Giardini et al. [12] propose a number of attachments for diagnosing visual acuity and glaucoma. Finally, Abdolvahabi et al. [1] discuss the possibility for digital photography to catch the early onset of rare eye cancers in newborns; they found that if the common "red-eye" effect in the pupils is replaced with a milky white color, it could indicate tumors in the back of the eye.

3 DATA COLLECTION

We collected images using the BiliScreen app with both the box and glasses accessories to train BiliScreen's models and evaluate their efficacy. Volunteers with normal bilirubin levels were recruited from the University of Washington. Volunteers with varying bilirubin levels (ranging from normal to elevated) were recruited from the University of Washington Medical Center. Below, we elaborate on the diversity of the participant pool. We then describe our data collection procedure, including the design of the BiliScreen accessories and our procedure for ground truth measurements. All facets of our study were approved by the University of Washington's Institutional Review Board.

Table 1. Participant demographics (N = 70)

BILIRUBIN CLASSIFICATIONS - N (mean ± std)	
Normal (<1.3 mg/dl)	31 (0.6 ± 0.2 mg/dl)
Borderline (1.3-3.0 mg/dl)	14 (2.1 ± 0.5 mg/dl)
Elevated (>3.0 mg/dl)	25 (9.7 ± 5.9 mg/dl)

3.1 Enrollment

Our study included 70 volunteers. From the university, 18 were male and 13 were female. From the medical center, 13 were male and 26 were female. Table 1 shows the distribution of the total serum bilirubin tests split across the two different populations. Note that the precision of the TSB is 0.1 mg/dl.

Thresholds classifying the concern warranted by a single bilirubin measurement can vary between clinics. For the purposes of BiliScreen, three classes are defined: normal (<1.3 mg/dl), borderline (1.3-3.0 mg/dl), and elevated (>3.0 mg/dl). The 1.3 mg/dl threshold is used by the University of Washington Medical Center as their upper limit for a normal TSB measurement, while the 3.0 mg/dl threshold is based on the findings of Ruiz et al. [31] concerning when jaundice is most apparent to clinicians. According to these thresholds, 31 participants had a normal bilirubin level, 14 had a borderline bilirubin level, and 25 had an elevated bilirubin level. Unsurprisingly, most of the university population had a normal bilirubin level. The lack of variation within that population was expected. Although the clinical upper threshold for normal bilirubin levels is 1.3 mg/dl, values near 0.6 mg/dl are the norm. The medical center population provided a much wider spread of bilirubin levels, ranging from normal to elevated.

3.2 Data Collection Procedure

The data collection procedure for the BiliScreen app was the same for both populations, but the methods of recruitment and collection of ground truth measurements were different. The participants from the university were volunteers recruited from emails on public mailing lists. After a research staff member collected data with the BiliScreen app (described in the next section), they were asked to undergo a TSB within 24 hours. Bilirubin can change over long periods of time but remains stable within a day barring any serious conditions.

The participants from the medical center were inpatients suffering from liver disease. A research staff member selected candidate participants on two criteria. The first criterion was a recorded TSB blood test within 24 hours. Again, this is to ensure that the patient's recorded bilirubin level matches closely with their level at the time of data collection. The second criterion relies on the Model for End-stage Liver Disease (MELD) [36], a scoring system for assessing the severity of chronic liver disease. The MELD score is a summary metric that combines three measures of a patient's liver condition - TSB, serum creatinine, and the international normalized ratio for prothrombin time (INR) - with a higher score indicating a higher three-month mortality rate. There is no guarantee that a patient with a high MELD score has an elevated bilirubin level since TSB is only one component of the MELD score; however, a high MELD usually includes an elevated TSB. The original recruitment criteria was a minimum MELD score of 14. This threshold was later lowered to 6 in order to recruit more patients with borderline levels (1.3-3.0 mg/dl) of bilirubin. If a patient satisfied the two recruitment criteria, they were approached by a research staff member and told about the study. Patients were enrolled in the study if and only if they understood the study and gave consent.

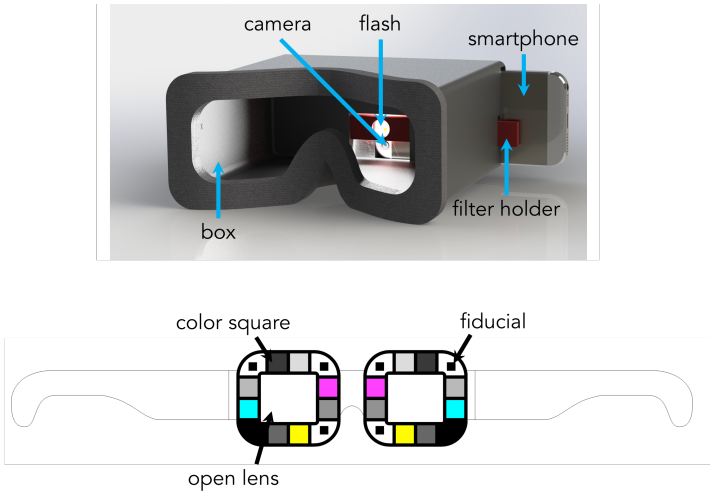

Fig. 2. **(top)** A 3D rendering of the BiliScreen box. The smartphone's flash lies in the horizontal center of the box. The flash is covered with a neutral density filter and a diffuser to make the light more comfortable. **(bottom)** A rendering of the BiliScreen glasses.

3.3 BiliScreen Accessories

Physics-based models for color information typically consider an object's visible color to be the combination of two components: a body reflection component, which describes the object's color, and a surface reflection component, which describes the incident illuminant [15]. When using digital photography, color information that gets stored in image files is also impacted by the camera sensor's response to different wavelengths. For our study, we examine the efficacy of two different accessories to isolate the sclera's body reflection component in different ways (Fig. 2).

The first accessory is a 3D-printed box reminiscent of a Google Cardboard headset[2]. There is no electrical connection between the phone and the box; the phone is simply slid into the box via a rectangular channel along the back. The channel at the back of the box also fixes the placement of the phone relative to the participant's face by centering the phone's camera and keeping it at a fixed distance. The box blocks out ambient lighting while allowing the phone's flash to provide the only illumination onto the eyes. From pilot studies, some participants found the flash to be overwhelmingly bright. A neutral density filter and a diffuser were placed in front of the flash using a filter holder to soften the light slightly. The box used in our study was 3D-printed, but it could be made with an even cheaper material like cardboard (provided that it is sturdy enough to support the weight of the phone). By using the flash as the only illumination source on the sclera, the surface reflection component is kept constant for all images. This leaves the body reflection component and the camera sensor's response as the only two components that affect the sclera's appearance. For the sake of this study, all images were captured using the same device, holding the camera sensor's response constant and leaving the body reflection component as the only variable left.

The second accessory (Fig. 2, bottom) is a custom pair of paper glasses, reminiscent of the 3D glasses found at movie theaters. The glasses have no lenses inside their frames. Along the rims of the glasses are various colored

[1]https://www.d-eyecare.com/
[2]https://vr.google.com/cardboard/

regions. The corners near the temples and the nose have smaller black squares surrounded by the glasses' white background. These squares act as fiducials, similar to those seen in QR codes. The rest of the regions along the rims are the following colors (in no particular order): cyan, magenta, yellow, 17% gray, 33% gray, 50% gray, 67% gray, 83% gray, and black. The use of the colored squares is inspired by color calibration target cards like the Macbeth ColorChecker [29]. Rather than keeping the surface reflection component and the camera sensor's response constant, the colored squares allow for all images to be normalized to the same references. The colors along the rims of the glasses are known *a priori*. This means that their body reflection component is known and any deviation between their appearance and their true color is due to the surface reflection component and the camera sensor's response. Section 4.3 explains the calibration procedure that is used to define a calibration matrix that best simulates the effects of the latter two color information components, which can later be applied to the sclerae themselves to reveal their true body reflection components.

From a usability perspective, the glasses are more convenient for the user and cheaper to manufacture. However, the colors along the rims of the glasses must always be consistent, both across time and different pairs. If the colors were to fade over time, the colors would become a changing reference that could lead to inaccurate results. Although the box is bulkier, its requirements are far looser. The box's main purpose is to block out ambient lighting; control over the precise placement of the smartphone is convenient for aspects of the automatic segmentation, but the box's dimensions do not require as strict precision as the glasses' colors.

From a technical perspective, the color calibration procedure for the glasses can incur its own inaccuracies. In BiliScreen's current state, though, the algorithm for the box accessory does not account for the camera sensor's response. If users were to use a phone with a camera different from that of the iPhone SE, we can make no guarantee that colors will appear the same between the two. Section 6.1 discusses ways for addressing this limitation. Even though the color calibration procedure for the glasses may introduce noise, it allows for any device to be used without issue. The calibration procedure captures the effects of both the surface reflection component and the camera sensor's response.

3.4 BiliScreen Application

All data was collected by a research staff member through a custom app on an iPhone SE. The images collected by the app were at a resolution of 1920×1080. The research staff member ensured that participants complied with the procedure and noted any difficulties that participants had with the app and its accessories.

The BiliScreen app developed for our study was designed to collect data for both accessories in a similar manner. Before the use of either accessory, the smartphone's flash was turned on. When using the box, the flash is necessary since it is the only way to make the eyes visible within it. Keeping the flash constantly on rather than bursting it at the time of the pictures was a consideration for participant comfort since the stark change in lighting can be unpleasant. When using the glasses, the flash was left on in case there was insufficient lighting in the room or the glasses created a shadow on the participant's face.

After the flash was turned on, the research staff member placed the smartphone in the BiliScreen box. A hole in the back of the box provided access to the screen for starting and stopping data collection. The app prompted the participant to look in four different directions - up, left, right, and straight ahead - one at a time while taking a picture after each. Having the participant look in different directions exposed different parts of the sclera, some of which may have exhibited more jaundice than others. The participant was not asked to look downward since doing so covers their eyes with their eyelids. Once the pictures were taken inside the box, the research staff member removed the smartphone and held it approximately 0.5 m away from the participant's face to take pictures with the glasses. This distance is roughly how far away we would expect participants to hold their smartphones if they were taking a selfie. The participant looked at each direction for two trials per accessory, yielding 2 BiliScreen accessories × 2 trials per accessory × 4 gaze directions per trial = 16 images per participant.

4 BILISCREEN ALGORITHM

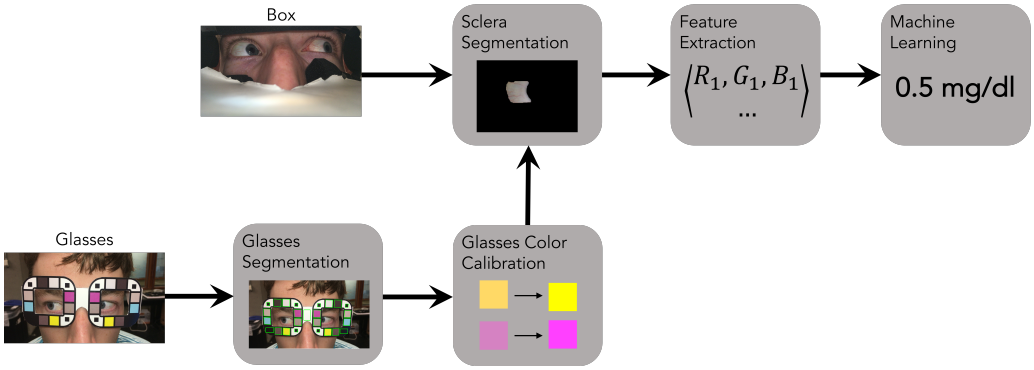

Fig. 3. The algorithm pipeline for both BiliScreen accessories. Images from both the box and the glasses go through the same sclera segmentation, feature extraction, and machine learning steps (with their own respective models and small parameter changes). Images gathered with the glasses must go through the extra steps of glasses segmentation and color calibration.

Fig. 3 outlines the high-level algorithm pipeline that transforms a BiliScreen image to a bilirubin estimate. We will provide further detail in this section on each of these steps, starting with the segmentation of various regions of interest, the transformation of those regions into feature vectors, and finally the machine learning itself.

4.1 Sclera Segmentation

The first step to segmenting the sclera from BiliScreen images is to define regions of interest where the sclera should be located. One way to logically identify these regions would be to use Haar feature-based cascade classifiers [20, 35] that are used in many applications that require eye detection. However, off-the-shelf eye detectors sometimes failed because features around the eyes (e.g., eyebrows) were obstructed by the BiliScreen box and glasses. To maintain consistency across images, regions of interest are defined through other methods depending on the BiliScreen accessory in use. Within the BiliScreen box, the regions of interest are defined as rectangular bounding boxes located on the left and right half side of the box using predetermined pixel offsets within the image. This is possible because the placement of the camera within the box is always the same. The offsets were defined such that the regions of interest would cover various face placements and inter-pupillary distances. For the BiliScreen glasses, the regions of interest are more precisely defined as the regions surrounded by the colored squares (refer to Section 4.2 for how those squares are identified).

Our approach to sclera segmentation relies on an algorithm called GrabCut [30], a technique for separating a foreground object from its background; the terms “foreground” and “background” do not necessarily refer to the perceivable foreground and background of the image, but rather a region of interest versus everything else in the image. GrabCut treats the pixels of an image as nodes in a graph. The nodes are connected by edges that are weighted according to the pixels’ spatial and chromatic similarity. Nodes in the graph are assigned one of four labels: definitely foreground, definitely background, possibly foreground, and possibly background. After initialization, graph cuts [6, 13] are applied to re-assign node labels such that the energy of the graph is minimized. Normally, GrabCut is an interactive technique that is typically initialized with a bounding rectangle and then followed with user-drawn strokes that further clarify the object of interest. BiliScreen uses GrabCut with a similar procedure, but without human intervention.

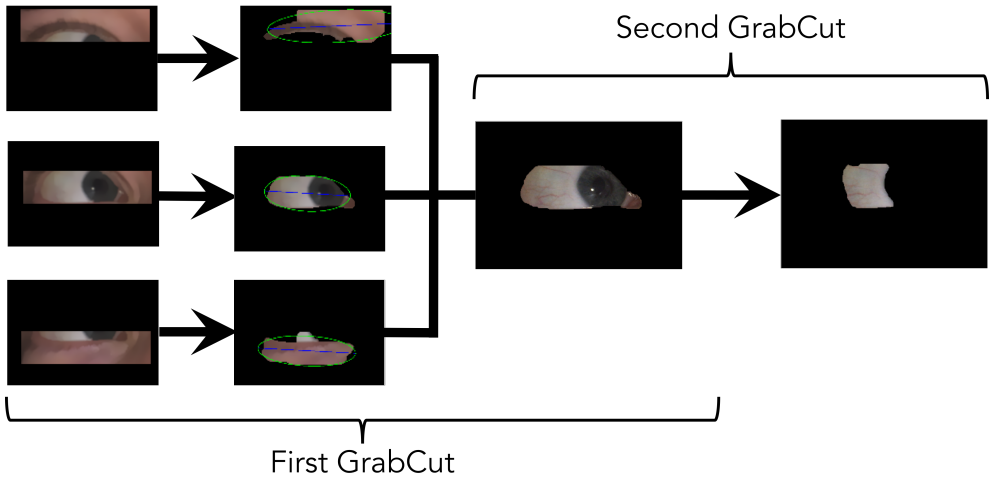

Fig. 4. The procedure for sclera segmentation. The first iteration of GrabCut is initialized with several translated rectangles in parallel. The one that leaves a region that most resembles the eye is used as the region of interest for the second iteration of GrabCut. The second iteration of GrabCut uses adaptive thresholds to select the brightest regions within the eye.

Before segmentation, bilateral filtering is applied to smooth local noise while maintaining strong edges. For the first iteration of segmentation, the eye is extracted using GrabCut with rectangles for initialization (Fig. 4, left). This not only limits the search space for the sclera, but also removes most of the skin around the eye, reducing any effects those pixels could have on color histograms or adaptive thresholds later in the algorithm. The location of the eyes within the image can vary, so rectangular initializations at different locations are tested. To determine which output is most likely to only contain the eye, the segmented regions from each initialization are described using the calculations listed in Table 2. As the second column indicates, some of the metrics are meant to be minimized, while others are meant to be maximized. Those that are meant to be minimized are negated so that higher values always imply that the region is more eye-like. The metrics are combined using the Mahalanobis distance relative to all of the other segmented regions. Overall, this calculation results in high distances for segmented regions that are small, elliptical, flat, and diverse in color, as well as rectangular initializations that likely do not crop out the eye. The segmented region with the highest distance wins out and is passed along to the second part of the sclera segmentation algorithm.

After the first iteration of GrabCut is applied, the pixels that are assigned to the foreground are considered to be part of the eye, regardless of whether they are labeled as "definitely foreground" or "possibly foreground". A second iteration of GrabCut is then used to extract the sclera from the eye (Fig. 4, right). The second iteration of GrabCut normally requires user interaction. In BiliScreen, however, the GrabCut initialization can be bootstrapped automatically using adaptive and pre-defined thresholds. After converting the image to the HSL color space, the four possible pixel assignments are initialized as follows:

- Definitely foreground: Top 90^{th}-percentile of L channel values
- Definitely background: Bottom 50^{th}-percentile of L channel values
- Possibly foreground: Otsu threshold [24] on L channel values
- Possibly background: Inverse Otsu threhsold on L channel values

Table 2. Metrics used to rate a result of GrabCut as an eye

Name	Min/Max	Description
Area fraction	Min	The fraction of the region's area over the total area of the region of interest
Ellipse area fraction	Max	The fraction of the region's area over the area of the ellipse that best fits the region
Incline	Min	The incline of the ellipse that best fits the region
Color variation	Max	The standard deviation of the color across the region
Variation over borders	Min	The standard deviation of the brightness values across the top and bottom borders of the rectangle used to initialize GrabCut

In cases when a pixel satisfies multiple assignments, the strongest assertion is prioritized (i.e., definitely foreground over possibly foreground). These assignments are based on the assumption that the brightest region in the eye should be the sclera. This assumption fails when glare appears within the eye, which is always the case with the BiliScreen box and sometimes the case with the BiliScreen glasses. Glare corresponds to high values in the lightness channel of the HSL image ($L > 230$). Pixels with glare are replaced using inpainting, a reconstruction process that re-evaluates those pixels' values via the interpolation of nearby pixels. Once GrabCut is run for the second time, the pixels that belong to the "definitely foreground" and "possibly foreground" labels are selected. The resulting mask is then cleaned by a morphological close operation to remove any tiny regions.

The distance between the smartphone and the person's face changes depending on which BiliScreen accessory is in use while the picture is being taken. The smartphone is at a fixed distance of 13.5 cm from the person's face when the BiliScreen box is in use and at a variable, farther distance when the BiliScreen glasses are in use. Changes in distance can have a modest effect on the lighting because the flash imparts more light on the eye when it is closer to the face. However, this effect is constant with the BiliScreen box and is canceled out by the color calibration procedure for the BiliScreen glasses. The distance does, however, have a greater effect on the parameters for segmentation. As the distance between the smartphone and the person's face increases, the effective size of the eye in the image shrinks. The size of the rectangle used to initialize the first iteration of GrabCut has fixed dimensions for the BiliScreen box (~600×200 px) and dynamic dimensions according to the size of the frames for the BiliScreen glasses (~90% of width $\times$ 60% of height).

4.2 Glasses Segmentation

The goal of the glasses segmentation is to identify the borders of the colored squares around the rims of the glasses and the white portion at the bridge of the nose so that their colors can be used for calibration. An example of correct segmentation is provided in Fig. 5. The process starts with identifying the fiducials at the corners of the glasses. The fiducials are designed to be square-shaped, but unless they are viewed straight on, they can appear more as quadrilaterals. Black quadrilaterals are found by converting the image to grayscale and filtering it so that only the contours with four corners and a brightness value less than 60 are kept. The small quadrilaterals correspond to the fiducials, while the others correspond to the outlines of the colored squares around the rims. The fiducials are roughly one-fourth the size of the colored squares. Therefore, quadrilaterals that are less than half of the average quadrilateral area are classified as fiducials; the other quadrilaterals are classified as colored squares. To confirm that the fiducials belong to the glasses and not something in the background, the algorithm checks that the pixels immediately outside of their borders are white. If any fiducials are not found because of glare or some other error, their locations are interpolated or extrapolated based on the locations of the discovered

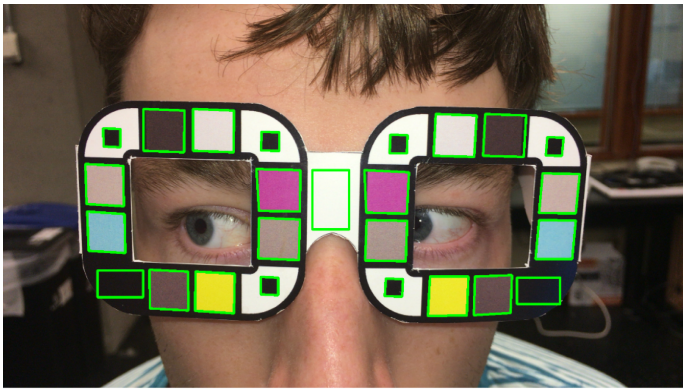

Fig. 5. An example of correct segmentation for the glasses. The region over the bridge of the nose is used as a white reference for both sides.

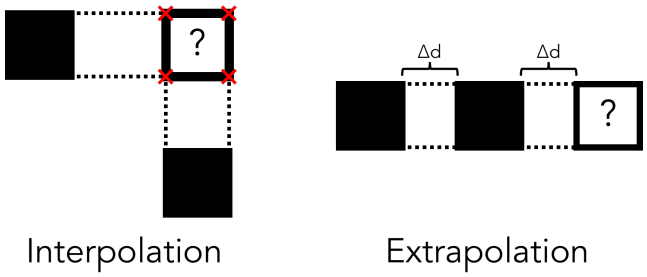

Fig. 6. Illustrations showing how the positions of known fiducials or colored squares can be used to **(left)** interpolate or **(right)** extrapolate the positions of missing ones.

fiducials and the known geometry of the glasses. The left side of Fig. 6 shows an example of interpolation. When there are known fiducials that are along the same vertical and horizontal axes as where the missing fiducial should be, the corners of the missing fiducial can be estimated by using the intersections of those lines. The right side of Fig. 6 shows an example of extrapolation. If there are not enough known fiducials to use interpolation, we rely on the known relative dimensions of the glasses to estimate where they would most likely lie.

The positions of the fiducials are then used to check the positions of the colored squares. The fiducials are connected with straight lines to provide guides on which the other squares should lie. Any quadrilaterals found outside of those bounds are discarded as the background. The fiducials are then used to develop a one-to-one mapping between the names of the colored squares (e.g., left yellow, right 33% gray) and their locations in the image. In the end, there should be two colored squares along each side of the lenses and black patches at the far bottom corners. The locations of the larger black-bordered quadrilaterals are compared to the expected positions of the colored squares. If the distance between a detected quadrilateral and the expected position of a colored square is less than a quarter of the expected square's width, the quadrilateral is matched with the corresponding label. There may not be enough detected black-bordered quadrilaterals to assign a border to every square label. This can be attributed to, among other reasons, glare from the camera or ambient lighting that obscures black outlines. Like the missing fiducials, the missing colored squares can be found using a combination of interpolation

and extrapolation. After the squares around the rims of the glasses are found, the white rectangle that rests on top of the bridge of the nose is selected using a specified offset from the rims to provide a white color reference.

Both interpolation and extrapolation in this algorithm assume that the squares are linearly arranged around the glasses. The glasses were designed to make interpolation and extrapolation straightforward, but there were cases when users had to bend them so that they would fit comfortably on their faces. In these cases, it can be difficult to find fiducials and colored squares when quadrilateral detection has already failed. That being said, the advantage of the BiliScreen glasses design is that there are squares with the same color on each side. It is preferable to detect the squares on the same side as the eye of interest since they better represent the lighting shone on that particular side, but if one of those squares cannot be found, the other side can be used as a contingency.

4.3 Glasses Color Calibration

By identifying the colored squares of the glasses, BiliScreen images can be normalized to a common reference. Doing so removes the effects of the ambient lighting and the camera sensor's response, both of which can change the appearance of the sclera.

The calibration procedure involves identifying the calibration matrix C that best maps the colors of the glasses' squares observed in the image to their actual colors. More formally, define O as the matrix of observed colors and T as the matrix of target colors, where each row contains an RGB vector that corresponds to a colored square. The matrix C defines the linear transform such that:

$$\begin{bmatrix} T_{R1} & T_{G1} & T_{B1} \\ T_{R2} & T_{G2} & T_{B2} \\ \vdots & \vdots & \vdots \\ T_{Rk} & T_{Gk} & T_{Bk} \end{bmatrix} = \begin{bmatrix} O_{R1} & O_{G1} & O_{B1} \\ O_{R2} & O_{G2} & O_{B2} \\ \vdots & \vdots & \vdots \\ O_{Rk} & O_{Gk} & O_{Bk} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \tag{1}$$

Because image files are gamma-encoded to optimize the usage of bits, gamma correction must be applied to the observed colors from the image so that linear operations on them are also linear. This is done by raising the values in O by a constant ($\gamma = 2.2$ for standard RGB image files). After a calibration matrix is applied, the gamma correction can be reversed by raising the values of the matrix to $1/\gamma$.

The calibration matrix C is calculated using an iterative least-squares approach detailed by Wolf [37]. The calibration matrix is first initialized under the assumption that the individual color channels are uncorrelated and only require a gain adjustment that would scale the mean value of the observed channel values to their targets:

$$C = \begin{bmatrix} \text{mean}(T_{Ri})/\text{mean}(O_{Ri}) & 0 & 0 \\ 0 & \text{mean}(T_{Gi})/\text{mean}(O_{Gi}) & 0 \\ 0 & 0 & \text{mean}(T_{Bi})/\text{mean}(O_{Bi}) \end{bmatrix} \tag{2}$$

For each iteration, the current calibration matrix is applied to the observed colors to produce calibrated colors. The colors represented by the rows are converted to the CIELAB color space so that they can be compared to the targets in T using the CIEDE2000 color error [32], the current standard for quantifying color difference. A new calibration matrix C is computed that reduces the sum of squared errors, and the process repeats until convergence.

For BiliScreen, the rows of the target color matrix T are defined as the expected RGB color vectors of the glasses' squares according to their specification. The rows of the observed color matrix O are computed by finding the median vector in the HSL color space of the pixels within the bounds of the squares found in Section 4.2 (excluding the fiducials) and converting the vector back to RGB. For a region R with N 3-dimensional colors, the median vector is defined as:

$$v_m = \underset{v_i \in R}{\operatorname{argmin}} \sum_{j=1}^{N} \|v_i - v_j\|_2 \tag{3}$$

The median vector is preferred over taking the mean or median across the channels independently because it guarantees that the result is a color that exists within the original image; by treating the channels independently, the combination of values in the three channels may not ever appear in the image. The difference between the two approaches is typically insignificant when the region is uniform (as is the case with the colored squares), but is a precaution taken nonetheless.

The calibration procedure is repeated for both eyes using the colored squares closest to them. This is done because the ambient lighting effects are not always uniform; there may be a shadow or beam of light that creates a gradient across the face, making one side look slightly different from the other. There can also be cases when a colored square is washed out by glare from the smartphone's flash. If the error between the colored square and the expected color is 5 units more than the error between the corresponding colored square on the opposite side and the expected color (color difference is unitless), the latter is used. We never encountered a case when squares of the same color on opposite sides of the face were simultaneously obstructed. If that were the case, however, that color could simply be thrown out of the calibration procedure.

4.4 Feature Extraction

Jaundice is characterized by yellow discoloration, so the features extracted from BiliScreen's images should summarize the color of pixels belonging to the sclera. The color of the sclera is described using the median vector over the pixels that belong to the sclera for the same reasons described at the end of Section 4.3. More often than not, the sclera contains other components like vessels or a gradient from the eye's curvature. In these cases, aggregating the color channels independently can lead to a color that is not present in the sclera. For example, if an otherwise pristine sclera contains many blood vessels, taking the mean of the color channels independently will represent the color of the sclera as a pinkish color; the median vector will represent it as white assuming there is more white area than there is red. The median vector is also useful for when sclera segmentation includes superfluous pixels outside of the sclera. Assuming most pixels belong to the sclera, those pixels do not factor in to the final sclera color.

Table 3. Variations for feature extraction

PIXEL SELECTION METHODS	
All pixels	All pixels
No glare	$L \leq 220$ in HSL space
No glare or vessels	$L \leq 220$ and $H \geq 15$ in HSL space
No glare or eyelashes	$5 \leq L \leq 220$ in HSL space
No glare, vessels, or eyelashes	$5 \leq L \leq 220$ and $H \geq 15$ in HSL space
COLOR SPACES	
RGB, HSL, HSV, L*a*b*, YCrCb	

There are two considerations that must be considered for feature extraction (Table 3). The first is which pixels are considered in the calculation. The most obvious answer is to use all the pixels that survived the sclera segmentation presented in Section 4.1. As mentioned earlier, though, not all pixels within the boundaries of the sclera actually represent the color of the sclera. Blood vessels and eyelashes can add undesired complications to the data. The median vector is meant to alleviate their effects, but as an extra precaution, BiliScreen uses the 5

different pixel selection methods described in Table 3. The thresholds for the different methods were selected empirically by examining images with prominent cases of glare, vessels, and eyelashes. They are by no means intended to capture all cases of non-sclera pixels; in fact, they are kept conservative on purpose to ensure that enough pixels remain in the calculation.

The second consideration for feature extraction is which color space is used. Images are saved from the smartphone camera in the RGB color space. Converting the image to a different color space is simply a calculation across the three channels that expresses those numbers in a different way, something that various machine learning models and feature transformation techniques can learn on their own. Nevertheless, explicitly carrying out color conversions can rearrange the color data in such a way that fewer features are needed. BiliScreen computes features for the 5 different color spaces listed in Table 3. Beyond features from the various color spaces, BiliScreen also computes the pairwise-ratios of the three channels in RGB. The intuition behind these features is that a yellower color will have low blue-to-red and blue-to-green ratios.

BiliScreen computes color representations of the sclera using every combination of pixel selection method and color space. Each color has 3 channels, resulting in 5 pixel selection methods × (5 color spaces × 3 channels per color space + 6 RGB ratios) = 105 features per eye. Not all of the features are used in the final model. Some pixel selection methods across the same regions can result in the same pixels, and some channels across color spaces represent the same information in similar manners. Automatic feature selection is used to select the most explanatory features and eliminate redundant ones. The top 5% of the features that explain the data according to the mutual information scoring function are used in the final models. Mutual information measures the dependency between two random variables [16]. We find that the features that best explain the data come from looking at the ratio between the green and blue channels in the RGB color space. A healthy sclera should be white, which produces high values across all three color channels. Blue is the opposite of yellow, so as the blue value of a white color is reduced, it becomes more yellow. This means that a high green-to-blue ratio implies a more jaundiced sclera.

4.5 Machine Learning

Separate models were developed for the two BiliScreen accessories to determine which would yield the better accuracy. The models use random forest regression and are trained through 10-fold cross-validation across participants. Note from Table 1 that the distribution of bilirubin levels is not evenly distributed; the healthy participants recruited from the university generally had similarly low values within 0.1 mg/dl, while the patients from the medical center had a far wider spread. The thresholds used in BiliScreen split the participants such that the normal and elevated classes have roughly equal sizes (31 vs. 25). The borderline class is roughly half as large (14), which is to be expected given that it is hard to catch such cases. To ensure that the training sets are balanced during cross-validation, splits are assigned using stratified sampling across the three bilirubin level classes. To be more specific, the typical fold for our dataset includes 3 participants with normal bilirubin levels, 1 participant with a borderline bilirubin level, and 3 participants with elevated bilirubin levels.

The data collection procedure resulted in 2 trials per accessory × 4 gaze directions per trial = 8 images per accessory. Note that each image contains 2 eyes, leading to 16 eye images per accessory. Each eye is summarized with a feature vector that leads to its own bilirubin level prediction. For the results that are presented in this paper, the estimates from the 8 images are averaged to produce a final bilirubin level estimate that is reported back to the user. In the future, we plan to examine methods for selecting the best subset of images and only using them in the calculations.

5 RESULTS

Our evaluation examines BiliScreen's two major components: the segmentation algorithms and the sclera color-to-bilirubin level regression. We first examine the performance of the glasses' and sclera segmentation. We then

show how accurate BiliScreen *can be*, assuming near-perfect segmentation of the glasses and sclera, as well as how accurate BiliScreen *is* with the current segmentation algorithms. We conclude by framing the accuracy of BiliScreen as a classification problem, showing how likely BiliScreen is to make the correct diagnostic decision.

5.1 Segmentation

All of the images were hand-annotated by the same researcher for ground truth. For the images taken with the BiliScreen box, the sclerae for both eyes were annotated; for the images taken with the BiliScreen glasses, the sclerae of both eyes, the colored squares, and the lenses were annotated. The performance of BiliScreen's segmentation algorithms can be described using precision and recall. The ground truth pixels annotated by the researcher are treated as targets. Precision defines the fraction of selected pixels that were correct, while recall defines the fraction of correct pixels that were selected. A low precision with a high recall would imply that the algorithm selects most of the pixels that belong to the target, but also includes several pixels outside of the target. A high precision with a low recall would imply that the algorithm only selects a fraction of the necessary pixels, but they are mostly within the target.

5.1.1 Glasses Segmentation. Finding the general region of interest for the sclera when the box is in use is trivial; it is based on rough rectangles on either side of the box. For the BiliScreen glasses, however, the region of interest is defined by the region within the glasses' lenses. We found that the glasses segmentation algorithm was able to locate the region of interest for the sclera with a mean precision of 94.0 ± 15.0% and a mean recall of 94.4 ± 15.1% across all images relative to the lens borders defined by the human annotator. Recall is more important than precision for this problem because, as a region of interest, it is okay for superfluous pixels to be included as long as those belonging to the lens are included. The first step of the sclera segmentation algorithm attempts to rule out pixels outside of the eye agnostic of whether they represent skin or something else.

The glasses segmentation algorithm is also important for locating the colored squares around the lenses for color calibration. On average, the algorithm found the squares with a mean precision of 83.5 ± 24.2% and a mean recall of 88.2 ± 24.1% across all images. Unlike the sclera region of interest, precision is more important than recall for the colored squares because superfluous pixels can add noise to the calculation that summarizes the pixel colors to a single color value. Nevertheless, that is the specific reason for why the median vector is used over other aggregation functions. BiliScreen can tolerate mediocre precision as long as most of the pixels belong to the colored squares, which is true even within a standard deviation of our results. BiliScreen also takes advantage of the fact that there is a copy of each colored square on both sides of the face. The expected colors of the squares are known beforehand, so if a square on one side appears significantly different from the other with respect to the expected color, BiliScreen prioritizes the one that is closer to expectations.

Many of the issues that arose for the glasses segmentation can be attributed to their deformability. The glasses were made from a thin cardstock that could bend if the glasses were not large enough to fit on the participant's head. If BiliScreen cannot find a square, the algorithm fits the squares it has found to linear rows and columns and uses their intersection to find the missing one. When the rows and columns are actually curves, lines do not properly infer the squares' locations. Higher-order polynomials could have been used to model the curvature, but most of the squares required extrapolation rather than interpolation. That is to say, the locations of the squares had to be inferred outside of the range of the available squares, so even higher-order functions would not always properly locate squares. In the future, we plan on improving the design of the BiliScreen glasses with a stiffer material and adjustable stems to avoid bending in the future.

5.1.2 Sclera Segmentation. As was the case for the glasses, ground truth for the sclera segmentation came from manual annotations. Pixel perfect labels are impossible by hand because of artifacts like eyelashes and blood vessels that encroach into the region. Nevertheless, those artifacts are handled post-hoc during feature extraction, so neither the ground truth annotations nor the segmentation algorithm are required to handle them.

Table 4. Sclera segmentation results per eye

	Precision	Recall
Box	74.8 ± 34.1%	56.9 ± 28.6%
Glasses	74.8 ± 35.0%	43.1 ± 27.1%

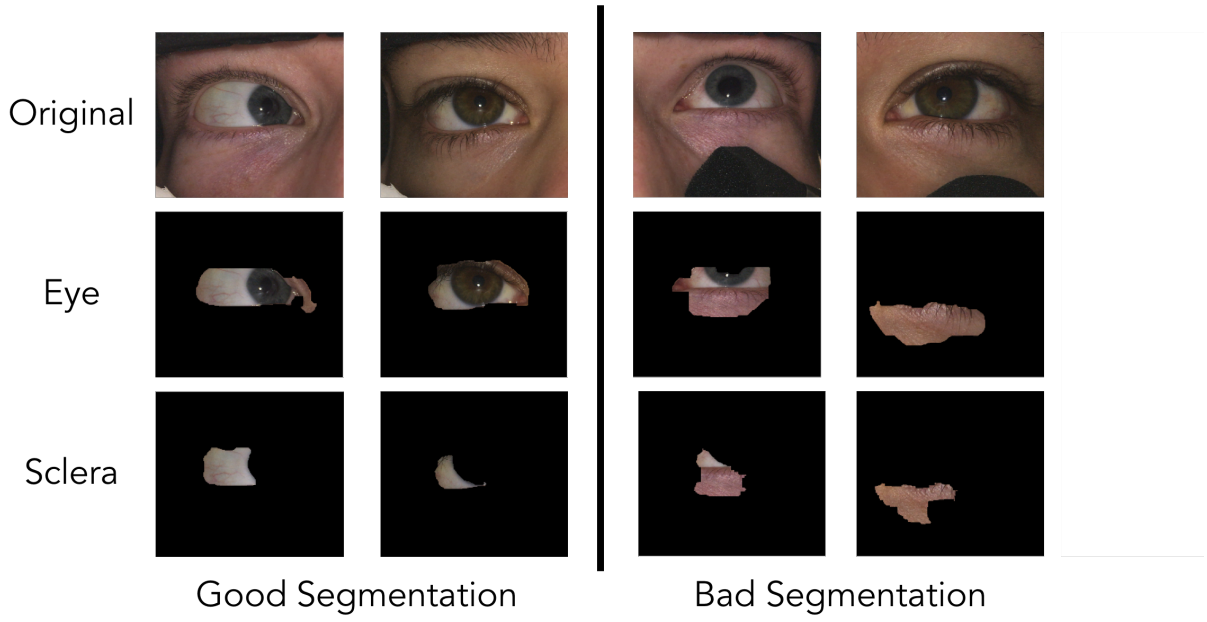

Fig. 7. Example cases of BiliScreen's segmentation working **(left)** correctly and **(right)** incorrectly while the BiliScreen box was in use. These images come from individuals who were not recruited for the study in order to protect the privacy of those participants.

For sclera segmentation, a high precision with a low recall is also preferred over a low precision and a high recall. During the feature extraction phase, the colors of the individual pixels are summarized into single color vectors that describe the entire region. Having a small but correct region is likely to result in a similar calculation outcome, but including pixels outside of the target region can contribute noise to the result. Table 4 shows the per-eye precision and recall for both BiliScreen accessories. The spread of these measures can be misleading since the performance of the algorithm is roughly binary; the segmentation algorithm either identifies a region that corresponds to the sclera and only the sclera, or it completely misses and identifies another region, though it is correct more often than it is not. Looking deeper into the results, we find that 57.1% of the images from the box were segmented with ≥90% precision and 56.5% of the images with the glasses were segmented with ≥90% precision. Failures were not evenly distributed amongst all users.

Fig. 7 shows successful and unsuccessful cases of sclera segmentation. For the most part, failures can be attributed to mistakes in the first half of the sclera segmentation algorithm, which uses GrabCut on the region of interest to locate the eye. In the first example of poor segmentation (third column of Fig. 7), a faint shadow is cast onto the top-right part of the sclera since its curves away from the smartphone's flash. The sclera is assumed to

be the brightest part of the image. Therefore, the algorithm prefers the rectangular initialization that includes the lower half of the sclera, which is bright, and the region just below the eye, where the flash reflects off of the skin and back to the camera. In the second example of poor segmentation (fourth column of Fig. 7), the sclera has a naturally darker tint. Again, the flash produces a reflection under the eye, so the algorithm completely fails to select any part of it.

The dataset includes some users who squinted or blinked during the study. No attempts were made to manually curate images, and there was usually still enough exposed sclera so that a human observer could barely pick out the correct region. Nevertheless, we plan on implementing quality checks in a future version of the BiliScreen app to handle such cases. For the sclera segmentation with the glasses, errors can also be attributed to incorrect regions of interest from the segmentation of the glasses themselves. If BiliScreen could not the properly locate the glasses, then the algorithm makes its best guess, which can hinder later parts of the pipeline. This is another quality check that we believe will be necessary in the next version of the BiliScreen app.

5.2 BiliScreen as a Measurement Tool

Fig. 8 shows the BiliScreen's optimal performance for estimating a person's bilirubin level when the exact boundaries of the sclera are known *a priori*. Of course, this claim assumes that the color-calibration procedure for the glasses and the feature extraction for both accessories properly capture the information needed to properly describe the color of the sclera. Although there are likely aspects of improvement in these regards, we suspect that automatic segmentation is the largest contributor of error since all calculations thereafter are dependent on its results.

The results are presented in two different arrangements. On the left, Fig. 8 shows the correlation of BiliScreen's predictions with the ground truth measurements gathered from TSBs. The points are shown on a log-scale for clarity since the distribution is biased towards lower values. The dotted lines on the correlation plots indicate the 1.3 mg/dl and 3.0 mg/dl thresholds that separate the three groups of measurements. With the optimal segmentation, the Pearson correlation coefficient between BiliScreen's predictions and ground truth are 0.86 with the box and 0.83 with the glasses. On the right, Fig. 8 shows the Bland-Altman plots of the same measurements. Again, the x-axis shows the ground truth measurements using a log-scale for clarity. With the box, BiliScreen estimates the user's bilirubin level with a mean error of -0.17 ± 2.81 mg/dl. With the glasses, BiliScreen estimates the user's bilirubin level with a mean error of -0.08 ± 3.10 mg/dl.

The optimal models in their current state are more accurate for lower levels (<1.3 mg/dl) than they are for higher levels (>3.0 mg/dl). This can be attributed to the underlying distribution of bilirubin measurements for our participants'. Two participants returned a TSB value greater than 20 mg/dl, far beyond the threshold between borderline and elevated values. Since these participants were not thoroughly represented in our dataset, the optimal models underestimates their bilirubin level to fall more in line with the rest of the distribution. In general, higher TSB values lead to larger prediction errors for this very reason. Comparing the box and glasses accessories, the box yields better results. The box eliminates the effects of ambient lighting on the appearance of the sclera. The glasses require the extra step of color calibration, which introduces its own errors into the pipeline.

The results shown in Fig. 9 are presented in the same manner as those in Fig. 8, but were calculated using BiliScreen's automatic segmentation algorithms for the sclera and glasses. We anticipated that BiliScreen's overall performance would degrade with the use of imperfect segmentation. Regarding the sclera segmentation, extra pixels almost always belonged to the skin surrounding the eye. Skin often appears more yellow than the typical white of the sclera, so significant patches of skin can improperly lead to overestimation. The median color vector is used during feature extraction to counteract such behavior, but it is not sufficient for cases when the majority of the extracted region belongs to the skin. The prediction results using BiliScreen's automatic segmentation algorithms confirm our hypothesis, particularly for the glasses. The Pearson correlation coefficient for pictures taken with the glasses drops to 0.78, and the mean error of that model widens to 0.15 ± 3.55 mg/dl. To our surprise,

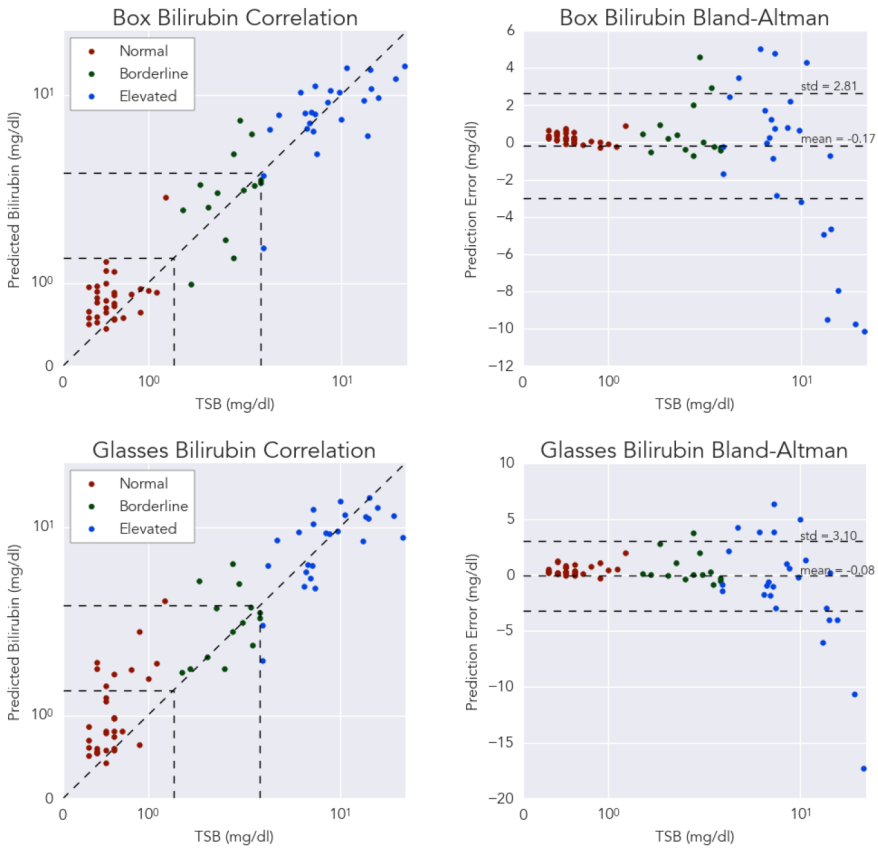

Fig. 8. The **(left)** correlation and **(right)** Bland-Altman plots for BiliScreen's bilirubin measurements with the **(top)** box and **(bottom)** glasses using the optimal sclera and glasses segmentation. Note that the axes of the correlation plot are both in log-scale, as is the x-axis of the Bland-Altman plot.

the Pearson correlation coefficient for the box rises to 0.89, and the mean error improves to -0.09 ± 2.76 mg/dl. A careful comparison of Fig. 8 against Fig. 9 reveals why this is the case. When the optimal segmentation is used to extract features, the model underestimates high TSB values. Because the addition of skin pixels can lead to overestimation, the underestimation is reverted and those predictions are improved. The model still overestimates all users, including those with normal and borderline bilirubin levels, but the improvement on the elevated levels outweighs the smaller errors that are incurred for those lower levels.

The results presented up until this point use all 8 images for each accessory, coming from the 4 gaze directions and the 2 trials. Asking the user to look in different directions provides different views of the sclera, some of which may exhibit more jaundice than others. Although these pictures take less than a minute to collect in total, we recognize that requesting users for 8 images can be burdensome. Using the optimal segmentation results, we find that there is little disadvantage to using the images from a single gaze direction; the Pearson correlation coefficient for the box and glasses accessories varies by no more than 0.05 for any given direction.

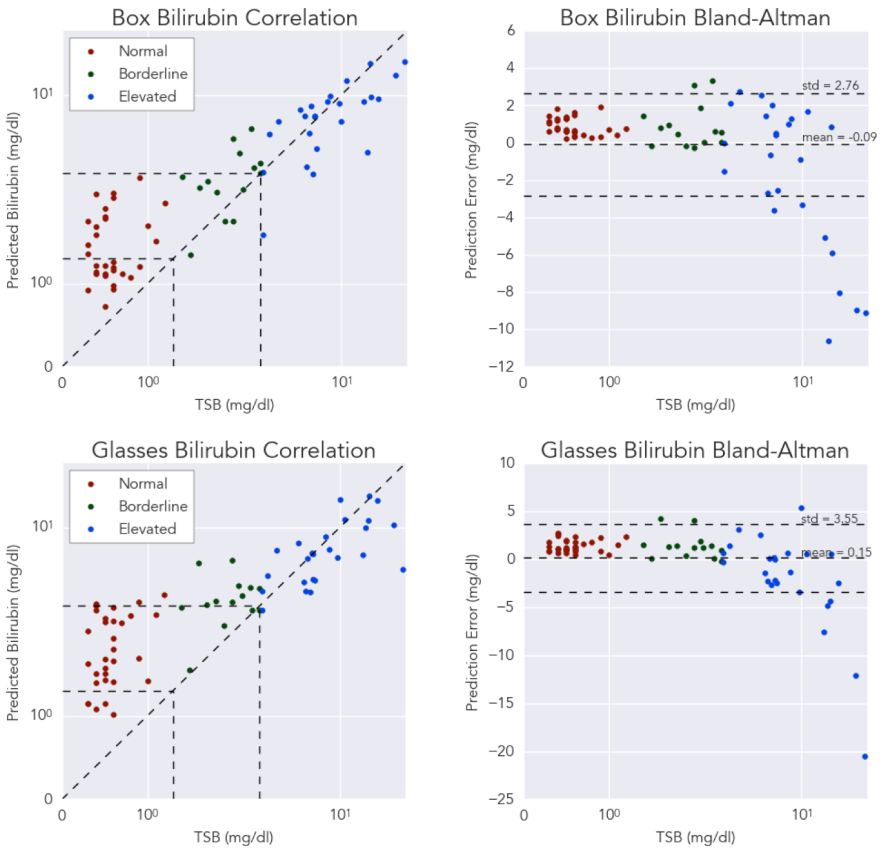

Fig. 9. The **(left)** correlation and **(right)** Bland-Altman plots for BiliScreen's bilirubin measurements with the **(top)** box and **(bottom)** glasses using BiliScreen's sclera and glasses segmentation algorithms. Note that the axes of the correlation plot are both in log-scale, as is the x-axis of the Bland-Altman plot.

Table 5 presents the Pearson correlation coefficient and error for BiliScreen's bilirubin measurements with the box and glasses accessories using the system's segmentation algorithm. Far more variation can be seen using the automatic segmentation, particularly when using the glasses and looking straight ahead. This could be because when the person looks straight ahead, the only parts of the sclera that are exposed are thin regions near the frames of the glasses. These regions are more likely to be covered in a shadow since they curve away from the camera and into the eye socket. The shadow not only affects segmentation, but also the color that is conveyed to the camera. Beyond this behavior, we do not believe there is any significant trend across different gaze directions. Incorporating more images into the final calculation allows BiliScreen to better tolerate an single image with incorrect segmentation. Sometimes, the results improved because incorrectly segmented images were removed from the final calculation. Other times, the results worsened because those same images were the only ones available for final calculation.

Table 5. BiliScreen measurement results across different subsets of images

BOX - Pearson correlation coefficient, mean error ± std error	
All images	0.89, -0.09 ± 2.76 mg/dl
Looking up	0.84, -0.06 ± 3.03 mg/dl
Looking left	0.85, -0.15 ± 2.89 mg/dl
Looking right	0.82, -0.13 ± 3.21 mg/dl
Looking straight ahead	0.87, -0.05 ± 2.78 mg/dl
GLASSES - Pearson correlation coefficient, mean error ± std error	
All images	0.78, 0.15 ± 3.55 mg/dl
Looking up	0.72, 0.06 ± 3.18 mg/dl
Looking left	0.82, -0.06 ± 3.22 mg/dl
Looking right	0.83, -0.31 ± 3.09 mg/dl
Looking straight ahead	0.51, 0.28 ± 4.72 mg/dl

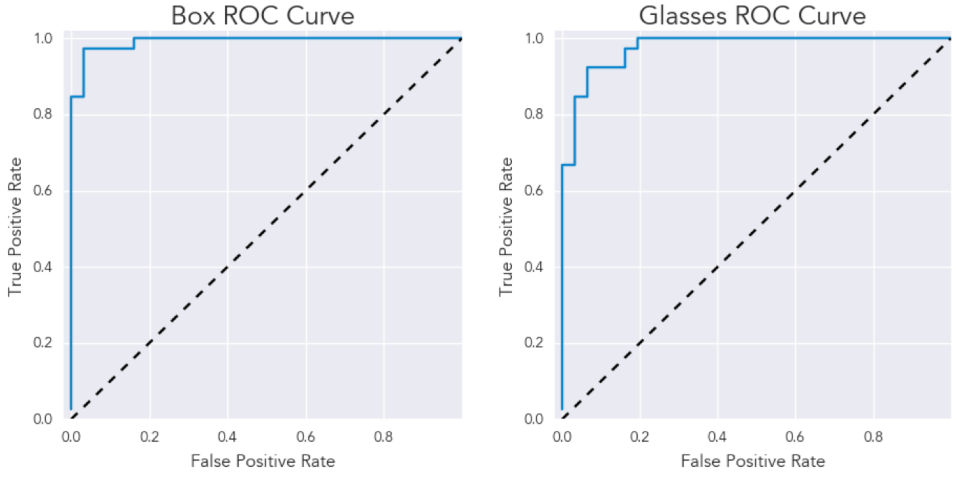

Fig. 10. ROC curves showing BiliScreen's efficacy as a screening tool using the **(left)** box and **(right)** glasses using the optimal sclera and glasses segmentation. For the purposes of this analysis, normal bilirubin levels are considered negative cases, while borderline and elevated levels are considered positive cases.

5.3 BiliScreen as a Classifier

The previous analyses have shown the accuracy with which BiliScreen can estimate a person's bilirubin level. Accuracy is always important, especially for capturing trends in the data. Nevertheless, the average user without a medical background is likely to be more concerned about how their estimated bilirubin level is classified rather than the value itself. In other words, if BiliScreen were to suggest that users with a borderline or elevated bilirubin level refer to a doctor for further tests, they would want assurances about BiliScreen's sensitivity (true positive rate) and specificity (true negative rate). From the perspective of the user, we group borderline and elevated bilirubin levels as positive cases when the user would be referred to a doctor and normal bilirubin levels as negative cases.

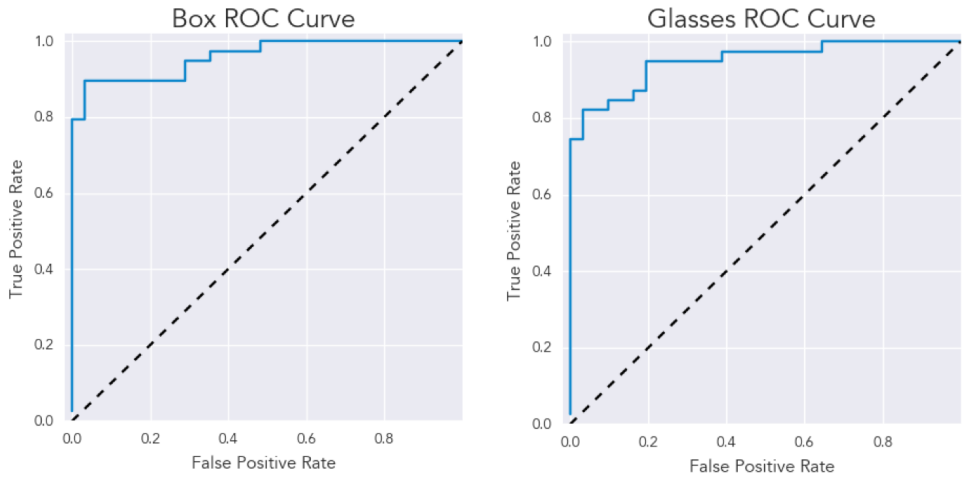

Fig. 11. ROC curves showing BiliScreen's efficacy as a screening tool using the **(left)** box and **(right)** glasses using BiliScreen's sclera and glasses segmentation algorithm. For the purposes of this analysis, normal bilirubin levels are considered negative cases, while borderline and elevated levels are considered positive cases.

Fig. 10 shows the ROC curves for BiliScreen as a classifier using the optimal sclera and glasses segmentation. The area under the ROC curve (AUC) is 0.99 for the box and 0.98 for the glasses. Using the pre-determined threshold of 1.3 mg/dl used by the medical center, BiliScreen with the box achieves a sensitivity of 95.7% and a specificity of 97.4%. The threshold that maximizes the accuracy is only 0.1 mg/dl higher, increasing the sensitivity to 97.4% without a change to the specificity. Since the BiliScreen model with the glasses is more prone to overestimating lower bilirubin levels, it achieves a sensitivity of 100% and a specificity of only 71.4%. The threshold that optimizes accuracy leads to a sensitivity of 92.8% and a specificity of 94.3%.

Fig. 11 shows the same curves for BiliScreen as a classifier using the system's segmentation algorithms. The AUC is 0.96 for the box and 0.95 for the glasses. Again, using the pre-determined threshold of 1.3 mg/dl leads to high sensitivity and low specificity since both models overestimate with the accidental incorporation of skin pixels. Using the optimal thresholds that maximize accuracy, BiliScreen with the box achieves a sensitivity of 89.7% and a specificity of 96.8%. With the glasses, BiliScreen has a sensitivity of 82.1% and a specificity of 96.1%.

6 DISCUSSION

Our goal was to develop a smartphone-based system that would estimate a person's bilirubin level by quantifying the extent of jaundice in their sclera. We believe that BiliScreen is the first step toward these goals. We designed and evaluated two different accessories - a box and glasses - that allowed for the sclera's true color to be measured. In our evaluation, we found that the box accessory performed better than the glasses, achieving a Pearson correlation of 0.89 and a mean error of -0.09 ± 2.76 mg/dl against a blood draw. As a classifier, the BiliScreen app and box accessory achieved a specificity of 89.7% and a sensitivity of 96.8%.

6.1 Hardware

The BiliScreen box was designed to block out ambient lighting, allowing the smartphone's flash to replace an otherwise varying surface reflection component with a constant one. However, the model associated with the box does not account for different camera sensors. All of the data for this study was collected using the same iPhone

SE device. Should another device be used, the BiliScreen model for the box would need to account for the camera sensor's response. This issue could be remedied in one of two ways. First, images could be gathered from the different cameras and separate models could be trained for each of them. This would clearly be a time-consuming endeavor, but lead to results like the ones presented in the paper. An alternative approach would be to perform a one-time calibration procedure as prescribed in Section 4.3 using a color calibration card within the box. The resulting calibration matrix would then be stored and applied on all images taken with the same device. This could be done offline by a researcher with a collection of devices, or the user could be asked to do it before using the BiliScreen app. The colored squares from the BiliScreen glasses could even be integrated into the BiliScreen box so that a separate color calibration card does not need to be purchased.

The latter approach assumes that a calibration matrix can perfectly correct an image's representation of color. Of course, this is the same assumption behind the BiliScreen glasses. If the assumption is not true, then the effects of the surface reflection component and the camera sensor's response cannot be fully eliminated. We believe that this assumption holds well enough that the lingering external effects on the sclera's color are negligible, but have yet to conduct a formal study on the matter.

6.2 Software

Prior work has proposed a variety of techniques for sclera segmentation. However, we found that they were not suitable for people with jaundice either because they assumed strict placement of the eye or that the color of the sclera would always be pale white. We developed a novel sclera segmentation algorithm that circumvents these issues while achieving a mean per-image precision of 74.8 ± 34.1% with the box and 74.8 ± 35.0% with the glasses. The medians are much higher than the means in both cases and the fact that the precision is greater than 50% is sufficient for BiliScreen's feature extraction, but we recognize that there may be other segmentation methods that would achieve even better results. The technique we are most interested in pursuing is fully convolutional neural networks (FCNs) [21]. FCNs take advantage of regular convolutional networks that have been trained to reach astonishingly high accuracy at identifying objects, only instead of the fully-connected layers at the end that produce object labels, FCNs use deconvolutions to achieve a label for every pixel. We believe we can train a similar network for our use case, though this would require far more labeled data than what has been acquired to this point.

As mentioned in various parts of the paper, optimizations can be made throughout BiliScreen's pipeline. We use the mutual information scoring function [16] to automatically select the top 5% of the features that best explain the sclera color. In the future, we plan on manually examining the contributions of the features and determining if certain feature calculations are redundant. The final bilirubin estimate is also based on all 8 images captured per accessory. Taking so many images can be burdensome for the user, but we also believe that getting the different views of the sclera ensures that any regions particularly affected by jaundice are captured. That being said, we have found that using all of the images only provides a small improvement to the final results. We plan on investigating this trade-off further.

6.3 Future Applications

BiliScreen does not directly assess a person's risk of pancreatic cancer; it examines the sclera for jaundice, one of pancreatic cancer's symptoms. Jaundice appears in other conditions, such as hepatitis and Gilbert's syndrome. Examining if there are differences between the visible symptoms of these diseases warrants further investigation.

The deployed implementation of BiliScreen depends on the target demographic for whom the app is designed. If BiliScreen were to be deployed as a screening application, we would prioritize notifying users about the possible risk of pancreatic cancer, even at the cost of extra false positives. This would be implemented by lowering the decision threshold for classifying a user's bilirubin level to increase sensitivity and decrease specificity; for example, lowering the decision threshold for BiliScreen with the box accessory improves its sensitivity from 89.7%

to 95.2% while degrading the specificity from 96.8% to 71.2% (Fig. 11, left). The downside to this change is that BiliScreen could induce a great deal of stress by falsely informing users that they may have a condition as serious as pancreatic cancer. To combat this issue, BiliScreen could require multiple consistent, high measurements before prompting the user to consult a clinician. If BiliScreen were to be deployed as a disease management tool, the trend of the data would be most important to clinicians.

Looking beyond jaundice, quantitative visual examination of the sclera can yield other fruitful observations. Osteogenesis imperfecta, a genetic disorder that results in brittle bones, produces a blue tinge in the sclera [33]. Diabetes results in fewer capillaries, dilated macrovessels, and changes in the curvature in the covering of the sclera [25, 26]. Hyperemia and conjunctivis can affect both the amount and contrast of the blood vessels on the scleral surface [14]. BiliScreen's sclera segmentation algorithm could be used as a starting point for a system that searches for these symptoms and others.

7 CONCLUSION

Ubiquitous bilirubin assessment could have a significant impact on the current state of pancreatic cancer diagnosis. To this end, we have presented BiliScreen, a smartphone app that analyzes digital photographs of the eyes for the degree of jaundice that presents in the sclera. We tested two different accessories to be used in conjunction with BiliScreen to allow the sclera's true color to be measured: a 3D-printed box that blocks out external lighting and a pair of glasses that provide references for color calibration. We evaluated BiliScreen in a study with 70 individuals with varying bilirubin levels and found that the box accessory led to better results. Using the box accessory led to a Pearson correlation of 0.89 and a mean error of -0.09 ± 2.76 mg/dl against a blood draw. As a classifier, BiliScreen with the box was able to screen participants for further consultation with 89.7% specificity and 96.8%. sensitivity. It is our hope to continue building on our initial prototype of BiliScreen with improved models, a more convenient and automated smartphone app, and a longer-term study that captures trends of bilirubin levels.

8 ACKNOWLEDGMENTS

We thank the National Science Foundation and the Coulter Foundation for their funding. We thank Ellie Roberts for helping collect data during the early stages of the project. We thank Mike Clarke for helping design the BiliScreen glasses. Finally, we thank Elliot Saba, Ruth Ravichandran, Mohit Jain, and Edward Wang for providing their feedback on the work.

REFERENCES

[1] Alireza Abdolvahabi, Brandon W. Taylor, Rebecca L. Holden, Elizabeth V Shaw, Alex Kentsis, Carlos Rodriguez-Galindo, Shizuo Mukai, and Bryan F. Shaw. Colorimetric and longitudinal analysis of leukocoria in recreational photographs of children with retinoblastoma. PloS one, 8(10):e76677, oct 2013.

[2] American Cancer Society. Cancer Facts & Figures 2016. Technical report, American Cancer Society, Atlanta, GA, 2016.

[3] Andrew Bastawrous, Hillary K. Rono, Iain A. T. Livingstone, Helen A. Weiss, Stewart Jordan, Hannah Kuper, and Matthew J. Burton. Development and Validation of a Smartphone-Based Visual Acuity Test (Peek Acuity) for Clinical Practice and Community-Based Fieldwork. JAMA Ophthalmology, 133(8):930, aug 2015.

[4] Vinod K Bhutani, Lois Johnson, and Emidio M Sivieri. Predictive ability of a predischarge hour-specific serum bilirubin for subsequent significant hyperbilirubinemia in healthy term and near-term newborns. Pediatrics, 103(1):6–14, 1999.

[5] Giles Bond-Smith, Neal Banga, Toby M Hammond, and Charles J Imber. Pancreatic adenocarcinoma. BMJ, 344, 2012.

[6] Yuri Y Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In Proc. ICCV '01, volume 1, pages 105–112. IEEE, 2001.

[7] Simona Crihalmeanu and Arun Ross. Multispectral scleral patterns for ocular biometric recognition. Pattern Recognition Letters, 33(14):1860–1869, 2012.

[8] Abhijit Das, Umapada Pal, Miguel Angel Ferrer Ballester, and Michael Blumenstein. A new efficient and adaptive sclera recognition system. In 2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM), pages 1–8. IEEE,

dec 2014.

[9] Abhijit Das, Umapada Pal, Miguel Angel Ferrer Ballester, and Michael Blumenstein. Sclera recognition using dense-SIFT. In International Conference on Intelligent Systems Design and Applications, ISDA, pages 74–79. IEEE, dec 2014.

[10] Lilian de Greef, Mayan Goel, Min Joon Seo, Eric C Larson, James W Stout, James A Taylor, and Shwetak N Patel. Bilicam: using mobile phones to monitor newborn jaundice. In Proc. UbiComp '14, pages 331–342, 2014.

[11] Maria Syl D De La Cruz, Alisa P Young, and Mack T Ruffin. Diagnosis and management of pancreatic cancer. American family physician, 89(8):626–32, apr 2014.

[12] Mario E Giardini, Iain A T Livingstone, Stewart Jordan, Nigel M Bolster, Tunde Peto, Matthew Burton, and Andrew Bastawrous. A smartphone based ophthalmoscope. Proc. EMBC '14, 2014:2177–2180, 2014.

[13] Dorothy M Greig, Bruce T Porteous, and Allan H Seheult. Exact maximum a posteriori estimation for binary images. Series of the Royal Statistical Society, 51(2):271–279, 1989.

[14] G. Heath. The episclera, sclera and conjunctiva. An overview of relevant ocular anatomy. Differential Diagnosis of Ocular Disease, 9(2):36–42, 2006.

[15] Gudrun J Klinker, Steven A Shafer, and Takeo Kanade. A physical approach to color image understanding. International Journal of Computer Vision, 4(1):7–38, 1990.

[16] LF Kozachenko and NN Leonenko. Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii, 23(2):9–16, 1987.

[17] Everett Lawson, Jason Boggess, Siddharth Khullar, Alex Olwal, Gordon Wetzstein, and Ramesh Raskar. Computational retinal imaging via binocular coupling and indirect illumination. In Proc. SIGGRAPH '12, page 51, 2012.

[18] Somsak Leartveravat. Transcutaneous bilirubin measurement in full term neonate by digital camera. Medical Journal of Srisaket Surin Buriram Hospitals, 24(1):105–118, 2009.

[19] Terence S Leung, Karan Kapur, Ashley Guilliam, Jade Okell, Bee Lim, Lindsay W MacDonald, and Judith Meek. Screening neonatal jaundice based on the sclera color of the eye using digital photography. Biomedical optics express, 6(11):4529–4538, 2015.

[20] Rainer Lienhart and Jochen Maydt. An extended set of Haar-like features for rapid object detection. In Proceedings. International Conference on Image Processing, volume 1, pages 0–3, 2002.

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 07-12-June, pages 3431–3440, mar 2015.

[22] Dan L Longo, Anthony S Fauci, Dennis L Kasper, Stephen L Hauser, J Larry Jameson, and Joseph Loscalzo. Harrison's Principles of Internal Medicine. 18th edition, 2006.

[23] Marco Marcon, Eliana Frigerio, and Stefano Tubaro. Sclera segmentation for gaze estimation and iris localization in unconstrained images. In CompIMAGE, pages 25–29, 2012.

[24] Nobuyuki Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1):62–66, jan 1979.

[25] Christopher G Owen, Richard SB Newsom, Alicja R Rudnicka, Sarah A Barman, and E Goeffrey Woodward. Diabetes and the tortuosity of vessels of the bulbar conjunctiva. Ophthalmology, 115(6):e27–e32, 2008.

[26] Christopher G Owen, Richard SB Newsom, Alicja R Rudnicka, and Tim J Ellis. Vascular response of the bulbar conjunctiva to diabetes and elevated blood pressure. Ophthalmology, 112(10):1801–1808, 2005.

[27] Vitor F Pamplona, Ankit Mohan, Manuel M Oliveira, and Ramesh Raskar. NETRA: interactive display for estimating refractive errors and focal range. ACM transactions on graphics (TOG), 29(4):77, 2010.

[28] Vitor F Pamplona, Erick B Passos, Jan Zizka, Manuel M Oliveira, Everett Lawson, Esteban Clua, and Ramesh Raskar. Catra: cataract probe with a lightfield display and a snap-on eyepiece for mobile phones. In Proc. SIGGRAPH '11, pages 7–11, 2011.

[29] Danny Pascale. RGB coordinates of the Macbeth ColorChecker. Technical report, 2006.

[30] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (TOG) '04, 23(3):309–314, 2004.

[31] Mario A Ruiz, Sammy Saab, and Leland S Rickman. The clinical detection of scleral icterus: observations of multiple examiners. Military medicine, 162(8):560–563, 1997.

[32] Gaurav Sharma, Wencheng Wu, and Edul N. Dalal. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research and Application, 30(1):21–30, feb 2005.

[33] DO Sillence, Alison Senn, and DM Danks. Genetic heterogeneity in osteogenesis imperfecta. Journal of medical genetics, 16(2):101–116, 1979.

[34] Audrey Vincent, Joseph Herman, Rich Schulick, Ralph H Hruban, and Michael Goggins. Pancreatic cancer. The Lancet, 378(9791):607–620, aug 2011.

[35] Paul Viola and Michael J Jones. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition (CVPR), volume 1, pages 511–518, 2001.

[36] Russell Wiesner, Erick Edwards, Richard Freeman, Ann Harper, Ray Kim, Patrick Kamath, Walter Kremers, John Lake, Todd Howard, Robert M. Merion, Robert A. Wolfe, Ruud Krom, and United Network for Organ Sharing Liver Disease Severity Score Committee. Model for end-stage liver disease (MELD) and allocation of donor livers. Gastroenterology, 124(1):91–96, jan 2003.
[37] Stephen Wolf. Color correction matrix for digital still and video imaging systems. pages 1–40, 2003.
[38] Zhi Zhou, Eliza Yingzi Du, N. Luke Thomas, and Edward J. Delp. A New Human Identification Method: Sclera Recognition. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 42(3):571–583, may 2012.

Received February 2017; accepted April 2017