This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

Evaluation of the Time Stability and Uniqueness in PPG based Biometric System

Dae Yon Hwang, *Student, IEEE,* Bilal Taha, *Student, IEEE,* Da Saem Lee, and Dimitrios Hatzinakos, *Fellow, IEEE*

***Abstract*—In this work, we demonstrates the feasibility of employing the biometric photoplethysmography (PPG) signal for human verification applications. The PPG signal has dominance in terms of accessibility and portability which makes its usage in many applications such as user access control very appealing. Therefore, we developed robust time-stable features using signal analysis and deep learning models to increase the robustness and performance of the verification system with the PPG signal. The proposed system focuses on utilizing different stretching mechanisms namely Dynamic Time Warping, zero padding and interpolation with Fourier transform, and fuses them at the data level to be then deployed with different deep learning models. The designed deep models consist of Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) which are considered to build a user specific model for the verification task. We collected a dataset consisting of 100 participants and recorded at two different time sessions using Plux pulse sensor. This dataset along with another two public databases are deployed to evaluate the performance of the proposed verification system in terms of uniqueness and time stability. The final result demonstrates the superiority of our proposed system tested on the built dataset and compared with other two public databases. The best performance achieved from our collected two-sessions database in terms of accuracy is 98% for the single-session and 87.1% for the two-sessions scenarios.**

***Index Terms*—Biometrics, Verification, Security, PPG, Deep Learning, CNN, LSTM.**

I. Introduction

THE dependency on the Internet and the rapid growth of number of users and devices connected to the Internet have increased in the last couple of years. People are connected to various online and off-line systems that facilitate their interactions and transactions performed in a daily basis. In all these activities, individuals use their credentials to create, share and pay which could be related to the individual's health, bank or personal information. Thus, these systems contain sensitive data which might include the address, social insurance number, and credit card information that should be highly secured. The main drawback of the proliferation of technology is the needed extra security measures to ensure the safety of the data.

Cyber criminals are more skilled where they continuously perform fraud, spoofing and hacking with the ultimate aim of stealing the individual's identity. The conventional method to insure the safety of your accounts is to generate a password which consists of alphabet letters, characters and numbers. However, no matter how much complexity and randomness are added to the password, it is still vulnerable and could be hacked because a normal password should be remembered or easily accessible by the user. As a result, alternative systems to increase the security of the information consist of biometric features of the individual which are used to verify the individual's identity. These include fingerprint [1], face [2], iris [3] and voice [4]. However, there are some disadvantages correlated with such modalities which include the simplicity of spoofing, and the low accuracy in an uncontrolled environment. Thus, researchers started to utilize a potentially reliable and easy to access modality which is the physiological signals.

Lately, many works focused on the developments of systems that incorporate the physiological signals for applications including biometric authentication, emotion recognition [5] and diabetes detection [6]. Examples of these physiological signals are the electrocardiograms (ECG) [7] and photoplethysmography (PPG) [8]. To this end, most of the works in the literature deploy the ECG signal for biometric verification and identification.

Although the ECG has shown to be more distinguishable compared to the PPG, it is inconvenient to use in practice and daily interactions. Compared to the ECG, the PPG signal can be acquired from a low cost, more accessible, and more portable devices. In addition, the PPG signal can be collected from different positions in the human body such as earlobes, fingertips or wrist. Further, the uniqueness and randomness characteristics that the PPG signal possesses make it hard to spoof. The advantages correlated to the PPG signal bring it more practical and appealing to be utilized in real time applications [9]. However, the physiological signals in general and the PPG signal in specific have many theoretical challenging aspects to be considered for user authentication which include time variability and inherent randomness. Therefore, it is essential to study the signal dynamics and time stability of the PPG signal.

To this end, this paper investigates the feasibility of employing the PPG signal for person verification by developing time-stable and distinguishable features. Previous works which have utilized the PPG signal relied on conventional, time-unstable characteristics. A system that consists of efficient, time-stable features is not yet presented. Thus, the main aim of this work is to develop robust, time invariant features by deploying Deep Neural Network (DNN) to build a completely data-driven approach based on Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM). The proposed network learns temporal biological features of each subject, which

D. Hwang, B. Taha, D. Lee and D. Hatzinakos are with the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (email: daeyon.hwang@mail.utoronto.ca).

eliminates the necessity of feature selection and extraction tasks.

The proposed model is a PPG-based personalized verification system, which solely employed for user verification. We foresee that our work can be extended and considered to verify individuals from others in daily transactions but it requires an additional experimentation considering the characteristics of PPG signals in diverse factors like pathological, emotional and daily activities. Currently, our paper focuses on developing robust, accurate and time-stable models for a given non-dynamic scenario. Even with these constraints, the proposed system can be implemented in real life scenarios due to the convenience and accessibility of the PPG signals.

The rest of the paper is divided as follows: Section II provides an overview of the research work done for user verification and classification from different modalities. Section III demonstrates the details for our developed dataset and the other databases used for testing. Section IV addresses the problem definition then followed by Section V which shows and explains the proposed methodology for user verification system. Section VI provides the simulation results and finally Section VII illustrates the conclusions and some future directions.

II. Related Research

There are several approaches done in the literature that involve the usage of physiological signals for individual authentication and classification. Here, the previous works with different methodologies are partially introduced with the ECG signal and mainly covered with the PPG signal. From an engineering point of view, both modalities have advantages and disadvantages compared to each other. A key point is that PPG and ECG signals share similar characteristics and are obtained from cardiac activity but the PPG signal is more suitable for real applications. However, the PPG is more challenging to attain similar performance compared to the ECG.

One of the most deployed physiological signals for user authentication is the ECG. Several works in the literature have been proposed and implemented that involve simple, hand crafted features, while others rely on deep learning solutions. Odinaka et al. proposed to use spectrogram and log-likelihood ratio (LLR) to perform user verification [10]. Louis et al. developed a continuous verification system from the ECG signal collected at two-sessions based on One-Dimensional Multi-Resolution Local Binary Patterns [11]. Furthermore, learned features are also implemented by various works in the literature. Guennec et al. proposed to use two layers CNN network trained in two different approaches to handle data size limitation [12]. Kachuee et al. developed a deep learning technique to classify five distinct arrhythmias [13]. The system relies on preprocessing the signal and zero padded signals are used to train a deep residual network.

Similar but less mature work compared to the ECG was done with the PPG signal. Many recent literatures proposed the usage of this signal for the authentication and other purposes. The earliest work that adopted the PPG signal for user identification is based on fuzzy logic classification [14]. They have utilized the characteristics of the signal namely number of peaks, downward and upward slopes, and the interval between the bottom and the peak as features for the fuzzy classification. Yao et al. employed a different set of features including local maximum, minimum, and inflection points of PPG pulses to perform user authentication [15]. Another work done by Spachos et al. demonstrates the usage of the PPG signal for biometric authentication by utilizing linear discriminant analysis (LDA) along with nearest neighbor (NN) classifiers [16]. Salanke et al. have proposed to use Fourier analysis to mitigate the noise effect, and semi discrete decomposition as well as Euclidean distance for feature extraction and selection [17]. Kavsaoglu et al. proposed a pipeline for PPG user recognition consisting of FIR filtering for preprocessing, time domain features, feature ranking using distance based criteria and classification using k-NN classifier [18]. Karimian et al. relied on non fiducial features to perform user verification [19]. They utilized wavelet analysis with Daubechies wavelet of order 4 which decomposes the signal into 4 levels. These levels are the features used to train two classifiers, Neural Network and Support Vector Machine (SVM). Another work was done by deploying the cycle detection and Karhunen-Loève transform to extract features from the PPG signal and the Manhattan distance for user classification [20].

On the other hand, there are several works in the literature that took advantage of deep learning methods to increase the effectiveness of the user authentication and classification with the PPG signal. Jindal et al. proposed to cluster the individual subjects into groups and then train a separate deep learning model for each cluster [21]. The system is based on a two-stage procedure which first involves clustering (with 11 hand-crafted features) using the Partitioning around Medoids and then deep learning models consisting of Restricted Boltzmann Machines and Deep Belief Networks are utilized to train each cluster. Shashikumar et al. proposed to detect the atrial fibrillation from PPG signal by developing a CNN network [22]. Wavelet power spectrum of the signal was utilized to classify each patient into one of two classes of atrial fibrillation or normal. On a similar frontier, Everson et al. proposed a biometric network with CNN and LSTM for user authentication [23]. Biswas et al. proposed to use the PPG signal for both heart rate estimation and user identification at the same time where they relied on the usage of CNN and LSTM [24].

III. Database

One of the paramount aims of this work is to learn time-stable features for each user in order to verify his/her identity in real time scenarios. Therefore, a key factor is to deploy a database which incorporates randomness and time variability within each subject. We have considered three datasets, two of them are collected at the BioSec. Lab at the University of Toronto while the third one is publicly available.

The first dataset, BioSec. PPG Dataset 1 (Biosec1) [25], consists of PPG signals acquired from Plux pulse sensor [26]. The sampling rate was fixed at 100 Hz while collecting the

signal. There are 31 participants in this dataset where the PPG signal was obtained from the subjects in two different sessions with at least 14 days gap in between. The signals were recorded in a relax condition for a continuous 3 minutes where the sensor was attached to the participant's fingertip. To resemble a real life case, the signals were collected in an uncontrolled environment where the participants were allowed to talk during data acquisition and the office environment was as usual. Finally, all participants have been chosen to have no history of any cardiac problems.

The second database was also obtained from the Biosec. Lab at the University of Toronto, Biosec. PPG Dataset 2 (Biosec2). This database was collected during this work as an enlarged and enhanced version of the first one where the data acquisition procedure was performed in a different manner. The same sensor, sampling rate, uncontrolled environment and health conditions were fixed. However, the data collection method and the size of the new dataset are different. First, the PPG signals were collected from the participants in a relaxed condition for 1.5 minutes from the fingertip in three different instants. Each time, the sensor was detached and attached again to the same fingertip (which was fixed to the index finger). The main reason for collecting the PPG signal at three different instants within the same session is to integrate the inherent randomness of the signal caused by the location of the sensor for the same participant within the same session. This is a key element that should be added and addressed to building a secure verification system. The other difference is the number of participants which is 100 consisting of 48 males and 52 females.

The final database studied in this work is PRRB [27], [28], which was collected during elective surgery and routine anesthesia. It contains only one session data where the PPG signal was recorded for 8 minutes with spontaneous or controlled breathing. Since the main goal of our work is to perform user verification, an 8 minutes recording is unrealistic for our aim. Therefore, we only used the first three minutes of the PPG signal for each subject. There are 42 subjects (29 pediatric and 13 adults) and the sampling rate was fixed at 300 Hz. This database was obtained while measuring the respiration rate under a controlled environment. Thus, it would contain a comparatively large noise levels caused by the respiration process. Even though this dataset consists of signals obtained from one session, it is valuable to test our proposed PPG verification system under such noisy condition. Table I demonstrates a summary of the databases used in this work.

TABLE I: The main characteristics of the employed databases for PPG verification. Avg time gap is the average time between the first and second data collection measured in days.

	Number of Subjects	Avg Time Gap	Age Range
Biosec1	31	36	19 - 35
Biosec2	100	17	18 - 44
PRRB	42	-	1 - 63

IV. Problem Definition

In this work, two distinct verification scenarios are considered. The first utilizes only one session data to perform user verification. In this case, the single session data is used for both training and testing without overlapping. The purpose of this experiment is to find unique features from the PPG signal for each subject. There have been many attempts in the literature to tackle this problem [19], [20], [29].

The second and the more realistic scenario is the two sessions case where the training data was taken only from the first session while the second session was used solely for testing. This case is more challenging and present a more practical approximation to the real life. This is mainly due to the inherent variability that the PPG signal possesses since it is based on heart activity. As a result, it is very sensitive to simple effects such as stress, exercise, diet etc. By having two sessions of data, we guarantee that the characteristics of the PPG signal are different even for the same participant. This is why this problem should be tackled by learning time-stable features that is robust enough to verify the participant from two similar, yet distinct PPG signals acquired from two sessions. There have been some papers employing two-sessions PPG signal [20], [25], [30] which we will be comparing our work with.

V. Methodology

Developing a verification system using physiological signal is a challenging task mainly due to the time-variability and randomness inherent in these signals. In this work, we focus on studying the feasibility of utilizing the PPG signal for verification systems and developing time-stable features with a performance comparable to other real life verification models.

In verification systems, a binary classification between the individual and other users is implemented to validate the user's identity. These systems are highly deployed in applications including access control. Our developed method consists of employing a novel input data coming from different stretching methods on the PPG signal then learning robust features using deep models. Stretching the signal is a key step to standardize the different length of the period in the PPG signal to maintain a consistent structure of deep learning models. Previous attempts include signal stretching using zero padding [13] or interpolation [20]. Here, we develop the different stretching methods along with investigating the current used ones. After that, we design deep learning models with employing CNN or CNN with LSTM to find the most suitable and robust models for our verification task. Fig. 1 demonstrates the main stages and pipeline of the proposed verification system.

A. Preprocessing

The first stage in our system is preprocessing the PPG signal by first filtering the signal and then, isolating the fundamental PPG signal pulse. These two steps are essential to successfully build the input PPG signals that will be used in the subsequent stages. For the filtering step, a 4^{th} order Butterworth filter with a cutoff frequency 0.5-18 Hz is used to mitigate the noises

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

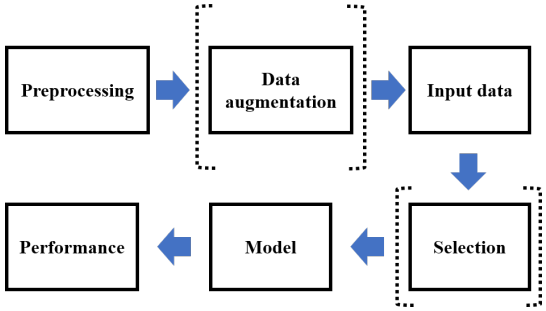
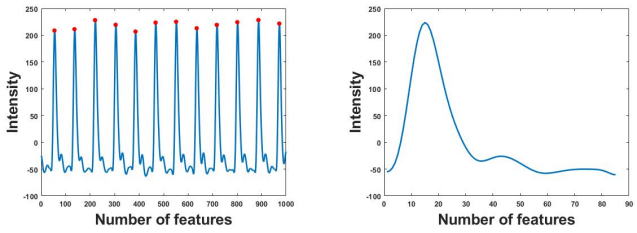


Fig. 1: Summary of algorithms. Some parts inside parentheses are only applied in certain cases. In Section V, all steps are explained as follows: Preprocessing → Input data → Selection → Data augmentation → Model.



(a) (b)

Fig. 2: Systolic peak detection and single pulse in Biosec1. (a) Systolic peak detection (red dots) on filtered PPG signal. (b) Single pulse from (a).

present within the PPG signal such as power line interference, baseline wanders and high frequency noise. The second step extracts the single PPG pulse which can be searched from systolic peak. We believe that the shape of the PPG pulse has the fundamental characteristics of the signal and thus, obtaining a more compact one will lead to higher information gain. The following procedure is implemented to extract the PPG pulse where Fig. 2 shows the PPG signal and the output extracted pulse.

- Detecting the systolic peak from filtered PPG signals using inverted second derivative [31]. The parameters in [31] are modified to provide more room for error to compensate the less clear shape of the PPG compared to the ECG signal. This landmark is selected because it is the most obvious and apparent one. The result is shown in Fig. 2 (a).
- Using the detected peak, the start and end points of each pulse in the signal are extracted by considering the range of each subject's heart rate.
- Then the average heart rate of each subject and the time difference of consecutive systolic peak are calculated.
- After that, if the time difference is 20 percent higher or 60 percent lower than the calculated average, the pulses between those detected peaks are removed.
- Finally, the start point is selected as the minimum point between 20 samples (for Biosec1 and Biosec2 which have 100 Hz sampling rate) or 60 samples (for PRRB which has 300 Hz sampling rate) in front of remained peak while the end point is equal to the start point in next systolic peak. Single pulse is demonstrated in Fig. 2 (b).

We evaluated the performance of the implemented algorithm to extract the PPG pulse by measuring the detection accuracy of the systolic peak, the start and end points of the single pulse. We achieved the average accuracy of 95.2%, 98.8% and 98% from all subjects in Biosec1, Biosec2 and PRRB, respectively. The databases considered in this paper are collected under stable, rest conditions where the participants were sitting in a chair or elective surgery and routine anesthesia. Therefore, the data possesses less dynamics and relatively small motion artifacts (MA). This is evident from the high accuracy achieved for the detection of the systolic peak, the start and end points without introducing any motion artifact suppression techniques to the system. Nonetheless, we implemented LMS based methods, which is one of the common methods for MA suppression [32], to investigate the presence of any noises coming from motion done by the participants. With the MA suppression, we attained 96.9%, 98.8% and 99.1% average accuracy for systolic peak detection in Biosec1, Biosec2 and PRRB, respectively. It is seen that there are slight improvements after applying the MA suppression for Biosec1 and PRRB datasets. However, since these improvements are insignificant and our model is mainly dedicated for data collected at stable condition, our subsequent stages will be independent of the MA suppression algorithm. Yet, the addition of the MA filtering at the preprocessing stage can be considered and easily integrated according to the degree of dynamics in the databases.

B. Input data

After preprocessing the raw PPG signal, we focus on building a consistent number of features to have a same length for all the PPG signals. This is done by selecting the proper heart rate (HR) when building the signal. The normal resting HR for pediatric/adults ranges from 50 to 120 beats per minute. Lower heart beats per minute require a greater number of features compared to the higher one. Therefore, we create each input signal assuming 40 beats per minute. As a result, the number of features for the Biosec1 and Biosec2 signals is 150 since the sampling rate is 100 Hz and for the PRRB is 450 because of the 300 Hz sampling rate. We consider three different ways of stretching the signal to investigate their effects on finding unique and time-stable features for our verification system namely Dynamic Time Warping (DTW), Zero padding in time (ZT) and Interpolation in frequency (IN). We then propose a new input data that is based on combining the different stretching methods at the data level.

1) Dynamic Time Warping (DTW): DTW is a famous time series comparison technique that has been applied for word and speech recognition. It has been used with physiological signals [33], [34]. DTW method overlaps two disjoint time series data either to measure significant changes in the signal shape or to find the existence of outliers. Compared to Euclidean distance-based matching, DTW has the advantage of matching up two-time series curves without synchronization.

The implementation of DTW method requires the presence of a reference signal pulse to the PPG one. Daubechies 4

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

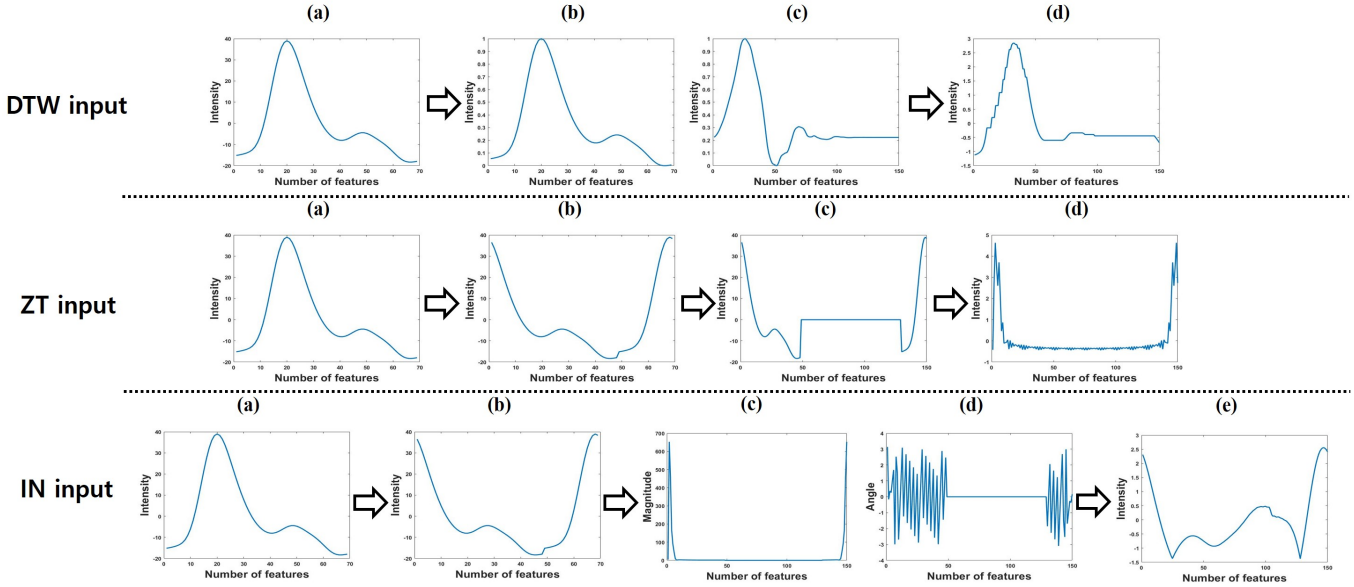


Fig. 3: The order of building input data in Biosec1. For DTW input: (a) Single pulse. (b) Normalized single pulse. (c) DB4. (d) Standardized DTW input data. For ZT input: (a) Single pulse. (b) Reshaped single pulse. (c) Zero padding on reshaped single pulse. (d) Standardized ZT input data. For IN input: (a) Single pulse. (b) Reshaped single pulse. (c,d) Zero padding on magnitude, angle after Fourier transform on (b). (e) Standardized IN input data.

wavelet (DB4) is considered because it has similar properties as the cardiac signal. Therefore, a scaling function of DB4 wavelet is utilized to build the DTW input data.

First row in Fig. 3 shows the process of building the DTW input data. After preprocessing, the pulses (Fig. 3 (a)) from the filtered PPG signal have different number of features and amplitude. Therefore, to build the DTW input, we first normalized (equation (1)) the single pulse to have intensities between 0 and 1 (Fig. 3 (b)) which has the same range as DB4 (Fig. 3 (c)). Then, we apply the DTW on our PPG pulse with DB4 as a reference to build the DTW input. Finally, the output pulse is standardized (equation (2)) with zero mean and unit variance (Fig. 3 (d)).

$$Y_{new} = \frac{Y - Y_{min}}{Y_{max} - Y_{min}} \tag{1}$$

$$Y_{new} = \frac{Y - mean(Y)}{std(Y)} \tag{2}$$

where Y is the original signal and min/max denotes the minimum and maximum values of original signal, respectively. $mean(Y)$ and $std(Y)$ are the average of and standard deviation of the original signal. Lastly, Y_{new} represents the final result after calculation.

2) Zero padding in time (ZT): Zero padding is a conventional way to extend the size of signals into a specific length. However, zero padding the physiological signal directly is risky task. Generally, new input data is built by appending zero values in certain location within the physiological signals but this should be handled carefully, especially in applications that require high security since it introduces bias information on the original signal. From this perspective, we consider zero padding with Fourier transform which yields less bias information. This will provide a smoother spectrum which is similar to non-zero padded signal.

Second row in Fig. 3 shows the sequence of making ZT input data. We want to do zero padding on certain locations where we can define well in each pulse and those firm locations are start, systolic peak and end points in single pulse. This is why single pulse (Fig. 3 (a), in sequence: start-systolic peak-end) is reshaped as Fig. 3 (b) (in sequence: systolic peak-end-start-systolic peak). After reshaping, we add zero values between end and start points to make consistent length (Fig. 3 (c)) which is 150 because it comes from Biosec1. Next, we do Fourier transform and only consider the magnitude of that result. In this paper, angle of Fourier transformed result is not used because it offers less distinctive features based on trial and error. Finally, we standardized it using equation (2) to make ZT input data with zero mean and unit variance (Fig. 3 (d)).

3) Interpolation in frequency (IN): IN input data comes from similar concept as ZT but it applies zero padding after Fourier transform. Third row in Fig. 3 shows the procedure of building IN input data.

Same as ZT input data, single pulse is reshaped in the sequence of systolic peak-end-start-systolic peak. Fourier transform is applied in reshaped single pulse and we divide the result into magnitude and angle. Zero padding is considered between the end and start points in magnitude and angle, separately (Fig. 3 (c) and (d)). After zero padding, magnitude and angle is merged using equation (3) and then, we do inverse Fourier transform with considering the scaling problem caused by different length from zero padding (equation (4)). Finally, we do standardization (equation (2)) to build IN input data with zero mean and unit variance (Fig. 3 (e)). It should be known that we only use the real value of final IN input data

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

because the imaginary part did not provide distinguishable features based on trial and error.

$$Y_{new\_ft} = mag(Y_{ft})e^{jang(Y_{ft})} \tag{3}$$

$$Y_{new} = \frac{N}{M} * ift(Y_{new\_ft}), \tag{4}$$

where Y_{ft} is signal after Fourier transform and mag/ang means the magnitude/angle of Fourier transformed signal. j is imaginary unit and Y_{new_ft} is the result from polar formula of Y_{ft}. N is the target length (150 for Biosec1 and Biosec2, while 450 for PRRB) and M is the length of original pulse. ift means the inverse Fourier transform.

4) Outlier removal: After building the input data (DTW, ZT or IN), we removed signal outliers. After building all potential inputs (blue curves in Fig. 4), we produce the average shape of those potential candidates which is red curve in Fig. 4. Then, we calculate the Euclidean distance between average shape and each potential input data. Using this information, we sort the potential candidates in ascending order and select sequentially. Removing outlier is important because it helps our verification model focusing on most certain target data and saving computational time.

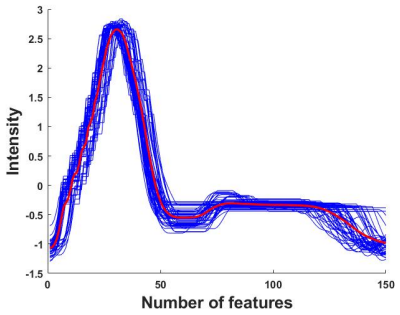


Fig. 4: Outlier removal. Blue curve: All potential signals which can be included as input data. Red curve: Average shape of all potential signals.

C. Selection

In this part, we cover how to assign the proper input data to each subject for improving the verification system performance. It should be noted that the selection stage is only considered when dealing with two-sessions' experiment. When the input data (DTW, ZT and IN input) is used solely to perform verification without the addition of the selection step, the performance is not promising as seen in the left sub-table in table III (a) from VI. Results.

Since the focus is on verification systems which is the binary classification between target and others, a personalized system is assigned to each subject. This means that the number of verification systems is equal to the number of subjects in each database. From this perspective, we found out that each subject prefers a different input data compared to others. This is shown in more details in Fig. 9.

Each subject's preference in Fig. 9 inspires us to assign the best input data for each subject and this can be determined by using the single-session's performance because uniqueness of the PPG signals should be held in different time session (time stability). Single-session data is same as training data in two-sessions' experiment which can be accessed during training the model. This perspective has a close relation with real world application because we never have a chance to get the future PPG signals which are testing datasets but can manipulate the current PPG signals which are training datasets. We will call this idea as "selection method" in the rest of paper. In detail, the performance in single-session used for deciding the input data is EER (Equal Error Rate. More covered in VI. Results).

Using the selection method alone among three inputs does not improve the performance significantly (left sub-table in table III (b)). Thus, we proposed a new input data by stacking previous input data as channels. This fusion of biometric data is addressed similarly in [35], [36]. In detail, there can be three instances of building new input data with 2 channel: DTW input with ZT, ZT with IN and DTW input with IN. We show in table VIII that the order of stacking has small effect on the model performance (i.e. DTW-ZT is similar to ZT-DTW), thus stacking order is considered the same in this work. We also tried 3 channel (DTW input with ZT, plus IN) as seen in table III. From table III, we choose the DTW input with IN as a new input data (written as 2 channel input in the rest) which was most promising with the selection method. This 2 channel input is competed with ZT input data (denoted as 1 channel input in the later) to decide the best input data for each subject using the selection method. All outcomes illustrated in VI. Results mostly consider these two inputs. Fig. 5 (a) visually explains the selection method using 2 channel and 1 channel input data. For example, if subject 1 prefers the 2 channel input in first session, it will only receive the training (generated from first session) and testing (come from second session) data from 2 channel input afterwards.

D. Data augmentation

Large annotated data is critical to improve the performance of deep learning model. Yet, it is usually hard to find, especially in the medical domain. Data augmentation increases the number of data from the original samples to enhance the generalization performance and this should be done without the affecting the labels. Various ways of data augmentation were already applied to the physiological signals [12], [37]–[39], but, in this paper, we focus on noise addition for data augmentation to investigate its effect on finding time-stable features from the PPG signals. To be clear, we apply different ways of augmentation in single-session and two-sessions. More detail will be covered in VI. Results.

Three different types of noises are considered as data augmentation. All these noises are applied on the single pulse before building the input data (DTW, ZT and IN input). First type of noise is the Gaussian noise [37], [38] with a zero mean and unit variance. The addition of Gaussian noise can be considered as a random jitter caused by the measuring device on the signal. Equation (5) shows how the Gaussian noise is added on the single pulse.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

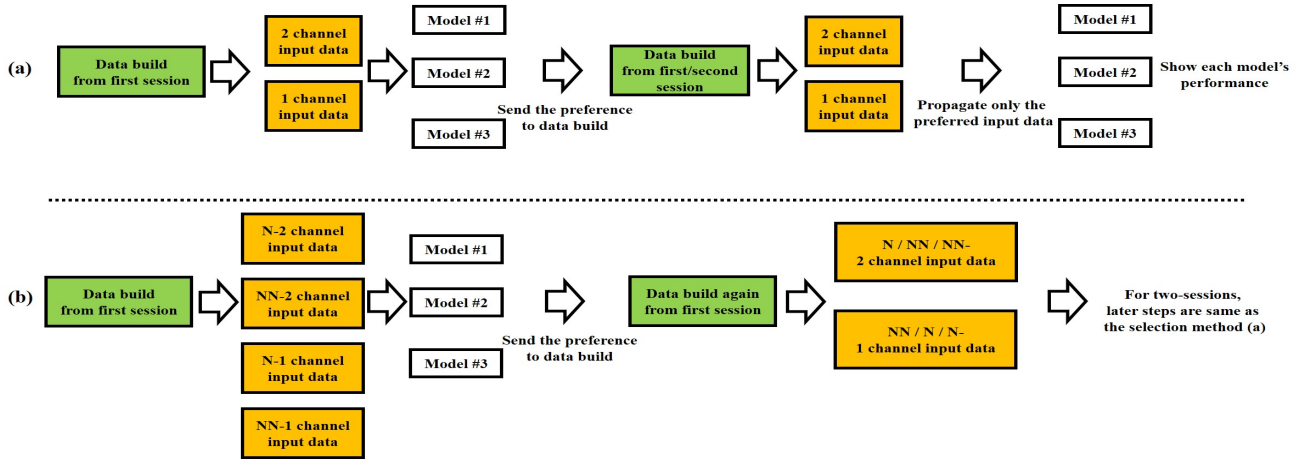


Fig. 5: Build preferred data for each subject through first session data. We example three subjects in this figure. (a) Selection method. Here, data from first session is training data and second session is testing data. (b) Process of data augmentation version 2. N and NN represent normalized and non-normalized input data.

$$Y_{gaussian} = Y + \frac{1}{3}Gaussian(0,1) \tag{5}$$

where Y is the original single pulse and $Gaussian(0,1)$ means the Gaussian noise with zero mean and unit variance. $Y_{gaussian}$ is the augmented single pulse after adding the Gaussian noise and $\frac{1}{3}$ is a factor multiplied to limit the amplitude under the original single pulse [37].

The second type is a sloping noise [37] which tries to emulate the envelope of the PPG signal originated from respiration. Equation (6) gives how the sloping noise is included in the single pulse.

$$Y_{slope} = Y + \frac{3}{M}Uniform(-1,1) \tag{6}$$

where Y is the original single pulse and $Uniform(-1,1)$ is the uniform distribution from -1 to 1. M is the length of original single pulse and Y_{slope} represents the augmented pulse with sloping noise. $\frac{3}{M}$ prevents sloping noise to exceed the usual breath of the PPG signal.

The final noise added is a combination between the previous two noises as shown in equation (7).

$$Y_{combined} = Y + \frac{1}{3}Gaussian(0,1) + \frac{3}{M}Uniform(-1,1) \tag{7}$$

where Y is the original single pulse and other noises are denoted in equation (5) and (6). $Y_{combined}$ is the pulse after adding combined noise. Fig. 6 visualizes the single pulse with/without data augmentation after filtering. The reason for adopting these noises for augmentation is because there can be different type of Gaussian and respiration noises in different recording time and thus, we utilize them to check its relationship with the time stability in the PPG signals. In addition, we consider stable databases with less dynamics and this is why we did not consider the noises coming from MA for augmentation.

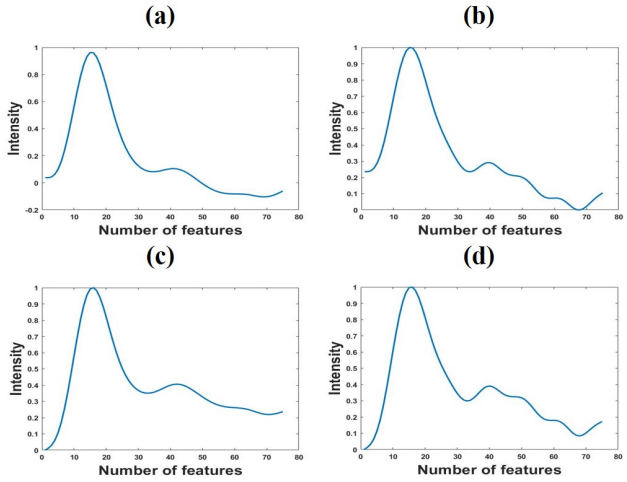


Fig. 6: Data augmentation on single pulse. (a) Original single pulse. (b) Gaussian noise on single pulse. (c) Sloping noise on single pulse. (d) Combined noise on single pulse.

1) Data augmentation version 1 (v1): There are two versions of data augmentation applied in this paper. The first version considers applying normalization (equation (1)) on the single pulse before adding the noise. The normalization has no effect on the DTW input, since it was already normalized before. However, normalizing the ZT and IN provides different shapes of the input data. In the rest of this paper, we call the input data with normalization as N-input data and without normalization as NN-input data. To clarify, data augmentation v1 uses N-single pulse to build input data (ZT and IN), which also be used for augmentation. Fig. 7 shows the difference between N/NN-ZT/IN input data.

2) Data augmentation version 2 (v2): Data augmentation v2 is an extension of v1. When considering data augmentation with selection method (C in Section V) in two-sessions, we can find that each subject prefers different types of input data, which are N-1 channel, NN-1 channel, N-2 channel and NN-2

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

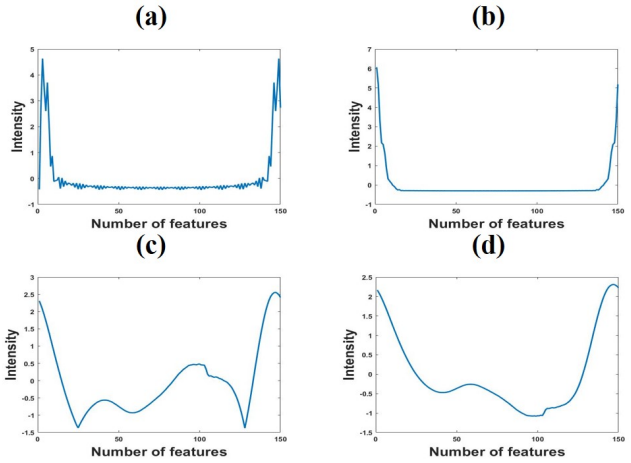


Fig. 7: N/NN-ZT/IN input data. (a) NN-ZT input data. (b) N-ZT input data. (c) NN-IN input data. (d) N-IN input data.

channel input (For 2 channel input, there are N-input and NN-input because of N/NN-IN). Using this preference information, we introduce an additional step before doing the selection method.

Fig. 5 (b) shows the procedure of data augmentation v2. Using the PPG signals in first session, we build four different kinds of input data (N/NN-1/2 channel input) and insert them into each subject's model to find the preference of N-input or NN-input in each of 2 channel and 1 channel. For example, Fig. 5 (b) explains when subject 1 prefers N-input for 2 channel and NN-input for 1 channel. Other subjects are vice versa. Same as selection method, preference is determined by single-session's performance (EER). Next, data build produces N/NN/NN-2 channel input data and NN/N/N-1 channel input data from first session data where subject's data is concatenated sequentially. Later, it follows the same step as selection method (Fig. 5 (a)) if it is two-sessions. Important thing is that this rebuilding information should be saved to reconstruct testing data in same structure. For single-session, the rebuilt inputs are used directly.

E. Model

Two types of models are developed in this paper to find unique and time-stable features. The first one is CNN and the second one is CNN with LSTM. Detail information for each model will be covered later.

In this paper, the followings ideas are implemented in all models. For controlling the overfitting problems, L2 regularization, bagging (with three models) and 10-fold cross validation are applied. Weighted cross entropy is used to offset the class imbalance problem (since in verification system, we discriminate one user from others) and Adam optimizer is considered to optimize the network. Learning rate is fixed at 0.0001 in all tests. All details about models are found in table II.

1) Convolutional Neural Network: CNN is a deep neural network that is useful to extract distinct features from signals and images. It is mainly considered to classify the data into

TABLE II: Details about CNN and LSTM models. Describe as follows: 2 channel/1 channel input. For 1 channel input, information in 3rd layer of CNN is not existed and thus, it is denoted as -. Red part is for CNN and blue part is for LSTM.

	Biosec1, Biosec2	PRRB
Number of Feature (Input)	150x2 / 150x1	450x2 / 450x1
Number of CNN Layers	3 / 2	3 / 2
Number of Filters (1st layer)	30 / 50	30 / 50
Size of Filter (1st layer)	30x2 / 60x1	90x2 / 180x1
Number of Filters (2nd layer)	50 / 70	50 / 70
Size of Filter (2nd layer)	40x30 / 60x50	120x30 / 180x50
Number of Filters (3rd layer)	70 / -	70 / -
Size of Filter (3rd layer)	50x50 / -	150x50 / -
Dropout Rate	50% / 50%	50% / 50%
Number of LSTM Layers	1 / 1	1 / 1
Number of Hidden Unit	32 / 32	32 / 32

specific labels and identify faces, tumors and even texts. Lately, CNN is also implemented for time series classification, especially in physiological signals [6], [12], [13], [22], [39], [40].

In this paper, each layer of the CNN is composed of convolutional filters, Rectified Linear Units (RELU) activation function and dropout. There are plenty of hyperparameters to be tuned carefully but, in this paper, we only cover the best hyperparameters from our exhaustive searching. Fig. 8 (a) and (b) show the structure of CNN models in 2 channel and 1 channel input data, separately. The detail information can be found in each step of models. Regularization parameters are 0.1 for 2 channel input and 0.005 for 1 channel input.

2) Long-Short Term Memory: Recurrent Neural Network (RNN) is a neural network that addresses sequential time series data and considers both the new input at the current time and the output of the neural network in the previous time. It works like a memory of computer and this great advantage leads RNN to implement in time series forecasting, natural language processing and speech recognition. RNN is already applied in physiological signals [6], [23], [24], [30], [38], [39], [41] but most of them applied with CNN to learn the temporal dependencies and recover the phase variant caused after CNN. LSTM is one of most favorable RNN because of its power of overcoming vanishing gradient problem where conventional RNN was not. In this paper, we apply 1 layer of LSTM with 32 hidden units for both 2 channel and 1 channel input data. These specific numbers are hyperparameters and they are decided from several trials. Fig. 8 shows the location of LSTM which is between CNN and fully connected layer. In the model of CNN with LSTM, regularization parameters are 0.05 and 0.005 for 2 channel and 1 channel input, respectively.

VI. Results

The simulation results aim at validating our proposed verification system under different databases. There are 6 different cases experimented in each part of the experimentation results. 1) No augmentation with CNN, 2) No augmentation with CNN and LSTM, 3) Augmentation v1 with CNN, 4) Augmentation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

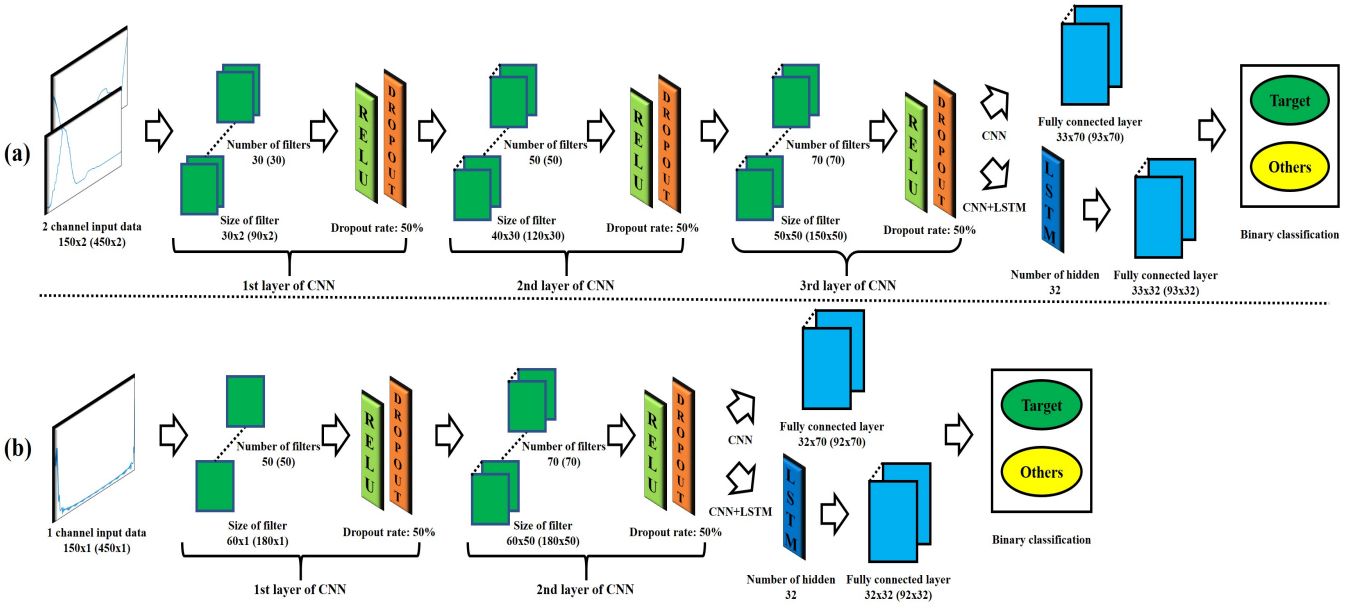


Fig. 8: Structure of models. (a) Model for 2 channel input. It possesses 3 layers of CNN (+1 layer of LSTM). (b) Model for 1 channel input. It has 2 layers of CNN (+1 layer of LSTM). The number inside/outside bracket is for PRRB/Biosec1, Biosec2.

v1 with CNN and LSTM, 5) Augmentation v2 with CNN, 6) Augmentation v2 with CNN and LSTM. In all cases, the training was performed using on TensorFlow 1.14.0 with Nvidia Tesla T4. For evaluating the performance, we considered four measurement methods: average of accuracy (ACC), EER, execution time and Receiver Operating Characteristic (ROC) curve. The ACC is the true predictions divided by total number of inputs. For verification system, ACC is not a true indicator of the system performance. Thus, we also choose the EER for measuring the system performance. EER is the certain point when false rejection rate (FRR) and false acceptance rate (FAR) are same or closest. The execution time is the largest training time needed for each case. ROC curve shows the performance of classification model at different thresholds between FRR and FAR. It is useful to compare the performance visually for different methods or cases.

The average ACC, EER and execution time are computed from the average of each verification system's performance as evaluation metrics. On the other hand, average ROC curve is determined from the average of the individual verification system's FRR and FAR at each classification threshold. Therefore, the former measures are useful for comparing the certain performance in single unit, while average ROC curve is used to measure the performance between different techniques. Furthermore, data shuffling is implemented every time before training the models and thus, all the reported results are the average of several trials.

The coming sections are organized as follows: First, we show the strength of selection method, compared to 2 channel and 1 channel input data. Then, we introduce the result of single-session and two-sessions in different databases to see whether our model finds the uniqueness and time-stability in the PPG signals.

TABLE III: Performance of all potential input data from Biosec2 in terms of ACC. (a) Without selection method. (b) With selection method. Best scores are in bold and underlined.

(a)

All possible inputs in 1 Channel	Average of ACC
DTW	79.9%
ZT	81.3%
IN	81.1%

All possible inputs in 2 Channel	Average of ACC
DTW+ZT	78.9%
DTW+IN	81.5%
ZT+IN	80.6%

All possible inputs in 3 Channel	Average of ACC
DTW+ZT+IN	81.2%

(b)

Selection with all possible 1 channel	Average of ACC
DTW or ZT or IN	81.4%

Selection with all possible 2 channel	Average of ACC
DTW+ZT or IN	81.3%
DTW+IN or ZT	**82.6%**
ZT+IN or DTW	80.9%

A. *Performance according to selection method*

In this experiment, the goal is to demonstrate the benefit of the selection method in the proposed system. This is done by showing the ACC at the system and individual levels. The selection method is only applied with two-session's databases which are Biosec1 and Biosec2. After performing the selection method, we will call that input data as "combination input" in the rest of the section.

Table III demonstrates the performance results for different input data in two-session's experiment with and without the selection method. The performance is illustrated from Biosec2 using CNN as classification model, where 3 channel input shares the same model structure as 2 channel input. From table III, we can notice the improvement in the performance when the selection method is implemented and find that DTW with IN and ZT are the best combination to show the maximum achievement (bold and underlined in table III (b)).

Fig. 9 shows the accuracy values for each subject using the different input data. The performance is computed using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

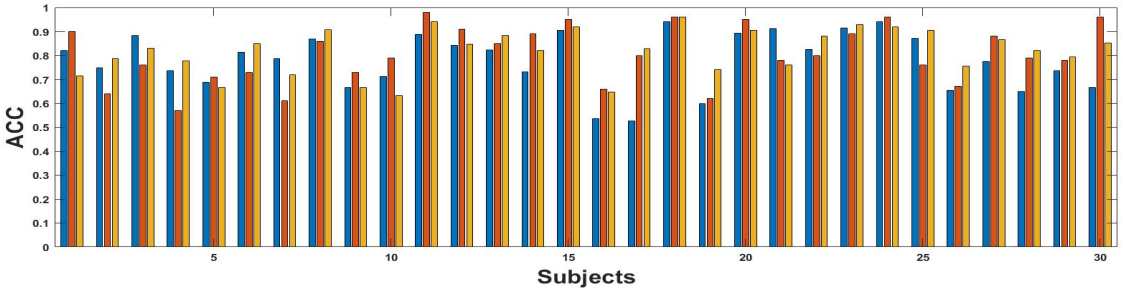


Fig. 9: Performance of input data for different subjects from Biosec2 in terms of ACC. Here, it only covers from subject 1 to 30 as example. Blue/red/yellow denotes the result of DTW/ZT/IN input.

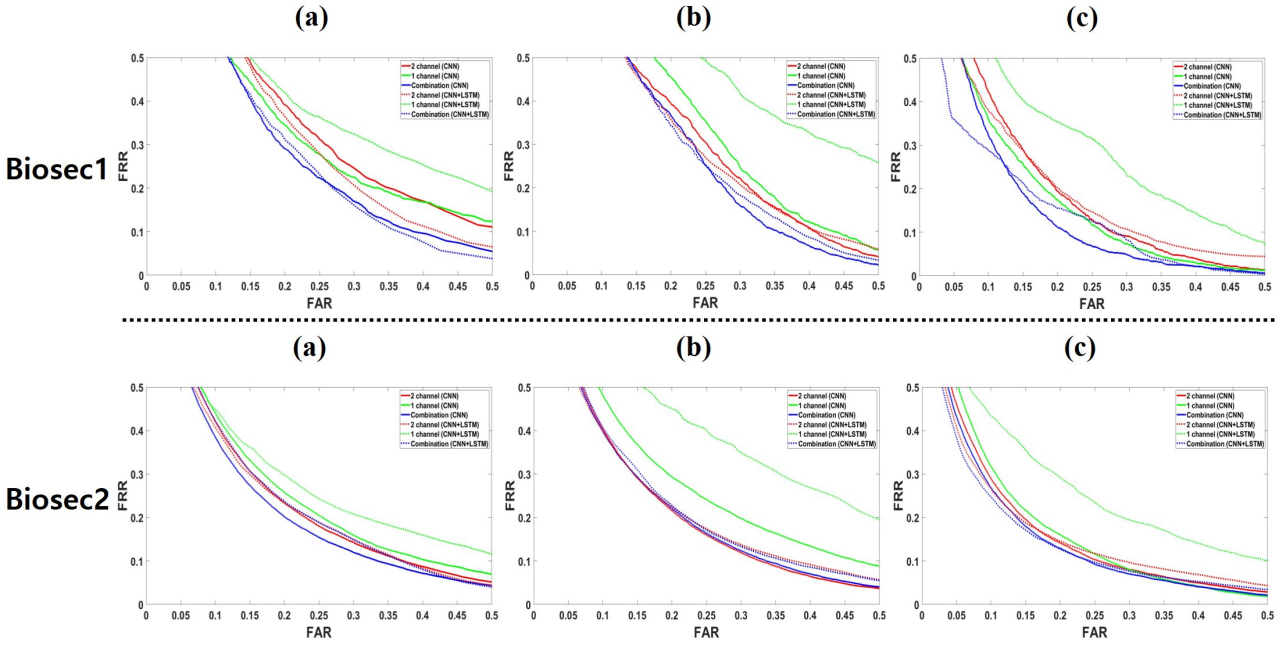


Fig. 10: Average ROC curves with selection method in Biosec1 and Biosec2. Red/green/blue color denotes the two-sessions performance of 2 channel/1 channel/combination input. Solid line describes the performance from CNN model and dotted line shows the performance from CNN with LSTM model. Every figures are zoomed-in (in range: [0-0.5]). (a) No augmentation. (b) Augmentation v1. (c) Augmentation v2.

Biosec2 two-session's data by employing the CNN model. It is evident that each subject prefers a different input data and thus, utilizing the selection method is suitable to improve the overall performance of the system.

The top part in Fig. 10 shows the average ROC curve among combination, 2 channel and 1 channel input data in Biosec1. When we consider two-sessions PPG signals in Biosec1, the result of the combination input outperforms the performance of 2 channel and 1 channel input as indicated in the smallest area under curve (AUC) even if we change the use of model. Thus, we can conclude that selection method is always useful to apply in Biosec1. The bottom part in Fig. 10 provides the average ROC curve among combination, 2 channel and 1 channel input in Biosec2. For two-sessions PPG data in Biosec2, we can conclude that the performance of combination input data is better than (or similar with) the achievements of 2 channel and 1 channel input in both CNN and CNN with LSTM models. In contrast with Biosec1, selection method has less influence since Biosec2 has diverse training data coming from measuring the signal three times while relocation of sensor, which helps to get better performance in 2 channel and 1 channel input. Thus, there is less room for improvement but still, selection method is valuable to apply for achieving the refinement in Biosec2.

B. Performance in single-session

In this part, we cover the results while having single-session PPG signal, depending on different databases. The performance in single-session is valuable since we understand whether our verification model learns the needed unique features from input data (both 2 channel and 1 channel). When we construct the input data before applying the models, we always bring the same number of input data from each subject to prevent the overfitting caused by class imbalance.

For two cases without data augmentation (which are no augmentation with CNN, and with CNN and LSTM), we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

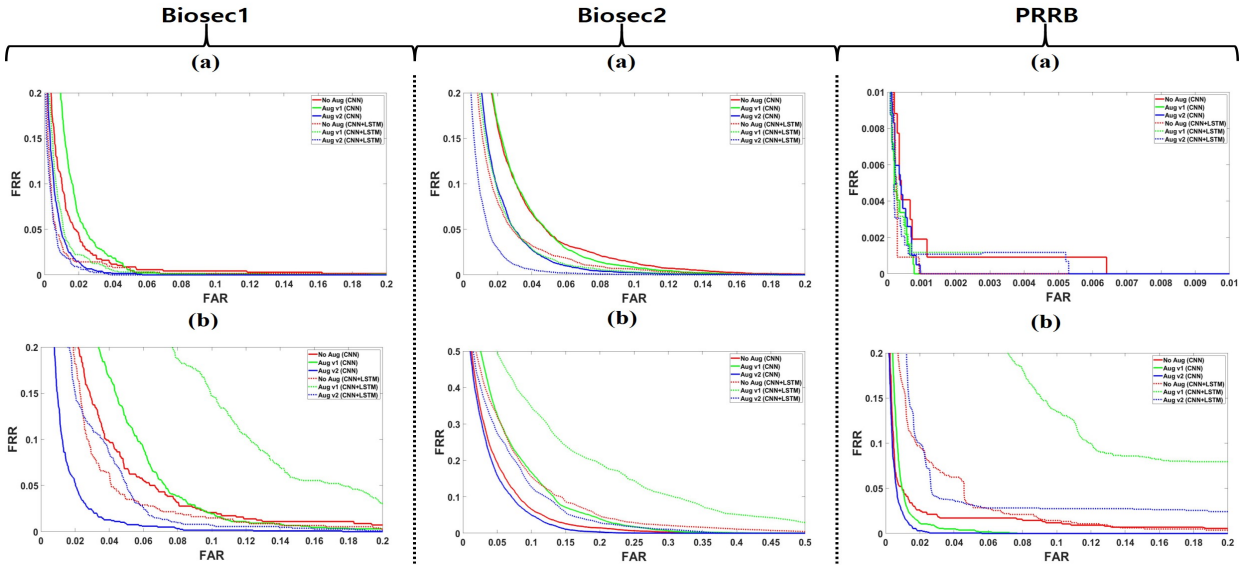


Fig. 11: Average ROC curves in single-session with zoom. Name of database is described on the top. Red/green/blue color denotes the single-session performance in no augmentation/augmentation v1/augmentation v2. Solid line means the performance from CNN model and dotted line describes the performance from CNN with LSTM model. (a) 2 channel input. (b) 1 channel input. For reference, Biosec2 (b) and PPRB (a) have different ranges, which are [0-0.5] and [0-0.01], to include all the cases and make it clear.

extract 50 input data from each subject through removing outlier. In other words, there is 1550 input for Biosec1 and 2100 input for PRRB, which are ready to apply in our designed model. As mentioned in Section III, Biosec2 has three measurements through relocation of the sensor at each session. We want to utilize all measurements to fix the problem from location of sensor, thus we collect 50 data from each trial which results in 150 for each subject in one session. Finally, there is 15000 input for Biosec2.

For the different four cases depending on the used data augmentation method (which are augmentation v1 and v2 with different models), we select more data from each subject. In single-session, we did not augment data by adding noise and thus, these cases are for testing the change of performance by the size of data. For Biosec1 and PRRB, we increase the size of the data twice which means the size of the Biosec1 becomes 3100 and for PRRB becomes 4200. For the Biosec2 dataset, we cannot increase the number of samples significantly since the PPG signals are collected for three different trials with 1.5 minutes each and in some cases the peaks are not obvious. Therefore, we take 54 input from each subject in each measurement which results in 16200 samples in total for Biosec2 dataset.

For building the verification system in single-session, inputs from the first session are divided into 60% and 40% applied for training and testing, respectively without overlapping between them. We selected these percentages to follow the same training protocol as two-sessions and to be able to compare the results with other works. It is true that the performance is improving if we increase the rate of training, but our main significance is the achievement in two-sessions. For Biosec1, the epoch for training is 60 while, for Biosec2 and PRRB, it is 50. Even we increase the number of epoch in Biosec2 and PRRB, the result was similar and thus, we decrease it to reduce the computational time.

Fig. 11 shows the average of ROC curve for different databases. First of all, we can find out that 2 channel input has better performance than 1 channel in every database. This is because 2 channel input has more information (or depths) in each data which helps to discern the true target from others. When we see the PRRB (a), there are differences between cases but this figure has small range ([0-0.01]); thus, 2 channel performance is always great. From Fig. 11, we can conclude that augmentation v2 mostly shows the best performance in every input data because of the smallest AUC in both CNN and CNN with LSTM models. This result is more obvious in 1 channel than 2 channel input.

For 2 channel input, augmentation v2 with CNN and LSTM gives the best performance, while, for 1 channel, augmentation v2 with CNN shows the best result. When we see no augmentation (red) and augmentation v1 (green) in both models, the performances are mostly getting worse (or small change) when we consider the augmentation v1. Thus, we can find that increasing the number of data by selecting more outliers is mostly not helpful.

Table IV shows the average of ACC, EER and execution time in each case with different databases. From table IV, we can see that PRRB always show the great performance compared to others. This is because the PPG signals in PRRB were collected in controlled environment with professional equipment and thus, it has high quality of signals with small noise. Even 1 channel input in PRRB has the change of performance in different cases, it is usually small, compared to Biosec1 and Biosec2. Biosec2 gives the low performance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

TABLE IV: Compare the single-session performance. Describe as follows: Avg of ACC (Avg of EER), Avg of execution time (seconds). (a) Performance with 2 channel. (b) Performance with 1 channel. Best results are in bold and underlined.

(a)

	Biosec1	Biosec2	PRRB
CNN	97.4% (2.5%), 66	95.3% (4.7%), 330	100% (0.1%), 200
CNN+LSTM	98.7% (1%), 472	96.5% (3.5%), 628	100% (0.1%), 1180
Aug v1, CNN	97.1% (2.8%), 98	95.3% (4.8%), 350	100% (0.1%), 320
Aug v1, CNN+LSTM	98.2% (1.8%), 488	96.5% (3.6%), 670	100% (0.1%), 1253
Aug v2, CNN	98.5% (1.4%), 98	96.5% (3.5%), 350	100% (0.1%), 320
Aug v2, CNN+LSTM	**99% (1%), 488**	**98% (2%), 670**	100% (0.1%), 1253

(b)

	Biosec1	Biosec2	PRRB
CNN	95% (4.8%), 53	91.6% (8.4%), 224	97.9% (1.8%), 150
CNN+LSTM	96% (4.1%), 450	90.2% (9.8%), 530	97.5% (2.3%), 1050
Aug v1, CNN	93.6% (6.3%), 76	88.7% (11.4%), 248	98.6% (1.4%), 220
Aug v1, CNN+LSTM	92.9% (7.2%), 465	84.3% (15.7%), 560	94.9% (4.9%), 1131
Aug v2, CNN	**97.3% (2.6%), 76**	**92% (8%), 248**	**99% (1%), 220**
Aug v2, CNN+LSTM	96.5% (3.5%), 465	91.4% (8.6%), 560	98.4% (1.5%), 1131

than other databases because it has more diverse input data from several measurements, which makes hard to get good achievement. As we explained on Fig. 11, we can find out the similar patterns in table IV but more certain performance with specific number. One more thing to be considered is that the execution time is getting much bigger when we use the LSTM. This is common story that LSTM requires more loops than CNN and thus, takes more computational time. Also, we can see that execution time increases when the number of data is getting bigger. This can be identified by comparing no augmentation and augmentation, or comparison Biosec1 with Biosec2 or PRRB. Computational time for CNN with LSTM in PRRB is the highest one because of the largest number of features.

Table V introduces the performance of PRRB database in our verification system and other papers. To be clear, verification is binary classification and identification means multi-class classification. In table V, all instances are verification but adopting different features and models. As mentioned earlier, PRRB holds good quality of the PPG signals that provides distinguishable features and thus, every results in table V are great. We can see that our verification system also works very well, meaning our algorithm is robust and precise.

TABLE V: Compare the performance of PRRB database. Some values are missing (denoted as -) because they were not reported in those papers. DWT: Discrete Wavelet Transform. DB: Daubechies wavelet. Coif: Coiflets wavelet.

	[19]	[29]	[20]	Proposed Method
Verification / Identification	Verification	Verification	Verification	Verification
Feature	DWT with DB	DWT with Coif	Multi-Cycles from PPG cycle detection	2 channel input (DTW+IN)
Model	Neural Network	SVM	Matching with Manhattan distance	CNN
ACC	100%	99.75%	-	**100%**
EER	-	1.46%	1%	**0.1%**

C. Performance in two-sessions

Here, we introduce the results of our experiments with two-sessions datasets. Even though the results from single-session is superb, it does not reflect the ability of designing time-stable features. Therefore, the experiments with two-sessions are important to understand whether our verification system finds the time-stable and unique features. This experiment resembles a real life application where we establish the verification system at one time but identifying the user at any other different time.

Before considering our approach, in table VI, we share the results of implementing simple classifiers (i.e. linear SVM, 3 nearest neighbors (3NN) and decision tree) with conventional features in two-sessions databases. For fiducial and non-fiducial features, we considered the systolic peak intensity, upward and downward slopes, autocorrelation and power spectrum density. In addition, we implemented linear interpolation in time with simple classifiers. Briefly, linear interpolation is a conventional way to find new data points within the range of known data points using linear polynomials. For comparison, we also added the performance of linear interpolation with CNN model. This CNN model has same structure as Fig. 8 (b). The results show that using the aforementioned classifiers limits the system performance where the use of CNN model with linear interpolation outperforms all the implemented methods.

In table VII, we show the results in two-sessions while considering usual linear interpolation in time and zero padding (without Fourier transform) in addition with the three stretching methods DTW, ZT and IN. Zero padding mentioned here simply adds zeros on data without Fourier transform, which is different from ZT and IN. The CNN model is used for this experiment and has the same structure as Fig. 8 (b). We illustrated the performances for two-sessions since our focus is finding time-stable features. From table VII, DTW, ZT and IN have the better performances in both databases; thus, we considered these features for further implementation.

Additionally, we simulate the performance according to the order of stacking in table VIII. In general, data manipulation can cause the changes in the output. However, in our case, the results in table VIII explain that there is a minimal difference in performance when the order of stacking is changed. Moreover, we demonstrate in Fig. 13 about the two-sessions accuracy of 2 channel input in 20 subjects from Biosec2 to further show the small changes in performance when stacking order is changed.

After implementing conventional classifiers along with different classical features, we evaluate our approach for different scenarios. For the two cases without data augmentation (which are with CNN, and with CNN and LSTM), configurations of input data are similar as single-session. The training data is coming from the PPG signals measured at the first time while the testing data is constituted from the PPG signals recorded in the second time (session). The size of training and testing data are same.

For the remaining four cases with data augmentation (which are augmentation v1 and v2 with each model), they have a dif-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIFS.2020.3006313, IEEE Transactions on Information Forensics and Security

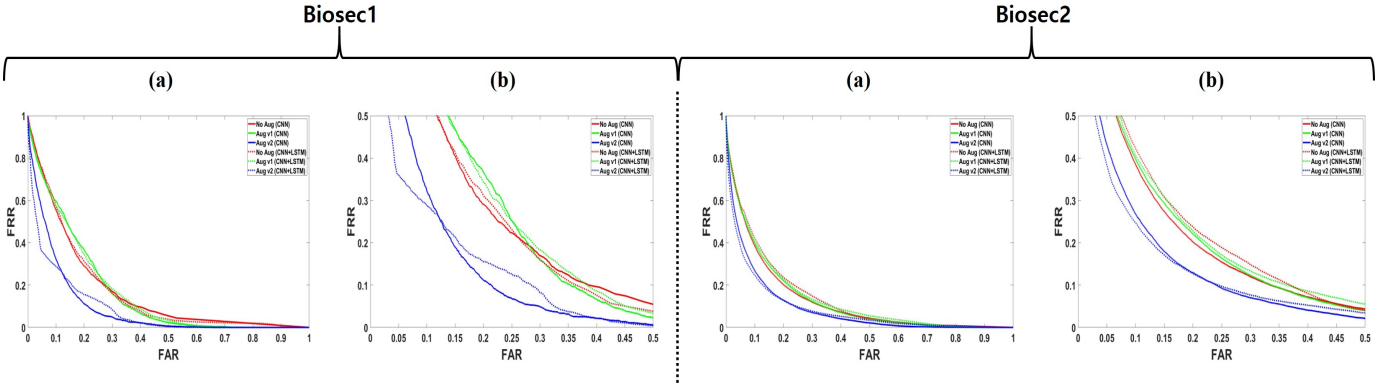


Fig. 12: Average ROC curves in two-sessions. Name of database is denoted on the top. Red/green/blue color means the two-sessions performance in no augmentation/augmentation v1/augmentation v2. Solid line denotes the performance from CNN model and dotted line describes the performance from CNN with LSTM model. (a) Average ROC curve with combination input. (b) zoomed in (a).

TABLE VI: Conventional features with simple classifiers and CNN in two-sessions. Denote as follows: Avg of ACC (Avg of EER).

	Biosec1	Biosec2
SVM-Linear (Fiducial+Non-fiducial)	66.4% (33.5%)	67.3% (32.8%)
3NN (Fiducial+Non-fiducial)	60.6% (39.3%)	58.8% (41%)
Decision Tree (Fiducial+Non-fiducial)	65.2% (34.8%)	61.5% (38.4%)
SVM-Linear (Linear interpolation)	68.3% (31.9%)	71% (29%)
3NN (Linear interpolation)	60.2% (39.7%)	60.3% (39.7%)
Decision Tree (Linear interpolation)	60.3% (39.7%)	61.9% (38.2%)
CNN (Linear interpolation)	**70.6% (29%)**	**78.5% (21%)**

TABLE VII: Different stretching methods with CNN in two-sessions. Describe as follows: Avg of ACC (Avg of EER).

	Biosec1	Biosec2
Linear Interp.	70.6% (29%)	78.5% (21%)
Zero padding	74.1% (25.8%)	78.7% (21.2%)
DTW	**79% (21%)**	**79.9% (19.8%)**
ZT	**75.5% (24.5%)**	**81.3% (18.8%)**
IN	**75.8% (24.1%)**	**81.1% (18.9%)**

TABLE VIII: Different stacking order with CNN in two-sessions. The performance is in terms of Avg EER.

	Biosec1	Biosec2
ZT+IN / IN+ZT	24.9% / 24.4%	19.5% / 19.3%
DTW+IN / IN+DTW	25% / 24.6%	18.6% / 18.4%
DTW+ZT / ZT+DTW	25.7% / 25.4%	21.1% / 21%
DTW+ZT+IN (there are 6 cases)	24.8% / 24.4% / 24% 24.2% / 24.4% / 24.2%	18.9% / 18.5% / 18.5% 18.3% / 18.3% / 18.2%

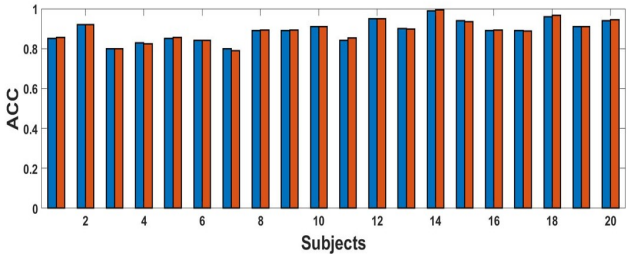


Fig. 13: Performance of DTW+IN vs IN+DTW for different subjects from Biosec2. They are similar even with different order of stacking. Blue color denotes DTW+IN while red color describes IN+DTW.

ferent training protocol from the single-session. For Biosec1, we extract 100 input data from each subject as single-session but also add the three different noises (Gaussian, sloping and combined noise) in each input to increase the size of the data for the training phase only. In other words, we obtain 400 input data from each subject and as a result, training data for Biosec1 is 12400. For Biosec2, we do not increase the size of input as much as Biosec1 because of the large execution time. As the case of data augmentation in single-session, we take 54 input in each measurement for one subject and then, add individual noise in each trial. To be clear, Gaussian noise is added on first measurement, sloping noise is included in second measurement and combined noise is considered in third measurement. Finally, there is 108 input for each measurement, which are 324 input for one subject and thus, training data for Biosec2 is 32400.

For creating the verification system in two-sessions, inputs from first session are used for training and inputs from second session are applied for testing without overlap. The number of epoch used for training is 80 and 60 for Biosec1 and Biosec2 datasets, respectively. Even we increase the number of epoch in Biosec2, the performance was similar and thus, we choose it to save the execution time.

Fig. 12 introduces the average of ROC curve in Biosec1 and Biosec2. When we compare the results in two databases, most cases (except for augmentation v2 with CNN) are better in

Biosec2. From this perspective, we can conclude that multiple data collections in each session are helpful to catch the time-stable and unique features in the PPG signals, even we also have various and large testing data. Also, it is obvious that the performance of augmentation v2 (blue) in each model overwhelms other cases because of the smallest AUC. Thus, rebuilding the data from normalized and non-normalized input is great to obtain time-stable and distinguishable features. From Fig. 12, other cases in each database give very similar results, meaning they are not beneficial for understanding the time stability.

TABLE IX: Compare the two-sessions performance. Denote as follows: Avg of ACC (Avg of EER), Avg of execution time (seconds). Top scores are in bold and underlined.

	Biosec1	Biosec2
CNN	79.4% (21.3%), 120	82.6% (17.4%), 660
CNN+LSTM	79% (21.4%), 680	82.5% (17.5%), 980
Aug v1, CNN	79.1% (20.8%), 725	82.4% (17.6%), 1405
Aug v1, CNN+LSTM	80.1% (19.8%), 1129	82.9% (17.2%), 1695
Aug v2, CNN	86.8% (13.2%), 725	86.3% (13.7%), 1405
Aug v2, CNN+LSTM	**87% (13%), 1129**	**87.1% (12.9%), 1695**

Table IX gives the average of ACC, EER and execution time in each case with two different databases. As we illustrated in Fig. 12, we can see the similar patterns with certain performance. Biosec2 mostly gives better performance than Biosec1 because of diverse input data. Using augmentation v2, Biosec1 has more improvement than Biosec2, compared to no augmentation. This seems because Biosec2 has more various testing data and thus, it is more hard to be refined but provides more general result. Same as single-session, we get large execution time for using LSTM and increasing the size of input data. In all databases, augmentation v2 with CNN and LSTM gives the best performance but requires large training time, compared to augmentation v2 with CNN. Thus, we need to choose the use of model according to the resource and application. From table IV and table IX, we can understand that performances in single-session (whose input is equal to training data in two-sessions) are almost irrelevant with results in two-sessions. Thus, we can carefully conclude that overfitting problem in usual machine learning has almost no relation with time stability problem in the PPG signals.

TABLE X: Comparison of the two-sessions performance with other methods. Some values are missing (described as -) because they did not report those values.

	Nonin [20]	Berry [20]	Biosec1		Biosec2	
Verification/ Identification	Verification	Verification	Verification		Verification	
Number of subjects	24	24	31		100	
Average time gap (days)	7	7	36		17	
Method	[20]		[20]	**Ours**	[20]	**Ours**
Accuracy	-	-	68.23%	**87%**	62.7%	**87.1%**
EER	23.2%	19.1%	31.77%	**13%**	37.33%	**12.9%**

Table X explains the two-sessions performance in different databases. Nonin and Berry database come from [20], and they were collected with Nonin WristOx2 pulse oximeter and Berry pulse oximeter. These databases consist of 24 subjects with 7 days gap between the first and second session. We also consider their approaches in our databases and, compare to our method, we get low performances which are 31.77% and 37.33% average EER on Biosec1 and Biosec2, respectively. Thus, we can see that our proposed model has great performances than others with more subjects and longer time gap, which means our verification model has a high potentiality to be applied in real devices.

VII. Conclusions and Discussions

This paper focuses on investigating the feasibility of deploying the PPG signal for user verification system by exploiting the unique and time-stable features. For protecting user's valuable information, highly secure verification system is necessary, and the PPG signal has great advantages to overwhelm the conventional encryption and biometrics. The proposed biometric system possesses novel inputs to hold consistent length with different characteristics and two deep learning models (CNN and LSTM) are utilized to find features for distinguishing user from others, both in single and two-sessions. In addition, the methods of assigning the preferred input to specific subject and data augmentation are applied to improve the performance in two-sessions.

In single-session, our verification system gives 99% and 98% average ACC on Biosec1 and Biosec2, respectively. In public database (PRRB), we attained 100% average ACC which is superior as other works. In two-sessions, our proposed verification system shows 87% and 87.1% average ACC on Biosec1 and Biosec2, separately. From these results, our developed verification system outperforms other state-of-art works and reveals a high potentiality to be utilized in real world.

Future exploration is to refine the proposed verification system to be robust and constant even when new user is coming in. For now, our system shows stable performance if it is trained again with new user's data. However, this is not proper in real world because we cannot train our model in any time. We infer that saving the learned features from model can be useful to overcome this problem. As usual deep learning model, training time is large and thus, we need a separate server or resource to train the model before implementing in real device. To decrease the training time, we will optimize the algorithm and change the code with efficient language. Also, we will experiment the multimodal verification system with proposed inputs to improve the performance greatly. Lastly, we are still collecting the PPG signals in Biosec2 database and thus, we will analyze our verification system to see whether it is robust and precise on the larger database.

Acknowledgment

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Royal Bank of Canada (RBC).

REFERENCES

[1] A. Jain, Lin Hong, and R. Bolle, "On-line fingerprint verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 302–314, April 1997.

[2] N. Werghi, C. Tortorici, S. Berretti, and A. Del Bimbo, "Boosting 3d lbp-based face recognition by fusing shape and texture descriptors on the mesh," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 964–979, May 2016.

[3] J. Chen, F. Shen, D. Z. Chen, and P. J. Flynn, "Iris recognition based on human-interpretable features," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 7, pp. 1476–1485, July 2016.

[4] A. Sizov, E. Khoury, T. Kinnunen, Z. Wu, and S. Marcel, "Joint speaker verification and antispoofing in the$i$-vector space," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 821–832, April 2015.

[5] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, Jan 2012.

[6] S. G, S. Kp, and V. R, "Automated detection of diabetes using cnn and cnn-lstm network and heart rate signals," *Procedia Computer Science*, vol. 132, pp. 1253 – 1262, 2018, international Conference on Computational Intelligence and Data Science.

[7] A. Lourenço, H. Silva, and A. Fred, "Unveiling the biometric potential of finger-based ECG signals," *Computational Intelligence and Neuroscience*, vol. 2011, pp. 1–8, 2011.

[8] Y. Y. Gu, Y. Zhang, and Y. T. Zhang, "A novel biometric approach in human verification by photoplethysmographic signals," in *4th International IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003.*, April 2003, pp. 13–14.

[9] M. Elgendi, "On the analysis of fingertip photoplethysmogram signals," *Current Cardiology Reviews*, vol. 8, no. 1, pp. 14–25, Jun. 2012.

[10] I. Odinaka, P. Lai, A. D. Kaplan, J. A. O'Sullivan, E. J. Sirevaag, S. D. Kristjansson, A. K. Sheffield, and J. W. Rohrbaugh, "Ecg biometrics: A robust short-time frequency analysis," in *2010 IEEE International Workshop on Information Forensics and Security*, Dec 2010, pp. 1–6.

[11] W. Louis, M. Komeili, and D. Hatzinakos, "Continuous authentication using one-dimensional multi-resolution local binary patterns (1dmrlbp) in ecg biometrics," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2818–2832, Dec 2016.

[12] A. Guennec, S. Malinowski, and R. Tavenard, "Data augmentation for time series classification using convolutional neural networks," in *in ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, vol. 2, Sep 2016.

[13] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "Ecg heartbeat classification: A deep transferable representation," in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, June 2018, pp. 443–444.

[14] Y. Y. Gu and Y. T. Zhang, "Photoplethysmographic authentication through fuzzy logic," in *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering, 2003.*, Oct 2003, pp. 136–137.

[15] J. Yao, X. Sun, and Y. Wan, "A pilot study on using derivatives of photoplethysmographic signals as a biometric identifier," in *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2007, pp. 4576–4579.

[16] P. Spachos, J. Gao, and D. Hatzinakos, "Feasibility study of photoplethysmographic signals for biometric identification," in *2011 17th International Conference on Digital Signal Processing (DSP)*. IEEE, 2011, pp. 1–5.

[17] N. G. R. Salanke, N. Maheswari, A. Samraj, and S. Sadhasivam, "Enhancement in the design of biometric identification system based on photoplethysmography data," in *2013 International Conference on Green High Performance Computing (ICGHPC)*. IEEE, 2013, pp. 1–6.

[18] A. R. Kavsaoğlu, K. Polat, and M. R. Bozkurt, "A novel feature ranking algorithm for biometric recognition with ppg signals," *Computers in biology and medicine*, vol. 49, pp. 1–14, 2014.

[19] N. Karimian, Z. Guo, M. Tehranipoor, and D. Forte, "Human recognition from photoplethysmography (ppg) based on non-fiducial features," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 4636–4640.

[20] J. Sancho, Á. Alesanco, and J. García, "Biometric authentication using the PPG: A long-term feasibility study," *Sensors*, vol. 18, no. 5, p. 1525, May 2018.

[21] V. Jindal, J. Birjandtalab, M. B. Pouyan, and M. Nourani, "An adaptive deep learning approach for PPG-based identification," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, Aug. 2016.

[22] S. P. Shashikumar, A. J. Shah, Q. Li, G. D. Clifford, and S. Nemati, "A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology," in *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Feb 2017, pp. 141–144.

[23] L. Everson, D. Biswas, M. Panwar, D. Rodopoulos, A. Acharyya, C. H. Kim, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte, "Biometricnet: Deep learning based biometric identification using wrist-worn ppg," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[24] D. Biswas, L. Everson, M. Liu, M. Panwar, B. Verhoef, S. Patki, C. H. Kim, A. Acharyya, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte, "Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 2, pp. 282–291, April 2019.

[25] U. Yadav, S. N. Abbas, and D. Hatzinakos, "Evaluation of ppg biometrics for authentication in different states," in *2018 International Conference on Biometrics (ICB)*, Feb 2018, pp. 277–282.

[26] pulsesensor.com. (Accessed on: Dec. 1, 2019). [Online]. Available: https://plux.info/sensors/42-pulsesensor.html

[27] W. Karlen, S. Raman, J. M. Ansermino, and G. A. Dumont, "Multiparameter respiratory rate estimation from the photoplethysmogram," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1946–1953, Jul. 2013.

[28] W. Karlen, M. Turner, E. Cooke, G. Dumont, and J. M. Ansermino, "Capnobase: Signal database and tools to collect, share and annotate respiratory signals," in *2010 Annual Meeting of the Society for Technology in Anesthesia*. Society for Technology in Anesthesia, 2010, pp. 25–25.

[29] N. Karimian, M. Tehranipoor, and D. Forte, "Non-fiducial ppg-based authentication for healthcare application," in *2017 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Feb 2017, pp. 429–432.

[30] D. Hwang and D. Hatzinakos, "Ppg-based personalized verification system - ppsnet," in *presented at 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, May 2019.

[31] J. Arteaga-Falconi, H. A. Osman, and A. E. Saddik, "R-peak detection algorithm based on differentiation," in *2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP) Proceedings*, May 2015, pp. 1–4.

[32] M. R. Ram, K. V. Madhav, E. H. Krishna, N. R. Komalla, and K. A. Reddy, "A novel approach for motion artifact reduction in ppg signals based on as-lms adaptive filter," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 5, pp. 1445–1457, 2012.

[33] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiological Measurement*, vol. 33, no. 9, pp. 1491–1501, Aug. 2012.

[34] B. Huang and W. Kinsner, "Ecg frame classification using dynamic time warping," in *IEEE CCECE2002. Canadian Conference on Electrical and Computer Engineering. Conference Proceedings*, vol. 2, May 2002, pp. 1105–1110 vol.2.

[35] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115 – 2125, 2003, audio- and Video-based Biometric Person Authentication (AVBPA 2001).

[36] A. A. Ross and R. Govindarajan, "Feature level fusion of hand and face biometrics," in *Biometric Technology for Human Identification II*, A. K. Jain and N. K. Ratha, Eds., vol. 5779, International Society for Optics and Photonics. SPIE, 2005, pp. 196 – 204.

[37] J. Lee, S. Sun, S. Yang, J. Sohn, J. Park, S. Lee, and H. C. Kim, "Bidirectional recurrent auto-encoder for photoplethysmogram denoising," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2018.

[38] I. Gotlibovych, S. Crawford, D. Goyal, J. Liu, Y. Kerem, D. Benaron, D. Yilmaz, G. Marcus, and Y. Li, "End-to-end deep learning from raw sensor data: Atrial fibrillation detection using wearables," 2018.

[39] M. Zihlmann, D. Perekrestenko, and M. Tschannen, "Convolutional recurrent neural networks for electrocardiogram classification," in *2017 Computing in Cardiology (CinC)*, Sep. 2017, pp. 1–4.

[40] J. Luque, G. Cortes, C. Segura, A. Maravilla, J. Esteban, and J. Fabregat, "End-to-end photopleth ysmography (ppg) based biometric authentication by using convolutional neural networks," *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 538–542, 2018.

[41] S. Singh, S. K. Pandey, U. Pawar, and R. R. Janghel, "Classification of ecg arrhythmia using recurrent neural networks," *Procedia Computer Science*, vol. 132, pp. 1290 – 1297, 2018, international Conference on Computational Intelligence and Data Science.

Dae Yon Hwang received the B.S. degree in Electronic Engineering from Hanyang University (2014) and M.S. degree in Electrical Engineering from Texas A&M University (2016). He was a research engineer in Hyundai Mobis working on vehicle camera for autonomous driving (2016 - 2018). Currently, he is a Ph.D. candidate in the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada. His research interests include biometrics, signal processing, computer visions and machine learning.

Bilal Taha received the B.Sc. and M.Sc. degrees (Hons.) in Electrical and Computer Engineering both from Khalifa University, UAE in 2016 and 2018, respectively. He was a research intern at the ICUBE laboratory, University of Strasbourg, France, in 2017. Bilal is currently a PhD candidate at the University of Toronto and affiliated to the Biometrics Security Laboratory. His research interests include deep learning, computer vision and biometrics.

Da Saem Lee received the B.S. degree in Electronic Engineering from Gangneung Wonju National University (2014) and M.S. degree in Electrical Engineering from Texas A&M University (2016). She was a research engineer in LG Electronics working on mobile camera for enhancing image quality (2016 - 2018). She also worked in Biometrics Security Lab for developing android application and PPG verification model (2018 - 2019). Currently, she is a machine learning engineer in Advanced Micro Devices (AMD), Markham, Canada.

Dimitrios Hatzinakos received the Diploma degree from the University of Thessaloniki, Greece, in 1983, the M.A.Sc degree from the University of Ottawa, Canada, in 1986 and the Ph.D. degree from Northeastern University, Boston, MA, in 1990, all in Electrical Engineering. In September 1990, he joined the Department of Electrical and Computer Engineering, University of Toronto, where now he holds the rank of Professor with tenure. He served as Chair of the Communications Group of the Department during the period July 1999 to June 2004. Between 2004 - 2014, he was the holder of the Bell Canada Chair in Multimedia, at the University of Toronto. Since 2007, he has been the director of the Identity, Privacy and Security Institute (IPSI) at the University of Toronto. His research interests are in the areas of Multimedia Signal Processing, Multimedia Security, Multimedia Communications and Biometric Systems. He is author/co-author of more than 300 papers in technical journals and conference proceedings and he has contributed to 18 books in his areas of interest. He is a Fellow of IEEE and a Fellow of the Canadian Institute of Engineering and a member of the Professional Engineers of Ontario (PEO).