Contactless Oxygen Saturation Detection Based on Face Analysis: An Approach and Case Study

Batol Hamoud, Walaa Othman, Nikolay Shilov
SPC RAS
Saint-Petersburg, Russia
{bkhamud, walaa_othman}@itmo.ru, nick@iias.spb.su

Alexey Kashevnik
ITMO University
Saint-Petersburg, Russia
Petrozavodsk State University, Petrozavodsk, Russia
alexey.kashevnik@iias.spb.su

Abstract—**One of the most essential physiological indicators for a human health is oxygen saturation level (SpO2). It is the primary determinant of how efficiently the body transfers oxygen from the lungs to blood cells. SpO2 is typically measured with a pulse oximeter, however, non-contact SpO2 estimate approaches based on face or hand videos have gained popularity in recent years. In this paper, we proposed a novel approach based on machine learning concepts to estimate SpO2 using facial videos. Our approach includes exploring several pre-trained convolutional neural networks (CNN) models to extract features from the consecutive images of different regions of interest (ROI), followed by the training of the XGBoost Regressor model, which in turn predicts SpO2 for three different test sets included in our research. We managed to determine the best three models through multiple stages of our testing process, which took into account three metrics: mean absolute error (MAE), Pearson's correlation coefficient, and the shape of the predicted samples distribution. However, our final models achieved contactless estimations of SpO2 with decent accuracy and high performance according to the results of the testing process (MAE of 1.17 and 0.84 when testing the models using VIPL-HR and UBFC-RPPG datasets, respectively).**

I. Introduction

Oxygen saturation level is a vital indicator for determining blood oxygen content and oxygen delivery. It measures the ratio of oxygenated hemoglobin (HbO2) to total hemoglobin, which reflects the amount of oxygen supply in the blood [1]. For patients who suffer from medical problems that can lower the amount of oxygen in the blood, measuring oxygen saturation is very crucial. These conditions include asthma, lung cancer, anemia, chronic obstructive pulmonary disease (COPD), pneumonia, chronic bronchitis, and other cardiopulmonary problems [1].

Pulse oximeters are typically used to assess SpO2 in a simple, painless, and non-invasive procedure that involves placing a probe on the fingertip or earlobe to detect the oxygen saturation level indirectly. Contact-based pulse oximeters, on the other hand, can be uncomfortable, and are not ideal or suitable for continuous monitoring. Non-contact SpO2 estimation techniques utilizing hand or facial recordings have recently drawn greater interest. These video-based techniques offer remote SpO2 monitoring in a more flexible, convenient, and simple way. Video-based SpO2 estimation methods are classified according to the type of camera used: either special cameras designed to capture certain wavelength bands or red/green/blue (RGB) cameras such as webcams and smartphone cameras [2]. Although special cameras can record wavelengths adequate for SpO2 estimation, they are not commonly employed. As a result, the approach of using RGB camera to estimate SpO2 is preferred. Earlier RGB camera-based approaches assessed SpO2 from facial videos using the ratio-of-ratios (RoR) [2], a principle similar to that implemented in pulse oximeters. The RoR approach computes the ratio of direct current (DC) to alternating current (AC) components of red and blue channel signals obtained from videos. Researchers aimed to enhance this method by engaging advanced algorithms such as Singular Value Decomposition (SVD) [3] or particular filters for more reliable extraction of the DC and AC components. Several authors used machine learning and deep learning approaches in their work, such as CNN [4], regression models, etc. Therefore, we proposed our novel approach of accomplishing a contactless assessment of SpO2 based on facial videos recorded by smartphone camera in the natural lighting conditions.

Our approach does not need any other equipment such as monochrome cameras or special lighting devices. It was constructed by the implementation of transfer [5] and ensemble learning [6] together. The transfer learning part was presented by using pre-trained CNN models to extract the features out of the frames series, then the XGBoost Regressor model [7] was trained using these features to output one value which is the SpO2 for each second of the video. This methodology is contactless and comfortable. It depends only on image analysis, and takes into account the three channels to process as much information as possible. This method is convenient to use, and only requires a video that can be taken by the frontal camera of the subject's smartphone, which is simple and easy procedure in most scenarios.

The rest of the paper is structured as follows. Section II provides an overview of the proposed approaches and techniques. Section III is divided into two subsections: the first one contains all the details about our experiments on building, training, and evaluating our models, and the second one provides information about the datasets we used. The results of the testing procedure are presented in Section IV. Section V completes the paper by outlining the conclusion, challenges, and future scope of the proposed approach.

II. Related Work

In recent years, many researchers focused on implementing contactless methods to predict SpO2, especially after the pandemic, which emphasized the danger of physical contact with people or devices that are considered ideal environments for germs, viruses, and fungi to reproduce, spread and attack human beings. They proposed many innovative and revolutionary methods based on machine learning algorithms or processing of the extracted photoplethysmography (PPG) signal. Sometimes, both concepts were used; therefore, in this section we reviewed the recent approaches and algorithms proposed to achieve contactless estimation of SpO2.

The authors of [8] used 1D CNN for estimating SpO2 using videos of the participant's finger. PPG signals were extracted from RGB frames by averaging the pixel values from candidate regions of interest (ROIs) to obtain the final PPG signal, which was the weighted average of the signals extracted from different ROIs, where each weight was calculated by signal to noise ratio (SNR) of a given ROI. Next, this signal was processed using a modified SVD to make it more robust against large motion artifacts. The output signal was decomposed into band-pass and low-pass filtered versions which were interpolated and scaled to be fed into 1D CNN to estimate SpO2.

The authors of [2] proposed a method to estimate SpO2 from facial videos using CNN. To achieve this, they came up with two methods. The first one started with extracting the direct current (DC) and alternating current (AC) components by applying low-pass and band-pass filters on the spatiotemporal map gained from the RGB signals of facial videos. Next, the extracted DC and AC components were input to two ResNet18 networks which were fused via intermediate and late fusion to predict SpO2. The second method they proposed was an end-to-end model that predicts SpO2 directly from the spatiotemporal map by extracting the DC and the AC components via convolutional layers, besides two ResNet18 networks that forecasted SpO2 from the estimated AC and DC components by CNN. The authors of [2] customized the loss functions used, for instance, they engaged the negative correlation with the mean square error (MSE) between the true and the predicted values of SpO2 for the first model, whereas for the end-to-end model, they kept the previous loss function with adding the MSE between the DC and AC components estimated by the convolutional layers and the components extracted by the filters.

The authors of [9] extracted the PPG signals from three facial ROIs (the forehead and the cheeks). Signals were extracted by separating each ROI into three channels: Red, Green and Blue, then each channel was averaged over all pixels. These signals were processed and underwent Power spectral density (PSD) through Welch's method [10] to select the most informative ROI. The AC_{RED} and AC_{BLUE} components of the selected signal were computed as the standard deviations of the red and blue signals, while DC_{RED} and DC_{BLUE} components were computed as the mean of red and blue signals values. Finally, by using Equation. 1 , the authors managed to estimate SpO2 according to the empirical evaluation of A and B mentioned in [11]. However, this approach was tested later in [12] to verify its robustness against the movement of the subject‘s face during the recording process. This study was performed using the PURE dataset [13], and it concluded that there were no significant differences among the SpO2 measurements in presence of different slight head movements and the authors believed that this was achieved due to the ROIs tracking mechanism using Kanade-Lucas-Tomasi (KLT) method [14].

$$SpO2 = A - B\frac{AC_{RED}/DC_{RED}}{AC_{BLUE}/DC_{BLUE}} \tag{1}$$

Many proposed methods shared the same stages of extracting the PPG signal from the videos, whether they captured the face or some parts of the hand, like the finger, the palm, or the back of the hand and this was achieved by spatial averaging of the ROIs over the time. However, the authors of [15] proposed to feed this extracted PPG signals from videos of the palm or the back side of the hand into three different deep learning models predicting SpO2. They tried in the first model to combine the color channels first using several channel combination layers followed by extracting the temporal features using convolutional and max pooling layers, while in the second one, they reversed the procedure and started with extracting the features first, then the color channel mixing was performed. The final model consisted of convolutional and max pooling layers only to explore the possibility of interleaving the color channel mixing and temporal feature extraction steps.

As was mentioned, machine learning concepts were used in many pioneering approaches that accomplished many satisfactory SpO2 estimation without any physical contact. For instance, The authors of [16] came up with a new methodology to estimate three human physiological bio-signals: heart rate (HR), breathing rate (BR), and SpO2. Their approach consisted of extracting raw time-series bio-signal data from the green channel of facial videos. These time-series signals got processed later to estimate the aforementioned vital signs using three types of machine learning models: Multi-layer Perceptron Algorithm (MPA), Long Short-Term Memory Algorithm (LSTM) [17], and Extreme Gradient Boosting Algorithm (XGBoost).

Moving away from machine learning, the authors of [18] decided to record PPG signals alternately at two specific wavelengths (611 nm and 880 nm) using a complementary metal–oxide–semiconductor camera (CMOS) with trigger control to record the area around the mouth. The SpO2 got estimated by the ratio of absorbance using the AC and DC components of the PPG signals at these wavelengths. Another usage for monochrome cameras is what the authors of [11] proposed in their paper, where the PPG signals at wavelengths of 520 and 660 nm got captured using two monochrome charge-coupled device (CCD) cameras, each of which had narrow bandpass filters mounted to the lens. The AC com-

TABLE I. COMPARISON BETWEEN THE PROPOSED METHODS

Reference number	Advantages	Disadvantages
[2]	Constructive combination of two models to create powerful loss function with productive implementation for fusion process	Can be computationally expensive due to the usage of two/three convolutional neural networks
[8], [15]	Simple, straightforward with effective artifact removal method for [8]	Not suitable to use in many contexts, e.g., driving, because of the recording of a portion of the hand (finger, palm, or the back side)
[9], [12]	Effective ROI tracking system	Computationally complicated because of the tracking system and the process of choosing the most informative signal
[11]	Can be implemented immediately with no need to process the signal	The constants were chosen empirically according to specific lighting conditions, which may not work in other environments. In addition, there was no elimination of motion artifacts
[16]	Efficient combination/comparison between deep learning models	Only green channel was used with ignoring the information of other channels
[18]	The innovative usage of ROI and the wavelengths with straightforward implementation	Custom illumination synchronized with the frames acquisition is needed
[19]	Considering the biggest amount of information carried by the channels	It is computationally complicated regarding the usage of SVD to find the weights in the calibration process

ponents were calculated as the peak-to-peak values obtained from PPG after de-noising and bandpass filtering, while the DC components got computed as the average value of the PPG signals at corresponding periods. On the other hand, the authors of [19] used the ratio of the intensities measured from the green and red channels to present their methodology, which is based on multiple linear regression (MLR) to take into account the deviation of the RoR caused by the change in light scattering besides the original ratio of the AC and DC components. SVD was used to determine the suitable weights in the calibration process of the MLR model to estimate the SpO2. A comparison between the proposed methods was listed in Table I, where we mentioned the strengths and weaknesses of each approach mentioned above.

So, based on the proposed methods and algorithms analysis and by taking into account their drawbacks and downsides, which were listed in Table I, we realized the need of a novel approach that addresses the weaknesses of the presented related work. In addition, there is a substantial increase in interest in engaging machine learning and artificial intelligence in our daily life. We intended to propose a completely convenient approach, based on machine learning concepts, has low computational cost, and does not need any equipment attached to the subject's finger, special lighting conditions, or uncomfortable scenarios to record.

III. APPROACH AND DATASETS

This section includes two main parts: the proposed approach and the used datasets. In the Proposed Approach subsection, we mentioned the details of constructing our models from the basics, starting with extracting the ROIs and pre-processing the obtained frames, followed by using pre-trained CNN models to extract the features from the consuctive frames, on which XGBoost regressor models were trained using VIPL-HR [20], [21] dataset. Finally, the best models according to specific criteria were used for further testing. Regarding the Datasets subsection, we outlined the used datasets for training and testing our models. We used three datasets: VIPL-HR and UBFC-RPPG Dataset-1 [22] which provide SpO2 for each second, and the Operators dataset which was used to test the stability of the models.

A. Proposed Approach

As was mentioned, ensemble learning was the best option to be included in our work, regarding the nature of the oxygen saturation values to be mostly between 90-100. We decided to use the XGBoost regressor model in our approach since it can be used directly for regression predictive modeling problems

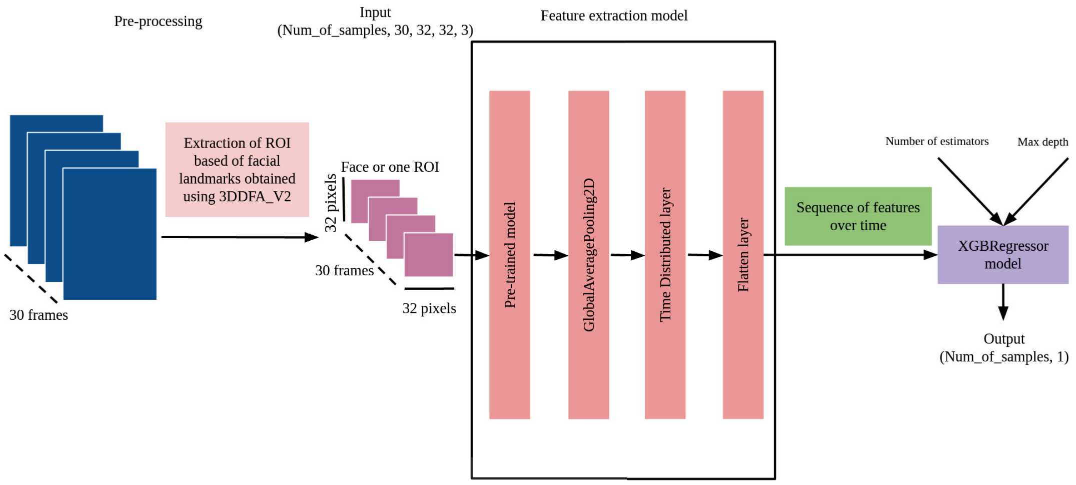


Fig. 1. The architecture of the proposed approach

that involve predicting a numerical value [23]. But to reach this point, we performed many stages to access the training phase, starting with detecting the face and cropping the ROIs, followed by pre-processing the obtained images before feeding them into our models that had the architecture shown in Fig. 1. In this subsection, we are going to talk about our approach step by step to achieve a contactless estimation of SpO2 for VIPL-HR dataset subjects. We should mention that the training and testing processes were run using NVIDIA T4 Tensor Core GPU provided by Google Colab.

1) Extracting the ROIs: At the beginning of the task, we needed to determine the best ROIs we can extract from the subject's face, since the proposed methods used different ROIs like forehead, cheeks, mouth area, or the lower part of the face, hence, we decided to extract the whole face, the forehead, and the cheeks out of the frames and implement our experiments to conclude which ROI is better to use for this task. We used 3D Dense Face Alignment (3DDFA) proposed in [24] for detecting the face and obtaining the facial landmarks that enabled us to crop the forehead, and the cheeks out of the consecutive images.

2) Pre-processing the images: The purpose of our work is to estimate SpO2 at each second, hence, we needed to feed 30 frames into our models for each estimation. We did not exclude any frame because we wanted to keep as much information as we can. We resized all the images of ROIs to 32x32x3, where the height and the width are equal to 32 and 3 is the number of the channels, we did not neglect any channel to preserve all the information provided by the changes of the pixels values over time. All the images underwent per-channel standardization of pixels by subtracting the mean from each pixel followed by division by the standard deviation.

3) Extracting the features: In our proposed approach that includes XGBoost Regressor, there was one huge obstacle, which was the fact that XGBoost Regressor requires structured or tabular datasets, so it can be used for classification and regression tasks [23]. On the other hand, we have images to input into the models, and they are considered nonstructural data. To get over this problem, we decided to extract the features out of each image in the sequence using a pre-trained model and stacking the output features over time, so each second of the video was represented by a stack of features extracted from 30 consecutive frames, so it can be considered as tabular data. However, there were many candidate pre-trained models to try in this stage, therefore, we tested the models, which have acceptable depth and number of parameters in addition to low time (ms) per inference step and high top 5 accuracy according to Keras official website [25]. The used models in this research were VGG16, VGG19, DenseNet169, Resnet50V2, EfficientNetV2B0, and EfficientNetV2B1. As shown in Fig. 1, the features extraction part consisted of 4 main blocks. It starts with feeding the images one by one to the pre-trained model using the default weights, which are imagenet weights. Next, the output of the pre-trained model is passed into the Global Average Pooling layer to reduce the computational cost. The operation continues to the Time Distributed layer. It is a wrapper that applies a layer to every temporal slice of the input, so the output of this layer would present the features extracted from 30 consecutive images. The process ends with a flatten layer to convert the output of each 30 frames (1 second) into a one-dimensional array, which is considered an acceptable input shape for the XGBoost Regressor.

4) Training the XGBoost Regressor model: XGBoost is a highly scalable decision tree ensemble based on gradient boosting. It minimizes a loss function to provide an additive expansion of the objective function [26]. Trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models [23]. However, XGBoost has many technological advantages, for instance, it provides more direct path to the lowest error, faster convergence with fewer steps, and simplified operations to enhance speed and reduce computational costs. These advantages motivated us to use it in our research, therefore, after extracting the features out of the consecutive frames of the train set, we fed them into the XGBoost Regressor, which required the number of estimators, and the depth of the decision trees. We chose the number of estimators and the depth from the intervals [50, 100] and [5,7], respectively. However, the models that achieved the best MAE for the different ROIs and pre-trained models were kept for further comparison to determine which models to test on UBFC-RPPG dataset-1 and Operators dataset.

5) Choosing the best models: Based on the results of testing the models using the test set from VIPL-HR dataset, we decided to consider three criteria to evaluate them to find out which models performed better than the others. We decided to pick, for each ROI, the model that had the lowest MAE, or the highest Pearson's correlation coefficient [27], or the model that predicted SpO2 with similar distribution to the test set distribution. The selected models were tested using the UBFC-RPPG dataset-1. However, we picked the top three models that achieved the lowest MAE and tested their performance on the Operator dataset to make sure that the models are well-grounded and able to estimate SpO2 with good accuracy and reasonable values even with different environment and dissimilar skin colors to what they were trained on.

B. Datasets

In this subsection, we included different datasets to train and test our models to verify their robustness against the changes in the lighting conditions or the subject's skin color. The first dataset is the VIPL-HR dataset, which we used to train and test our models using different ROIs cropped from the facial videos to determine the best models to be used in the further stage. The second dataset is the UBFC-RPPG dataset-1, which was used to test the best models from the previous step to prove that our models can estimate SpO2 regardless of the changes in lighting or the shades of the subject's facial skin. However, we tested the generated models using our Operators dataset, which does not provide the true SpO2, but the experiments were run to confirm the validity of the models by analyzing the output SpO2 to make sure that the models would predict values from the normal range (usually above 95% [8]), considering that the subjects are healthy, young, and sitting on a chair without any exhausting exercises that may affect on SpO2 values.

1) VIPL-HR Dataset: The VIPL dataset was developed by the Key Laboratory of Intelligent Information Processing of the Chinese Academy of Sciences. It contains nine scenarios recorded by three distinctive devices for 107 subjects. There are a total of 2,378 visible light videos (VIS) and 752 near-infrared (NIR) videos. This dataset provides the SpO2, HR, and blood volume pulse (BVP) signals for all the videos. Since we intended to facilitate the usage of our approach with available devices and usual lighting conditions, our research was limited to the videos recorded by the frontal camera of a HUAWEI P9 smartphone. The frame rate was 30 fps with a resolution of 1920*1080. The face area only was retained. The distribution of the samples in this dataset is shown in Fig. 2 which was taken from [16].

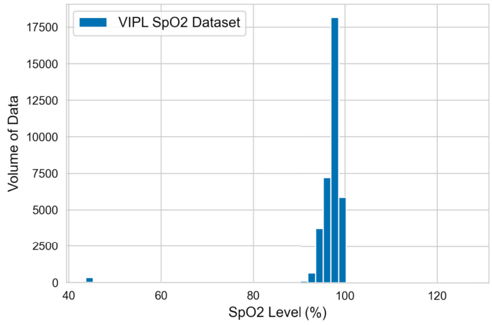


Fig. 2. Distribution of the samples in VIPL-HR dataset

Considering the huge dataset and the long time of processing and cropping the ROIs from all the videos, we included 55 subjects (2 or 3 videos for each) in our current research to have in total 162 videos. We split these videos into 136 videos for training and 26 videos regarding the testing process. However, we made sure that the distribution of our training set samples is similar to the distribution of the whole dataset samples. The distribution of our training set is shown in Fig. 3 In addition, Fig. 4 represents the distribution of the test set samples.

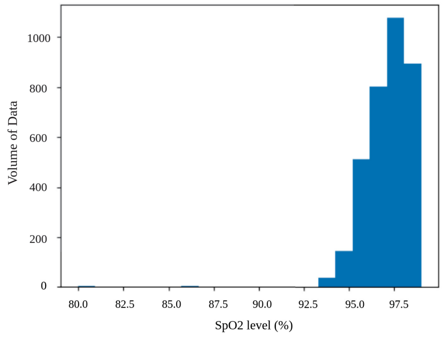


Fig. 3. Distribution of the samples in the training set

Majority of the samples' SpO2 values are concentrated between 95% to 98%, which was not helpful to use traditional deep learning terms like CNN and LSTM due to the high probability of creating models that generate a constant optimal value that achieves minimal error for all the samples. We run many experiments and implemented our proposed approach in [28], and our concerns did occur, hence, ensemble learning was used in our approach due to its flexibility with this kind of issues with datasets [29].

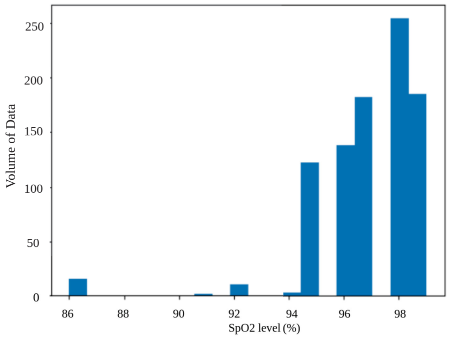


Fig. 4. Distribution of the samples in the test set

2) UBFC-RPPG Dataset-1: This dataset was created using a Logitech C920 HD Pro webcam at around 30 fps and 640x480 resolution in uncompressed 8-bit RGB format. It contains eight videos for six different subjects with distinctive skin colors, which we found useful to test our models subjectively. A basic Matlab implementation was provided to read ground truth data acquired with a pulse oximeter since the true values were saved in files of format .xmp. The whole 8 videos were used in the second stage of the testing process after training and testing the models using the VIPL-HR dataset. As we mentioned before, we believed that it would be a challenge for the models to be tested using videos with different lighting circumstances and new shades of facial skin since the majority of VIPL-HR dataset subjects had light pale skin color.

3) Operator Dataset: We present our dataset, which includes 60 videos of operators sitting in front of their computers, either reading or working. The videos were shot on almost consecutive days, twice or three times a day. They cover the morning and evening hours (sometimes, the afternoon period got included). The videos last from 16 to 20 minutes and have a resolution of 640x480. The fps is 30 for all the videos except one video has an fps of 7, and another has 15 frames per second. We handled the 7 fps video by duplicating the first and last frame from each second five times and four times for the other five frames. The other video was operated by repeating each frame one time so we get 30 frames in a second instead of 15.

IV. Results

In this section, we included the results obtained through the testing process, which consisted of three stages, starting with testing the models using the test set from VIPL-HR dataset, followed by using UBFC-RPPG dataset-1 to test the best models from previous step and determining the top three models according to the lowest MAE to test them on Operators dataset in the final stage to analyse the robustness of our models.

1) Testing the models using the VIPL-HR dataset: Several XGBoost Regressor models were trained on the extracted features from the training set obtained from the VIPL-HR dataset. Each model was trained using one of the ROIs, which were the face, the forehead, the left cheek, and the right cheek. In addition, we used in each model one pre-trained model from the following list: VGG16, VGG19, DenseNet169, Resnet50V2, EfficientNetV2B0, and EfficientNetV2B1. We tested the models using three criteria, MAE, Pearson's correlation coefficient, and the shape of the estimated SpO2 distribution. We listed the MAE and Pearson's correlation coefficients of the models in Table II and Table III, respectively.

TABLE II. Results of testing the models using VIPL-HR test set / MAE

Model	Face	Forehead	Left cheek	Right cheek
VGG16	1.21%	1.28%	1.30%	1.28%
VGG19	**1.17%**	**1.22%**	**1.29%**	**1.27%**
Resnet50V2	1.27%	1.27%	1.40%	1.43%
DenseNet169	1.36%	1.29%	1.47%	1.37%
EfficientNetV2B0	1.65%	1.75%	1.29%	1.71%
EfficientNetV2B1	1.68%	1.72%	1.35%	1.75%

TABLE III. Results of testing the models using VIPL-HR test set /Pearson's Correlation Coefficient

Model	Face	Forehead	Left cheek	Right cheek
VGG16	0.24	0.22	0.17	0.40
VGG19	0.016	0.16	-0.007	**0.64**
Resnet50V2	**0.60**	0.009	0.38	0.47
DenseNet169	0.15	0.28	0.17	0.35
EfficientNetV2B0	0.001	**0.36**	0.23	-0.038
EfficientNetV2B1	0.13	0.31	0.20	0.15

For each ROI, we marked the best model according to the standards. Regarding Table III, all the marked models achieved p-value$<$0.05, which indicates that there is a relationship between the real and expected values. We should note that no model with left cheek as ROI could achieve p-value$<$0.05, hence, we did not include any model from left cheek column in the following experiments. However, we also took the averaged estimations of left cheek and right cheek models that share the same pre-trained model to explore if using multiple ROIs would improve the results, then we tested the output according to the same standards mentioned in Section III-A5. We also implemented the same procedure for left cheek, right cheek, and forehead models. The results of this operation are listed in Table IV and Table V. Similarly to the previous tables, we marked the best models according to the MAE and Pearson's correlation coefficient.

TABLE IV. Results of testing the averaged models using VIPL-HR test set / MAE

Model	Cheeks	Cheeks + Forehead
VGG16	1.27%	1.24%
VGG19	**1.26%**	**1.20%**
Resnet50V2	1.38%	1.31%
DenseNet169	1.40%	1.34%
EfficientNetV2B0	1.46%	1.53%
EfficientNetV2B1	1.52%	1.55%

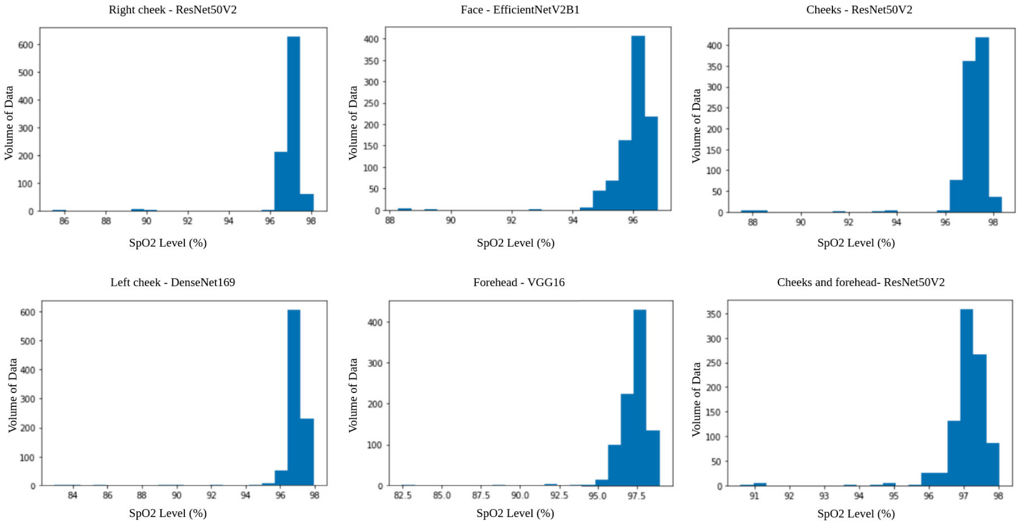


Fig. 5. The Distributions of the best models have similar distribution to the VIPL-HR test set distribution

TABLE V. RESULTS OF TESTING THE AVERAGED MODELS USING VIPL-HR TEST SET / PEARSON'S CORRELATION COEFFICIENT

Model	Cheeks	Cheeks + Forehead
VGG16	0.39	0.41
VGG19	0.2	0.23
Resnet50V2	**0.51**	**0.47**
DenseNet169	0.38	0.38
EfficientNetV2B0	0.06	0.21
EfficientNetV2B1	0.21	0.27

Regarding the best distributions, we decided to show the models that achieved the best results according to this criterion only in Fig. 5, rather than including the distribution of all the models, which would take much space.

2) Testing the models using the UBFC-RPPG dataset-1: From the previous stage, we obtained the best models that performed well according to one of the three standards. However, we tested these models using UBFC-RPPG dataset-1 to determine the best three models among the others to be tested on the Operator dataset. We only chose the models that had the lowest MAE to reduce the complexity in general since we need to test the candidate models from this stage on 60 videos in the next step. The results were shown in Table VI.

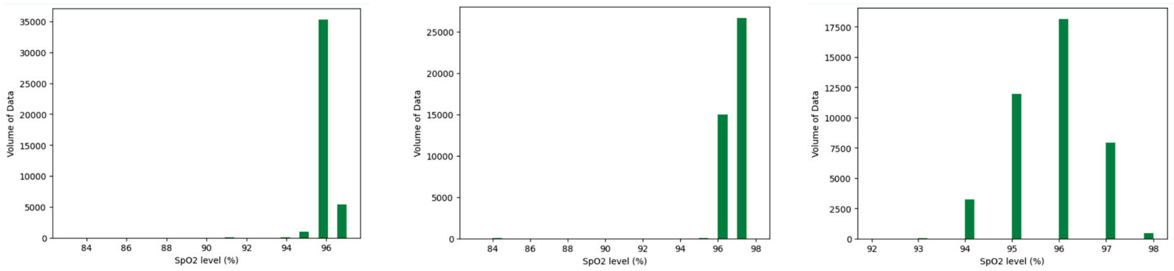
3) Testing the models using the Operators dataset: Our goal in testing the best models we obtained from the previous step on Operators dataset is to make sure that our models would predict plausible values even with the differences in the recording conditions and the subjects' skin shades. Fig. 6 shows the histogram of the estimated SpO2 values for the whole dataset by Left cheek–DenseNet169, Right cheek–ResNet50V2, and Face-VGG19 models. As we can see, the majority of the predicted values are concentrated above 95%, which looks normal and expected since the subjects are young, healthy, and sitting on a chair with no exhausting activities. By these results, we can say that Face–VGG19 model has achieved good performance on the three datasets used in this research, therefore, it is valid to be used to estimate SpO2 of people who are not able to check their Oxygen saturation level in the clinic for any reason.

TABLE VI. RESULTS OF TESTING THE BEST MODELS FROM PREVIOUS STAGE USING UBFC-RPPG DATASET-1/ MAE

Model	MAE
Right cheek - VGG19	1.1%
Right cheek - ResNet50V2	**0.97%**
Left cheek - DenseNet169	**0.89%**
Left cheek - VGG19	1.12%
Forehead - VGG19	1.23%
Forehead - VGG16	1.26%
Face - ResNet50V2	1.07%
Face - VGG19	**0.84%**
Face - EfficientNetV2B1	1.58%
Cheeks - VGG19	1.06%
Cheeks - ResNet50V2	1.08%
Cheeks + Forehead - VGG19	1.1%
Cheeks + Forehead - ResNet50V2	1.03%

V. DISCUSSION

This approach started with detecting the face and cropping the selected ROIs. We considered in our experiments four regions: face, forehead, left cheek, and right cheek. After obtaining the ROIs, we fed them into the pre-trained models after excluding the fully connected layers with preserving the default weights, since the goal of this process was to extract the spatial features, we believed that the ImageNet weights were qualified to cover this step, because of the diversity and size



Left cheek–DenseNet169

Right cheek–ResNet50V2

Face–VGG19

Fig. 6. The estimated SpO2 distributions by the best three models using the Operators dataset

of ImageNet, which enabled the models to learn from many variations that may include camera angles, lighting conditions, and so on [30].

After extracting the features of the consecutive frames of the train set, we organized them as time series presenting the features over each 1 second (30 frames). These arranged features were input to the XGBoost Regressor model, which required the number of estimators and the max depths of the learners. However, after many experiments, we picked the most accurate models according to their performance on the VIPL-HR test set. The standards we used to evaluate our models in this stage were MAE, Pearson's correlation coefficient, and the shape of the estimated SpO2 distribution. The selected models were tested further using the UBFC-RPPG dataset-1, which contained subjects with distinctive skin color, in addition to the changes in the lighting conditions. By testing our models on this dataset, we found out which models are more well-founded and reliable to use according to the lowest MAE. Finally, these candidate models were tested using our Operators dataset, which does not provide the true SpO2. Instead, we considered the performance on the latter dataset as proof that these models keep predicting reasonable and convincing values, regardless of the skin shade or the recording circumstances. Nevertheless, by taking into account the fact that the subjects of the Operators dataset did not include any health condition, we found that the final models performed in very convincing ways, which led us to consider all of them reliable to be employed for the task of contactless estimation of SpO2. However, the biggest challenge was that there were not enough samples that present SpO2 <85, which would create an obstacle for the models when the subject is suffering from a special health condition associated with low SpO2.

VI. Conclusion

In this paper, we introduced a novel, low-cost, and time-efficient approach to estimate SpO2 using only a smartphone camera. It combined transfer and ensemble learning and employed these two important machine learning concepts in order to accomplish an accurate estimation of SpO2 that requires only a recorded video for the facial area.

In the future, we would try to extend the training set with subjects/samples that have unfamiliar SpO2 values due to some medical conditions that the VIPL-HR dataset did not cover. In addition, our approach can be easily modified into an application for SpO2 monitoring anywhere and anytime people need it.

Acknowledgment

This work was supported by the Russian Science Foundation (project 18-71-10065).

References

[1] S. S. Dutta, "What is oxygen saturation?" Jun 2020. [Online]. Available: https://www.news-medical.net/health/What-is-Oxygen-Saturation.aspx

[2] Y. Akamatsu, Y. Onishi, and H. Imaoka, "Blood oxygen saturation estimation from facial video via dc and ac components of spatio-temporal map," *ArXiv*, vol. abs/2212.07116, 2022.

[3] V. Klema and A. Laub, "The singular value decomposition: Its computation and some applications," *IEEE Transactions on Automatic Control*, vol. 25, no. 2, pp. 164–176, 1980.

[4] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015. [Online]. Available: https://arxiv.org/abs/1511.08458

[5] R. Saha, "Transfer learning - a comparative analysis," 12 2018.

[6] M. Zadane, "Ensemble learning its methods in machine learning," Feb 2022. [Online]. Available: https://blog.knoldus.com/ensemble-learning-its-methods-in-machine-learning/

[7] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. [Online]. Available: https://doi.org/10.1145\%2F2939672.2939785

[8] X. Ding, D. Nassehi, and E. C. Larson, "Measuring oxygen saturation with smartphone cameras using convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2603–2610, 2019.

[9] G. Casalino, G. Castellano, and G. Zaza, "A mhealth solution for contact-less self-monitoring of blood oxygen saturation," in *2020 IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1–7.

[10] O. M. Solomon, "Psd computations using welch's method," 1991.

[11] L. Kong, Y. Zhao, L. Dong, Y. Jian, X. Jin, B. Li, Y. Feng, M. Liu, X. Liu, and H.-C. Wu, "Non-contact detection of oxygen saturation based on visible light imaging device using ambient light." *Optics express*, vol. 21 15, pp. 17 464–71, 2013.

[12] G. Casalino, G. Castellano, and G. Zaza, "Evaluating the robustness of a contact-less mhealth solution for personal and remote monitoring of blood oxygen saturation," *Journal of Ambient Intelligence and Humanized Computing*, 01 2022.

[13] R. Stricker, S. Müller, and H.-M. Gross, “Non-contact video-based pulse rate measurement on a mobile service robot,” vol. 2014, 08 2014, pp. 1056–1062.

[14] C. Tomasi and T. Kanade, “Detection and tracking of point features,” 1991.

[15] J. Mathew, X. Tian, C.-W. Wong, S. Ho, D. K. Milton, and M. Wu, “Remote blood oxygen estimation from videos using neural networks,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2023.

[16] I. Al-Zyoud, F. Laamarti, X. Ma, D. Tobón, and A. El Saddik, “Towards a machine learning-based digital twin for non-invasive human bio-signal fusion,” *Sensors*, vol. 22, no. 24, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/24/9747

[17] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[18] D. Shao, C. Liu, F. Tsow, Y. Yang, Z. Du, R. Iriya, H. Yu, and N. Tao, “Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system,” *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1091–1098, 2016.

[19] Z. Sun, Q. He, Y. Li, W. Wang, and R. Wang, “Robust non-contact peripheral oxygenationsaturation measurement using smartphoneenabled imaging photoplethysmography,” *Biomedical Optics Express*, vol. 12, 02 2021.

[20] X. Niu, S. Shan, H. Han, and X. Chen, “RhythmNet: End-to-end heart rate estimation from face via spatial-temporal representation,” *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2020. [Online]. Available: https://doi.org/10.1109\%2Ftip.2019.2947204

[21] X. Niu, H. Han, S. Shan, and X. Chen, “Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video,” 2018. [Online]. Available: https://arxiv.org/abs/1810.04927

[22] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, “Unsupervised skin tissue segmentation for remote photoplethysmography,” *Pattern Recognition Letters*, vol. 124, 10 2017.

[23] J. Brownlee, “Xgboost for regression,” Mar 2021. [Online]. Available: https://machinelearningmastery.com/xgboost-for-regression/

[24] J. Guo, X. Zhu, Y. Yang, Y. Fan, Z. Lei, and S. Li, *Towards Fast, Accurate and Stable 3D Dense Face Alignment*, 11 2020, pp. 152–168.

[25] K. Team, “Keras documentation: Keras applications.” [Online]. Available: https://keras.io/api/applications/

[26] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of xgboost,” 11 2019.

[27] L. Statistics, “Pearson product-moment correlation,” 2018. [Online]. Available: https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[28] B. Hamoud, A. Kashevnik, W. Othman, and N. Shilov, “Neural network model combination for video-based blood pressure estimation: New approach and evaluation,” *Sensors*, vol. 23, no. 4, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/4/1753

[29] C. Makhijani, “Advanced ensemble learning techniques,” Oct 2020. [Online]. Available: https://towardsdatascience.com/advanced-ensemble-learning-techniques-bf755e38cbfb

[30] s. gurumoorthyP, arvindpdmn, “Imagenet,” Jul 2019. [Online]. Available: https://devopedia.org/imagenet