Stool Image Analysis for Precision Health Monitoring by Smart Toilets

Jin Zhou[1] JIN.ZHOU@DUKE.EDU
Nick DeCapite[1] NICK.DECAPITE@DUKE.EDU
Jackson McNabb[1] JACKSON.MCNABB@DUKE.EDU
Jose R. Ruiz[2] JOSE.RUIZ@DUKE.EDU
Deborah A. Fisher[2] DEBORAH.FISHER@DUKE.EDU
Sonia Grego[1,3] SONIA.GREGO@DUKE.EDU
Krishnendu Chakrabarty[1] KRISHNENDU.CHAKRABARTY@DUKE.EDU
[1] *Department of Electrical and Computer Engineering, Duke University, Durham, NC, United States*
[2] *Division of Gastroenterology, School of Medicine, Duke University, Durham, NC, United States*
[3] *Center for WaSH-AID, Duke University, Durham, NC, United States*

Abstract

Precision health monitoring is facilitated by long-term data collection that establishes a health baseline and enables the detection of deviations from it. With the advent of the Internet of Things, monitoring of daily excreta from a toilet is emerging as a promising tool to achieve the long-term collection of physiological data. This paper describes a stool image analysis approach that accurately and efficiently tracks stool form and visible blood content using a Smart Toilet. The Smart Toilet, can discreetly image stools in toilet plumbing outside the purview of the user. We constructed a stool image dataset with 3,275 images, spanning all seven types of the Bristol Stool Form Scale, a widely used metric for stool classification. We used ground-truth data obtained through the labeling of our dataset by two gastroenterologists. We addressed three limitations associated with the application of computer-vision techniques to a smart toilet system: (i) uneven separability between different stool form categories; (i) class imbalance in the dataset; (ii) limited computational resources in the microcontroller integrated with the Smart Toilet. We present results on the use of class-balanced loss, and hierarchical and compact convolutional neural network (CNN) architectures for training a stool-form classifier. We also present results obtained using perceptual color quantization coupled with mutual information to optimize the color-feature space for the detection of stool images with gross (visible) blood content. For the classification of stool-form, we achieve a balanced accuracy of 81.66% using a hierarchical CNN based on MobileNetV2. For gross blood detection, the decision tree (DT) classifier provides 74.64% balanced accuracy.

1. Introduction

Biomedical imaging is one of the cornerstones of medical diagnostics and it is being enhanced by sophisticated machine-learning techniques (Halicek et al., 2017). Recent applications of machine learning for health applications have focused on the analysis of physiological data collected over a prolonged period of time. This analysis provides individualized risk assessment and early warning of disease onset that can be used to trigger interventions. As

opposed to a snapshot of the visit to a doctor's office, such precision health monitoring is empowered by time-dense health data that establishes a health baseline and enables the detection of deviations from it (Collins and Varmus, 2015).

Long-term adherence to precision health monitoring is facilitated by not requiring the user to personally collect the data. Human excreta (urine and stool) are readily available specimens, regularly deposited in toilets. With the advent of the Internet of Things (IoT) paradigm, the monitoring of physiological functions from a toilet during bathroom visits is emerging as an active area of research for precision health monitoring (Park et al., 2020; Wald, 2017; Bae and Lee, 2018; Ghosh et al., 2020; Ra et al., 2018; Kim and Allen, 2016).

Research on "smart toilet" for health monitoring has thus far mainly focused on urine analysis (Bae and Lee, 2018; Ghosh et al., 2020); however, important health information is also found in feces. Specifically, stool physical characteristics such as form (i.e., consistency) and color contribute to the diagnosis and management of many acute and chronic gastrointestinal (GI) conditions. Stool appearance is one of the early diagnostics indicator for evaluation of irritable bowel syndrome (IBS), (as much as 10-15% of the world population is estimated to suffer from IBS) (Halmos et al., 2018), inflammatory bowel disease (IBD), malabsorption syndromes, and upper and lower GI bleeding (Tanaka et al., 2018). The impact of GI diseases on patients and the health care system is substantial; for example in the US, GI healthcare cost is higher than the cost associated with heart disease (Peery et al., 2019).

In clinical practice, patient self-reporting on bowel movement is limited by subjectivity, poor recall accuracy, and the burden of constant tracking (Halmos et al., 2018). Discrepancies have been documented between patient self-reports and standardized stool descriptors of color and frequency (Zuckerman et al., 2005), and between patient report and clinician assessment of diarrhea (Majid et al., 2012), despite the development of paper- and digital-based diaries (Halmos et al., 2018). Considerable variability and inconsistency has been found in how patients describe the color of gross (i.e., visible) blood in stool, a symptom associated with GI bleeding (Zuckerman et al., 2005). The color of blood provides information to help ascertain the severity and the likely anatomic site of bleeding and helps direct the initial diagnostic and therapeutic evaluation. Blood color ranges from bright red (usually distal intestinal bleeding) to dark red/maroon (colonic bleeding or rapid upper GI bleeding) to black, tarry stool, termed melena (usually gastric or proximal small intestinal bleeding).

There is no approved clinical method that can reliably and consistently monitor stool frequency, form, and color, either in the home setting or in the hospital. To address this limitation, image capture of the content of toilet bowl either by the user (Hachuel et al., 2019) or without user intervention (Park et al., 2020) has been proposed.

Our group is developing a smart toilet that enables discreet imaging of stool in the toilet plumbing, after flushing and outside the purview of the user, thereby not changing the user experience. Stool image analysis is a key enabler of smart toilets for the monitoring of bowel movement. In this paper, we present two functions for a smart-toilet module that provide clinically relevant information: gross blood detection and stool-form classification. We use color quantization, feature selection, and a decision-tree (DT) classifier framework for gross blood detection. For stool-form classification, we use a compact architecture, such

as MobileNetV2 as a typical example, and class-balanced softmax cross-entropy loss for model training.

The main contributions of this paper are as follows:

1. We construct a stool image dataset containing 3,275 stool images spanning all seven Bristol Stool Form Scale (BSFS) types.
2. We present the design of a hierarchical CNN architecture for stool-form classification over seven BSFS values and three consolidated categories (constipated, normal, and loose).
3. We present results for stool form classification using class-balanced softmax cross-entropy loss based on two compact CNN models, namely MobileNetV2 and ShuffleNetV2, for training the CNN classifier.
4. We present results for gross blood detection with two machine-learning techniques, namely decision tree (DT) and k-nearest-neighbors (KNN). Perceptual color quantization coupled with mutual information are employed for optimizing the color-feature space.

Generalizable Insights about ML in the Context of Healthcare

We demonstrate a machine learning (ML) approach for clinically relevant stool characteristics that is both accurate and computationally efficient. The ML solution enables classification of stool characteristics and provides objective data to inform improved clinical care. This computational tool will be most impactful if implemented as edge computing near the image data source. We describe an approach that addresses challenges that are commonly faced by computer-vision techniques being applied to medical imaging. First, we utilize approaches such as hierarchical CNN architecture to overcome the issue of uneven separability between different categories and color quantization to achieve blood detection. Second, by training the CNN using class-balanced loss based on effective number of samples, we can address the problem of class imbalance in the dataset. Third, by evaluating several recent CNN designs, we select a design that enables image classification that is computationally efficient as defined by metrics of the number of float-point operations ($FLOPs$) and memory requirement, so that it will be easier to deploy the ML model in a resource-constrained environment, such as the physical smart toilet hardware. Overall, this combination of machine learning and stool specimen imaging enables a new form of physiological monitoring that may provide early warning of disease for timely intervention and improved clinical outcomes.

The rest of this paper is organized as follows. Section 2 describes related prior work and provides further motivation for this research. Section 3 describes the background of the smart toilet and our system design for stool analysis. Section 4 describes the proposed methods for stool-form classification and blood detection. Section 5 presents experimental results. Section 6 describes the limitation of this paper. Finally, Section 7 concludes the paper.

2. Related Prior Work

In this section, we describe some representative prior work on stool-form classification, and then focus on recent work on using machine learning (ML) for stool image assessment.

2.1. Bristol Stool Form Classification

The BSFS scale (Lewis and Heaton, 1997) is a standard medical diagnostic tool for categorizing adult stool based on its physical appearance. Normal stool consistency is defined as BSFS Type 3, 4 and 5 (Markland et al., 2013). Constipation is defined as a Type 1 (separate hard lumps, like nuts) or Type 2 (sausage-like, but lumpy). Diarrhea is defined as a minimum of three loose stools (Type 6 and Type 7) per day. The BSFS stool chart is shown in Figure 1.

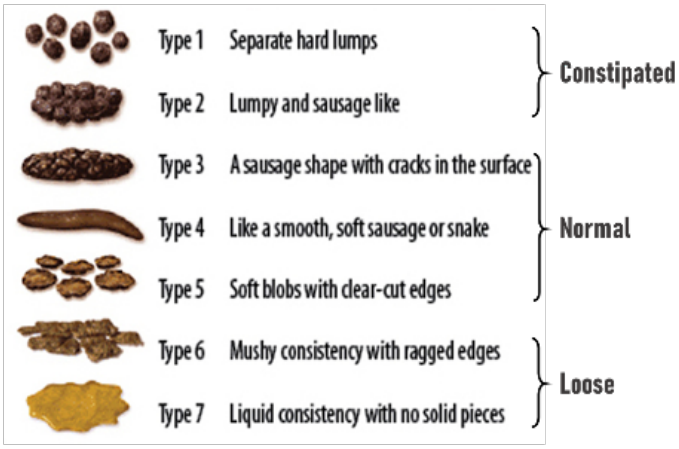


Figure 1: Illustration of the BSFS chart (adapted from http://cdn.intechopen.com/pdfs-wm/46082.pdf)

A 2019 study validated the use of the BSFS by having participants use a printed card tool with graphics to assess the properties of their bowel movements (Ohno et al., 2019).

2.2. Machine Learning Approaches for Stool Image Assessment

Yang et al. (Yang et al., 2019) introduced StoolNet, which combines the region of interest (ROI) detection and a shallow CNN for color classification of stool images. Park et al. (Park et al., 2020) utilized transfer learning to train a classifier on top of a trained deep learning architecture. While these studies provide key insights into automated stool classification, three major challenges have yet to be addressed, namely, uneven separability, class imbalance and model complexity.

Uneven Separability. Visual separability between different BSFS categories is uneven. For example, it is difficult to distinguish type 3 from type 4, while it is easy to tell a type 1 from type 3. The traditional CNNs (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015) use the flat structure to train a N-way classifier and do not consider such uneven separability, which often leads to sub-optimal performance in the task of fine-grained classification. A common strategy to address this problem is to predefine a hierarchy or

taxonomy of classifiers so that a given testing image can be first evaluated by a coarse classifier and use the corresponding fine classifier to make the fine prediction (Murthy et al., 2016; Yan et al., 2015).

Class Imbalance. Medical diagnostic data may have a normal distribution (bell-shaped curve) or a skewed distribution. For instance, in the stool image dataset collected by Park et al. (Park et al., 2020), only a few images report constipated stool while most images indicate normal stool. A number of solutions have been proposed in the literature to address the problem of class imbalance. The first approach is re-sampling, which aims to alter the training data distribution, usually by random under-sampling and over-sampling techniques (Oquab et al., 2014; Chawla et al., 2002). The second approach is cost-sensitive learning, which assigns higher misclassification costs to the minority classes compared to the majority classes (Wang et al., 2017; Cui et al., 2019; Zadrozny et al., 2003).

Model Complexity. In practice, state-of-the-art CNNs models (Simonyan and Zisserman, 2015; He et al., 2016) incur significant compute overhead, which imposes a barrier to their deployment on devices with limited computational power, e.g., a micro-computer (Raspberry Pi). Many approaches have been proposed to address this challenge, which can be categorized on the basis of techniques that use either model compression or compact architectures. Model compression techniques include parameter pruning and weight quantization (Han et al., 2015; Denton et al., 2014; Cheng et al., 2017). However, these methods require dedicated hardware or software customization for practical implementation (Han et al., 2015). In contrast, compact architecture design methods target more efficient and compact neural network architectures (Iandola et al., 2018; Ma et al., 2018; Howard et al., 2017).

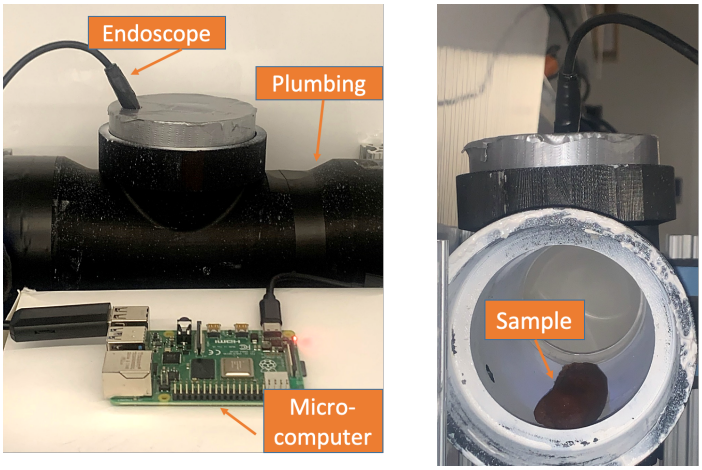
3. Smart Toilet System

In this section, we provide an overview of the smart toilet system and formulate stool analysis as a real-time computer-vision problem.

Smart-toilet approaches have been proposed to obtain health-related information from different configurations, e.g., devices snapped on the toilet bowl (Hall et al., 2020) or integrated in the toilet seat (Park et al., 2020; Conn et al., 2019). Notably, Park et al. (Park et al., 2020) introduced a defecation-monitoring module that uses sensors and computer vision to acquire basic properties of human excreta from sensors integrated in a commercially available electronic bidet. The acquired images are fed offline to machine-learning algorithms for analysis. However, cameras and illumination device in the toilet seat create an uncomfortable environment for the user, as highlighted by the results of a user survey regarding the technology (Park et al., 2020).

An alternative approach, which avoids the adoption barrier due to user discomfort, is a technology that integrates sensors in the toilet plumbing where they are not visible to users. A toilet manufacturer reported such a configuration for urine analysis (Tsang et al., 2017).

We have developed a novel approach to image feces in the plumbing at the outlet of the toilet, after the user has flushed. Our design offers a unique opportunity for real-time inline sensing approach specific to feces without engendering user discomfort. Using fingerprint recognition on the toilet flush button, users within a residential setting can



(a) (b)

Figure 2: (a) The set-up for stool image analysis. (b) Cross-section view of the plumbing with surrogate specimen.

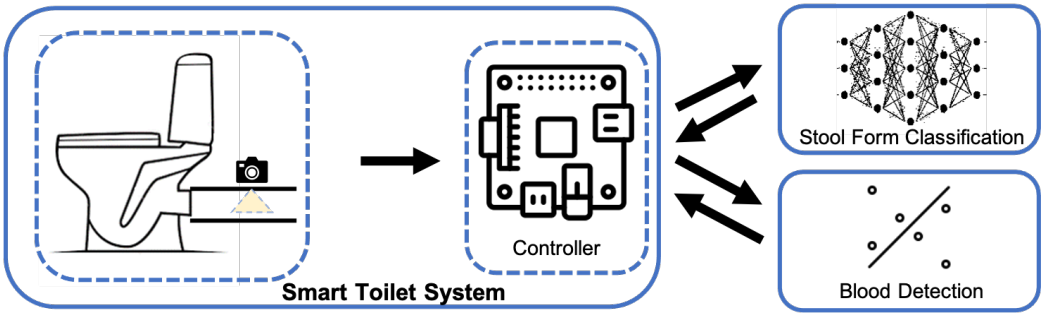


Figure 3: Framework for stool image analysis in a smart toilet system. (a) Images are captured when the stool is immobilized. (b) Images are processed by the controller and fed to machine-learning algorithms for gross blood detection and stool-form classification.

be individually tracked. The hardware setup used for stool image analysis is shown in Figure 2(a). The endoscope featuring six illuminating LEDs is installed in a viewing port in the toilet plumbing and connected to a micro-computer (Raspberry Pi 3B+) for real-time image collection and analysis. Figure 2(b) is a cross-sectional view of the immobilization region in the plumbing.

The image analysis approach reported in this paper enables real-time categorization of bowel movements according to two dimensions: stool form classification and blood detection. Figure 3 provides an illustration of the overall system.

4. Stool Image Analysis

In this section, we describe the data set used for analysis, as well as the machine learning techniques used for classification.

4.1. Stool Image Data Set Preparation

No publicly available stool image dataset exists, thus we developed our own image dataset. Of note, the stool image dataset used in (Yang et al., 2019) for their StoolNet model only contained 110 images (each rotated and used four times) and, furthermore, those images are not publicly available.

Our work leverages a dataset of 3,629 stool images spanning all seven BSFS types obtained from two sources: the web and anonymous image uploads from the general public. A total of 2,720 online unique images were obtained through search engines such as Google and platforms such as Reddit with keywords such as 'feces in toilet' and 'bristol'). Additionally, to collect images representing a wide range of bowl movement, we developed an institutional review board (IRB) approved protocol (Duke University IRB 2020-0569) to request the general public to take images of stool in their toilet and upload anonymously to a secure repository site. We created a website, then advertised on social media and leveraged the physician collaborators' professional networks. Potential personally identifying information unintentionally contained in images metadata, such as GPS coordinates, were removed from the images uploaded to the depository prior to further analysis. We obtained 909 public-uploaded images. A total of 256 online images and 98 public-uploaded images were rejected because the image quality did not allow accurate annotation; therefore, the study used a total of 3,275 images.

A total of 552 images were annotated by two gastroenterologists and an additional 2723 were rated by one of them. We used the online platform Labelbox (Labelbox, 2019) and assigned to each image a BSFS score from 1 through 7 and, importantly, indicating the presence/absence of blood. Despite being a clinical standard, the BSFS score does not capture the full variety of stool forms and does not account for the presence of stools of more than one BSFS category in the same image. From a clinical point of view, the important information is whether the bowel movement is normal (types 3,4,5) or abnormal, which is further classified as constipated (types 1,2) and diarrhea (types 6,7). The correlation between the labels assigned by the two gastroenterologists was measured using Cohen's Kappa statistic (Banerjee et al., 1999), which ranges from 1 (perfect classification) to -1 (extreme misclassification). The Cohen's Kappa statistic metric was calculated to be $k = 0.435$ for BSFS and $k = 0.540$ for the consolidated categories. These values show a satisfactory agreement between the two gastroenterologists. Due to heterogeneity of feces, large inter-rater variability is not surprising and our k = 0.540 is similar to k = 0.584 reported by (Hachuel et al., 2019).

4.2. Stool Form Classification

The first part of our classification problem is to determine the bowel movement type based on the stool form.

4.2.1. Hierarchical Architecture

Visual separability between different BSFS categories is uneven. For example, it is difficult to distinguish type 3 from type 4, while it is easy to tell a type 1 from type 3. As introduced in Section 2.1, type 3 and 4 are defined as the same consolidated category 'normal'. To

leverage the hierarchical structure of stool-form categories, we deploy a tree-like hierarchical architecture (Yan et al., 2015; Seo and Shin, 2019), as shown in Figure 4.

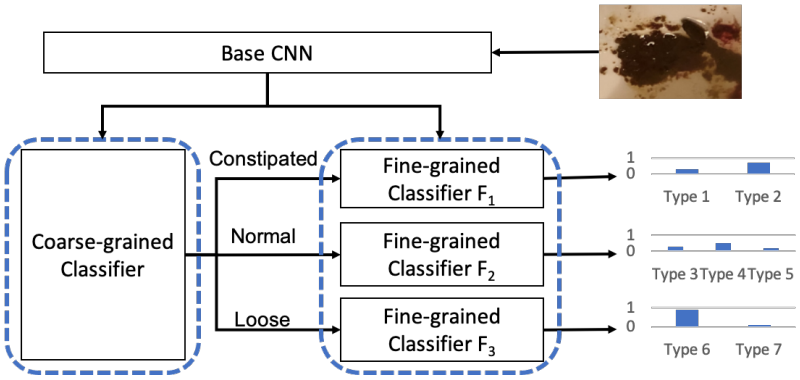


Figure 4: Hierarchical architecture for stool-form classification.

The main components in our architecture include the base CNN, a coarse-grained classifier, and three fine-grained classifiers. For an input image, the base CNN extracts low-level features. The extracted features are fed to the coarse-grained classifier and which produces a consolidated prediction over three categories ('loose', 'normal', and 'constipation'). The consolidated prediction enables the corresponding fine-grained classifier to reuse the extracted features and make the prediction over the seven BSFS types. For example, the consolidated prediction 'constipated' triggers the fine-grained classifier F_1 to make the prediction over BSFS type 1 and type 2. In the proposed architecture, both coarse-grained classifier and fine-grained classifiers are configured with a two-layer CNN, as described in Table 1. The depth for Layer-2 is three for coarse-grained classifier and fine-grained classifier classifier F_2, and is two for fine-grained classifier F_1 and classifier F_3.

Table 1: Configurations used in the fine-grained classifiers.

Layer	Type	Depth	Activation	Stride	Padding
1	Convolution	320	ReLU	3	1
2	Fully-Connected	2 or 3	N/A	N/A	N/A

4.2.2. Base CNN Design

Various CNN designs have been proposed over the past few years for a wide range of applications (Simonyan and Zisserman, 2015; Howard et al., 2017; Gatys et al., 2015). As introduced in Section 3, we consider a single-board computer, i.e., Raspberry Pi, to load our CNN models. The computational resources available on Raspberry Pi is limited compared to a server, therefore deep CNNs such as VGG16 are not feasible in this application scenario. In this paper, we explore two compact CNN designs, namely MobileNetV2 (Sandler et al., 2018) and ShuffleNetV2 (Ma et al., 2018).

MobileNetV2. To reduce computation cost, MobileNetV1 (Howard et al., 2017) replaces the standard convolutional filters by two layers: depthwise convolution and 1×1

pointwise convolution, where depthwise convolution only extracts spatial features for each independent channel and pointwise convolution extracts channel-wise information. Furthermore, MobileNetV2 uses an inverted bottleneck structure to increase representational power.

ShuffleNetV2. ShuffleNet (Zhang et al., 2018) employs group convolution, which splits the input into different groups and processes each with regular convolution. The outputs from different groups are concatenated. ShuffleNet also introduces channel shuffle to enable cross talks between channels from different group. ShuffleNetV2 further increases the model efficiency by introducing channel-split operator which split feature channels into branches and concatenate them after convolution.

The base CNN is used for feature extraction. We utlize these two compact CNN designs as the base CNN by removing their last classification layers. Specifically, we remove the last two layers (one dropout layer and one fully-connected layer) for MobileNetV2, and remove the last one layer (fully-connected layer) for ShuffleNetV2.

4.2.3. Loss Function

The softmax cross-entropy loss has been commonly used for CNN training (Goodfellow et al., 2016; Murphy, 2012). For an input sample x with class label y, assuming that the predicted output from the model for all classes is $z = [z_1, z_2, ..., z_C]$, where C is the total number of classes. The softmax cross-entropy (CE) loss for this sample is defined as:

$$\mathrm{CE}(\mathrm{z}, y) = -\log(\frac{\exp(z_y)}{\sum_{j=1}^{C}\exp(z_j)}) \tag{1}$$

Considering the problem of class imbalance in the training data, the network is trained with a class-balanced softmax cross-entropy loss (Cui et al., 2019), which re-weights loss inversely with the effective number of samples per class. The class-balanced softmax cross-entropy (CB) loss for this sample is defined as:

$$\mathrm{CB}(\mathrm{z}, y) = -\frac{1-\beta}{1-\beta^{n_y}}\log(\frac{\exp(z_y)}{\sum_{j=1}^{C}\exp(z_j)}) \tag{2}$$

where n_y is the number of training samples in class y and $\frac{1-\beta}{1-\beta^{n_y}}$ is the weighting term for the loss function, with hyperparameter $\beta \in [0, 1)$.

4.3. Gross Blood Detection

We use perceptual color quantization to address the challenge of detecting blood in stool and classifying blood color (which ranges from bright red to black) in images with different size and illumination. In preliminary study we evaluated color thresholding to classify images containing blood, but this approach was found non-selective. This may be due to two reasons: first, blood presence in stool has a heterogeneous presentation, from mixed with brown stool, to blood clot, or in stool colors which are dark red or black; Second, healthy stools of reddish-brown uniform have a red color component. We therefore adopted machine-learning techniques and propose a blood detection approach that consists of 3 steps: 1) color quantization; (2) color feature selection; (3) blood detection by machine learning algorithm.

4.3.1. Perceptual Color Quantization

We use perceptual color quantization (Crandall and Luo, 2004), which employs the CIE LAB color space (Giorgianni and Madden, 1997) and the standard ISCC-NBS Color Names Dictionary (Kelly and Judd, 1976). The CIE LAB color space was designed to approximate human vision. The ISCC-NBS system defines 267 standard color partitions. Each color partition has a standard color name and a centroid color. The color names are designed as basic colors with one or more adjectives, such as "Vivid Red" and "Light Grayish Yellowish Brown". For an input image I, each pixel is assigned the closest ISCC-NBS centroid color based on the Euclidean distance in CIE LAB space.

4.3.2. Color Feature Selection

The purpose of color feature selection is to carefully select most representative features for comparing the blood-stool images to normal-stool images. This choice can significantly affect the performance of the subsequent steps in gross blood detection. We use the frequency of 'red' and 'black' colors as the features. Here, 'red' and 'black' colors can be defined as the ISCC-NBS names that contain the strings 'red' and 'black'.

We use Mutual Information (MI) for feature selection because MI is able to capture both linear and non–linear dependencies and is invariant under invertible and differentiable transformations in the feature space (Beraha et al., 2019). Let F be the full set of features and Y be the target variable. The mutual information MI between a single feature input and the target is defined as:

$$MI(F_i; Y) = K(p(f_i; y) \| p(f_i)p(y)) \tag{3}$$

where $p(f_i; y)$ is the joint probability density function and $p(f_i)$ and $p(y)$ are marginal density functions of feature f_i and label y. Note that $K(p(x)\|q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)}$ refers to the K-L divergence of two probability distributions p and q. A greedy search is performed to select the desired numbers of features, as shown in Algorithm 1.

Algorithm 1: MI-based feature selection
Input: F: set of features;
y: labels;
n: number of features to select;
Output: S: set of selected features;

$S = \emptyset$;
while $n > 0$ **do**
  $f_{MI} = \arg\max_{f_i \in F-S}(MI(f_i; y))$;
  $F = F - f_{MI}$;
  $S = S + f_{MI}$;
  $n = n - 1$;

4.3.3. Blood Detection

For blood detection, we investigate two machine-learning techniques, namely DT and k-nearest neighbors (KNN). An advantage of using DT and KNN is that the analysis results are easy to interpret and explain, which facilitates interactions with the medical team.

A DT is a tree-like model that consists of two types of nodes, leaf (terminal) nodes and decision (internal) nodes (Quinlan, 1986). Leaf nodes refer to the nodes that do not branch and contain prediction information and each leaf node holds a class label. Decision nodes refer to the nodes that can branch to multiple child nodes or leaf nodes. Class labels are denoted by the elements of the set $Y = \{Y_0, Y_1\}$, where Y_0 indicate images of healthy stool, and Y_1 indicate images containing blood content. These class labels are the leaf nodes in the decision tree. Our selected color features $C = \{C_0, C_1, ...C_i\}$ are encoded in the decision nodes.

KNN is a distance-based technique (Altman, 1992). Assuming that the training dataset is C_{train} and the test dataset is C_{test}. Then for each instance C^j_{test} in C_{test}, we calculate its distance to all instances C^i_{train} in C_{train}. The list of distances obtained in this manner is sorted in ascending order, and the most common label among the first k elements in the sorted distance list is assigned to instance C^j_{test}

5. Experiments and Results

Experiments were conducted to evaluate the effectiveness of the proposed approach for classifying stool form and detecting images of stool that contains blood. We preprocessed the stool images by cropping them to remove noise, e.g., due to the toilet seat.

5.1. Results on Stool Form Classification

The balanced accuracy (BA) metric was utilized to evaluate different model architectures trained with various loss functions. The BA metric is defined as the average recall obtained on each class Brodersen et al. (2010), as shown in Equation (4).

$$BA = \frac{1}{n}\sum_{i=1}^{n}\frac{tp_i}{tp_i + fn_i} \times 100\% \tag{4}$$

where n is the number of classes, and tp_i and fn_i are the number of true positive and false negative predictions for class i, respectively.

5.1.1. Hierarchical Architecture

The training process for the hierarchical CNN includes three steps. We first initialize the base CNN with pretraining on ImageNet (Deng et al., 2009). After initialization, we train the coarse-grained classifier and the base CNN together over three consolidated categories. In the last step, the base CNN is kept fixed and fine-grained classifiers are trained over seven BSFS scales. We utilized 552 images annotated by two gastroenterologists for testing and 2,723 images rated by one of them for training. We resized the images to 224×224 pixels.

We used Pytorch (Paszke et al., 2019) to implement and train the CNN using stochastic gradient descent with momentum. Experiments were executed on a Linux platform integrated with a 11 GB-memory GPU (Nvidia GeForce RTX 2080 Ti). The training was performed with mini-batches of size 14. In each batch, the numbers of samples for each class were restricted to be the same. We used the best BA during the training process as the quality metric.

We consider two compact CNN designs described in Section 4.2 and two traditional CNN designs, namely, VGG16 (Simonyan and Zisserman, 2015), ResNet50 (He et al., 2016) as the base CNN in the hierarchical architecture. Furthermore, we train the models with flat and hierarchical architectures and with both cross-entropy loss and class-balanced cross-entropy loss for comparisons. We evaluate the performance of coarse-grained classification over three consolidated categories (constipated, normal, and loose) and fine-grained classification over the BSFS scale with seven values. The results are shown in Table 2 and summarized as follows:

1. For all models except ShuffleNetV2, CB loss improves the performance in terms of the BA metric. One possible explanation of the performance decrease in ShuffleNetV2 with CB loss is that its more complex architecture and small size lead sensitivity to the re-weighting strategy in CB loss.
2. Hierarchical architectures improve the performance of both coarse-grained and fine-grained classification in terms of the BA metric for ResNet50 as well as for MobileNetV2. The hierarchical architecture brings slight increase in required memory and $FLOPs$ for inferencing, because it has three more classifiers than the flat architecture.
3. Hierarchical architectures with MobileNetV2 as the base CNN achieve the best performance (81.66% BA in coarse-grained classification and 54.58% in fine-grained classification). Moreover, MobileNetV2 only requires 0.35 $GFLOPs$ for inferencing and the memory requirement is only 15.6 MB .

Table 2: Balanced accuracy of flat and hierarchical architectures with various base CNN designs on the stool image dataset.

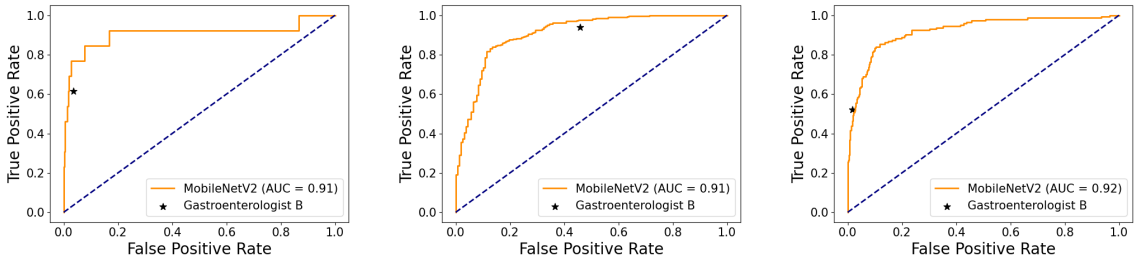
Models	Architecture	Coarse-grained		Fine-grained		$FLOPs$	Memory Required
		CB Softmax	Softmax	CB Softmax	Softmax		
VGG16	Hierarchical	73.50%	65.78%	50.03%	37.58%	15.39 G	61.5 MB
	Flat	75.17%	65.83%	43.41%	39.70%	15.38 G	59.5 MB
ResNet50	Hierarchical	80.15%	74.66%	51.32%	44.75%	4.16 G	104.8 MB
	Flat	71.92%	62.95%	41.33%	39.36%	4.13 G	96.9 MB
MobileNetV2	Hierarchical	81.66%	69.61%	54.58%	42.10%	0.35 G	15.6 MB
	Flat	74.55%	61.45%	44.25%	41.01%	0.33 G	10.7 MB
ShuffleNetV2	Hierarchical	33.33%	74.12%	22.86%	43.59%	0.17 G	10.4 MB
	Flat	47.60%	64.22%	26.53%	36.59%	0.16 G	6.5 MB

5.1.2. Prediction Analysis

As illustrated in section 5.1.1, MobileNetV2 outperforms ShuffleNetV2, VGG16 and ResNet50 in terms of the metrics of BA and computational costs ($FLOPs$ and required memory).

We first analyze the performance of MobileNetV2 for the task of coarse-grained classification over three consolidated categories. Figure 5 shows the receiver operating characteristic (ROC) analysis. Here, ROCs were reduced to a dichotomous classification based on the one versus rest approach, treating the corresponding class as positive and all of the other classes as negative. Table 3 shows the analysis based on the metrics of sensitivity and specificity and MobileNetV2 achieves satisfactory performance in coarse-grained classification.

Our model for stool form coarse classification achieves 81.7% BA, superior to values (73.9%) reported by (Hachuel et al., 2019), which was trained on a smaller dataset and also did not correct for class imbalance resulting in the inability to predict abnormal classes. Our model accuracy as measured by AUC=0.91 is comparable with AUC ranging from 0.89 to 0.98 obtained by (Park et al., 2020) using computationally heavy GoogleNet Inception v3 CNN architecture while we achieve this accuracy with a very efficient model suitable for deployment in a microcomputer.



(*a*) Constipated (*b*) Normal (*c*) Loose

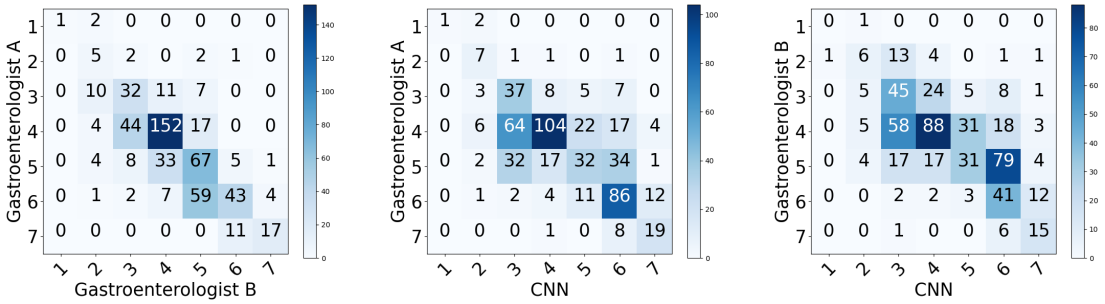
Figure 5: ROC analysis for the prediction results of our hierarchical architecture with MobileNetV2 as the base CNN. Using the labels assigned by the first gastroenterologist (A) as the ground truth, all AUCs are greater than 0.91 and comparable with the second gastroenterologist (B)

Table 3: Sensitivity and specificity of our hierarchical architecture with MobileNetV2 as the base CNN. C, N and L stand for Constipated, Normal and Loose respectively.

	Coarse-grained			Fine-grained						
	C	N	L	1	2	3	4	5	6	7
Sensitivity	0.77	0.81	0.87	0.33	0.70	0.62	0.48	0.27	0.74	0.68
Specificity	0.98	0.87	0.84	1.00	0.97	0.80	0.91	0.91	0.85	0.97

We further analyze the performance of our model for the task of fine-grained classification over the BSFS scale with seven values. Figure 6 shows the confusion matrices, which describe the classification conformance among our CNN model and the two gastroenterologists regarding the BSFS. The agreement between our CNN model and Gastroenterologist A is measured by Cohen's Kappa statistical metric and has the value of $k = 0.388$, which is close to the agreement between the two gastroenterologists ($k = 0.435$, as discussed in Section 4.1). The agreement between the CNN and Gastroenterologist B has $k = 0.263$,

and this can be explained by the fact that in 2,723 training images, 1,928 were annotated by Gastroenterologist A and only 248 were annotated by Gastroenterologist B. The analysis based on the metrics of sensitivity and specificity is shown in Table 3. For BSFS 1, the metric is not meaningful because we have only three examples for BSFS 1. The reason behind the poor performance for BSFS 4 and 5 is that these two scales are often misclassified with each other, as shown in Figure 6.



(a) (b) (c)

Figure 6: Confusion matrices for comparison between the stool classifications made by (a) the two gastroenterologists, (b) Gastroenterologist A and and the CNN, (c) Gastroenterologist B and and the CNN. The values on the axes indicate the BSFS.

5.2. Results on Detecting Stool Images with Blood Content

Experiments were performed on two datasets, namely Dataset A with 1798 online images and Dataset B with 561 crowdsourcing images. Relevant information about the datasets is provided in Table 4. We used a m-fold ($m = 5$) cross validation method to evaluate the performance of the MI-based feature-selection method and the DT classifier. A m-fold cross validation method randomly partitions experimental dataset into m groups. In each round of an experiment, one group is regarded as the test dataset while all the other groups are used for training.

Table 4: Class distribution in the two datasets used for gross blood detection.

	Dataset A	Dataset B
Blood Content	70	29
Healthy Stool	1728	527
Total	1798	556

Table 5 shows the balanced accuracy for the two different classifiers. The DT classifier provides 74.64% BA on Dataset B and 62.82% BA on Dataset A. The relatively low performance on Dataset A can be explained by the fact that the collected online images tend to have lower resolution and this can introduce undesirable noise in the training process. Also, DT consistently outperforms KNN on both two datasets. One possible explanation is that KNN is more sensitive to class imbalance. Prior work on classification of stool color images (StoolNet (Yang et al., 2019))) used a shallow (2-layers only) convolutional neural

network but did not report on blood detection nor red color accuracy so we are unable to benchmark our results.

Table 5: Balanced accuracy of two classifiers for blood detection on two stool image datasets.

	Dataset A	Dataset B
KNN	58.45%	73.00%
DT	62.82%	74.64%

6. Limitations and Discussion

The development of an ML-AI program to automatically classify stool images for form (Bristol scale) and the presence of gross blood requires a large number of annotated photos of stool in a toilet. There was no publicly available database available to us, therefore, we developed our own with over 3000 images. A limitation of this approach is that the photos had no clinical data associated with them, and, while they spanned the full spectrum of the Bristol scale, the representation of associated gastrointestinal conditions or symptoms unknown. Additionally, while the use of the Bristol scale helps standardize stool evaluation, there remains some variability in assessment even among gastrointestinal specialists. However, the agreement between the two gastroenterologists in this study was satisfactory.

We envision that with the future deployment of the Smart Toilet hardware prototype for use by human subjects, we will be able to collect time-series data from individual subjects. We expect that stool image data collection from the controlled environment will result in more consistent lighting and even background that will enhance the model accuracy. A smart toilet with machine-learning image analysis capability to determine stool frequency, form, presence of visible blood will provide important diagnostic data that can help identify specific food intolerance (e.g., foods that exacerbate IBS or chronic diarrhea) and effects of medication (e.g., medications taken for diarrhea or constipation), and can trigger timely evaluation (e.g., IBD flare with bloody diarrhea). Additionally, ongoing development of the technology include stool specimen sampling for biochemical marker analysis that will provide highly specific disease data. We envision that the Smart Toilet time-series data collected from individuals will be integrated with machine learning predictive models and provide a valuable diagnostic and surveillance tool for GI, infectious disease, and other specialties.

7. Conclusion

We have described an automated technique for stool classification and gross blood detection using a combination of a Smart Toilet and machine learning. We have developed a comprehensive stool image dataset for assessing the classification approach. We have utilized hierarchical and compact CNN architectures that can be used for stool image analysis in a resource-limited computational environment. Specifically, we showed that the hierarchical CNN based on MobileNetV2, trained with class-balanced softmax entropy loss, can achieve a balanced accuracy of 81.66% in coarse-grained classification and 54.58% in fine-grained

classification, with the memory requirement of only 15.6 MB and 0.35 $GFLOPs$ for inferencing. For detecting stool images with blood content, we obtained a balanced accuracy of 74.64% using perceptual color quantization coupled with mutual information to optimize color-feature space and using DT as the classifier. Our results open up an interesting new research direction on privacy-preserving and real-time stool classification for single time-point and longitudinal health assessment.

8. Acknowledgements

his research was funded in part by the Duke Bass Connection Program. We thank Ms. Mara Shurgot for helping with the recruitment of participants through social media.

9. Data Availability Statement

The data of this study are available to other researchers upon request to sonia.grego@duke.edu. Fulfillment of the request will be subject to limitations due to ethical restrictions.

References

N. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

J. Bae and H. Lee. User health information analysis with a urine and feces separable smart toilet system. *IEEE Access*, 6:78751–78765, 2018.

M. Banerjee, M. Capozzoli, L. McSweeney, and D. Sinha. Beyond kappa: a review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1):3–23, 1999.

M. Beraha, A. Metelli, M. Papini, A. Tirinzoni, and M. Restelli. Feature selection via mutual information: New theoretical insights. In *International Joint Conference on Neural Networks*, 2019.

K. Brodersen, C. Ong, K. Stephan, and J. Buhmann. The balanced accuracy and its posterior distribution. In *International Conference on Pattern Recognition*, 2010.

N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357, June 2002.

Y. Cheng, D. Wang, P. Zhou, and T. Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

F. Collins and H. Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

N. Conn, K. Schwarz, and D. Borkholder. In-home cardiovascular monitoring system for heart failure: comparative study. *JMIR mHealth and uHealth*, 7(1), 2019.

D. Crandall and J. Luo. Robust color object detection using spatial-color joint probability functions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

Y. Cui, M. Jia, T. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

E. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, 2014.

L. Gatys, A. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2015.

P. Ghosh, D. Bhattacharjee, and M. Nasipuri. Intelligent toilet system for non-invasive estimation of blood-sugar level from urine. *IRBM*, 41(2):94 – 105, 2020.

E. Giorgianni and T. Madden. *Digital Color Management: Encoding Solutions*. Addison-Wesley, Reading, MA, 1997.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

D. Hachuel, A. Jha, D. Estrin, A. Martinez, K. Staller, and C. Velez. Augmenting gastrointestinal health: a deep learning approach to human stool recognition and characterization in macroscopic images. *arXiv preprint arXiv:1903.10578*, 2019.

M. Halicek, G. Lu, J. Little, X. Wang, M. Patel, C. Griffith, M. El-Deiry, A. Chen, and B. Fei. Deep convolutional neural networks for classifying head and neck cancer using hyperspectral imaging. *Journal of Biomedical Optics*, 22(6):1–4, 2017.

D. Hall, D. Crismon, J. Larsen, K. Campbell, J. Reynolds, and J. Blake. Us20200225121a1: Toilet equipped to provide fecal analysis, 2020.

E. Halmos, J. Biesiekierski, E. Newnham, R. Burgell, J. Muir, and P. Gibson. Inaccuracy of patient-reported descriptions of and satisfaction with bowel actions in irritable bowel syndrome. *Neurogastroenterology & Motility*, 30(2), 2018.

S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural networks. In *Advances in Neural Information Processing Systems*, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

F. Iandola, M. Moskewicz, K. Ashraf, S. Han, W. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1 mb model size. In *International Conference on Learning Representations*, 2018.

K. Kelly and D. Judd. *Color Universal Language and Dictionary of Names.* National Bureau of Standards, 1976.

H. Kim and D. Allen. Using digital filters to obtain accurate trended urine glucose levels from toilet-deployable near-infrared spectrometers. *Journal of Analytical & Bioanalytical Techniques*, 7:338, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Labelbox. https://labelbox.com, 2019.

S. Lewis and K. Heaton. Stool form scale as a useful guide to intestinal transit time. *Scandinavian Journal of Gastroenterology*, 32(9):920–924, 1997.

N. Ma, X. Zhang, H. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *European Conference on Computer Vision (ECCV)*, 2018.

H. Majid, P. Emery, and K. Whelan. Definitions, attitudes, and management practices in relation to diarrhea during enteral nutrition: a survey of patients, nurses, and dietitians. *Nutrition in Clinical Practice*, 27:252–260, 2012.

A. Markland, O. Palsson, P. Goode, K. Burgio, J. Busby-Whitehead, and W. Whitehead. Association of low dietary intake of fiber and liquids with constipation: evidence from the national health and nutrition examination survey. *The American Journal of Gastroenterology*, 108(5):796, 2013.

K. Murphy. *Machine learning: a probabilistic perspective.* MIT Press, 2012.

V. Murthy, V. Singh, T. Chen, R. Manmatha, and D. Comaniciu. Deep decision network for multi-class image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

H. Ohno, H. Murakami, K. Tanisawa, K. Konishi, and M. Miyachi. Validity of an observational assessment tool for multifaceted evaluation of faecal condition. *Scientific Reports*, 9(1):1–9, 2019.

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

S. Park, D. Won, B. Lee, D. Escobedo, A. Esteva, A. Aalipour, T. Ge, J. Kim, S. Suh, E. Choi, A. Lozano, C. Yao, S. Bodapati, F. Achterberg, J. Kim, H. Park, Y. Choi, W. Kim, J. Yu, A. Bhatt, J. Lee, R. Spitler, S. Wang, and S. Gambhir. A mountable toilet system for personalized health monitoring via the analysis of excreta. *Nature Biomedical Engineering*, 4:624–635, 2020.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

A. Peery, S. Crockett, C. Murphy, J. Lund, E. Dellon, J. Williams, E. Jensen, N. Shaheen, A. Barritt, S. Lieber, B. Kochar, E. Barnes, Y. Fan, V. Pate, J. Galanko, T. Baron, and R. Sandler. Burden and cost of gastrointestinal, liver, and pancreatic diseases in the united states: Update 2018. *Gastroenterology*, 156(1):254–272.e11, 2019.

R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, Mar 1986.

M. Ra, M. S. Muhammad, C. Lim, S. Han, C. Jung, and W. Kim. Smartphone-based point-of-care urinalysis under variable illumination. *IEEE Journal of Translational Engineering in Health and Medicine*, 6:1–11, 2018.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Y. Seo and K. Shin. Hierarchical convolutional neural networks for fashion image classification. *Expert Systems with Applications*, 116:328–339, 2019.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

S. Tanaka, K. Yabunaka, M. Matsumoto, N. Tamai, H. Noguchi, M. Yoshida, G. Nakagami, J. Sugama, and H. Sanada. Fecal distribution changes using colorectal ultrasonography in older people with physical and cognitive impairment living in long-term care facilities: a longitudinal observational study. *Healthcare*, 6(2):55, 2018.

W. Tsang, Y. Liong, L. Raman, Z. Wai, D. Consigliere, and E. Chiong. Validation of the toto flowsky uroflowmetry device. *European Urology Supplements*, 16(3):e1959–e1960, 2017.

C Wald. Diagnostics: A flow of information. *Nature*, 551, 2017.

Y. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, 2017.

Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu. Hd-cnn: Hierarchical deep convolutional neural network for large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

Z. Yang, L. Leng, and B. Kim. Stoolnet for color classification of stool medical images. *Electronics*, 8:1464, 2019.

B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *IEEE International Conference on Data Mining*, 2003.

X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: an extremely efficient convolutional neural network for mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

G. Zuckerman, D. Trellis, T. Sherman, and R. Clouse. An objective measure of stool color for differentiating upper from lower gastrointestinal bleeding. *Digestive Diseases and Sciences*, 40:1614–1621, 2005.