IU ST
Iran University of Science and Technology

Iranian Journal of Electrical and Electronic Engineering

Journal Homepage: ijeee.iust.ac.ir

# BiliBin: An Intelligent Mobile Phone-based Platform to Monitor Newborn Jaundice

Eisa Zarepour*(C.A), Mohammad Reza Mohammadi*, Morteza Zakeri-Nasrabadi*, Sara Aein*, Razieh Sangsari**, Leila Taheri***, Mojtaba Akbari**, Ali Zabihallahpour*

**Abstract:** Using mobile phones for medical applications are proliferating due to high-quality embedded sensors. Jaundice, a yellow discoloration of the skin caused by excess bilirubin, is a prevalent physiological problem in newborns. While moderate amounts of bilirubin are safe in healthy newborns, extreme levels are fatal and cause devastating and irreversible brain damage. Accurate tests to measure jaundice require a blood draw or dedicated clinical devices facing difficulty where clinical technology is unavailable. This paper presents a smartphone-based screening tool to detect neonatal hyperbilirubinemia caused by the high bilirubin production rate. A machine learning regression model is trained on a pretty large dataset of images, including 446 samples, taken from newborns' sternum skin in four medical centers in Iran. The learned model is then used to estimate the level of bilirubin. Experimental results show a mean absolute error of 1.807 mg/dl and a correlation of 0.701 between predicted bilirubin by the proposed method and the TSB values as ground truth.

**Keywords:** Health Sensing, Image Processing, Internet of Things, Machine Learning, Neonatal Jaundice.

## 1 Introduction

SMARTPHONES have been widely used as assistive tools in various human daily life applications such as navigation, shopping, and homecare. Using a mobile phone for medical applications has also been rapidly growing due to the range of high-quality sensors embedded in the new generation of smartphones.

Jaundice is one of the most common causes of the need for medical care in newborns and is the most common reason for neonatal hospitalization [1]. Nearly all newborns experience serum bilirubin levels above 1 mg/dL, which is the highest acceptable level for adults. About 84% of them experience apparent jaundice in the first week of their life [2].

Neonatal jaundice occurs due to the high rate of bilirubin production (because of the high turnover of red blood cells), the immaturity of bilirubin excretory pathways, and the high rate of the enterohepatic cycle. Bilirubin is a toxic substance for nerve cells. High serum levels, significantly above 25 mg/dL, may damage nerve cells, leading to irreversible neurological complications such as hearing loss, cerebral palsy, and cognitive impairment if not diagnosed and treated in time [3]. These phenomena highlight the importance of early diagnosis and treatment of neonatal hyperbilirubinemia in infants.

For the time being, the medical gold standard for measuring neonatal jaundice is total serum bilirubin (TSB) [4]. Despite its high accuracy, it is an invasive method that requires blood sampling from the newborn and is also not accessible everywhere, including at most outpatient clinics, doctors' offices, and at home. Therefore, TSB often requires visits to health centers. These issues, along with the fact that serum bilirubin peaks usually occur after newborns are discharged from the hospital, have made the TSB an ineffective screening

Iranian Journal of Electrical & Electronic Engineering, 2024.
Paper first received 21 July 2024 and accepted 31 July 2024.
* The authors are with the School of Computer Engineering, Iran University of Science and Technology (IUST), Tehran, Iran.
E-mail: zarepour@iust.ac.ir.
** The authors are with the Children Medical Center, Tehran University of Medical Sciences, Tehran, Iran.
*** The author is with the Department of Pediatric Nursing, Faculty of nursing and midwifery, Qom university of medical sciences, Qom, Iran.
Corresponding Author: Eisa Zarepour.

neonatal jaundice approach. An alternative method for screening neonatal jaundice is to use a transcutaneous bilirubinometer (TcB) [5]. This device estimates the amount of serum bilirubin by analyzing the reflected wavelength received from the skin's surface [6]. The medical community confirms that TcB values above 14.5 mg/dl of bilirubin are unreliable [7]. They are also expensive medical devices requiring frequent calibration, so their usage is limited [7].

The most common method of screening neonatal jaundice by parents and therapists is the amount of skin jaundice that appears through visual assessment [8]. Studies reveal that although parents and therapists can diagnose neonatal jaundice by observing the skin, they cannot detect the extent and severity of jaundice. The correlation coefficient of this method with TSB is between 0.36 to 0.75 [7]. Even the most experienced medical staff underestimate bilirubin levels in situations where the amount of bilirubin is high [9]. The importance of early detection of neonatal jaundice, on the one hand, and the presence of such defects in existing screening methods, on the other, demonstrates the need for creating an efficient tool. Such a tool should screen neonatal jaundice with high accuracy, low costs, and high accessibility (even at home and doctors' offices).

One solution is to use a smartphone to take images of the body's skin and then analyze it to estimate the extent and severity of jaundice. Greef et al. have proposed an application called BiliCam, based on the smartphone digital camera, to measure the bilirubin level and screen jaundice [10], [11]. They have estimated the serum bilirubin level of newborns with a maximum rank-order correlation of 85% by applying existing image processing and machine learning algorithms. BiliCam employs an ensemble of the five different regression models which only uses a set of 21 features obtained from primary images. Training and evaluating are performed on a dataset of 100 samples, which seems pretty small for learning. BiliCam only works on iPhone 4S and suffers from many parameters that should be configured manually before use.

We hypothesize that we could increase the model performance in predicting bilirubin by adding more samples and improving feature engineering. This paper proposes a tool, BiliBin, to estimate the newborns' bilirubin serum level from their sternum skin images. BiliBin uses image processing with machine learning regression. Creating an accurate regression model requires a large and high-quality image with known bilirubin levels. To overcome this challenge, we gathered a large dataset of sternum skin images from newborns in health centers where their serum bilirubin level was determined by total serum bilirubin (TSB). The TSB values were used as labels for training samples. Our dataset is larger than the dataset in [10] regarding the size and variation of the samples. BiliBin uses different smartphones to generalize the results and mitigate previous limitations. In summary, the following contributions have been provided by this paper:

1. We design and implement BiliBin, a smartphone application for determining neonatal jaundice with machine learning techniques. Both newborns' parents and clinicians can use our proposed software, BiliBin, on various Android-based mobile phones available in the market, which significantly increases the reusability of BiliBin compared to similar applications.
2. We propose a new feature engineering scheme by extracting an extended set of features from two types of newborns' skin photos, one photo with flash and another without flash. Our experiments confirm that combining the statistical features of these two images for each sample alleviates learning difficulties caused by data limitations in this area and leads to better prediction performance.
3. We introduce a large dataset of newborns' skin images, including 446 samples taken from four medical centers in Iran, with known bilirubin levels for learning the relationship between skin color and bilirubin level with machine learning techniques. Our learned model on the proposed dataset predicts the amount of neonatal jaundice from skin images with a mean absolute error of 1.807 and a correlation of 0.701.

The remainder of this paper is organized as follows. In section 2, smartphone-based medical applications and jaundice detection are discussed. Section 3 describes the data collection process, BiliBin's design, feature engineering, and learning approach. Section 4 deals with descriptive statistics of our collected dataset and evaluation of BiliBin performance. The conclusion and future work will be discussed in Section 5.

## 2 Related Work

Modern smartphones contain high-quality sensors enabling them to measure numerous medical cases. In recent years, they have been used for monitoring the eye, skin, heart, mental health, and human activity [12]. We focus on the applications of smartphone cameras in healthcare or vision-based health sensing. Engineering advancements in designing high-resolution cameras for

mobile phones have made them suitable for medical applications that require body images from the patient's body, especially those related to skin color and appearance, such as skin cancer and neonatal jaundice.

The first attempts at using the smartphone camera for retinal fundus imaging were proposed by Lord et al. [13], where the image of the retinal fundus was taken with a smartphone camera. However, their system was not user-friendly and could not ensure image quality. A smartphone-based system, DERMA/Care, has been proposed by Karargyris et al. [14] to support melanoma screening. Existing portable solutions for skin disease detection, such as skin cancer, mainly rely on conventional image processing techniques with RGB color imaging, feature extraction, and finally, training machine learning models [15], [16]. Smartphone cameras can estimate heart rate and heart rate variation from the photoplethysmogram signal derived from the video of the bare skin, such as the fingertip [17]. However, near-infrared (NIR) red light sources are used in most commercial cardiovascular systems [18]. Lee et al. have used a smartwatch-based mobile system to monitor the user's scratching behaviors in severe itching conditions such as eczema or atopic [19].

In the literature, there are few studies about jaundice estimation for newborns with completely known materials and data. One of the early works is transcutaneous bilirubin measurement in the full-term neonate, which simply uses the difference between the image's yellow color and a control paper in CIELAB color space [20]. A recent study also uses Pearson's correlation coefficient (r) to evaluate the relationship between each channel intensity in RGB color space and TSB levels [21]. More advanced methods used machine learning algorithms. BiliCam [10], [11] is a smartphone application to estimate neonatal jaundice based on the image which is taken from the newborn's sternum and forehead. Machine learning techniques are used to create a model that can predict neonatal jaundice by learning from numerous image samples with known serum bilirubin levels previously collected in different medical centers. The current version of the application can only work on the iPhone. The authors argue that they have chosen the iPhone because it has the most standardized hardware of the available smartphone platforms. However, it limits the use of BiliCam for a large portion of people who do not access Apple smartphones. Hence, designing a phone-agnostic application is required to overcome this issue. It can be achieved by increasing the number of training samples, improving the feature extraction, and selecting the most suitable learning algorithms.

A neonatal jaundice detection system has been proposed by Aydın et al. uses advanced image segmentation with support-vector regression (SVR) [22] and k nearest neighbors (kNN) [23] machine learning regressions on a dataset consisting of 40 healthy and 40 jaundiced babies' images taken from baby's abdomen [24]. They have not reported any smartphone application for their proposed method. neoSCB [25] uses a color measurement of the sclera to make a screening decision based on 87 samples from a clinical setting in Ghana. Outlaw et al. [26] have used the sclera to screen jaundice. They have discussed that the sclera is free from the melanin and hemoglobin chromophores found in the skin, so its color provides a more direct indication of the TSB. However, they manually select a region of interest (ROI) in the sclera and process it. BiliScreen [27] is a smartphone application that takes pictures of the eye and estimates the person's bilirubin level for early detection of Pancreatic cancer. Aune et al. [28] have reported the results of bilirubin measurement on 302 full-term, normal-weight infants and compared them with a digital image recognition software used to estimate the value of bilirubin. They have not proposed the engineering details of their image-processing approach. Compared to their study, we propose a larger dataset including 446 samples along with a new image processing approach tuned to measure bilirubin. We introduce a new application, BiliBin, for neonatal jaundice estimation, which works on any smartphone with a pretty high-quality camera.

There are other active research areas in vision-based health sensing, including automated point-of-care diagnostics [29], [30], where visual analysis of test results from blood or urine samples on specialized materials is required, portable eye examination [31], and visual impairment recognition in children [32]. The common denominator of all these applications is image processing techniques, primarily relying on machine learning algorithms for reporting final human-readable results. Deep learning techniques [33] dramatically outperform conventional machine learning algorithms, specifically image processing and machine vision algorithms. For instance, convolutional neural networks (CNNs) have been used for human activity recognition with smartphone sensors [34], [35].

Despite their high capacity to learn difficult problems, deep learning approaches have been solely used in vision-based health sensing with a smartphone. A recent study can only classify images into two classes, jaundice, and non-jaundice [36], which is not suitable for use in practice. Indeed, there are two challenges with using deep learning approaches. The first is that they require many training samples to train adequately, which are not accessible for most applications, specifically in medical applications, where data collection is a time-consuming and expensive process, even restricted by privacy laws [37]. The second reason is that their

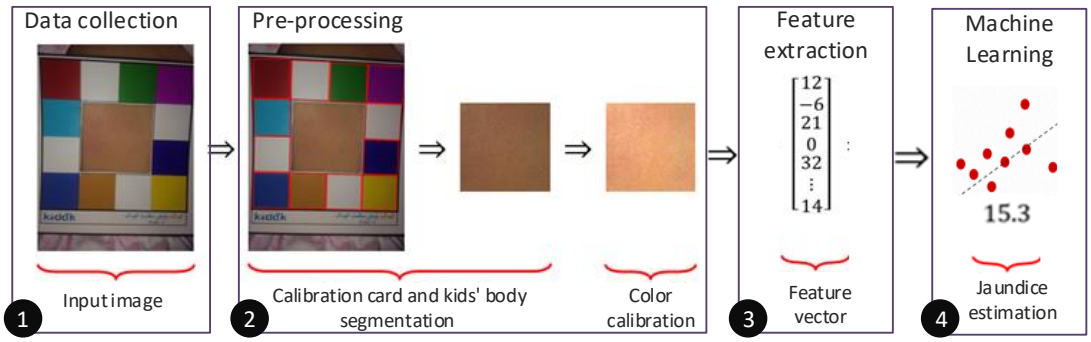



**Fig. 1** BiliBin operations.

learning process is computationally intensive [38], and for example, it cannot be performed on smartphones. Therefore, creating an application for collecting data and processing them on cloud platforms helps address the aforementioned challenges. Our jaundice monitoring application, BiliBin, also provides a data collection module and application programming interface (API) to connect to the cloud, enabling us to use deep learning techniques.

## 3 Methodology

Our proposed method operations consist of four steps: data collection, data preprocessing, feature extraction, and jaundice estimation using machine learning regression. **Fig. 1** illustrates BiliBin operations. In the following subsections, we describe each step in detail.

### 3.1 Data collection

The data collection process was conducted at four hospitals in Iran (Tehran, Qom, Jahrom, and Lamard). After the approval of the joint proposal with the Tehran University of Medical Sciences, the relevant licenses were obtained. One of our researchers referred to these sites and obtained the approval of one or more samplers from each center and gave them the necessary training. Newborns who were referred to the medical center for jaundice and had admission requirements were selected. The parents' consent to participate in the research was obtained through clear and complete explanations to the parents of selected infants. The most important criteria for entering the study included newborns who were healthy and had not received phototherapy in the past 24 hours. Criteria for leaving the study included parents' reluctance to continue collaborating, loss of essential data, and photos that lacked the quality needed for analysis.

The blood of these newborns was taken by medical staff to measure bilirubin according to the routine method (blood sampling of the newborn's veins or soles of the feet). The sampler then used a specific mobile application designed for imaging within a few minutes after blood sampling and takes photos from the newborn's sternum skin. The pre-designed calibrated paper was placed on the lower part of the sternum without bending. The smartphone camera was held without an angle, at a suitable distance from the baby, and four clear photos were taken from this area: two photos without flash and two photos with flash.

After the successful uploading of photos, a complete set of additional information related to the newborns, including birth week, age, weight, gender, as well as maternal age, type of delivery, type of breastfeeding, and city of birth, were recorded in this system. Such additional metadata can be used in further analysis of jaundice. During two years, 493 successful newborn registrations were recorded within the system, 446 of which had the necessary quality to enter the study and use in subsequent analysis.

### 3.2 Data preprocessing

Before the data samples (*i.e.*, the photos taken from newborns' sternum skins) can be fed to the machine learning models, each sample must be preprocessed and converted to a vector of numerical features. The preprocessing operations are shown in step 2 of **Fig. 1**. Sections 3.2.1 and 3.2.2 describe details of the preprocessing operations.

#### 3.2.1 Identifying the coordination of the calibration card and the newborn's body

The calibration card designed for use with BiliBin is shown in **Fig. 2**. It is worth noting that we come to this colored pattern experimentally and there is no standard in designing the calibration card. The card consists of 12 small-colored squares that are placed around a large empty square (related to the newborn's body). Colors used in this card are selected in such a way that every two adjacent squares have a significant color difference, which allows color-based image segmentation algorithms to distinguish these squares with good quality. An example of the segmentation task is shown in **Fig. 3**. A graph-based image segmentation algorithm is used in our design.

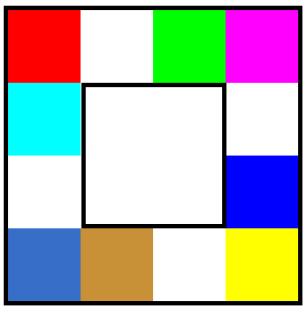


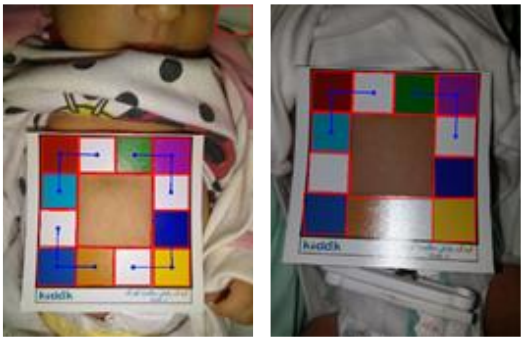
**Fig. 2** BiliBin calibration card.

The segmentation algorithm's output is a large number of areas, most of which are not related to the calibration card (**Fig. 3**.b). Non-square areas are discarded to reduce the number of areas for further processing. As can be observed in **Fig. 3**.c, using this idea, only two false-positive squares remained such that one of them is related to the newborn's body image. After identifying the square segments in the image, it is necessary to identify the calibration card's corresponding segments by considering its geometric characteristics. Segments of the calibration card may not be detected correctly in the image due to poor lighting conditions and light reflection. Twelve squares on the card are divided into four groups of three segments; each one contains a corner square and two adjacent squares to address this problem. For instance, the red, cyan, and white segments (at the top left corner of the card) form a group. Recognition of areas related to the card begins with the detection of one of these four groups.

(a) (b) (c)

**Fig. 3** Graph-based image segmentation. (a) original image, (b) the output of the segmentation algorithm, (c) removing non-square segments.

The center of the obtained segments is used to identify a group. The two neighbor centers with the shortest distances are identified for each center, and then the angle between the lines obtained from these three centers is calculated. If this angle is close to 90 degrees, these three segments are considered as a group. **Fig. 4** illustrates the result of this algorithm on two samples, which confirms the successful performance of the proposed algorithm.



(a) (b)

**Fig. 4** The result of detecting triplet groups on two images.

After identifying all triplet groups in the image, it is necessary to determine to which corner of the calibration card each group belongs. The color of the calibration card segments is used to perform this distinguishing. The first corner of the calibration card is the top left corner containing red, cyan, and white colors. The following three criteria are defined to specify the corresponding colors forming this corner:

$$r = R - \max(G, B) \quad (1)$$

$$c = \min(G, B) - R \quad (2)$$

$$w = 1 - \frac{\max(R,G,B) - \min(R,G,B)}{\max(R,G,B) + \min(R,G,B)} \quad (3)$$

In Equation 1, $r$ determines the redness of a color, which is equal to the difference between the value of the red channel ($R$) and the maximum value of the green and blue channels. Ideally, the red color equals $[1, 0, 0]$, and therefore the maximum value for $r$ is one. The minimum value for $r$ is $-1$. Criterion $c$ in Equation 2 determines the cyan level of the color, which equals $[0, 1, 1]$ for pure cyan color. Criterion $w$ in Equation 3 determines the whiteness of a color. In pure white color, all three channels are one, and therefore, the maximum and minimum values of this color are equal. The sum of these three criteria is calculated for all the groups. The segments of the calibration card have a predefined order,

and for each segment, the color criterion of that segment is computed. The group with the maximum value of $c + r + w$ is selected as the first corner. However, the value of this expression must be greater than a threshold to be considered a correct detection. If no groups can satisfy this condition, the above steps are repeated for the next corners. An image in which none of the corners are correctly identified will be omitted at this step.

In the previous step, one of the corners of the calibration card (which is equivalent to three segments) is identified. The calibration geometry features of the calibration card are used to identify the remaining nine segments. The centers of the first three identified segments are used to estimate the center of the fourth valid segment. Then, the segment whose center is closest to the estimated center is selected as the fourth segment. The center of the fifth segment is estimated based on the center of previous segments, and the above process will be repeated until all segments are identified. The proposed algorithm for detecting the segments of a given image is shown in **Fig. 5**.

It is worth noting that in Figure 5, the border widths and border colors are not important. The color is only used to distinguish the identified segments by the segmentation algorithm. Identified segments are marked with green borders, unidentified segments are marked with red borders, and the segment that has the most overlap with the predicted center is marked with a blue border and a dot indicating the predicted center in that step.

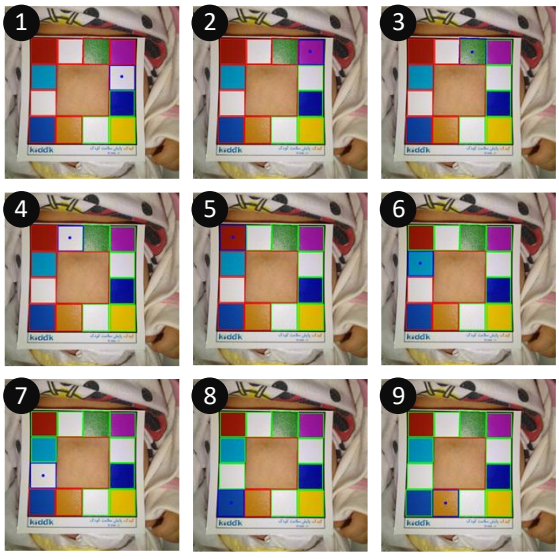
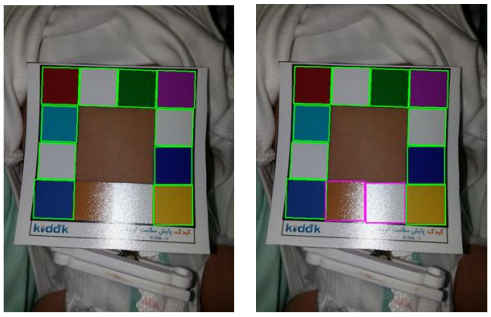



**Fig. 5** Steps to identify calibration card segments.

It is possible that some of the squares inside the calibration card failed to detect in the segmentation step. **Fig. 6**.a illustrates an example of this condition. If the number of not detected squares is more than four, the input image will be excluded from further processing. Otherwise, using the calibration card's geometry and the adjacent segments, these squares can be interpolated, shown in **Fig. 6**.b.



(a) (b)

**Fig. 6** Interpolating segments that are not detected. (a) detected segments of the calibration card, (b) interpolating missing segments.

After identifying the segments of the calibration card, the segment that belongs to the newborn's body, *i.e.*, the large square in the center of the card, can be easily estimated. Since the areas close to the border of this square may be erroneous due to unwanted shadows and lights, 20% of the image is removed on each side. The remaining square is cropped as a body area and converted to an image with a size of $100 \times 100$ pixels to be used in the subsequent steps. **Fig. 7** shows samples of newborns' images along with the final segments of their bodies.


**Fig. 7** Examples of cropped newborns' bodies. First row: original images with the identified segment of the body; second row: the cropped areas; third row: the calibrated areas.

### 3.2.2 Body-color calibration

Skin color is used to determine the extent of the newborn's jaundice. The color recorded by the camera for the skin and the actual color of the newborn's skin depends on other variables such as the amount of ambient light and the camera's technical features, which can be observed in Figure 7. For this reason, it is necessary to reduce the effect of other variables from the

image as much as possible before it can be used in the jaundice estimation algorithm. A calibration card, whose segmentation and detection were discussed in Section 3.3.1, is placed on the newborn's body. However, the lighting conditions and the camera's quality also affect the colors recorded within segments of the calibration card. By knowing the correct values of the colors on the calibration card segments, the effects of other variables can be modeled to some extent and removed from the image.

Experiments in this work and the results reported in related work show that sophisticated modeling has failed to improve the jaundice estimation accuracy. Similar to [10], [11], the proposed method utilizes white color balancing to calibrate the skin color. In this method, the amount of white color is extracted from the image, and calibration is performed based on this value. There are four white segments on our calibration card. Since some segments may be noisy, these four areas are sorted by light intensity. Then, the brightest and darkest white segments are left out, and the average white color value of the other two segments is calculated as the white color value recorded by the camera. The color of the newborn's skin pixels is calibrated using the following equation:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1/R'_w & 0 & 0 \\ 0 & 1/G'_w & 0 \\ 0 & 0 & 1/B'_w \end{bmatrix} \begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} \quad (4)$$

where $(R', G', B')$ is the color of a raw pixel, $(R'_w, G'_w, B'_w)$ is the raw white color in the original picture, and vector $(R, G, B)$ is the value of the calibrated color. After this processing, the white color in the image becomes $(1,1,1)$. The third row of Figure 7 illustrates examples of the effect of white balancing using the proposed method.

### 3.3 Feature extraction

The previous steps resulted in two $100 \times 100$ images with calibrated colors of the newborn's body: One image with flash and one without flash. Since newborns' skins color are different, the effect of bilirubin level discoloration on their skins are also different, and it is not possible to get a simple relationship between the values of $(R, G, B)$ vector and the bilirubin level. Besides, there may be disturbing factors in the images, such as the newborn's blood vessels, making it difficult to calculate $(R, G, B)$ for each newborn correctly. According to these arguments, it is necessary to extract more suitable features from these images, discussed in this section.

**Color spaces.** RGB space is not only the best color space for measuring a newborn's jaundice from the skin color. HSV, YCbCr, and LAB color spaces have been used in addition to RGB color space when extracting statistical features.

**Statistical characteristics.** For each pixel of the image, 12 numerical values are calculated (3 channels in 4 different color spaces). The mean and standard deviation of each channel is then calculated as two statistical features, which result in 24 features. Besides these two features, the median of each pixel is also computed as the third statistical feature. The advantage of the median over the mean is that it is robust to outliers. If a portion of the pixels in the newborn's body area have different colors due to the poor lighting conditions or the presence of blood vessels, they cannot change the median.

**Image gradient.** Skin surface roughness can be a proper criterion for diagnosing jaundice. The amount of local variation of each pixel compared to its neighborhood is calculated using the gradient operator. Then, the median of the gradient values in the whole image is calculated as a new feature.

**Combine two images.** As mentioned, at least one image with flash and one without flash are recorded for each newborn. The combination of statistical features of these two images can help diagnose jaundice. For this reason, in addition to the 74 features of the two images, the ratio of the features of the two images is calculated as 37 features, and a total of 111 features are extracted from the pair images.

### 3.4 Jaundice estimation

The output of Section 3.2 is 111 distinguishing features for each pair of images of each newborn. Collecting many images with known bilirubin levels (used as data labels) is necessary to estimate neonatal jaundice from extracted features. We discussed the process of data collection in Section 3.1. The mapping between features and the amount of jaundice can be learned through regression algorithms. Various regression models can be used depending on the number of training data and the complexity of the problem. We trained different regression algorithms on our dataset, including but not limited to kNN [23], SVR [22], Random Forest [39], and Bayesian regression [40], to find the best algorithm for predicting bilirubin levels. Finally, Gaussian process regression (GPR) [41] was selected as the best algorithm based on evaluation metrics discussed in the following subsection. GPR calculates the probability distribution over all admissible functions that fit the data. It has several benefits, such as working well on small datasets, with few numbers of features, and having the ability to provide uncertainty measurements on the predictions. We, therefore, will only report the result of the best model in the evaluation section.

### 3.5 Evaluation process and metrics

The $k$-fold cross-validation method is applied to evaluate the efficiency of the proposed algorithms. In this method, the data set is divided into k groups, and $k$ experiments are performed. At each experiment, one group's data is used to evaluate the model, and the rest of the data is used to train the model. After all of the experiments have been performed, the predicted values are stored in a vector, $p$, to compare with the vector, $t$, containing the actual value for each sample. We report various metrics that are available for evaluating regression models. Root mean squared error (RMSE) and mean absolute error (MAE) are respectively defined by Equation 5 and Equation 6.

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(p_n - t_n)^2} \tag{5}$$

$$MAE = \frac{1}{N}\sum_{n=1}^{N}|p_n - t_n| \tag{6}$$

where $N$ is the number of samples in the test set. The normalized correlation coefficient between the predicted values, $p_n$'s, and the target values, $t_n$'s, is also used as an appropriate metric for evaluating regression algorithms, which can be calculated using Equation (7).

$$\rho = \frac{\sum_{n=1}^{N}(p_n - \bar{p})(t_n - \bar{t})}{\sqrt{\sum_{n=1}^{N}(p_n - \bar{p})^2 \sum_{n=1}^{N}(t_n - \bar{t})^2}} \tag{7}$$

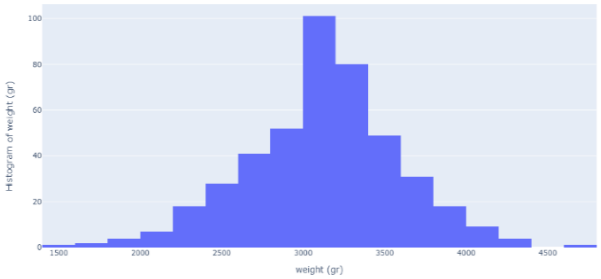
The rank correlation coefficient is similar to the normalized correlation coefficient, but instead of using the values $p$ and $t$, it uses their rank. The vectors $p$ and $t$ are arranged in ascending order to compute the rank correlation coefficient. Then, the correlation is calculated between the ranks instead of the values.

## 4 Experiments and Results

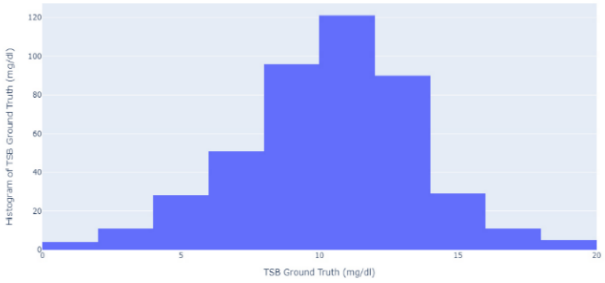
We first report the result of the data collection, described in Section 3.1. Then, we evaluate the performance of Gaussian process regression [36] on the collected dataset by comparing the predicted value of bilirubin for each sample with its grand-truth value obtained by blood sampling.

### 4.1 Dataset

Table 1 shows the data obtained from each medical center (C1 to C4) by gender. The histogram diagram of weights and TSB values of the collected datasets are shown in Figure 8. As can be observed, the recorded TSB values are in the range of 1.7 mg/dl to 19.9 mg/dl, but most of the TSB values are around 10 mg/dl (10.3 ± 3.2 mg/dl). According to the literature [42], healthy adults have a total serum bilirubin (TSB) level of less than 1 mg/dL. In neonates, normal TSB levels are comparatively higher, with age-dependant levels. Healthy, full-term newborns typically have peak serum bilirubin concentrations of 5 to 6 mg/dL compared to adult levels of < 1 mg/dL. Severe hyperbilirubinemia is commonly defined as a TSB>25 mg/dL and it occurs in approximately 1 out of 2500 live births [42].




(a)




(b)

**Fig. 8** Histogram of weights and TSB values in the dataset.

**Table 1.** Number of samples in each hospital or medical center.

| | | Medical center | | | | All |
|---|---|---|---|---|---|---|
| | | C1: Jahrom (Motahari Hospital) | C2: Lamard (Valiasr Hospital) | C3: Tehran (Children's Medical Center) | C4: Qom (Bouali Medical Laboratory) | |
| Samples | Male | 37 | 14 | 165 | 17 | 233 |
| | Female | 35 | 9 | 154 | 15 | 213 |
| **Total** | | **72** | **23** | **319** | **32** | **446** |

**Table 2.** Result of the regression model.

| Center | RMSE (mg/dl) | MAE (mg/dl) | Corr. | Rank corr. | Most negative difference (mg/dl) | Most positive difference (mg/dl) | Mean of difference (mg/dl) | Deviation standard of difference (mg/dl) |
|---|---|---|---|---|---|---|---|---|
| C1 | 2.206 | 1.765 | 0.703 | 0.708 | -5.154 | 4.872 | -0.376 | 2.189 |
| C2 | 2.495 | 2.056 | 0.466 | 0.470 | -4.849 | 5.153 | -0.863 | 2.393 |
| C3 | 2.258 | 1.777 | 0.718 | 0.688 | -5.645 | 5.626 | 0.198 | 2.253 |
| C4 | 2.491 | 2.030 | 0.641 | 0.607 | -5.261 | 4.671 | -0.495 | 2.481 |
| **All** | **2.280** | **1.807** | **0.701** | **0.678** | **-5.645** | **5.626** | **0.001** | **2.283** |

### 4.2 Model evaluation

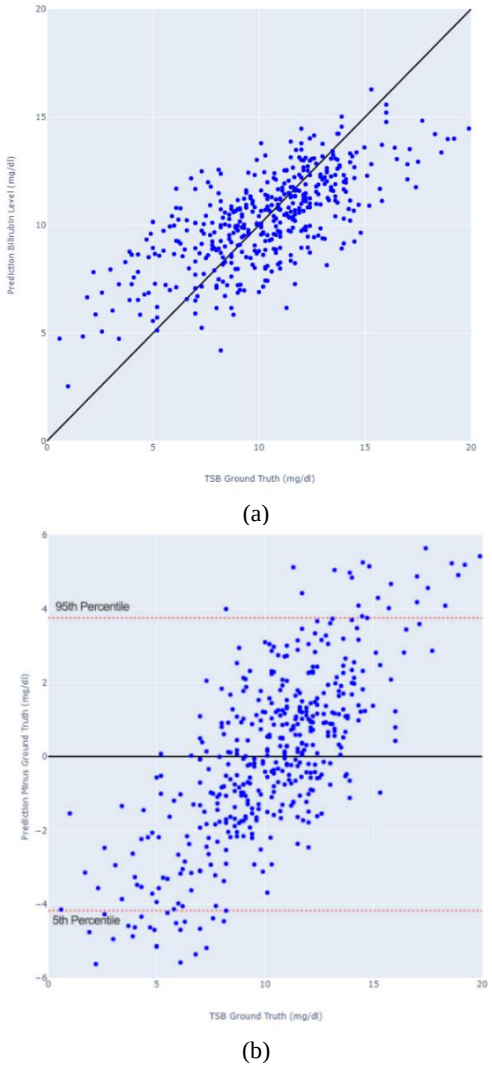
Table 2 shows the results of the Gaussian process regression in terms of evaluation metrics discussed in Section 3.4. The following results are observed:

- The mean error in centers C1, C2, and C4 is negative. In other words, we have underestimated the bilirubin, and the estimated values are a little less than the actual values on average. On the other hand, the mean error is positive for center C3. In other words, we have overestimated in this center, and the estimated values are slightly higher than the actual values, on average. This observation reveals that the equipment in these centers is somewhat different from each other.
- In total, our model can predict the bilirubin level with an RMSE of 2.280 mg/dl, MAE of 1.807 mg/dl, and correlation of 70.1%, which shows promising results regarding the applicability of the BiliBin level for use in practice. The achieved correlation is comparable with state-of-the-art works [10], [11] on our newly collected dataset. Adding more training samples by enlarging the dataset can improve the model performance, and we are planning to perform it in our future works. We are concerned about the absolute value of TSB. A threshold is then applied to indicate whether the subject should be considered healthy or not.

### 4.3 Predicting bilirubin levels

Fig. **9**.a illustrates a scatter plot of BiliBin prediction of bilirubin levels compared to the TSB values. **Fig. 9**.b illustrates a modified Bland-Altman plot [43], where residual ($BiliBin - TSB$) is plotted against the TSB as the ground-truth. As can be observed, the residuals are highly correlated by the TSB values. In other words, we usually overestimate the small TSB values while we underestimate the large TSB values. As described in Section 4.1, only a small fraction of the collected samples have small or large amounts of TSB, and most of them are around 10.3 mg/dl.

The current results are based on data from four hospitals, which might have controlled environments and standardized procedures. Home settings can vary significantly, which could affect the app's performance. Factors such as lighting conditions, the quality of the smartphone camera, and user proficiency can influence the accuracy of the app. While the current results are promising, replicating them in a home setting requires careful consideration of various factors. With proper training, continuous validation, and complementary tools, our application has the potential to be an effective tool for home-based jaundice diagnosis.




(a)

(b)

**Fig. 9** Estimated bilirubin level vs. the TSB values.

## 5 Conclusion

It is possible to estimate the bilirubin level from newborns' skin images and predict the risk of neonatal jaundice. Our proposed tool, BiliBin, is a smartphone application to monitor the newborn's jaundice non-invasively with a minimum cost. Bilibin, trained on 446 data samples, predicts the bilirubin level with a mean absolute error of 1.807 mg/dl and a correlation of 0.701 compared to the bilirubin level measured by TSB values. BiliBin's design goal is to be simply used at home by the newborn's parents as an early jaundice screening tool instead of a visual assessment, which requires experienced healthcare providers. We believe that it is more acceptable to have a false-positive alarm than a false-negative one that misses a case with potentially high bilirubin. Therefore, it can be replaced with a transcutaneous bilirubinometer (TcB) at outpatient clinics. The mean absolute error of BiliBin can be enhanced by providing more data samples (especially with small or large TSB values) and utilizing new learning methods for regression. In future work, we are planning to use a convolutional neural network (CNN) [44] on a larger dataset to improve the performance of BiliBin. It is also recommended to consider more healthy subjects along with unhealthy ones to distinguish among them and also consider more subjects with a variety of skin colors to further generalize the proposed application.


### Acknowledgments

The authors would like to thank the data collectors and medical collaborators.


### Ethical approval

All parents of newborns gave their informed consent for inclusion before their newborns participated in the study.

### Competing interests

All of the authors declares that they have no conflict of interest.

### Authors' contributions

**Eisa Zarepour:** Conceptualization, Methodology, Project administration. **Mohammad Reza Mohammadi:** Supervision, Methodology, Writing - Review & Editing, Resources. **Morteza Zakeri-Nasrabadi:** Methodology, Validation, Writing - Original Draft. **Sara Aein:** Formal analysis, Visualization, Writing - Original Draft. **Razieh Sangsari:** Data Curation, Resources. **Lila Taheri:** Investigation, Data Curation. **Ali Zabihallahpour:** Software, Formal analysis.


### Funding

This study has received no funding from any organization.


### Availability of data and materials

The dataset used in this paper is available from the authors upon request.

### References


[1] S. Mitra and J. Rennie, “Neonatal jaundice: aetiology, diagnosis and treatment,” Br J Hosp Med, vol. 78, no. 12, pp. 699–704, Dec. 2017, doi: 10.12968/hmed.2017.78.12.699.

[2] P. A. Dennery, D. S. Seidman, and D. K. Stevenson, “Neonatal hyperbilirubinemia,” New England Journal of Medicine, vol. 344, no. 8, pp. 581–590, Feb. 2001, doi: 10.1056/NEJM200102223440807.

[3] V. K. Bhutani and R. Wong, “Bilirubin-induced neurologic dysfunction (BIND),” Semin Fetal Neonatal Med, vol. 20, no. 1, p. 1, Feb. 2015, doi: 10.1016/j.siny.2014.12.010.

[4] S. Randev and N. Grover, “Predicting neonatal hyperbilirubinemia using first day serum bilirubin levels,” The Indian Journal of Pediatrics, vol. 77, no. 2, pp. 147–150, Feb. 2010, doi: 10.1007/s12098-009-0335-3.

[5] F. Ebbesen, L. Rasmussen, and P. Wimberley, “A new transcutaneous bilirubinometer, BiliCheck, used in the neonatal intensive care unit and the maternity ward,” Acta Paediatr, vol. 91, no. 2, pp. 203–211, Jan. 2007, doi: 10.1111/j.1651-2227.2002.tb01696.x.

[6] G. Nagar, B. Vandermeer, S. Campbell, and M. Kumar, “Reliability of transcutaneous bilirubin devices in preterm infants: a systematic review,” Pediatrics, vol. 132, no. 5, pp. 871–881, Nov. 2013, doi: 10.1542/peds.2013-1713.

[7] N. C. C. for W. and C. H. (UK), “Neonatal jaundice,” RCOG Press. Accessed: Oct. 15, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK65113/

[8] M. J. Maisels, V. K. Bhutani, D. Bogen, T. B. Newman, A. R. Stark, and J. F. Watchko, “Hyperbilirubinemia in the newborn infant > or = 35 weeks’ gestation: an update with clarifications,” Pediatrics, vol. 124, no. 4, pp. 1193–1198, Oct. 2009, doi: 10.1542/peds.2009-0329.

[9] A. Riskin, A. Tamir, A. Kugelman, M. Hemo, and D. Bader, “Is visual assessment of jaundice reliable as a screening tool to detect significant neonatal hyperbilirubinemia?,” J Pediatr, vol. 152, no. 6, pp. 782-787.e2, Jun. 2008, doi: 10.1016/j.jpeds.2007.11.003.

[10] L. de Greef et al., “Bilicam,” in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp ’14

Adjunct, New York, New York, USA: ACM Press, 2014, pp. 331–342. doi: 10.1145/2632048.2632076.
[11] J. A. Taylor et al., "Use of a smartphone app to assess neonatal jaundice," Pediatrics, vol. 140, no. 3, p. e20170312, Sep. 2017, doi: 10.1542/peds.2017-0312.
[12] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," Sensors (Basel), vol. 19, no. 9, p. 2164, May 2019, doi: 10.3390/s19092164.
[13] R. K. Lord, V. A. Shah, A. N. San Filippo, and R. Krishna, "Novel uses of smartphones in ophthalmology," Ophthalmology, vol. 117, no. 6, pp. 1274-1274.e3, Jun. 2010, doi: 10.1016/j.ophtha.2010.01.001.
[14] A. Karargyris, O. Karargyris, and A. Pantelopoulos, "DERMA/Care: An advanced image-processing mobile application for monitoring skin cancer," in 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, Nov. 2012, pp. 1–7. doi: 10.1109/ICTAI.2012.180.
[15] S. Kim et al., "Smartphone-based multispectral imaging: system development and potential for mobile skin diagnosis," Biomed Opt Express, vol. 7, no. 12, pp. 5294–5307, Nov. 2016, doi: 10.1364/BOE.7.005294.
[16] Q. Li, X. He, Y. Wang, H. Liu, D. Xu, and F. Guo, "Review of spectral imaging technology in biomedical engineering: achievements and challenges," J Biomed Opt, vol. 18, no. 10, p. 100901, Oct. 2013, doi: 10.1117/1.JBO.18.10.100901.
[17] S. A. Siddiqui, Y. Zhang, Z. Feng, and A. Kos, "A pulse rate estimation algorithm using PPG and smartphone camera," J Med Syst, vol. 40, no. 5, p. 126, May 2016, doi: 10.1007/s10916-016-0485-6.
[18] S. Majumder, T. Mondal, and M. J. Deen, "Wearable sensors for remote health monitoring," Sensors (Basel), vol. 17, no. 1, p. 130, Jan. 2017, doi: 10.3390/s17010130.
[19] J. Lee et al., "Itchtector: a wearable-based mobile system for managing itching conditions," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, New York, NY, USA: ACM, May 2017, pp. 893–905. doi: 10.1145/3025453.3025569.
[20] S. Leartveravat, "Transcutaneous bilirubin measurement in full term neonate by digital camera," 2009.
[21] S. B. Munkholm, T. Krøgholt, F. Ebbesen, P. B. Szecsi, and S. R. Kristensen, "The smartphone camera as a potential method for transcutaneous bilirubin measurement," PLoS One, vol. 13, no. 6, p. e0197938, Jun. 2018, doi: 10.1371/journal.pone.0197938.
[22] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," Stat Comput, vol. 14, no. 3, pp. 199–222, Aug. 2004, doi: 10.1023/B:STCO.0000035301.49549.88.
[23] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," Am Stat, vol. 46, no. 3, pp. 175–185, 1992, [Online]. Available: http://www.jstor.org/stable/2685209
[24] M. Aydın, F. Hardalaç, B. Ural, and S. Karap, "Neonatal jaundice detection system," J Med Syst, vol. 40, no. 7, p. 166, Jul. 2016, doi: 10.1007/s10916-016-0523-4.
[25] F. Outlaw et al., "Smartphone colorimetry using ambient subtraction," in Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers - UbiComp/ISWC '19, New York, New York, USA: ACM Press, 2019, pp. 172–175. doi: 10.1145/3341162.3343805.
[26] F. Outlaw, J. Meek, L. W. MacDonald, and T. S. Leung, "Screening for neonatal jaundice with a smartphone," in Proceedings of the 2017 International Conference on Digital Health - DH '17, New York, New York, USA: ACM Press, 2017, pp. 241–242. doi: 10.1145/3079452.3079488.
[27] A. Mariakakis, M. A. Banks, L. Phillipi, L. Yu, J. Taylor, and S. N. Patel, "BiliScreen: Smartphone-based scleral jaundice monitoring for liver and pancreatic disorders," Proc ACM Interact Mob Wearable Ubiquitous Technol, vol. 1, no. 2, pp. 1–26, Jun. 2017, doi: 10.1145/3090085.
[28] A. Aune, G. Vartdal, H. Bergseng, L. L. Randeberg, and E. Darj, "Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels," Acta Paediatr, vol. 109, no. 12, pp. 2532–2538, Dec. 2020, doi: 10.1111/apa.15287.
[29] L. Shen, J. A. Hagen, and I. Papautsky, "Point-of-care colorimetric detection with a smartphone," Lab Chip, vol. 12, no. 21, p. 4240, 2012, doi: 10.1039/c2lc40741h.
[30] N. Dell and G. Borriello, "Mobile tools for point-of-care diagnostics in the developing world," in Proceedings of the 3rd ACM Symposium on Computing for Development - ACM DEV '13, New York, New York, USA: ACM Press, 2013, p. 1. doi: 10.1145/2442882.2442894.
[31] M. E. Giardini, I. A. T. Livingstone, N. M. Bolster, S. Jordan, and A. Bastawrous, "Phone-based ophthalmoscopy for Peek, the Portable Eye Examination Kit," 2014.
[32] H. K. Rono et al., "Smartphone-based screening for visual impairment in Kenyan school children: a cluster randomised controlled trial," Lancet Glob Health, vol. 6, no. 8, pp. e924–e932, Aug. 2018, doi:

10.1016/S2214-109X(18)30244-4.
[33] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[34] C. A. Ronao and S.-B. Cho, “Human activity recognition with smartphone sensors using deep learning neural networks,” Expert Syst Appl, vol. 59, pp. 235–244, Oct. 2016, doi: 10.1016/j.eswa.2016.04.032.
[35] C. A. Ronao and S.-B. Cho, “Deep convolutional neural networks for human activity recognition with smartphone sensors,” 2015, pp. 46–53. doi: 10.1007/978-3-319-26561-2_6.
[36] A. Chakraborty, S. Goud, V. Shetty, and B. Bhattacharyya, “Neonatal jaundice detection system using CNN algorithm and image processing,” International Journal of Electrical Engineering and Technology (IJEET), vol. 11, no. 3, pp. 248–264, 2020, doi: 10.34218/IJEET.11.3.2020.029.
[37] M. I. Razzak, S. Naz, and A. Zaib, “Deep learning for medical image processing: overview, challenges and the future,” 2018, pp. 323–350. doi: 10.1007/978-3-319-65981-7_12.
[38] S. Rajbhandari, Y. He, O. Ruwase, M. Carbin, and T. Chilimbi, “Optimizing CNNs on multicores for scalability, performance and goodput,” ACM SIGARCH Computer Architecture News, vol. 45, no. 1, pp. 267–280, May 2017, doi: 10.1145/3093337.3037745.
[39] L. Breiman, “Random forests,” Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[40] K. W. Fornalski, “Applications of the robust Bayesian regression analysis,” Int J Soc Syst Sci, vol. 7, no. 4, p. 314, 2015, doi: 10.1504/IJSSS.2015.073223.
[41] C. K. I. W. Carl Edward Rasmussen, Gaussian processes for machine learning. MIT Press, 2006. [Online]. Available: http://www.gaussianprocess.org/gpml/
[42] Betty Ansong-Assoku, Sanket D. Shah, Mohammad Adnan, and Pratibha A. Ankola, Neonatal Jaundice. StatPearls, 2024.
[43] J. M. Bland and D. G. Altman, “Statistical methods for assessing agreement between two methods of clinical measurement,” Lancet, vol. 1, no. 8476, pp. 307–310, Feb. 1986.
[44] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org/.

**Eisa Zarepour** is an Assistant Professor in the School of Computer Engineering at Iran University of Science and Technology. His areas of research are wireless sensor networks and nanobiotechnology.

**Mohammad Reza Mohammadi** is an Assistant Professor in the School of Computer Engineering at Iran University of Science and Technology. His areas of research are computer vision and machine learning.

**Morteza Zakeri-Nasrabadi** is a Ph.D. graduate from the School of Computer Engineering at Iran University of Science and Technology. His research interests are software engineering and machine learning applications in medicine and biomedical engineering.

**Sara Aein** is an M.Sc. graduate from the School of Computer Engineering at Iran University of Science and Technology. Her research interests are computer vision and machine learning.

**Razieh Sangsari** is an Assistant Professor in the School of Medicine at Tehran University of Medical Sciences. Her research interests are jaundice and assistant ventilation in neonatal medicine.

**Leila Taheri** received her Ph.D. degree in nursing from the School of Nursing and Midwifery at Tehran University of Medical Sciences. She is a faculty member of Nursing and Midwifery College at Qom University of Medical Sciences. Her areas of research are pain management, identity, transition, nursing care, and FCC.

**Ali Zabihallahpour** received his M.Sc. degree in computer engineering from the School of Computer Engineering at Iran University of Science and Technology. He is a full-stack Android and web application developer with experience in embedded system design and development.

**Mojtaba Akbari** received MD degree from Tehran university of medical sciences. He is currently a senior radiology resident at Iran university of medical sciences. His research interests are MS, infertility and radiation.