The biometric authentication using photoplethysmography (PPG): a twenty years systematic literature review

Benjamin Vignau*
Patrice Clemente
Pascal Berthomé
benjamin.vignau@insa-cvl.fr
patrice.clemente@insa-cvl.fr
pascal.berthome@insa-cvl.fr
Laboratoire d'Informatique Fondamentalle d'Orléans, INSA Centre Val de Loire,
Bourges, Cher, France

Abstract

In this paper, we made a systematic literature review of the authentication systems based on PPG. We collected and filtered more than 700 papers, giving us 44 relevant papers. For each of these papers, we analyzed the employed methodology developed by authors to authenticate people from their PPG record. We compared all the major phases: signal recording, noise filtering, feature extraction, and classification. The main observation is the heterogeneous conditions limiting the ability of researchers to compare their work on a common basis. Thus, in this survey, a common methodology is proposed to the community. Upon adoption, this could enable the community to compare their methods uniformly. To the best of our knowledge, we are the first to provide a systematic literature review that gathers all the papers talking about biometric authentication with PPG published between 2003 and late 2022 to analyze them with the prism of data science.

Keywords: Human authentication; PPG recognition; Deep Learning, Review, Biometric Continuous Authentication

ACM Reference Format:
Benjamin Vignau, Patrice Clemente, and Pascal Berthomé. 2023. The biometric authentication using photoplethysmography (PPG): a twenty years systematic literature review. In *Proceedings of (Journal of Systems Software).* ACM, New York, NY, USA, 40 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Journal of Systems Software, ,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Internet of things (IoT) and smart devices have grown to be ubiquitous in our daily life. Nowadays, smartwatches, smart fridges, smart toys, smart intimate devices, etc. [76] are widespread among the population. The goal of these objects is to improve our lives, and medical connected devices are more and more present. In the past decade, we saw the democratization of biometric authentication, mainly through our smartphones and their fingerprint sensors. Nowadays, passports also use fingerprint authentication. For example, France started to deliver them in 2006 [1] and Canada in 2013 [19]. Those sensors present many advantages but can be fooled with latex forgery [60]. Moreover, this authentication method is punctual, it authenticates someone once at the beginning of the session and never again. In the past few years, researchers showed the need to develop continuous authentication [75]. The main problem with static authentication is the impossibility to remedy a hijacked session. Continuous Authentication aims to re-authenticate the user multiple times during the session while keeping the process transparent for the user [52]. Many methods have been explored during the last decade, such as behavioral biometrics (keystroke, mouse movement, etc) [23]. Recently, the usage of IoT as wearables to enforce continuous authentication has been studied [67]. The two main advantages of wearable systems are the possibility to wear them discretely, without causing any discomfort to the user, and the possibility to continuously measure a physical signal (temperature, light, sound, force, etc.). The usage of IoT for biomedical technologies is evolving and Aledhari et al. [7] made a full description of the enabling technologies and the remaining challenges. IoT, such as smart wearables, can be used to monitor many physiological signal such as blood pressure, heart rate, glucose level etc. to improve medical monitoring of people. But we know that most of the physiological signals are unique to people. Thus we may use these signals, first measured for medical purpose, to recognize people and develop more ergonomic and robust authentication systems.

In this survey, we focused on the usage of plethysmography [8] (or PPG) sensors, also called PulseOxymeter sensors for the authentication of individuals in a computer system. PPG can be defined as a cardiac signal, measured with a LED and photo-optical sensors [3]. PPG is a method for measuring the amount of light that is absorbed or reflected by blood vessels in living tissue. Since the amount of optical absorption or reflection depends on the amount of blood that is present in the optical path, the PPG signal is responsive to changes in the volume of the blood, rather than the pressure of the blood vessels. In other words, PPG detects the change of blood volume by the photoelectric technique, whether transmissive or reflective, to record the volume of blood in the sensor coverage area to form a PPG signal. This signal represents the variation of blood pressure in veins, induced by heartbeats [27]. These sensors are used by many smartwatches to provide heart rate, or by medical devices to provide oxygen saturation (SPO2) [56]. It is worth mentioning that PPG is a non-invasive technique and it does not require direct contact with the skin, which makes it more comfortable for users and less prone to contamination.

Human heartbeat signals is a biological trait, which can be easily measured, such as voice, iris or fingerprint all used to recognize human. [2]. For heart authentication, two main methods are used: one with electrocardiogram (ECG) [57] and one with PPG. The ECG signal gives more information and is more precise, however it's harder to measure. To measure ECG multiple electrodes need to be stuck on individuals, whereas only one sensor is needed to be attached to the finger or the wrist to measure PPG. Moreover, PPG sensors are cheaper and widely used in hospitals and in commercial systems, which can measure your heart rate.

During the last few years, many research teams worked on this problem and many methods were developed. In this work, we review and compare these methods. We provide a systematic state of the art and identify challenges for future works in this domain. Our goal is to answer to the main research question (MRQ): *Can we use the PPG signal of a smart watch to build a continuous authentication system?*

To answer this question, we draw the evolution of the community of this research topic. Our paper is divided in the following sections:

- Section 2 presents the problem definition, why use the PPG and how to measure a good biometric authentication system.
- Section 3 presents our methodology used to collect and filter papers for systematic literature review. We also briefly present the most commonly used methodologies in biometric authentication with PPG.
- Section 4 briefly summarizes year by year all the collected papers, describing their methodologies and their results.
- Section 5 explains the conditions used to measure and obtain a PPG signal from subjects.
- Section 6 presents in-depth the methods described in the literature to reduce the noise in the PPG signal.
- Section 7 provides the methods used by researchers to extract and select features for authentication.
- Section 8 presents the most used algorithms to recognize individuals with PPG.
- Section 9 presents a short comparaison of multiple studies that used the same datasets.
- Section 10 gathers all our analyses of the studied works and provides challenges and recommendations for future works.
- Section 11 concludes the paper.

2 Problem definition and related works

2.1 Related Works

In this paper, we focus on the study of identification and authentication of people using PPG signals. The main advantages of this technology is its cost (few dollars for a PPG sensor), its difficulty to be counterfeited and the possibility to add this sensor inside wearable devices (watches, T-shirt etc). This leads to the ability to provide a new ergonomic, simple and non invasive form of continuous authentication. Finally, this technology also provides medical data that can be exploited to provide a medical monitoring to users.

The technology description, its advantages and disadvantages are described in most of the papers that we studied. However the authors from [44] made a full description of use case scenarios. To the best of our knowledge, they are the only ones to provide a survey on this problem. But they have studied only 14 papers, mainly between 2016 and 2021, and their study lack of a methodology section. This is why we have made this study, gathering 44 papers over 20 years and providing a full dataset of all the experiences realized on this topic.

2.2 Problem definition

First, we need to define the differences between authentication and identification and illustrate it in our PPG context. The authentication is the action to prove the identity of someone. The user gives the claimed identity with a proof and the system only checks the proof. In an authentication system with PPG an user could claim an identity and give its PPG signal that will be used by the system to check the identity.

The identification process is quite similar, but only provides the proof, and the system has to find the associated identity. Authentication is just a proof check or proof validation, while during identification the system have to check the proof with all available proof of identity in order to find the good identity.

Both identification and authentication rely on a proof check based on PPG, also called PPG-based biometric recognition method. These methods ressemble a template matching problem, or a classification problem. The goal is to separate the proof of each user and when a new one is provided, the system needs to find the right class (the identity of user) or match with the already known template of a user (in case of authentication). The heart of the problem is to be able to recognize one person with only its PPG signal.

From this definition, we understand that the processes of identification and authentication needs an enrollment phase where the user gives a first proof of its identity. Then, during the verification phase, the system has to check if the second proof given by the user matches the one used for enrollment. The enrollment phase consists in the creation of a database of templates for each authorized user.

An identification or authentication system is used to prevent identity thieves and impostors. Thus we need to have a good attention on the false positive match (where user A matches the identity of User B) and false negative match (in which user A presents a valid proof but is not recognized by the system). The false positive matches are a big problem in term of security because it lets people take the identity of others. The false negative matches are a problem for usability and ergonomic, because a user may need to authenticate multiples times before having one good authentication. A bad ergonomic of the system lead to the abandonment of technology.

To do this survey, we developed 10 research questions splitted in three mains axis: the robustness of the system, the ergonomy of the system and key factors of the biometric PPG recognition. Our research questions are described in Figure 1 and summarized here:

- 1.1 What are the performances in short term scenarios?
- 1.2 What are the performances in long term scenarios?
- 1.3 Can the system scale up with the number of users?
- 1.4 Are the performances stable against biological changes?
- 2.1 How many users can not use the system?
- 2.2 How many tries a user needs to be authenticated?
- 3.1 How much architectures have been tested?
- 3.2 How much architectures need to be tested?
- 3.3 Are some pieces of architectures more efficient than others?
- 3.4 Are some pieces of architectures more popular than others?

To answer these questions, we will describe our methodology to collect, analyse and classify the papers of the literature. Then we analyse the collected papers, and exhibit the main elements to answer our questions. We will also focus on the data used to build the proposed systems and the validation methodologies. This is part of a study comparison. In order to answer our research questions, we need to provide a clear and robust comparison methodology of the experiences. At last, we discuss the methods used by each paper to compare their work, and provide a first works on the metrics that can be used to compare the proposed methods.

#	Request
0	Personal Identification with PPG
1	Personal recognition with PPG
2	Signature with PPG
3	Biometric identification with photoplethysmography
4	Personal Identification with photoplethysmography
5	Personal recognition with photoplethysmography
6	Signature with photoplethysmography
7	PPG signal for biometric personal identification system
8	Photoplethysmography signal for biometric personal identification system

Table 1. Request made on Scholar and PubMed

#	Exclusion criteria
0	Does not use the PPG technology
1	Does not authenticate or identify human
2	Creation of a database but not using it to build a system to authenticate patients
3	Only explain the PPG technology
4	Only list the application of the PPG technology.
5	Identify actions, emotions, movement etc. but not human.
6	Multi modal authentication (ex PPG and ECG in one system).
7	Vitals monitoring.
8	Only keep the extended version of a paper

Table 2. Exclusion criteria

3 Methodology

In this section, we present the methodology we used to gather, filter and analyze papers about human authentication with PPG. Then we present the common methodology used by researchers to achieve PPG biometric recognition. We used the main phases of this common methodology to structure our paper.

3.1 Papers collection

To make an efficacious systematic literature review, we followed the guidelines provided by Wohlin [78]. Thus, we defined requests for two search engines: Google Scholar and PubMed. All of our requests were made on these two engines, for two periods: one with no time limits and the second on papers from 2017 to 2021. This step was done in April 2021. We took the first 10 results (when available). The 9 requests are given in Table 1. Then we did the same thing in September 2022 to add the last published papers.

This first collection gave us 360 papers. However, many papers appeared in multiple requests and multiple search engines. So we made a python script to merge duplicate papers resulting in 136 different papers. For each of these papers, we analyzed the title, abstract, and if needed the introduction and conclusion. We excluded all the papers which matched at least one of the criteria defined in Table 2.

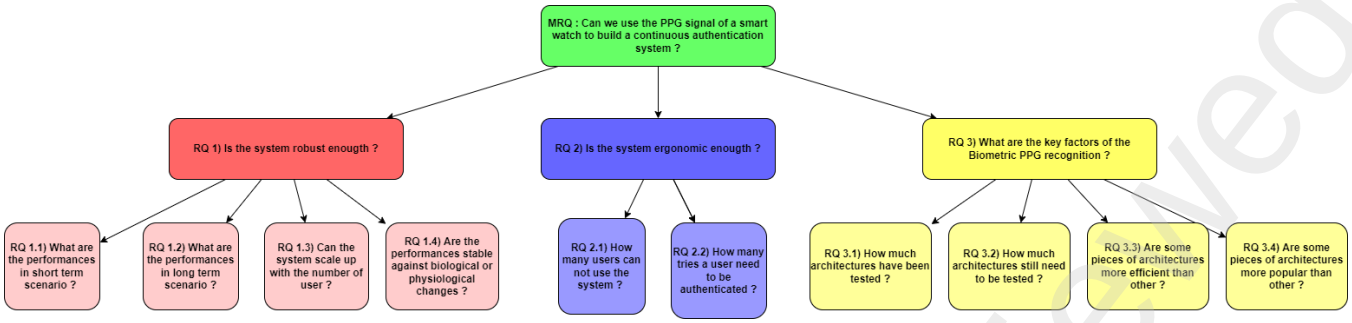


Figure 1. Research questions of this paper

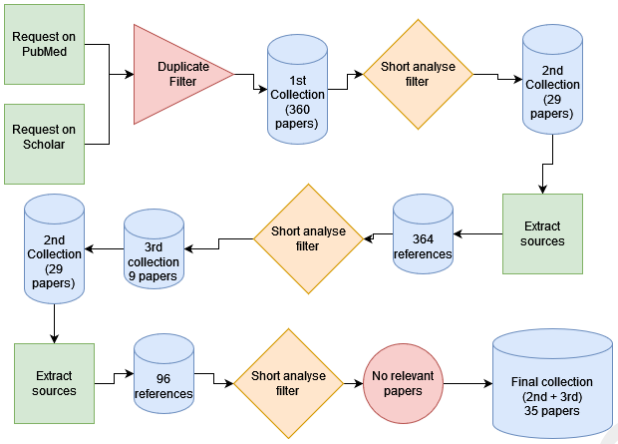


Figure 2. Methods for paper collection in 2021

At the end of the first filtration, we kept 29 relevant papers. Then, we extracted all the references for each of these 29 papers. This gave us 364 references to analyze. We applied the same process to analyze and filter papers. We added new exclusion criteria: if we found a paper and its extended version (ex $paper_1$ published in a conference and $paper_2$ published one or two years after, in a journal, to extend $paper_1$) we only kept the extended version. This first snowball allowed us to add 6 papers. Then we do the same process again, giving us 96 new references. After analysis, no papers in those 96 references were new or relevant. It resulted in no new references and, therefore, the collection was halted with 35 papers. This process is depicted in Figure 2.

Then we did the same process again at the end of September 2022, but only with papers published between 2021 and 2022. This give us 9 more relevant papers, and 4 papers which investigate fusion of PPG with other signals to make a biometric authentication. Finally we keep in our study 44 papers, from 2003 to 2022.

3.2 Analysis

To make a good analysis, we first read once all the papers. This allow us to extract the main phases of the design of PPG-bio metrics recognition methods. All methods can be segmented in four main phases:

- Signal acquisition
- Signal pre-processing
- Feature extraction and selection
- Classification or Matching

We used these majors phases as sections for our paper. Then in each of these phases, we identified specific criteria. The value of these criteria change for all papers. For example, in the classification phase, the algorithm used to authenticate patient changes over the years. In the oldest papers, simple metric calculus was used to classify the subjects, such as distance. In the most recent papers, deep learning algorithms are used (such as CNN for example). Our criteria applies for each phase given in Table 3. We describe each criterion and their possible values in each dedicated section.

Phase	Criteria
Signal Acquisition	Sampling Frequency
	Acquisition time
	Condition
	Total subjects
Signal Pre processing	Noise filtering
	Signal Segmentation
	Signal Normalization
Feature extraction & selection	Total extracted features
	Total fiducial features
	Total non fiducial features
	Extraction Method
	Selection Algorithm
Classification	Classification Algorithm Type
	Training dataset
	Evaluation dataset
	Validating method
	Accuracy
	Lowest False Matching Rate
	Lowest False Rejecting Rate
	Equal Error Rate

Table 3. Analysis criteria for each phase

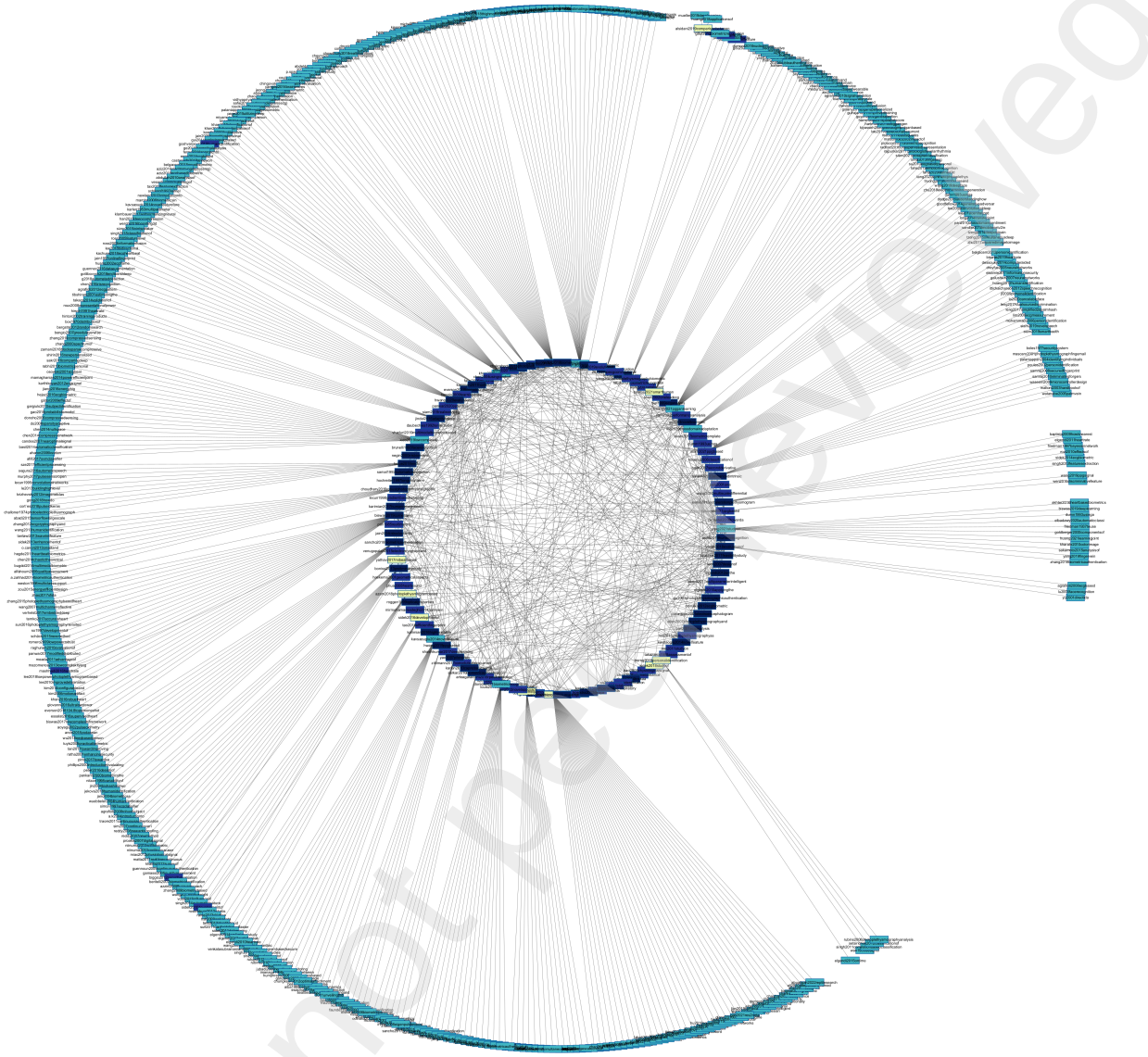


Figure 3. Representation of all the papers collected and their references

3.3 General statistics on the collected papers

In all the studied papers, we extracted all the experiences and the criteria for each one. For example, if a paper made an architecture for the signal pre-processing, feature extraction and selection and classification we counted it as one experience. If the same architecture is tested over two different datasets (as in two publicly available databases of PPG), we counted two experiences. Each time where the value of one criterion changed, we counted a new experience. All the gathered data are stored in a csv file hosted on our Github repository: https://github.com/bvignau/PPG_SLR_dataset

Figure 4 represents general statistics of the number of papers and experiences done from 2003 to 2022. We can observe two main periods: before 2014 and after 2014. We can see very few experiences done before 2014. Fewer than 5 experiences per year were made during this period. With the histogram, we can observe that before 2013 only three papers were published, one in 2007 and two in 2003. After 2013, at least one paper was published each year. We can observe a big increase in the number of the publications from 2016 with at least 5 paper per year from 2016 to 2021 (except in 2017). Moreover we can see that the number of publications is not really correlated to the number of experiences, showing that few papers made most of the experiences. For example, in 2014 [42] was the only published paper but referenced 120 different experiences, which represents 19.9% of the total experiences conducted from 2003 to 2022.

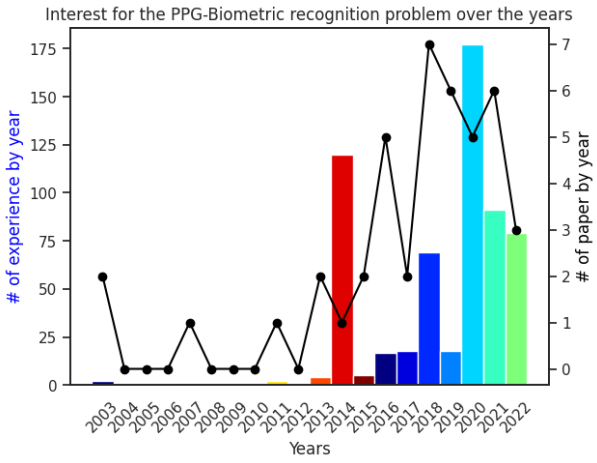


Figure 4. Number of experience and number of published paper per year

4 Previous works summary

In this section we will explain the main works done between 2003 and 2013. Then in the next we will explain the main works done between 2014 and 2022.

4.1 Works summary from 2003 to 2013

During this period, 7 papers were published, gathering 9 experiences.

4.1.1 2003. The first paper published about PPG biometric recognition was the one made by Gu et al. [31] where the authors made the first experience to recognize people with PPG. They collected data over 17 people. To recognize people they extract four fiducial features from the raw PPG signal, and stored them as a template. Then they computed a ratio for each variable to maximize inter-class variation and minimize intra-class variation. Then they used a classical distance metric to recognize people. Next they made a second experience published the same year [30]. Here the authors used a fuzzy logic on four fiducial features to recognize the subjects. They used a Gaussian function to make a template matching between the signal recorded in enrollment and the provided signal. They achieved to recognize people in 82.3% of the total tests. As for the previous works, they tested only true identity, they did not test impostors. These first works were good enough to start the research on this topic; with that said, many things are lacking: testing on impostors, compute the accuracy, false matching rate (FMR) and false non matching rate (FMNR) and Equal Error Rate (EER).

4.1.2 2007. No papers were published for 4 years, until the one made by Yao et al. in 2007 [84]. In this work, the authors extracted fiducial features from filtered PPG and its two derivatives. They collected data on 3 patients. Then they showed the correlation between the features extracted from different pulse for each patient and the poor correlation between features extracted from different patients. They concluded that it was feasible to identify people with the PPG. However, no identification nor authentication metrics were provided.

4.1.3 2011. In 2011, Spachos et al. [73] published a study made on 29 subjects taken in two public datasets: OpenSignal PPG Dataset and Biosec 1 [54]. The Biosec1 dataset is still available while the OpenSignal PPG Dataset is not available anymore. This work is the first to clearly define a methodology for all the steps: single pulse segmentation, normalization, feature extraction and classification. Here the authors used a Linear Discriminant Analysis (LDA) to compute weight for each pulse and create a template for each user. In the verification stage, they computed LDA weight for the input signal and use a KNN and a major vote to class the input signal and match the identity of the user. They achieved a 0.5% of EER for the Opensignal PPG dataset and a 25% EER for the Biosec1 dataset. This work provided a good improvement in the methodology for the biometric-PPG recognition and is a good feasability study. However the parameter for each stage were lacking (number of weigth use for LDA, K, etc.)

4.1.4 2013. In 2013 three papers were published [17, 62, 63], gathering 4 experiences. Salanke et al. published two papers [62, 63]. In the first one they split the signal in single pulse then used the Kernel Principal Component Analysis (KPCA) to reduce the dimensionality and used a Mahalanobis distance to compute intra and inter subject variation. No parameters and no metrics are given (number of feature, accuracy etc.) In the second one, they introduced the signal decomposition and recomposition using the FFT to reduce noise in the signal. Then they used the Semi Discrete Decomposition (SDD) method to reduce the dimensionality of the filtered signal. Finally, they have tested two feature selection methods. In the first one they only took the first 5 coefficients after SDD for each subject. In the second one they took the q first coefficients where q changes for each subject. Finally, they computed the Euclidean Distance between stored template and input signal to identify subjects. They drew the intra subject variation and inter subject variation for two subjects but did not try to use the system to identify people and compute metrics about accuracy EER etc. They just showed that their techniques may be usable to identify people. In both papers, they used the same dataset, collected on 9 subjects from their university. These two papers did not provide much interest for the community due to the lack of methodology and the lack of metrics about the developed system. On the opposite, Bossini et al [17] made a study on 44 subjects in which they filtered the signal with a high-pass Butterworth filter. Then they computed a template for each subject using a fixed number of single pulse. For each pulse

they computed the correlation with all others. If the correlation value was too low the pulse was removed from the dataset. Then to identify a subject, they computed an input matrix with the same method using the same number of pulses and computed the correlation between the template and input matrix. They tested 6 different ways to merge the data and obtained a matching score but only presented the result for one: The maximum value of the correlation. Finally they provided multiple metrics on this system. They tested identification with genuine and impostors. They achieve a 5.29% EER wich is quite adequate.

In conclusion, between 2003 and 2013 few papers were published with few experimentations. Most of the papers provided simple studies, with poor metrics and poor methodology. Most of the datasets are not publicly available and most of the studies concluded to the feasibility of using the PPG to identify people with their PPG.

4.2 Works summary from 2013 to 2022

In this second period, research about PPG-biometric recognition increased a lot.

4.2.1 2014. In 2014 Kavsaoglu et al. [42] provided 120 different experimentations on this topic. They extracted 40 different time domain features on the raw PPG, first and second derivative. Next they ranked from the most important to the least using a Z-score. Finally they used a subset of the extracted features to compute a template and a KNN and major voting with Euclidean distance to identify subjects. They tested multiple values for K (1 ; 3 ; 5 ; 7; 10) and for the number of extracted features (5 - 10 - 15 - 20 - 25 - 30 - 35 - 40). They collect data on 30 healthy subject, 15 cycles in two sessions (no precision on the delay between the two sessions). They tested their methods in a sub-dataset containing only the first session cycle (CUSTOM 1), a sub-dataset containing only the second session cycles (CUSTOM 2) and the full dataset (CUSTOM 1 + CUSTOM 2), thus leading to 5 * 8 * 3 = 120 different experiences to test one single architecture. They computed accuracy, recall, specificity and f-measure for each subject and in mean for all experiences. This allowed them to find the best parameter combination for the KNN algorithm. Their results showed that the ranking process significantly increased the accuracy. However the optimal number of extracted feature change from one dataset to another. They achieve good accuracy, over 90%.

4.2.2 2015. In 2015 two papers were published, gathering 5 experiences [37, 45]. In the first one, the authors studied the impact of using only the APG (second derivative of the PPG) to authenticate people. To do so, they used the MIMIC dataset, split signal in single pulse and derive the signal two times. They extracted 5 fiducial points in the APG and used it in a classifier. To classify the people, they used the Naive Bayes and KNN. They used the 10-cross fold validation methods to avoid overfitting problem. They also compared their system with the same fiducial features extracted from the raw PPG signal. They showed that the accuracy is better when they used the features extracted on the APG signal. The Naive Bayes classifier seemed to be better than the KNN algorithm and provide 97.5% accuracy vs 90% for the KNN. The results were pretty goods and used a public dataset, their architecture must now be tested on bigger datasets.

The second one extracted 22 physical features from single pulse (different lengths and angles in the signal) and used a CNN to classify them. They achieved a 4.2% FMR and 3.7% FNMR wich is quite good. However the test were made on only 10 subjects, using a custom dataset. They are the first one to use deep learning methods to identify people with PPG.

4.2.3 2016. In 2016 17 experiences were made for 5 published papers [20, 21, 39, 65, 71].

Sidek et al. [71] used the public dataset MIMIC II to study the usage of the APG to identify people. They used a Butterworth filter to delete the high frequencies in the signal. Then they segmented the signal and created a new representation called "cardiod representation". To create this representation they extracted waves from the signal and plot them in a circular diagram.

They used the main parameters of this representation to feed multiple deep-learning algorithms. They tested the Multi-Layer Perceptron (MLP) and the Naive Bayes Classifier (NBC). They achieved a 95% accuracy for the two classifiers. However, they achieved 45% and 55% accuracy when using raw PPG and not APG.

Choudhary et al [21] used a public dataset MIT-BIH Polysomnographic Database and built a simple architecture. They split the signal in single pulse, normalized it in time and amplitude. Then they filtered the signal with a Gaussian derivative filter (GDF) and used an ensemble average technique to build template. To match the identity of subjects, they tested three methods: Normalized Cross Correlation with averaged pulsating waveform, Wavelet weighted-based PRD with averaged pulsating waveform and Wavelet distance measure with averaged pulsating waveform. They achieved 29% of EER for the best score, which is not very conclusive in comparison with the 5% of EER achieved by less recent experiences.

Chakraborty et al [20] made one experience, using a custom dataset composed of 3-minute signals for 15 subjects. They segmented the signal in two pulse cycles, normalized the signal in amplitude and used a Butterworth filter to reduce the noise. Then they extracted 12 fiducial features from the raw PPG and use the Linear Discriminent Analysis (LDA) to class subject. They claimed to achieve 100% accuracy with this architecture. However they did not test any impostors scenario cases, and the used datasets were too small. This may be over-fitting and the experience should be reproduced using publicly available dataset.

Jindal et al [39] made one experience using the publicly available dataset TROIKA. They segmented the signal in single pulse, then used the standard deviation to normalize the signal in amplitude and fixed a input size of 125 points. Then they used a Buterworth filter of the 7th order and a moving average filter to reduce noise. Next, they extracted 11 statistical feature from the time domain and used a Deep Belief network (DBN) to classify the subjects. They achieved 96.1% accuracy using a 10-cross fold validation method. The results seems consistent, and this architecture should be selected for experiences with a higher number of participants.

Sarkar et al [65] made 8 experiences using the DEAP dataset, where subjects had to watch different emotional video while bio-metrics signals were recorded. They were the first to use this type of dataset. They first segmented the signal in single pulse and use an angular transformation to map the value of the signal from 0 to 2π. Then they used the Gaussian Decomposition to generate features. They have done so to extract 3 and 5 features. After that, they tested LDA and QDA to classify subjects. Finally, they tried to train their classifier with 75 or 100 pulses. In the end, they achieved good accuracy scores up to 95.67%. In general QDA seems to be more efficient than LDA and the increase of features and training set increase the performances.

4.2.4 2017. In 2017 18 experiences were done and two papers were published [40, 41].

Both papers were published by the same team. In the first one, the authors use the Capnobase IEEE TBME dataset to test two feature types and three algorithms. They segmented the signal in single pulse then used a Butterworth band pass filter of the 2nd order to reduce noise. Then, in the first set of experiences they extracted fiducial features and used the KS-test and the KPCA to reduce the number of features to 10. In the second set of experiences they used the Discrete Wavelet Transform (DWT) to extract features. They were the first one to use this technique to extract features. Then they used the same technique to select the best 10 features, after which, they tested the SVM, SOM and KNN algorithms. They achieved good accuracy scores with 99,84% in the maximum, using KNN and the DWT features. From their results, the differences in performances between the algorithm is somewhat low and may not be significant. However the usage of DWT features seem to really improve the performances compared to fiducial features.

In their second paper, the authors made 12 experiences, using the same datasets and similar filtering methods. However, in this paper, they added a Zero mean normalization of the signal before filtering. They tested two different kinds of features: Fiducial and Wavelet decomposition domain (DWT). To select the best features from the morphology and the DWT they used a Genetic Algorithm. They tested with and without selection algorithm but never provided details about the number of features used. Finally they tested the SVM and MLP algorithms to identify subjects. It seems that the differences between SVM and MLP are not significant. In both case, when using the DWT or morphology features combined with a genetic algorithm, they hit 100% accuracy. However, no details were provided about training and testing set, no validation methods were used. Thus this may have been over-fitting. The experiences should be reproduced with better validation methodology.

4.2.5 2018. In 2018 69 experiences were made, over 7 published papers [6, 24, 32, 51, 64, 70, 81].

Sidek et al [70] made 8 experiences with a custom dataset. They used a single cycle segmentation and zero mean normalization with a Butterworth filter to pre process the signal. Then they tested the extraction of fiducial features, 2 on raw PPG or 5 on APG signal. Finally they tested KNN, Bayes Networks, MLP and SOM algorithms. No details about training and testing were provided. In all cases, the usage of 5 features on APG provided better performances. The SOM algorithm seemed to be the best with 96% accuracy.

Horng et al [32] made 5 experiences, using a custom dataset. They used a single cycle segmentation and used a Butterworth high pass filter and a Low Pass filter and a polynomial decomposition and a Savitzky-Golay filtering to reduce the noise. They extract 30 fiducial features and test multiple algorithms: Fuzzy logic, KNN, Naive Bayes, Random Forest, MLP. They used 66% of the available data for training and the rest for testing. They used a 10-cross fold validation. The results were quite similar for all the algorithms, except for Random Forest and Naive Bayes which were less efficient. All the others achieved 94% accuracy which is in the norm compared to others experiences.

Yadav et al [81] made 5 experiences using the Capnobase IEEE TBME. They segmented the signal in 3 cycles, used the zero mean normalization and a Butterworth band pass filter to pre-process the signal. Then they used the continuous wavelet transform (CWT) to extract features and test LDA, DLDA, KDDA, KPCA and PCA to select features. But they did not said how many features they have extracted. Then they used the Pearson's distance to achieve template matching. In the best case they achieved an EER of 0.46% which is relevant. However the results may be hard to reproduce due to the lack of details about the number of features extracted and the precision about the training and testing set.

Everson et al [24] made one experience on the TROIKA dataset. They did not provide any details about the pre processing methods. They simply built a CNN called BiometricNet and feed it with the raw signal. They said to have achieved a 96% accuracy score but they did not provide any details on training, testing and validation methods.

Sancho et al [64] made 49 experiences. In this study, the authors gatherd 4 publicly available datasets to study the usage of PPG signal for authentication. They study two main problems: the authentication in short terms, where they used

signals collected within the same session to enroll and test an user. The long term authentication study the usage of two distinct signals for enrollment and verification. For example using one signal for enrollment and using another, acquired one week later for testing. Moreover, they tested multiple feature extraction methods and two distances metrics for a template matching architecture. They results were interesting, they showed a big increase of EER with the long term study, whilst EER went over 20% while in short term it stayed around 10%.

Most of the architecture provided likewise results. We can say that the augmentation of cycles for the training improve the performances. But the standard deviation between the algorithms are overlapping, thus, we can not conclude to that one outperforms the other.

Luque et al [51] made one experience using 1s of signal per subject with a dense neural network to identify users. No other details are provided.

4.2.6 2019. In 2019 18 experiences were made over 6 papers [16, 25, 33, 36, 47, 80].

Xiao et al [80] made one experience on a custom dataset. They segmented the signal into single cycles, applied Wavelet transform decomposition and recomposition to reduce the noise. They extracted 12 fiducial features, that are given to a SVM-RBF classifier. They used the 10-cross fold validation and achieved a 91.31% of accuracy which is positive and corresponds to other papers levels.

Lee et al [47] made 12 experiences using the Capnobase IEE TBME dataset. They tested multiple segmentation methods: 10, 30, 50 and 100 cycles. In all case, they used the Zero mean normalization and extracted features using the Discrete Cosinus Transform (DCT). The number of features was not given. Then they tested Decision Tree, KNN and Random Forest algorithms. Each algorithm showed similar accuracy score with all the segmentation methods. In all cases, Random Forest seemed to be the best with 99% accuracy. This showed that using more than 10 cycles does not improve the performances.

Al-sidani et al [5] made one experience using the VORTAL dataset. Very few details were given about the architecture. They only said to extract 40 fiducial features from raw PPG and its two derivatives and used KNN for classification. They claimed to achieve 100% accuracy. Then they compared it to the SVM algorithm using the same features. The SVM show lower results. They only used 23 patients on the 100 available. The KNN score is quite high compared to all other studies and the lack of details in the paper did not allow us to conclude to an good enough architecture.

Farago et al [25] made only one experience. They collected a custom dataset of 5 subjects, used a Butterworth band pass filter, extracted only one fiducial feature (peak to peak interval) and used the cross correlation to identify people. They achieve 98% of accuracy.

Hwang et al [33] made two experiences on the Biosec1 and Biosec 2 dataset. They split the signal in 1000 point samples, then reduced the noise with a Butterworth band pass filter. They used the raw signal to feed a CNN+LSTM algorithm. With Biosec 1 they used 75% of the datasets to train the algorithm and the rest to test it. They used the 10-cross fold validation and achieve 99.8% of accuracy which stands as strongly relevant. It was the first time that natural language algorithms were used in this domain. For the second experience, they used the first session to train the algorithm and the second one to test it, using once more the 10-cross fold validation. They achieved 99.8% showing the robustness of their architecture for long time stability.

Biswas et al [16] used the TROIKA dataset to make one experience. They did not split the signal and used the zero mean normalization and a Butterworth band pass filter for the preprocessing stage. Then all the signals are used to feed a bi-layer 1D CNN, which extracts the best features. Then the output of this CNN fed 2 LSTM which provide the classification. They achieved 96% accuracy.

4.2.7 2020. In 2020, 177 experiences were made, across 5 papers [9, 35, 43, 46, 82].

Yang et al [82] made 80 experiences using 3 publicly available dataset: BIDMC, MIMIC II and Capnobase. They segmented the signal using a Sliding window. Next they transformed the signal using a soft-max vector of the sparse representation. Then they defined 3 different layers of feature extraction. They tested all the possible combinations of feature extractors. Finally they compared the KNN, RF, LDC and NB algorithms as classifiers. In all experiences, they used 80% of the available data for training and 20% for testing. However they did not use a validation technique such as 10-cross fold validation. They achieved satisfactory accuracy scores, with most of them between 85% and 100%. The feature extractor only influenced the scores of the NB and LDC classifier, with a signification loss of accuracy when using only the layer 3 extractors.

Using a combination of multiple extractors seems to improve the performances for all the algorithms, expect for the KNN algorithm. It is the best algorithm from these experiences and the only one to hit 100% accuracy. However, they reproduced the experience using the combination of all the feature extractor 5 times and show the variability of the results. The 100% accuracy of KNN was hit 3 times out of 5 on the BIDMC dataset and 2 times out of 5 on the Capnobase dataset. The experiences provided by this paper were representative enough but must be reproduced with a better validating method, such as 10-cross fold validation.

Khan et al [43] made 7 experiences using a custom dataset. They used an empirical mode decomposition and recomposition to reduce the noise in the signal. Then they extract 20 temporal and frequency domain features. Finally, they tested 7 different algorithms using the 10-cross fold validation. The

tested algorithm are: QDA, Linear SVM, Quadratic SVM, Cubic SVM, Medium Gaussian SVM and Naive Bayes. Their results showed a better performance from the quadratic SVM with 93.1% of accuracy. This was satisfactory and the result should be reproduced using a public dataset.

Lee et al [46] made 4 experiences using two datasets: the TROIKA and a custom one. In this paper, the authors derived the MobileNet Neural network to work with PPG in one dimension. They filtered the noise with a Butterworth band pass filter of 5th order. Then they fed the raw signal with the classifier. They tested the PPG-MobileNet and BiometricNet classifier. They indicated the standard deviation of their accuracy which is quite interesting. Thanks to that the comparison of their experience was more robust. For example, we can see that there is no difference in accuracy between the two tested models, when tested with the TROIKA binary class, while in all other experiments their model surpassed the BiometricNet. They reached 95.68% of accuracy showing good performances. The results must be reproduced with a validation technique.

Alotaiby et al [9] made 5 experiences using the Capnobase dataset. To reduce noise, they used a moving median filter. Then the signal was split in different frame lengths (1 , 3, 5, 7, 10 or 15s). To extract the desired feature, they created multiple vectors of statistical features extracted from the raw PPG signal with its first derivative and on the signal filtered with DWT. Then they made multiple experiences with the usage of one or multiple vectors. However the results were not clear. Moreover they did not provide all the results for all the experiences. Most of the results can not be exploited. Finally they tested multiple classifiers like KNN, SVM or RF. with only 15s vectors and the vector from DWT decomposition. They achieved 99.3% of accuracy with a 0.02% of EER, which is relevant. The results should have been reproduced with a validation technique and using all the publicly datasets to check if the model were stable in long time and works well with a significant increase of the number of subjects.

Hwang et al [35] made 80 experiences. In this paper, the authors collected PPG signals over 100 subjects and used it with two other public datasets to develop a model based on CNN and LSTM. The two other publicly available datasets are Biosec1 and Capnobase. They wanted to study the time stability of the PPG authentication, using signals collected in two distinct sessions, seperated by a duration of 17 days. They filtered the signal with a Butterworth 4th order filter. Then they computed mean HR for each people and used it to split the signal in single cycles. Then they removed bad cycles in which the heart rate was too low or too high. They tested multiple feature extractions: DTW, Zero Padding in Time or Interpolation in frequency. After treating the signals, they computed the mean shape for all of them and removed the outliers. Next they used a data augmentation to select the best features for each subject. They developped two kinds of models: CNN and CNN+LSTM. To prevent overfitting problem, they used 10 fold cross validation with L2-regularization. All the details for each layer were provided, which is much appreciable. Concerning the experiments on the selection methods, the differences in average accuracy are very low, and the standard deviation should have been computed, which means we cannot draw conclusions about the efficiency of the selection method. They also made experiences to compare the architecture and the selection method, depending on the database. With the two channel data (DTW + IN) the results are better compared to the one channel usage. Moreover, in both cases, the feature selection method using two channels provided better results. However in the two channel scenario it was the CNN architecture which performed best while in one channel experience it was the CNN + LSTM architecture that dominated. In both cases we can observe that the performance on Capnobase dataset are higher and hit 100% accuracy with 0.1% EER in the two channel scenarios. The authors explained this difference by the fact that the Capnobase dataset used a better PPG sensor in a controlled environment, providing a better signal with less noise.

Next, the authors experimented on the time stability of the PPG by using signal collected in one session for training and signal collected in a 2nd session for testing. They showed a significantly drop of the performances when using a two session scenario. The best performances in two session scenarios are 81.3% of accuracy and 18.8% EER which is still too low for a real case usage.

4.2.8 2021. In 2021, 91 experiences were made and 6 papers were published [14, 22, 34, 69, 83, 85].

Donida et al [22] made two experiences using the Capnobase dataset. In this paper, the authors tried to identify users using a template matching method based on the spectogram of the PPG. They first normalized the signal, then computed a pseudo image from the signal using a spectrogram feature extraction. Then they used a PCA to reduce the dimensionality and finally identified the user with a KNN or an package of SVM. They hit 99.16% accuracy. A quick remark is that they used all the available data in the training dataset to compute the PCA coefficients; with this technique authors need to recompute all the PCA coefficient for each new user enrollment, leading to the impossibility to use this model for real world cases.

Bastos et al [14] made two experiences over the MIMIC II and the Capnobase datasets. In that paper, the authors tried to create a new authentication system for human based on ECG or PPG. They wanted to use IoT to collect those signals and they were willing to store the ID of people inside the device and make a template matching system. Their algorithm used six layers: signal filtering, peaks detection, specific waves correlation, correlation mean extraction, correlation between media and specific waves and template

generation. By specific waves correlation, the authors created a matrix of multiple single filtered cycles. They assumed that this step returned a matrix, where each column meant a sample of the wave, and each line meant to represent a wave. Following this procedure, they calculated the mean of each wave and created a template with it. Then they used a simple correlation between input signal and mean template to identify people. The last layer was just storing the mean waves template inside the IoT. To test the process, they used the two first minutes of each available signal from each patient to compute their mean wave template. Next they used the last minutes of the signal to test the model and define accuracy. We can see here that the testing method could have been more effective. Indeed, all subjects were enrolled and no impostors were used, as it would have been done in a k-fold validation. The authors said that they achieved good accuracy, however, it can be seen that their system produced a lot of false positives compared to true positives. For example, they have 50 true positives for users in the MIMIC database but 85 false positives. Therefore, it can not be used for real-life events.

Yang et al [83] produced 42 experiences over the BIDMC, MIMIC II and Capnobase datasets. In this paper the authors studied the PPG biometric recognition based on multiple classifiers. They extracted 17 time domain features, 4 frequency domains and 4 features from wavelet decomposition. The time domain features are classical fiducial points such as min value, max value, peak value and other metrics such as mean value, square root amplitude, Skewness, Kurtosis. For the frequency features they used Gravity frequency, Mean frequency, RMS frequency and frequency standard deviation. The four features extracted from wavelet packet decomposition are: frequency band energy ratio, energy entropy, scale entropy and singular entropy. They experimented multiple models: Linear discriminant classifier and Naive Bayes classifier. Then they tried the Euclidean distance on their feature vector. For each they computed the recognition rate (accuracy) and FAR and FRR. They enrolled all subjects at each test and use 80% of the available signals of each subject for training. They obtained interesting results, showing how adding frequency and wavelet domain features improves accuracy. However, each kind of features alone is not enough; for example, using only frequency domain feature with the BIDMC dataset led to 38.28% accuracy for LDC and 28.18% for NBC meanwhile they hit 87.33% for LDC and time features and 96,73% with NBC. However, they extracted only 4 features in the frequency and wavelet domain against 17 on the time domain feature. This may explain the difference in performances. For all cases, the NBC shows better performances against the LDC. Next they tested the Euclidean distance as classifier and got similar results for all every type of feature vectors and database (between 96% and 98% of accuracy). This paper proved to be quite interesting and showed the effect of the different types of extracted features on the classification.

Ye et al [85] made only one experience using the BIDMC dataset. In that paper, the authors defined a new model for biometrics authentication using the PPG. First they used a Butterworth band pass filter from 0.5 to 5Hz to reduce the noise, then they use a Zero mean normalization. The raw PPG signal is segmented in unit cycles using the pan Tompkins algorithm. For feature extraction, they used CNN + LSTM. The architecture was composed of two 1-D CNN composed of batch norm layer, max pooling, drop-out and RELU layer. They fed two LSTM and output 32 features. They used a KNN with Mahalanobis distance to classify the extracted features. They used the BIDMC and only took 12 users on the 53 available. Then they tried to identify new users on the system and showed the training time and percentage of discovery or identify a new user. The training time exploded, it went from 18 minutes with 6 user to 589 minutes for 18 users. This showed that this system could not be used in real life condition.

Siam et al [69] made 20 experiences using a custom dataset. Here, authors collected raw PPG signal with a custom material. They collect 50 to 60 raw signals of 6s from 35 users. To extract features, they used a FFT on a windowed frame. The magnitudes of the resulting spectra were mapped with the Mel scale [29] (a perceptual scale of music notes). Then they used the DCT on the results. They extracted 24 values used as features. Then they used these features in a MLP fed forward with one single hidden layer. The activation function of hidden neurons was the hyperbolic tangent sigmoid (tansig). The output layer contained 35 neurons, one for each class. Regarding our objective to build an authentication system because, we would have to modify and retrain the whole system when adding a new person, which is not acceptable. 66% of the dataset was used for training and 34% for testing. They achieved good performance with an accuracy (called recognition rate) between 92.14% and 100%. However multiple metrics were lacking such as EER, FAR and FRR. Moreover, the system has been trained on custom data not publicly available. We need to test this model on publicly available data and compute other performance metrics.

Hwang et al [34] made 25 experiences. In this paper, the authors focused on the usage of generative adversarial networks to improve performances of PPG based authentication systems. They were the first to test the adversarial networks. They used the GAN to generate synthetic data only for true users to reduce subject specific variation and help to mitigate adversary attacks. It was a promising and useful idea, but if it had worked, it would have been used to attack the system by generating synthetic signals of the user. They used multiple databases, such as TROIKA, PRRB, Biosec 1, and Biosec 2. In this paper, two scenarios are tested: one with a single session and one with two sessions. In the two-session scenarios, the signal for registration was taken in Session 1, and the signal

for testing was taken in Session 2. Their model was based on:

- Noise filtering with band pass filter between 0.5Hz and 18Hz.
- Single pulse segmentation.
- Size signal fixation and normalization using zero mean and unit variance plus DTW or Time Padding (TP) or frequency padding (FP) or cubic interpolation.
- Outlier removal: this step was done only during the registration phase. The authors computed the euclidean distance between the average shape of all signals and each one. Then they removed the ones with the most important distance.
- For the authentication part, two models are tested and compared: one called Wide-Shallow and one called Narrow-deep. Both are based on CNN, but they had a different kernel shape and used different functions.
- They used L2 regularization and 10-fold cross validation.

Once the tests have been done in one session scenario, they tried to improve the performances of their model in a two session scenarios with GAN. To do that, they tested 3 GAN. The GAN are only used for improving two session scenarios. They try GAN, DCGAN, WGAN and LSGAN, then developped a new GAN: PBGAN. To do that, they searched for the best PPG features using a linear SVM with exhaustive search methods and narrowed the usage of 7 features. The selected features were: area, max upward and downward slope near peak, AC value at 0.25, 0.5 and 0.75 lag of length and area from PSD. Then they tested the traditional GAN and two adapted versions of the traditional GAN using PGBAN. This led to the creation of six different versions of PGGAN. We can see that the best algorithm depended on the database. In two session case, the PGBAN-DC outperformed the other on Biosec3 while it was PGBAN-LS that outperformed the other with Biosec one. Another problem was the fact that different metrics were used in the two experiments. However this paper is excellent and showed how GAN can help to increase time stability for PPG biometric recognition.

4.2.9 2022. In 2022, 79 experiences were made over 3 papers [50, 59, 77].

Pu et al. [59] made 2 experiences. They combined the Capnobase and Biosec 1 datasets to make their experiences. In this paper, authors developed a new authentication system based on PPG. They filtered the raw signal and removed Motion Artifact. Then they segmented the signal in single cycle and removed outlier (bad cycles). After this procedure, they extracted a template and normalized it by meaning all pulses. They used wavelet decomposition then an auto encoder to create a new representation of the signal in a compress way that will be used with a L2 norm to authenticate the user. They used different metrics to test their model: EER, ROC and AUC. They achieved 97.9% accuracy with 5.5% of EER with the multi wavelet features. The system seemed efficient.

Wang et al [77] made 24 experiences. In this paper the authors tried to create a new authentication system based on PPG. They used 3 public databases to test their model: Vital DB, Capnobase & BIDMC. Their first step was to pre-process the signal with a complex pipeline involving re-sampling wavelet decomposition and recomposition, segmentation quality assessment with Skewness and zero mean normalization. This step reduced the noise and filter the non usable signals. Then they compute the first and second derivative (VPG and APG) of the signal. These signals were used as input in a 1-D CNN which produced a template for authentication. They tested 8 different models. For the Capnobase and BIDMC datasets, the ROCKET extraction feature gives the best results, but for the VITAL DB it was the SKNET and SNL which gave the best results. However the difference between the two algorithms was only 0.05% in accuracy and EER. Next they evaluated the computational performances for the algorithms using the number of train Epochs, the train time, run time, FLOPs number and total parameters. From their results, it seemed that the ROCKET algorithms provided the best performances in all criteria. Finally, they showed why they took 3 channels PPG and not only one. The performances of all their algorithms can decrease a lot (it could be halved for some) when using only one channel. This is probably due to the diminution of data when using one channel comparing to using three, and the fact that the three channels have a different noise sensibility due to the usage of different wavelengths.

Liu et al [50] made 53 experiences. In this paper, the authors tested a new method based on non negative matrix factorisation. The extracted features are based on fiducial points (min, max etc) and on frequency domains (SFTF, DWT etc.) They decomposed their features matrix in two matrixes named U and V, one based on the features and the other on the sample. The V matrix represents the common features of one subject. Their algorithm tried to find one common matrix for all features matrix of one subject. Then they used a distance metric to match or no the template of each people. They tried multiple combinations of time domain features and frequency features. They ran all their experiments on three databases: CapnoBase, BIDMC and MIMIC. Their best results were achieved with a combination of 1DLBP and DWT. They achieve 98.78% ; 97.86% and 99.82% accuracy on respectively BIDMC, MIMIC and CapnoBase.

4.3 Representation of the most used pipeline

With all the experiences data extracted, we build a graph to represent the used pipelines of each experience. The full graph and multiple figures are available on our Github repository: https://github.com/bvignau/PPG_SLR_dataset.

In Figure 5, we have represented the most common fully integrated pipeline (from signal segmentation to classification).

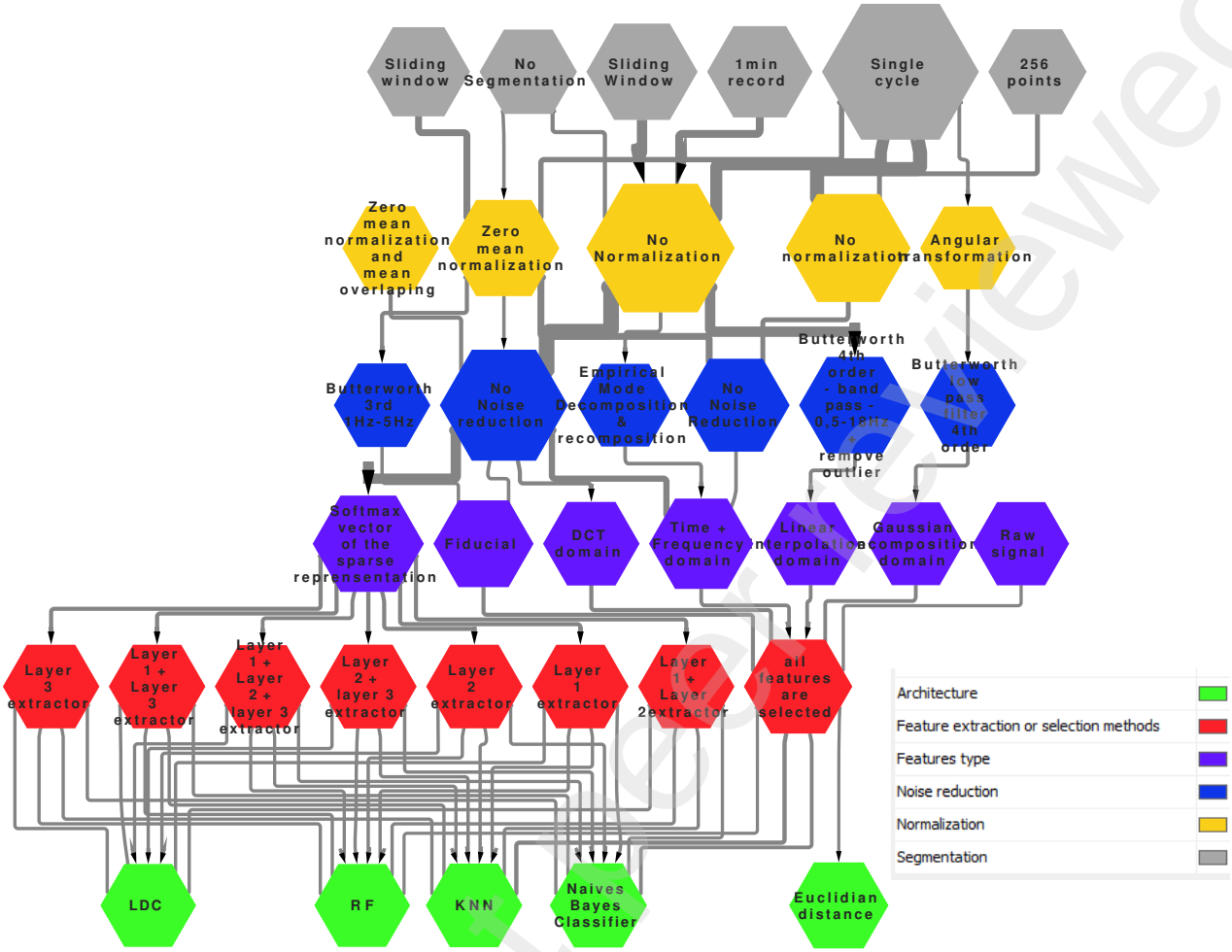


Figure 5. The representation of the most popular models pipelines

To create our graph, we computed a graphical representation of the experiments. In our representation, each step (e.g., signal segmentation in a single cycle) is a node, weighted by its total usage in the experimental dataset. Each connection is also weighted in the same way. For example, if the transition 'Single Cycle Segmentation' to 'No Normalization' is observed two times in the experimental dataset, then the edge is weighted as 2. In Figure 5, we only represent the nodes with a weight of 6 or more, and with a fully integrated pipeline (from signal segmentation to classification).

This representation shows the most used pipeline in number of experiences however, this truncated representation may be biased. Indeed, papers with many experiences can largely influence this kind of representation. This is why we let in open access the raw data and unfiltered figures that are difficult to include and read in a paper. The algorithm to exploit the data and create the graph are also given in our Github repository.

5 Signal Acquisition

This section focuses on signal acquisition and aims to answer the major question: how are the data collected? To answer this question we defined 4 criteria: the number of subjects, the sampling frequency, the acquisition time, and the general conditions (is the subject in rest, activity, etc.). Many papers have their methods to acquire data, many built their own datasets and did not share them, but still gave the parameters of said datasets. In a dedicated subsection, we will present the dataset usage over the experiences and the years, the evolution of the sampling sampling frequency over the years, the evolution of the number of patients and the general conditions of the acquisition of signals. As a conclusion, a new metric to measure the contribution of each dataset to the community will be defined.

paper	Year	Sampling frequency (Hz)	Acquisition time	Number of patients	Conditions	Time interval
[69]	2021	50	6s x 50	35	rest, wrist	NA
[42]	2014	2000	15 cycles x 2 sessions	30	Finger PPG; in relaxation	NA
[80]	2019	500	NA	23	Finger PPG; in relaxation	NA
[45]	2015	NA	NA	10	Finger PPG	NA
[84]	2007	300	70s x 3	3	Finger PPG; in relaxation	NA
[30]	2003	1000	1min x 1	17	Finger PPG, in relaxation	NA
[70]	2018	NA	1min x1	10	NA	NA
[63]	2013	37	30s x2	8	At rest for 1; with motion artifact for 2	NA
[32]	2018	5	90 cycle x 2	50	Finger PPG, in relaxation	NA
[10]	2018	360	15s x 1	36	Finger PPG, in relaxation	NA
[43]	2020	200	30min x 1	20	Finger PPG, in relaxation	NA
[20]	2016	1000	3min x 1	15	Finger PPG, in relaxation	NA
[17]	2013	75	2min x 1	44	Finger PPG, in relaxation	NA
[31]	2003	1000	1min x1	17	Finger PPG, in relaxation	NA
[25]	2019	10	NA	5	NA	NA
[62]	2013	37	60s x 4 x 2	9	Relax & stress	NA
[64] Nonin	2018	75	60s x 1 x 3	24	NA	7 days
[64] Berry	2018	100	60s x 1 x 3	24	NA	7 days

Table 4. Characteristics of the custom datasets

Dataset	Year	Sampling frequency (Hz)	Aquisition time	Number of patients	Conditions	Time interval
BIDMC	2018	125	8min x 1	53	Finger in intensive care	NA
Capnobase IEEE TBME	2013	300	8min x 1	42	NA	NA
VITAL DB	2016	500	NA	6388	Intra operative (30min-10h)	NA
MIMIC II	2008	125	60s x1	56	Finger in intensive care	NA
TROIKA	2014	125	5min x 1	12	Efforts on treadmill	NA
Biosec 1	2011	100	3min x 2	15	Finger PPG; in relaxation	14 days
Biosec 2	2020	100	1,5min x 3	100	Finger PPG; in relaxation	few seconds
VORTAL	2014	500	10min x 2	130	1st session in bed; second session in exercice	few seconds
DEAP	2012	512	1min x 40	32	record signal while watching different emotional videos	few seconds
MIT-BIH Polysomnographic Database	2000	250	NA	18	Night at hospital (2-7hours)	NA
OpenSignal PPG Dataset	2011	NA	NA	14	Finger PPG; in relaxation	NA
PulseID	2018	200	1min x 5	43	Finger PPG; in relaxation	few seconds

Table 5. Characteristics of the public datasets

5.1 Dataset usage

In this section, we present the evolution of dataset usage in Figure 6, which illustrates the number of experiments that use each dataset, and Figure 7, which categorizes the number of experiments per years and displays how each dataset has been utilized over time.

We identified 20 different datasets categories. Each category corresponds to the combination of the public datasets used for the experiences. A special category, named CUSTOM, consider all the custom datasets.

A custom dataset signifies a collected dataset not published by the authors in any single study. Consequently, it may not be possible to reproduce all the experiments conducted with a custom dataset. Detailed information about each custom dataset can be found in Table 4.

Furthermore, you can find the characteristics of the public datasets in Table 5. Here, the acquisition time is represented as "time x number." The first part denotes the time required for one acquisition session, during which researchers collect PPG signals. The second part corresponds to the number of different sessions. For certain datasets, there is an additional third number, indicating the number of recorded channels.

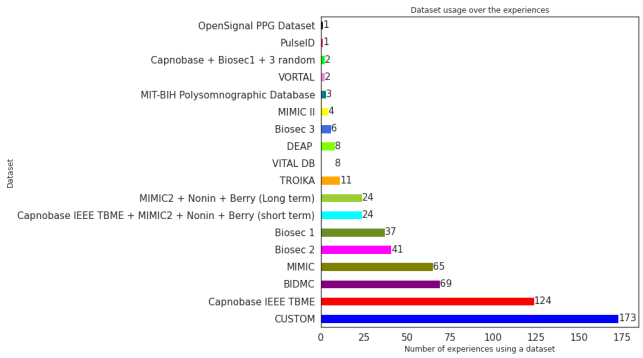


Figure 6. dataset usage over the experiences

PPG signals can be measured using green light, red light, and infrared light. Some sensors provide all three channels, and some research teams recorded more than one channel. The time interval represents the duration between two recorded sessions.

One thing to point in Figure 6 is the massive usage of custom datasets over the total experiences. Indeed, 28% of the experiments are made with a custom dataset. However, all

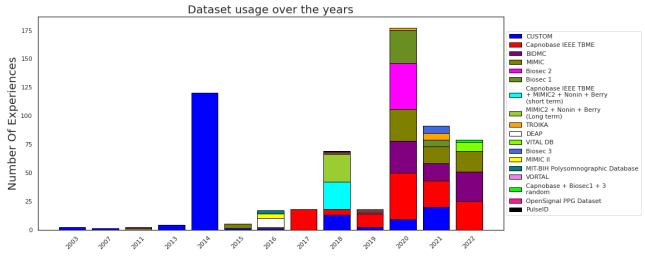


Figure 7. dataset usage over the years

of the experiences cannot be reproduced. Multiple publicly available datasets are used. Each experience uses a different subset of these datasets, and in an ideal world, each experience should be done with every available dataset. For future works, each paper should produce the same number of experiences with each dataset. Note that some works have started to implement this, mainly using the Capnobase, BIDMC, and MIMIC II datasets, as seen in [14, 83].

With Figure 7, we can observe that the custom datasets were mainly used in 2014 and before. Few usages persist across the experiences after, but they are very low compared to all the other dataset usages. In 2015 and after, we can observe a big diversity in the dataset usages. In 2020 and after, most of the published papers made experiences across two or three public datasets, making the experiences robust and easier to reproduce.

5.2 Sampling frequency

In this section we are studying the evolution of the sampling frequency over the experiences. The sampling frequency is crucial as it plays a major role in the signal quality. The Shanon's sampling theorem said that the sampling frequency should be at least two times higher than the highest frequency in the signal; most of the studied papers in this literature review shows that the PPG signal ranges from 0.5 to 20 Hz. So, we need to use at least 40 Hz as sampling frequency. Using higher sampling frequency add noise in the signal because it will measure electromagnetic perturbations.

Figure 8 represents the distribution of the sampling frequency used by the datasets and the evolution over the years.

On Figure 8a we can observe 15 different values for 30 datasets. For this parameter we have 10% of missing values. We can observe that the frequency values ranges from 5Hz to 2000Hz. This shows a big gap in disparity regarding the dataset, which will influence its quality and the final performances of each algorithm.

On Figure 8b we can observe that the sampling frequency does not follow any particular law. Their is no clear augmentation or reduction of the sampling frequency and the distribution over the years and datasets seems random.

5.3 Acquisition time

The acquisition time is another important parameter of a dataset, which we represent with two or three numbers (a x b x c). The first number corresponds to the length of the measured signal, which can range from short signals of just a few heartbeats to long signals of several minutes. The second number represents the number of sessions or the number of available signals per subject. The final number, present only in three datasets, indicates the number of available channels. To measure the PPG signal, sensors can use three different lights: red, green, and infrared. Most datasets only have one channel, but some have kept all three.

Increasing the acquisition time increases the amount of available data, which can be helpful for building better algorithms. Having multiple sessions with long time intervals between them can help create more realistic scenarios and allow for studying the stability of algorithms over time.

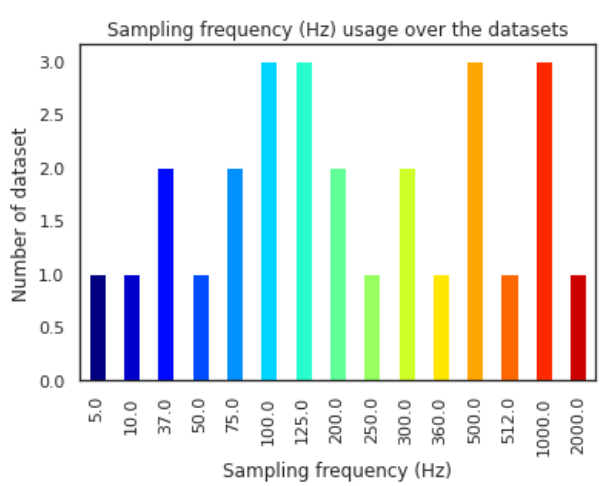
In Figure 9 we plot the distribution of the different acquisition time over the datasets and the years. For this parameter we observe 20% of missing values. Some datasets do not provide this parameter. For others such as the VITAL DB, the acquisition time is too heterogeneous and too different for each patient. For this dataset, the acquisition time range from few minutes to ten hours. On Figure 9a we can see that most of the datasets have their own acquisition time thus leading to heterogeneous datasets.

Figure 9b confirms our observation, as it plots the evolution of acquisition time over time. We define the acquisition time in seconds by multiplying all three numbers given in the dataset definition for each dataset. From this figure, we can see that most datasets provide less than 500 seconds of signal per subject, indicating a challenge in acquiring long PPG signals over multiple sessions.

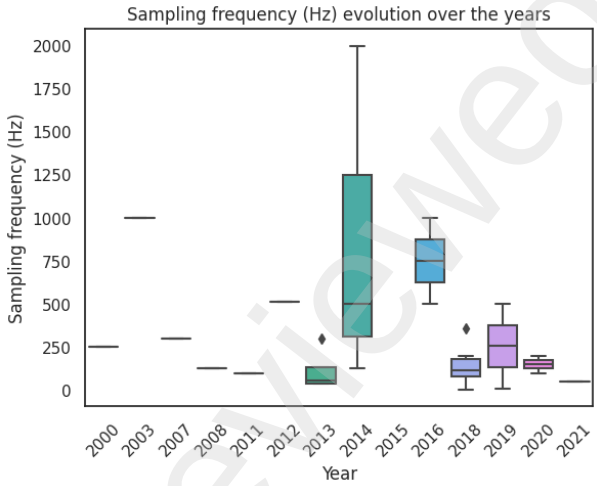
5.4 General conditions

The general conditions of signal acquisition, such as the environment and the positioning of sensors, can greatly influence algorithm performance. Most datasets were recorded in controlled environments where subjects sat with PPG sensors on their index finger. Additionally, many datasets were collected for medical purposes, such as the MIMIC II dataset which gathers signals from patients in intensive care units. The VITAL DB also collects signals from people, providing generally good-quality signals but potentially far from real-world conditions.

One major goal of PPG-based biometric recognition is to provide continuous authentication, requiring the collection of real-time PPG data over long periods. To create such a system, it is crucial to collect PPG signals in various conditions, including rest, physical activity, and different emotional states. The TROIKA dataset provides an excellent source of signals acquired during a test effort on a treadmill, which is highly relevant for building a robust system. Similarly, the

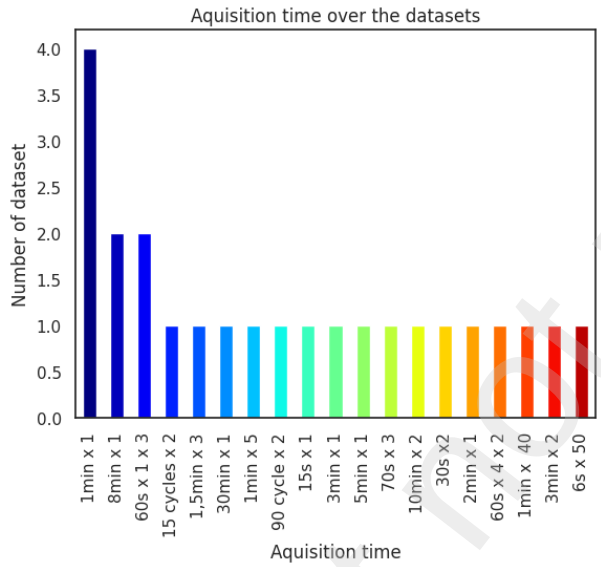


(a) Sampling frequency used by dataset

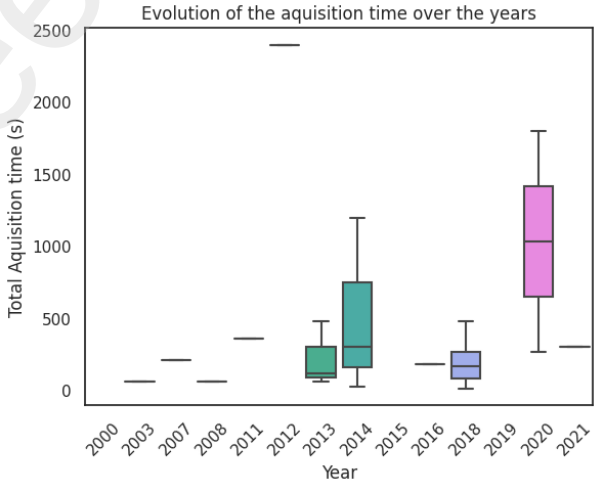


(b) Evolution of the sampling frequency over the years

Figure 8. Sampling frequency distribution and evolution



(a) Acquisition time over the datasets



(b) Evolution of the acquisition time over the years

Figure 9. Time acquisition distribution and evolution

DEAP dataset was collected with patients watching different emotional videos, providing valuable data for building a system robust to emotional variations.

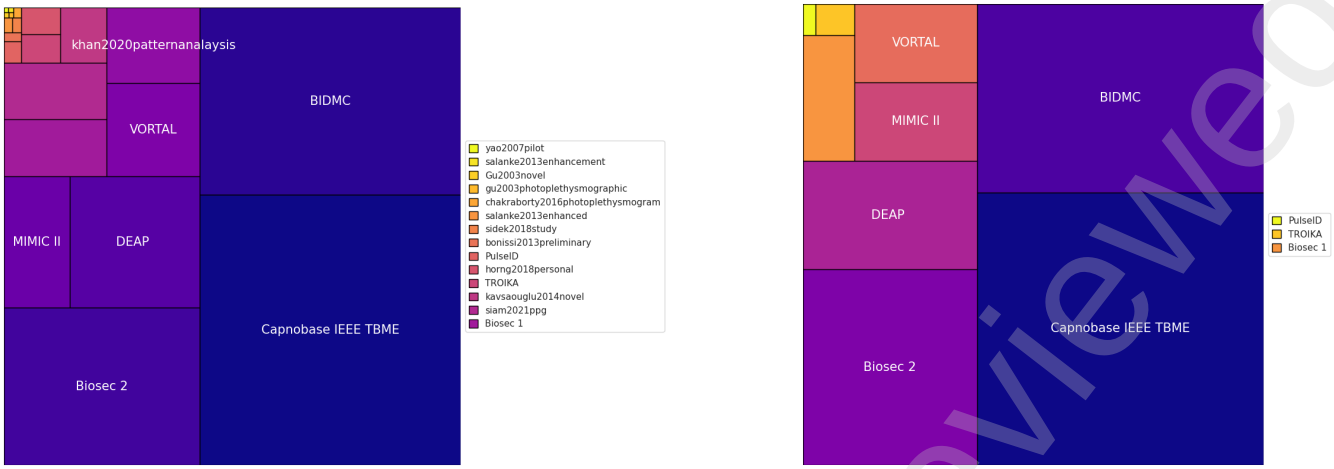
5.5 Datasets contribution to the community

To demonstrate the significance of a dataset to the research topic, we have defined a new metric called "data consumption" or "Dataset contribution" which is calculated as the product of the total acquired time, the number of subjects, and the number of experiences that used the dataset. So a dataset with more data (more users and more aquisition time) will contribute more to the community. Also a dataset massively used will also contribute more to the community.

$$D_{contribution} = N_{subjects} \times T_{aquisition} \times N_{expreriences}$$

In Figure 10, we present a tree map showing the total consumed data for each dataset. Similarly, Figure 10a illustrates the contribution of all datasets, while Figure 10b focuses only on publicly available datasets.

In Figure 10, we can see that custom datasets provide relatively little contribution to the research topic, with their total consumption being lower than that of the BIDMC dataset alone. However, some custom datasets contribute more than



(a) Datasets contribution **(b)** Public datasets contribution

Figure 10. dataset usage over the experiences

public datasets such as TROIKA. For example, the Kavsaoglu et al.[42] and Khan et al.[43] datasets have higher numbers of experiences and longer acquisition times, resulting in a greater overall contribution.

In Figure 10b, we focus on publicly available datasets and can see that the Capnobase and BIDMC datasets contribute the most, representing around two-thirds of all public dataset contributions. They are also among the most widely used datasets, as shown in Figure 10. Additionally, the Capnobase provides a high number of patients (42) with a good time record (8 minutes), while the BIDMC dataset has 53 patients and the same time record.

The Figure 11, represents the evolution of the contribution of all the datasets over the studied time period. This new representation allow us to quickly find the dataset which are the most used in time, so the one with possibly the best quality.

In Figure 11 we can see that there were few contributions to the research topic before 2014, likely due to a lack of publications and limited experience using custom datasets with small numbers of subjects and short acquisition times. However, in 2014, Kavsaouglu et al.[42] provided 120 experiences, marking the beginning of more significant contributions to the field. While custom datasets remain relatively rare, publicly available datasets such as Capnobase and BIDMC have contributed significantly to the research topic over the years.

In Figure 12, we can see that the Capnobase dataset is the only one to contribute every year from 2016 to 2022 and is consistently among the most contributing datasets each year, indicating high popularity within the community. The MIMIC II dataset is also very popular but provides little contribution due to its short acquisition time (60 seconds). The BIDMC dataset has gained in popularity since 2019 and has contributed significantly to the research topic.

5.6 Proposed methodology to build future datasets

From all the observations we made we propose a new methodology to build a dataset that will help the community to build PPG continuous authentication systems. The two majors key points are: the number of patients and the recorded condition. We need to have a maximum of subjects to see if a PPG-authentication still work with scaling up. Currently, the maximum of subject used in one experience is 100. Even Wang et al [77] who used the VITAL DB, only keep 100 subjects over the 6000 available. So we need to build a dataset with at least 1000 peoples of all ages.

The materials used to measure the signal should be a smartwatch or similar device, as people may not or cannot wear a pulse oximeter on their right index finger continuously. However, people may wear smartwatches that already measure PPG to provide heart rate; it is thus highly probable that continuous authentication systems using PPG will use smartwatches.

The signal should be acquired over multiple days, with a dedicated protocol performed at least once per measured day. This protocol should include a subject resting in a silent and dark room, who will then perform an effort test on a treadmill, a bike, or other cardiovascular machines. The tested individual should then rest again and meditate, for example, to reduce BPM optimally. Ultimately, the protocol should include activities that induce various emotional and physical states, such as watching different emotional videos or playing a defined video game. The goal here is to measure the change in PPG signal induced by different emotional states. After that, the subject should return to their daily life. Collecting such datasets would be interesting and should be done over three consecutive days, two or three times per day, with at least one month between each three-day session.

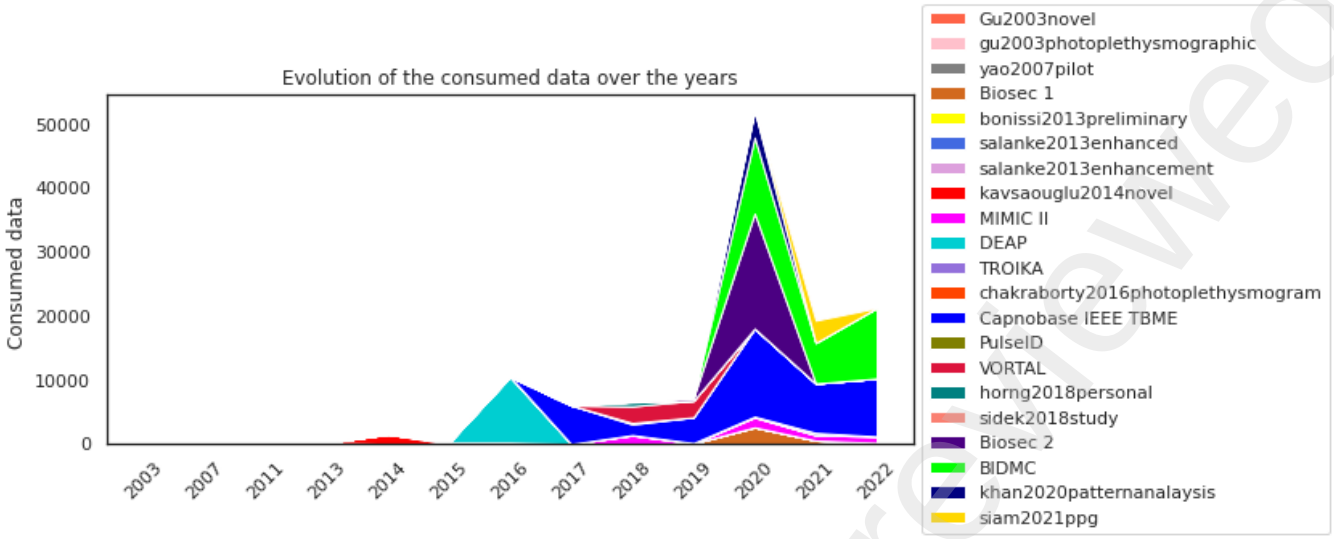


Figure 11. Datasets contribution over the years

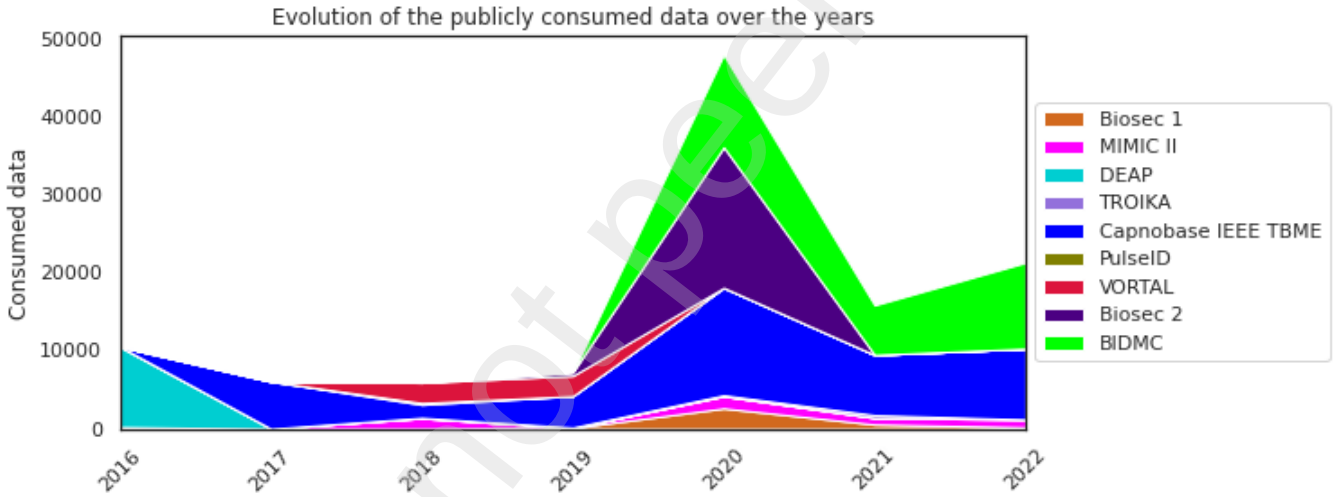


Figure 12. Public datasets contribution over the years

Constituting such a dataset can prove to be difficult. This is why a contribution between universities or laboratories could be of the utmost value. Such a dataset will greatly benefit the community of PPG-Biometric recognition but also the medical community and the Human-Machine Interface community. The PPG represents the blood volume variation and such a dataset can be used to develop new medical monitoring systems. Moreover, the variation in PPG induced by the emotional state can help to build new HMI. The DEAP dataset was collected to achieve that goal.

6 Pre-processing

Here, we will discuss the key components of signal preprocessing. We will explain how noise reduction techniques, segmentation, and normalization are applied to the heartbeat signals. These steps are crucial for improving the performance of PPG-biometric recognition algorithms.

6.1 Noise reduction

PPG signals are recorded by measuring the amount of light absorbed or reflected by blood vessels, which measures the variation in blood volume. These signals are subject to various types of noise, such as motion artifacts and electromagnetic perturbations. Additionally, the PPG signal frequency

range is typically between 0.5 and 5 Hz [11], but most studies recorded their PPG signals with a sampling rate above 100 Hz. This provides more detailed information about the signal but can also introduce noise. The Shannon sampling theorem states that the minimum required sampling frequency is twice the highest frequency of the signal we want to represent [68]. Therefore, the PPG signal should be sampled at least 10 Hz and possibly more than 300 Hz may introduce unnecessary noise. This is why all teams had to filter out the noise to clean the PPG signal.

Figure 13 illustrates the various noise reduction techniques employed by each experiment. We observe approximately 25% missing values for this parameter, indicating a high level of heterogeneity in the methods used to reduce noise in the PPG signal. A total of 33 different methods were used across the 44 papers, further highlighting the lack of consensus on this issue.

Figure 14 illustrates the usage of noise reduction techniques over time. We can observe a high level of heterogeneity in the methods used, with many studies employing Butterworth filters [18]. Each team tested different parameters, such as different filter orders and cut-off frequencies. In this figure, we have color-coded all the methods using a Butterworth filter with different shades of green (green, olive, forestgreen, etc.). The usage of Butterworth filters started to increase in 2016 and was the most commonly used technique until 2021. Teams also used low-pass, high-pass, and band-pass filters, but each selected different kinds of filters and ranges. The most common filtering range is between 0.5 and 15 Hz.

In rare cases, some teams used Gaussian filters [21] or Discrete Wavelet Transform (DWT) [9] to reduce noise in the signal. Additionally, the team of Kavsaouglu et al. [42] used Finite Impulse Response filters (FIR), while Salanke et al. [63] kept only the most important coefficients in Fast Fourier Transform (FFT) by selecting the first 8 coefficients for each pulse and reconstructing the signal using an inverse function and the chosen coefficients.

Some studies do not provide enough details about their noise reduction techniques or filter frequencies, making it difficult to assess their impact on the results. Additionally, some papers lack precision in describing their filtering processes ([5], [6], [10], [17], [30], [37], [47], [62], [65], [71], [73], [80]). As these methods are part of larger experiments with different data and classification techniques, it is difficult to determine the impact of each filtering method.

To provide a clear comparison between each method, multiple filtering techniques should be tested on the same dataset and then on heterogeneous datasets using a fixed algorithm and feature extraction methods. This will help the community determine the pros and cons of each filtering technique. Every team used a filtering method without explaining why they chose it, making it difficult to compare their results. The filtering phase is critical because it removes data from the signal (idealy, noise) and influences all subsequent phases. Finally, other techniques such as deep learning [36] may be explored for noise reduction. A filtering framework should be created to help the community develop generalized and robust algorithms for authenticating individuals.

6.2 Segmentation and normalization

Most studies use two additional techniques in their preprocessing phases: signal segmentation and normalization.

First, there are 18 different methods to segment the signal, which will be used for classification. There are virtually infinite ways to segment a continuous signal, as there are numerous possibilities for dividing it into segments. However, since a PPG signal is cyclical in nature, we can assume that most of the relevant information is contained within one cycle.

Generally, PPG signals are segmented into single beats or cycles. A modified version of the Pan-Thomkins algorithm [55] is commonly used to detect the systolic peak, allowing researchers to split the signal from one peak to another. Another technique used to detect the systolic peak is by computing the first derivative of the signal and finding the zero-crossing points.

Figure 15 illustrates the different techniques used by the community to segment their signals. We present the data in raw values and as a proportion. We observe 2.65% missing values for this parameter, indicating good usage and description of this stage within the community. Approximately 50% of the studies used single-cycle segmentation, showing a consensus on this criterion. However, further research on optimizing this parameter may be beneficial.

Figure 16 illustrates the usage of segmentation techniques over time. We can see that single-cycle segmentation is the only method used consistently across all years and multiple papers, indicating the beginning of a consensus on this criterion.

The only five studies which used more cycles are [62], [81], [47], [33] and [51]. Finally, only [16] and [43] used all the signal, and [82] used a sliding windows decomposition. The three remaining studies [31], [6] and [46] do not provide details about segmentation. Only Lee et al [47] experiences the same architectures using 10, 30, 50 and 100 cycles. They do show that using more than 10 cycles was not improving the performances. However they did not test the most used technique, which is a single cycle.

The normalization process is less common. Generally speaking, it is to define a new amplitude space, between 0 and 1, for example, by dividing the signal by its maximum. For this parameter we observe 41.79% of missing values, showing few uses and a lack of knowledge for this process.

Figure 17 draws the usage of the normalization techniques over the experiences. We can observe that in all the experience, the most used methods are to provide raw signal, with no normalization (40% of the provided experiences). Then

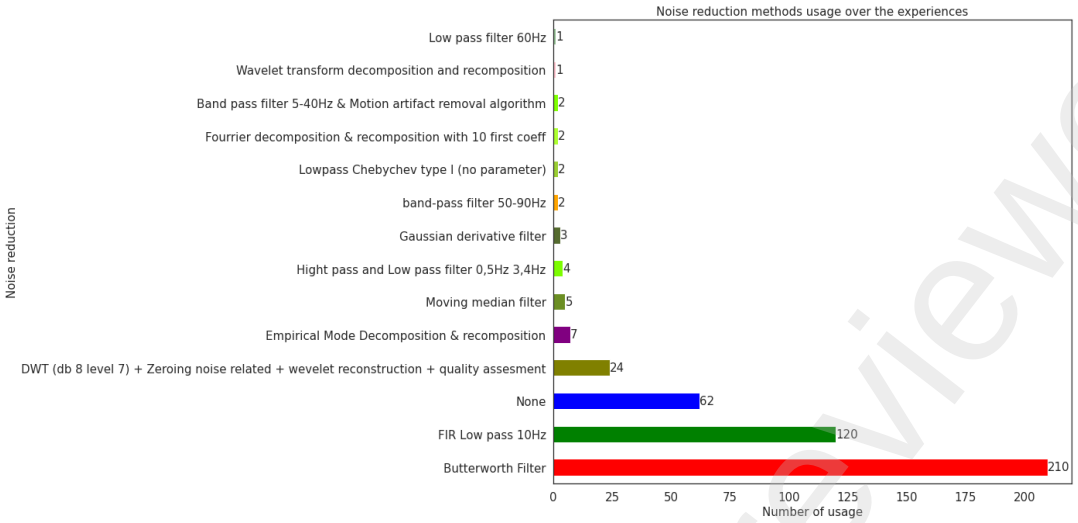


Figure 13. use of noise reduction techniques

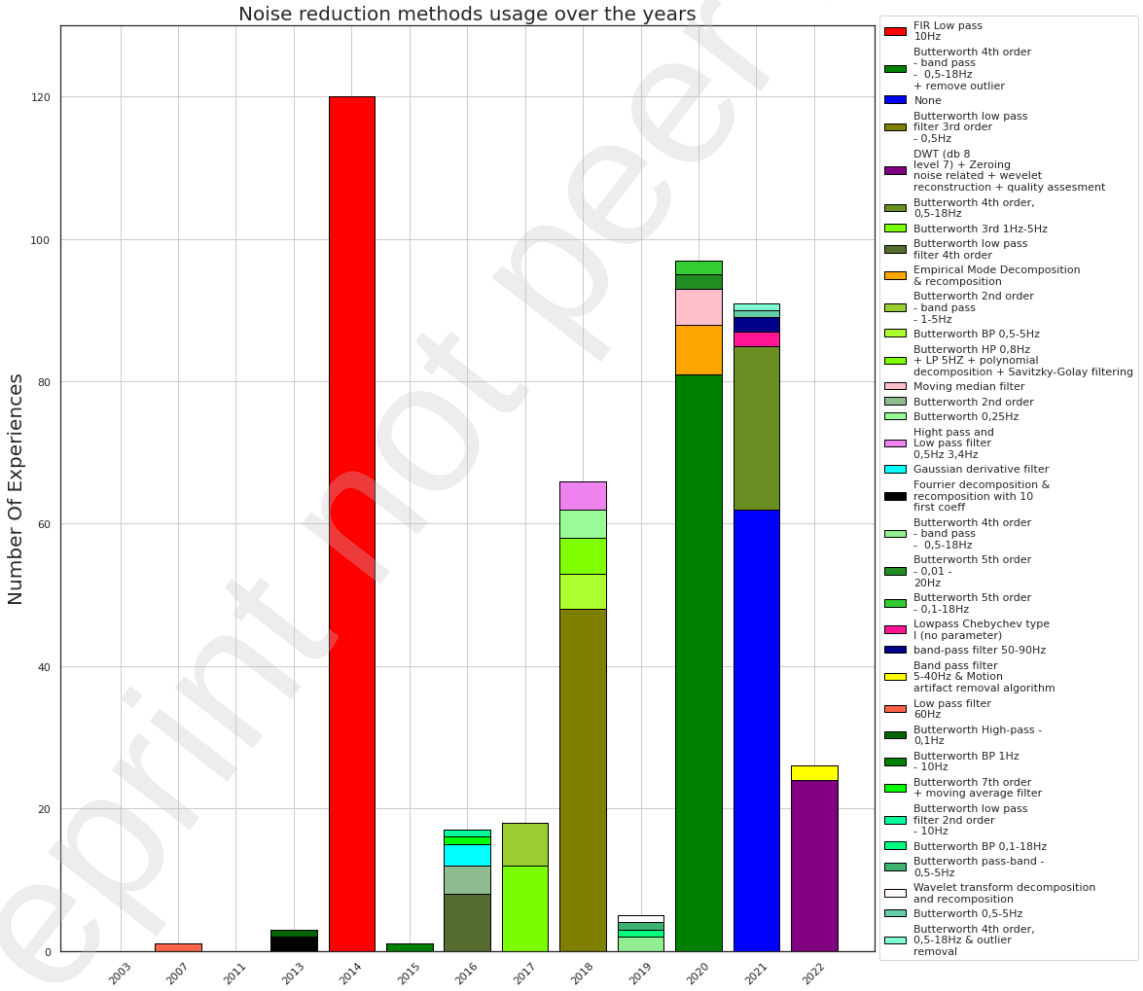


Figure 14. Noise reduction usage over the years

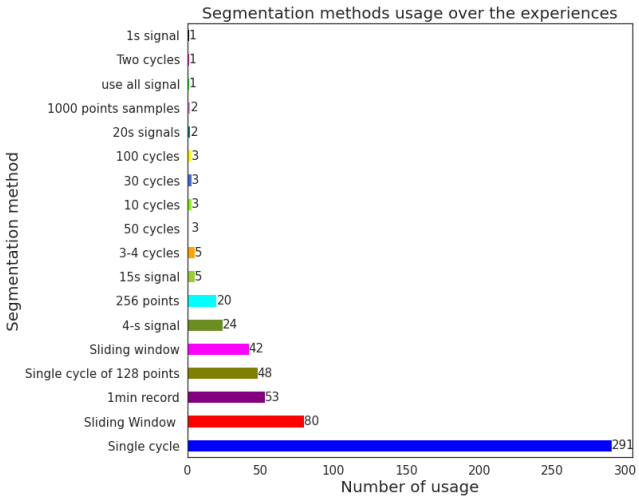


Figure 15. Use of segmentation techniques

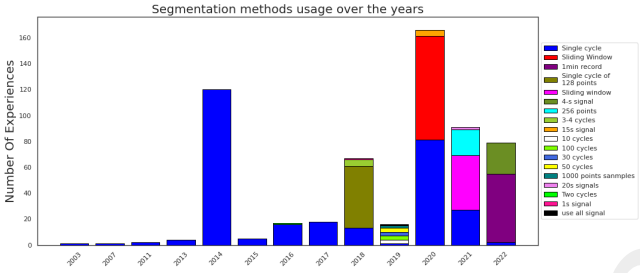


Figure 16. Segmentation usage over the years

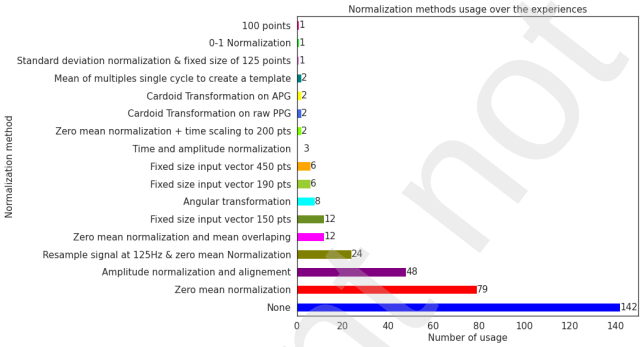


Figure 17. Use of Normalization techniques

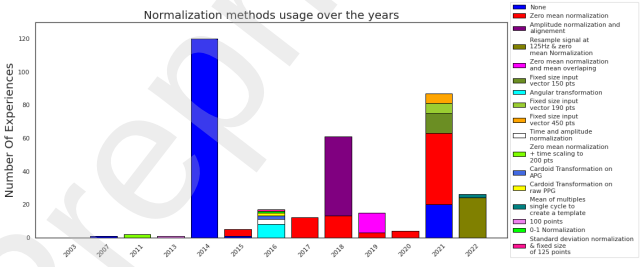


Figure 18. Normalization usage over the years

the Zero Mean normalization is the second most employed method. We can observe 17 different methods but most of them are not used recurrently, only in one paper, in fact.

Figure 18 demonstrates the different usage of the normalization techniques over the years. We can observe that the only methods to have been reused multiple times is the Zero Mean normalization. Others seem to have been tested only once.

However, every normalization technique is known in statistics and machine learning (zero mean normalization, amplitude normalization, alignment etc.) except for one: cardioid normalization [71]. These techniques have not been used by other research teams, and we cannot say if they improve identification or not, due to the difficulty in comparing their results with the literature. More than half of the research teams do not use the normalization technique, and we need to quantify the effects of these phases. It may be removed if it does not provide a real improvement.

As for noise filtering, it is impossible, with the actual studies, to compare the impact of each method. Moreover, we cannot say if this phase has an impact or not, as no clear justification is given to carry out this phase. These phases induce a computation usage and may not be useful, in the IoT world, where computational power and energy are limited; following this logic it can be worth deleting a phase that has limited impact.

7 Features extraction and selection

In this section, we explain the different methods used to extract and select features to train classifiers. There are two major sorts of features: fiducial and non-fiducial. The first one takes as a feature, physiological points used in medicine, such as systolic and diastolic peak, heart rate, heart rate variability, mean of the signal, etc. Non-fiducial features [41] can be extracted by numerous techniques, such as Discrete Wavelet Transform (DWT) [38], Fast Fourier Transform (FFT) [13], [53], Discrete Cosinus Transform (DCT) [4] etc.

The feature extraction is a key step to build an efficient biometric recognition system, so most papers provide details about this stage, including the types of features, the number of features, and the methods used for extraction or selection. We have relatively few missing values for feature types, with only 0.16% missing. Figures 19 and 20 show the usage of feature types over experiences and years, respectively.

The fiducial domain refers to historical landmarks on the signal, which are typically taken from biology and include features such as the systolic peak. The time domain features describe statistics about the signal in the time domain, including minimum, maximum, kurtosis, skewness, and so on. Finally, the transformed domain includes all features extracted through signal transformations like FFT, DWT, or DCT.

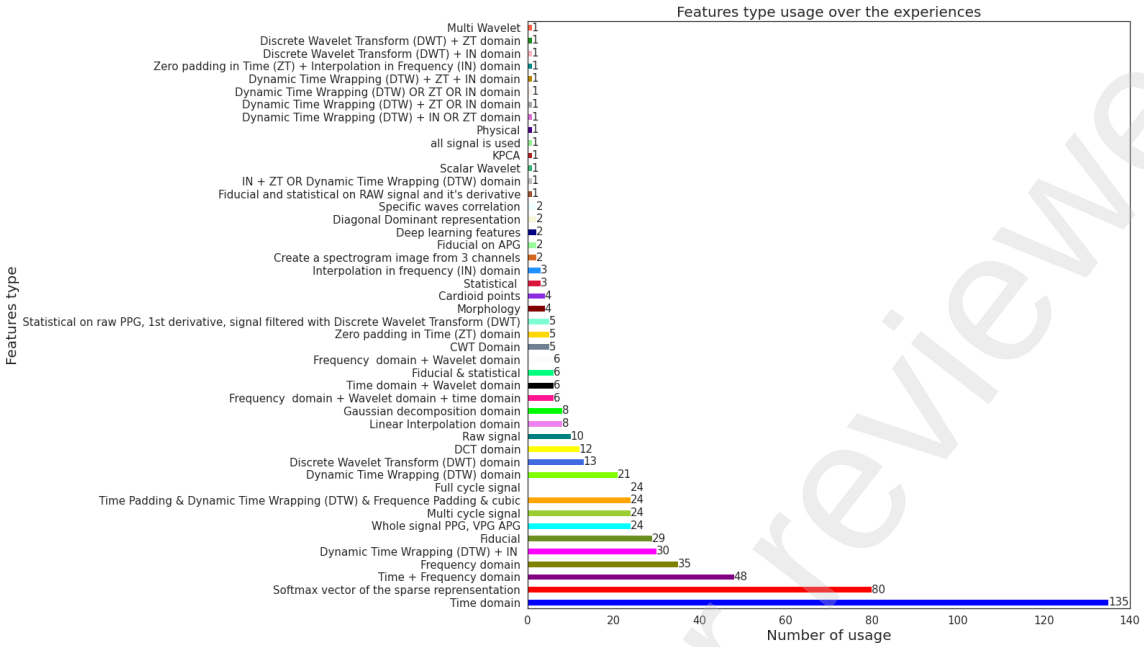


Figure 19. Feature type usage over the years

In our study, we observed that there were 46 different methods used in 46 different papers. Some feature types were used by multiple papers, such as time domain, fiducial, or statistical features. Additionally, some papers compared the efficiency of different types of feature extraction, like interpolation in frequency, zero padding in time, dynamic time wrapping, and all possible combinations of these feature types. Furthermore, we found that feature types were extracted on multiple versions of the signal, such as the raw signal, first and second derivatives. Before 2013, most papers used fiducial features, which were less commonly used after 2014 in favor of other non-fiducial features. This can make it challenging to compare methods and evaluate their pros and cons.

Physiology	Systolic peak
	Diastolic peak
	Dicrotic notch
	Heart rate
statistics	Local maximums
	local minimums
	Distance between fiducial points
	Mean of the signal
	Energy of the Signal
	Standard deviation of the signal

Table 6. Most common fiducial features

7.1 Fiducial features

Fiducial features are landmarks or points of interest on the PPG signal or its derivative, which are determined by standards of physiology and statistics. There are over 40 different standard fiducial features, all in the time domain. The most commonly used include mean, standard deviation, systolic peak, and others, as shown in Table 6. An example of fiducial points on a PPG signal and its derivatives is given in Figures 21 and 22, from [42].

The first derivative of the photoplethysmogram (PPG) signal can be used to determine certain points, such as those found in the systolic peak or diastolic peak. Some studies have utilized a range of features extracted from these signals, including up to 200. In one study, Azam et al. defined a window of 200 points centered around the systolic peak, which were then combined into a feature vector of dimension 200. Alternatively, the maximum number of fiducial features is approximately 40 when all time differences of fiducial points are taken as a feature. This approach may result in repetitive information and one study has demonstrated that using fewer features can improve the accuracy of the system.

For example, [42] extracted 40 fiducial features and compared the accuracy of their classification algorithm depending on the number of extracted features. They also provided a ranking algorithm between fiducial features. They demonstrated that using 20 features over the 40 available maximizes the accuracy of their algorithm but they did not provide a clear list of the 20 best features. Adding to this, the “best” features are dependent on the used algorithm, its configuration and the signal used. For example, using KNN with k=3

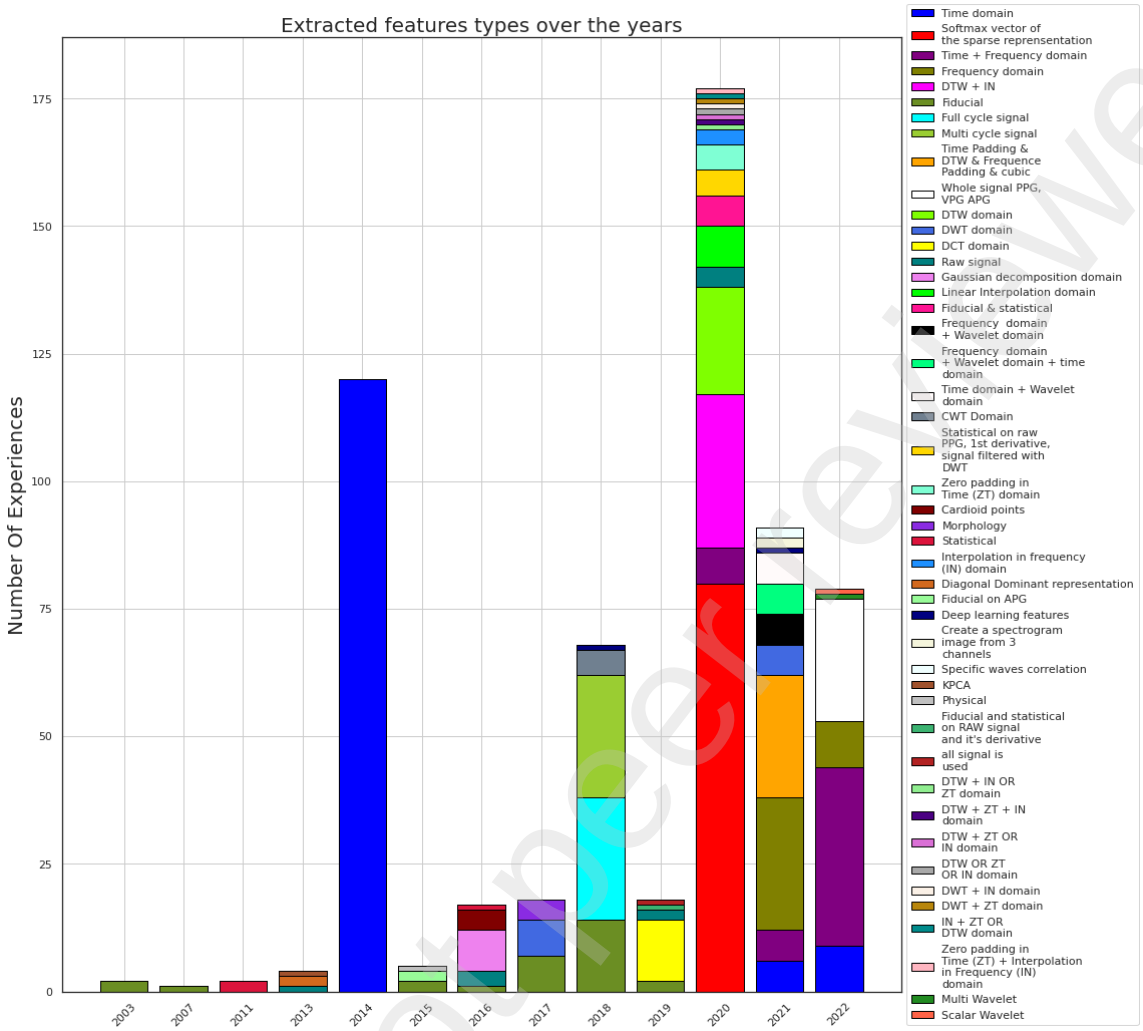


Figure 20. Feature type usage over the years

and k=5 changes the order of the best features. If we look at the 10 best features for one signal, with k=3 and k=5, most of the features are the same, but their rank is different. But if we compare the 10 best features between two signals, some features are replaced by others. However, they are the only ones to make this experience, all other teams used all the extracted features and do not provide a ranking on the best features.

It appears that there is no clear consensus on the optimal number of features to extract from PPG signals for classification tasks. The specific choice of features may also depend on the signal being analyzed and the classification algorithm used. It would be beneficial to conduct further research on this topic to determine the most effective approach for feature extraction from PPG signals.

7.2 Non-fiducial features

Non-fiducial features can also be extracted from PPG signals using various signal transformations [41], such as the Fast Fourier Transform (FFT) [13] [53], Discrete Cosine Transform (DCT) [4], and Discrete Wavelet Transform (DWT) [38]. These features are not points in the time domain or statistical values but are instead given by the transformation itself. Different types of information can be obtained from each transformation, resulting in a different way of describing the signal. Some studies have suggested that non-fiducial features may be less sensitive to noise compared to fiducial features. Further research is needed to determine the most effective approach for feature extraction from PPG signals, taking into account both fiducial and non-fiducial features.

There are several types of signal transformations that can be used to extract non-fiducial features from PPG signals, including the Discrete Cosine Transform (DCT), Discrete

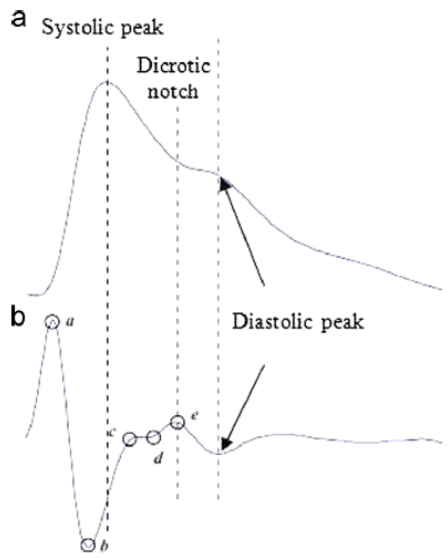


Figure 21. An example of fiducial points from PPG raw signal and APG. From [42]

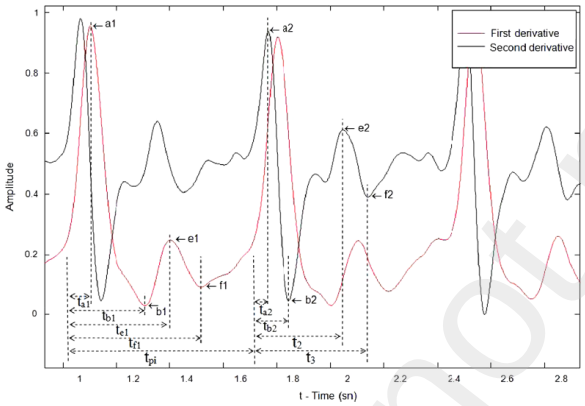


Figure 22. An example of fiducial points from PPG 1st and 2nd derivative. From [42]

Wavelet Transform (DWT), Fast Fourier Transform (FFT), and Gaussian decomposition. Other methods such as Continuous Wavelet Transform (CWT) are less commonly used in the literature. Further research is needed to determine the most effective approach for feature extraction from PPG signals, taking into account both fiducial and non-fiducial features.

There are two other ways of extract non-fiducial features: through classical dimensional reduction such as Linear discriminant analysis (LDA) [12] or principal component analysis (PCA) [79] or with a deep learning algorithm. For example, [73] used the LDA, [62] used a modified version of PCA, called KPCA [66] as feature extraction. For the deep learning feature extraction, teams used in general a Convolutional Neural Network (CNN) [28]. For example [24] dedicated the first layer of its CNN to extract features, [82] and [46] used multi-layer CNN for feature extraction. In this case, we cannot exactly say what kind of features are extracted. In general, studies using these techniques give their algorithm the whole signal. For the dimensional reduction methods, they are also used as a feature selection methods [41], and in few studies as the classifier [65] [20].

7.3 Features selection

Using many features increases the computational needs of a classification algorithm. Additionally, using highly correlated features does not add any information and can actually hinder the performance of the classification algorithm. Therefore, dimensional reduction can improve the efficiency of the system by reducing the computational needs while maintaining or even improving the performance.

Common techniques to achieve dimensional reduction are Linear Discriminant Analysis (LDA) [12], Principal Component Analysis (PCA) [79] and their variations (DLDA [81] and KPCA [66]). They can be used with all fiducial features and some non-fiducial features.

For example, studies have used DWT to extract time-frequency domain features and LDA for feature selection [81]. However, this step is generally unnecessary with deep learning since the features are automatically extracted and summarized by the algorithm. This is the case with works that use CNNs, such as [16], where the number of "neurons" in the first layers and the number of outputs determine the selected features.

Figure 23 shows the different techniques used over the experiences to select features. On this parameter we observed 1.49% of missing values, indicating a good description of this stage in the literature. A total of 78 different feature selection techniques were used across 44 papers, demonstrating a good exploration of multiple methods. Some studies test multiple methods or a single method with different parameters, such as Sancho et al [64] using KLT averages on 10, 20, and 30 cycles. However, most of the techniques were not tested in comparison with others, making it difficult to determine which are the most efficient.

Figure 24 shows the usage of multiple feature selection techniques over the years. We observe that the only method used across multiple papers is the selection of all extracted features, indicating that there is no consensus in the community on using dimensional reduction methods to reduce the number of features. This suggests that researchers may not be aware of the benefits of dimensional reduction or may prioritize other aspects of their models over feature selection.

Kavsaoglu et al. [42] demonstrated that reducing the number of features can significantly improve accuracy. It would be interesting to use both fiducial and non-fiducial features and apply classical dimensional reduction methods such as

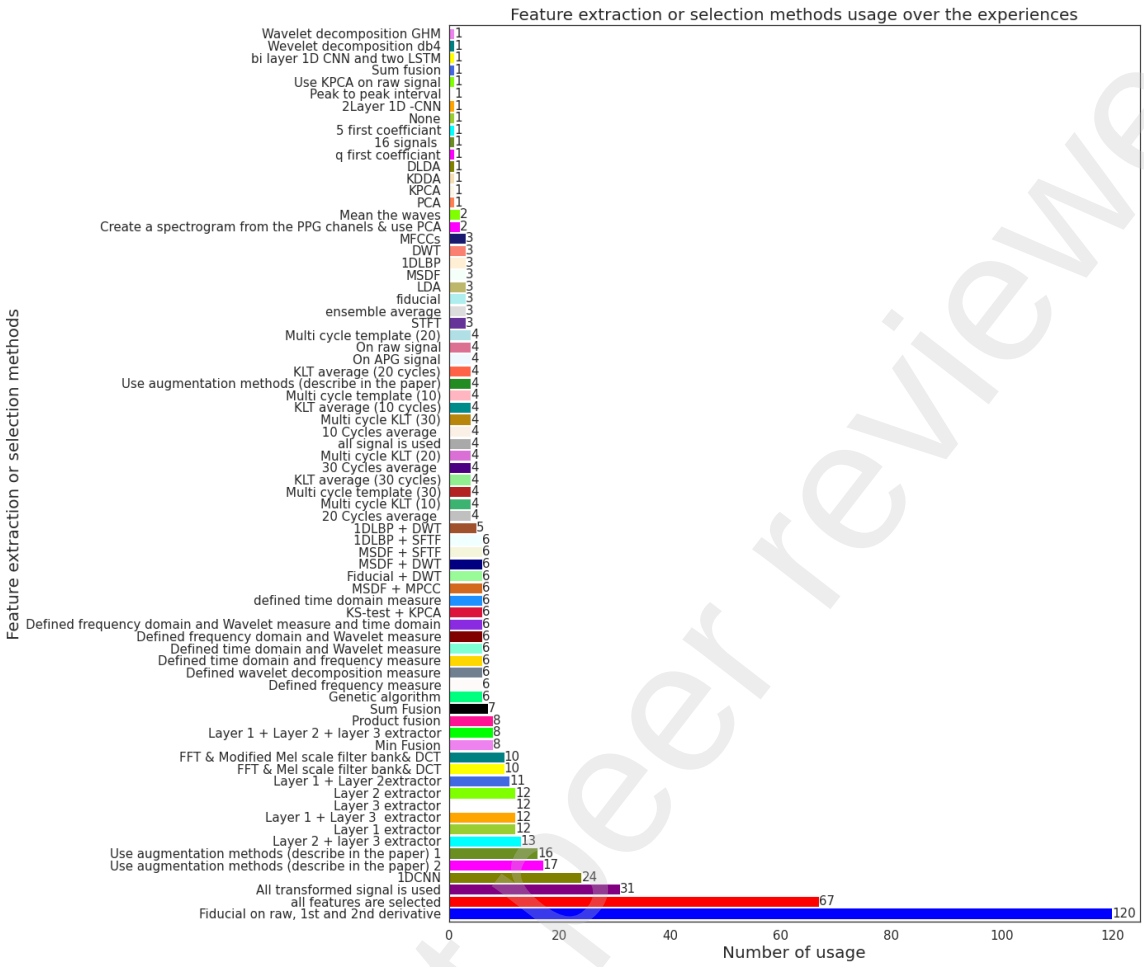


Figure 23. Methods to select or extract features over the experiences

LDA or PCA, or deep learning algorithms like CNNs [24], to find the best features to extract. Comparing these extraction methods with an extraction done using a Deep Neural Network, on the same dataset and with the same algorithms, would also be informative.

In conclusion, comparing methods and studies can be challenging, and some lack important details. For example, Jaafar et al. [37] explain research that used algorithms such as KNN and Naive Bayes but do not specify which features were extracted or how. Additionally, some studies do not use learning methods to create their systems. For instance, Choudhary et al. [21] propose a system in which they create a signal template by combining several aligned pulses. When a user is presented to the system, a PPG pulse is extracted and various distance metrics are used to compare it to all templates. In this system, there is no need to extract or select features.

7.4 Features length

The final parameter on the features is the "feature length," which represents the information that will be used by the algorithm to recognize subjects. As previously mentioned, the feature length will also impact the performance and computational needs of the system. In our dataset, we represent the feature length using one-, two-, or three-dimensional vectors (a x b x c) when possible.

Figure 25 shows the different usage of feature lengths over the experiences. The version with the number of experiences for each length is available on our Github repository https://github.com/bvignau/PPG_SLR_dataset. We observe that around 30% of data is missing for this parameter. A total of 60 different feature lengths were tested across the studies, indicating a good exploration of multiple lengths. Some studies test multiple feature lengths or use one-dimensional vectors when using only a few features. For example, Kavsaoglu et al. [42] extracted 40 fiducial features and tested 5, 10, 15, 20, 25, 30, 35, and 40 feature(s) in the input vector. They show that increasing the number of features used generally

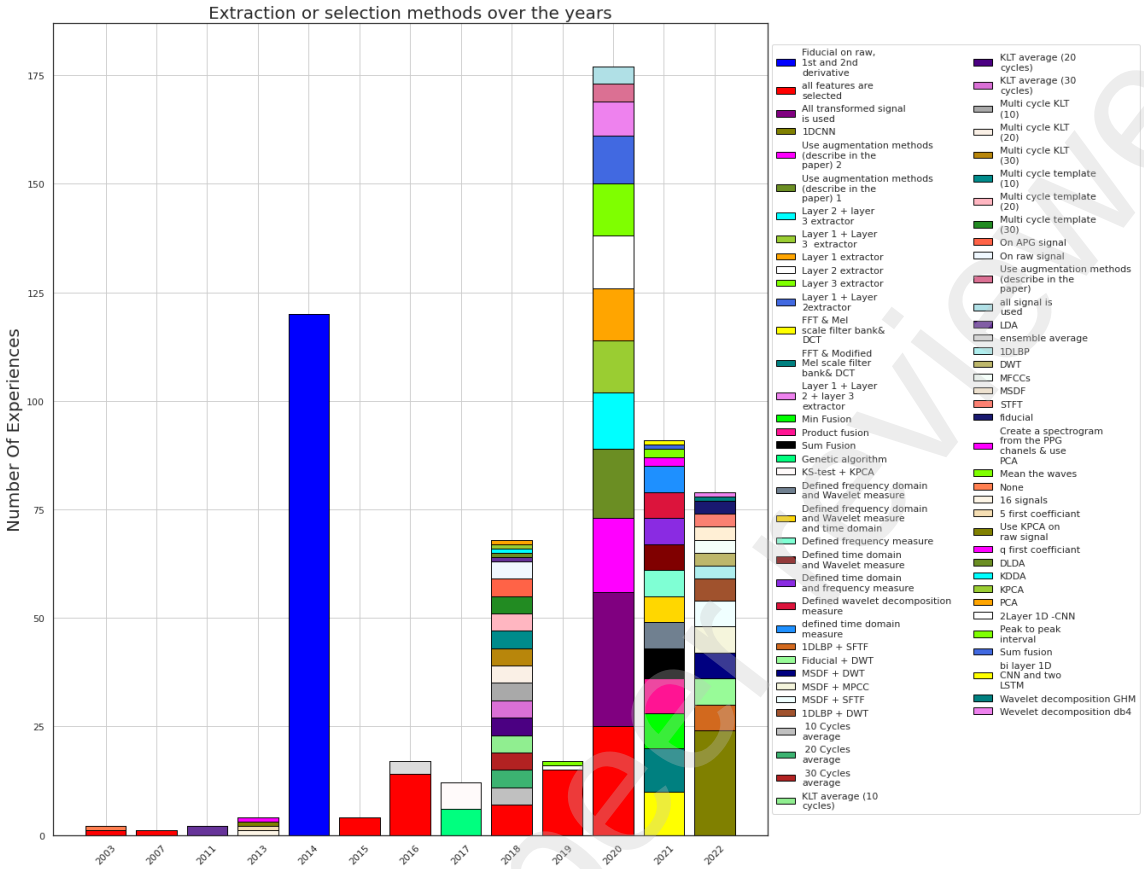


Figure 24. Methods to select or extract features over the years

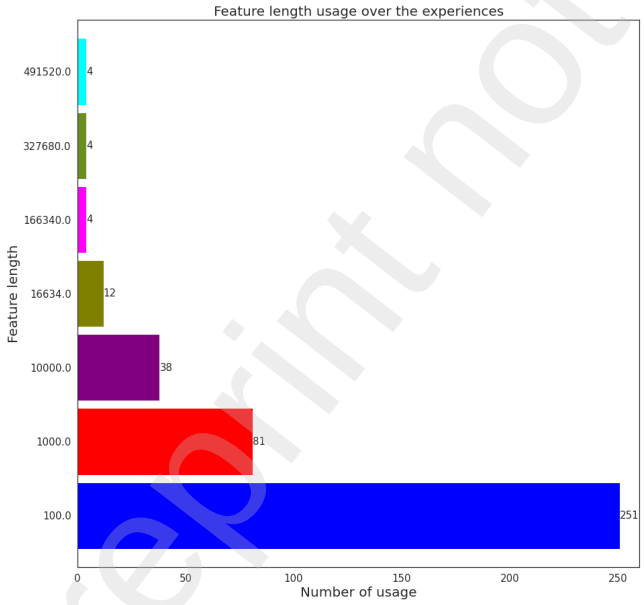


Figure 25. Different feature length usage over the experiences (we grouped in one category the experiences with less than 100 ; 1 000 and 10 000 features for readability)

improves performance until a limit is reached. In their experiments, they found that this limit is usually reached at around 25 features. Using more features does not improve performance. Some studies only use one-dimensional vectors, for example when using only five fiducial features.

Figure 26 shows the usage of feature lengths over the years. We observe that each paper uses its own parameter, and there is no consensus in the community on this stage. To compare the length used by the community, we have converted all multi-dimensional vectors to a single value by multiplying each dimension. We also seldom kept the numerical values of the dataset for this calculation.

Figure 27 shows the evolution of the feature vector length over the years. We can observe an overall increase in the feature length over time. Before 2018, most feature vectors had a length between 10 and 100. In 2018, the feature length increased significantly, with a mean size around 10^5. After this point, the size decreased and stabilized around 1000 features.

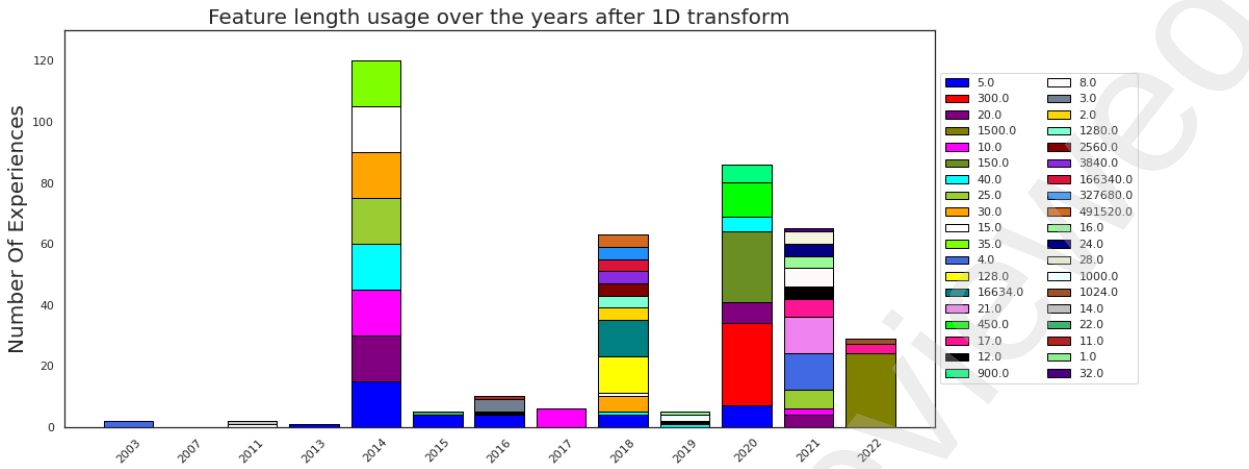


Figure 26. Feature length type usage over the years

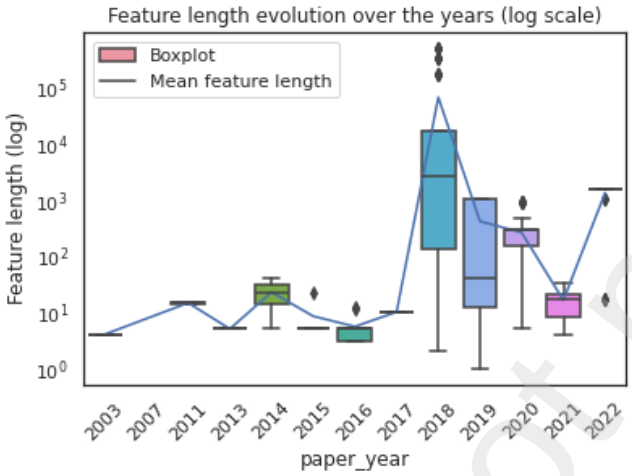


Figure 27. Representation (box plot and mean evolution) of the feature length evolution over the years (log scale)

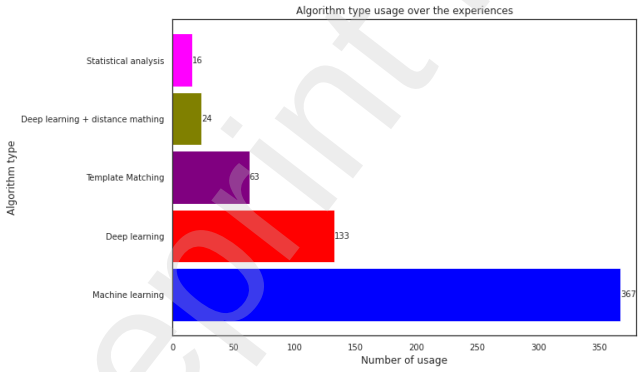


Figure 28. Algorithm type usage over the experiences

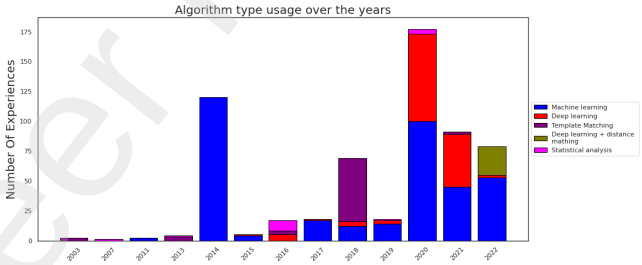


Figure 29. Algorithm type usage over the years

8 Recognition algorithm

The last part of an authentication system is the classification algorithm. It is this piece of software that recognizes registered people and rejects unknown people. Two metrics are important to determine the performance of an algorithm: accuracy and equal error rate. Accuracy represents the ability to accurately identify a subject, which is the true positive rate, and we want to maximize it. Additionally, two metrics are used to measure errors in an authentication system: false match rate (FMR) and false non-match rate (FNMR). The FMR is the probability that the algorithm recognizes someone different from the one being tested, which represents the probability of an intruder being treated as a genuine user. The other metric, FNMR, represents the probability of a genuine user being treated as an impostor. Since biometric authentication is a template matching problem, an algorithm provides a matching probability, and we need to apply a threshold value. The value of the threshold deeply influences the FMR and FNMR. It is also important to note that a threshold value where FNMR and FMR are equal always exists; this point is called Equal Error Rate (ERR). This point is traditionally used to measure the performance of a biometric authentication system, which we will use in the same way in this study.

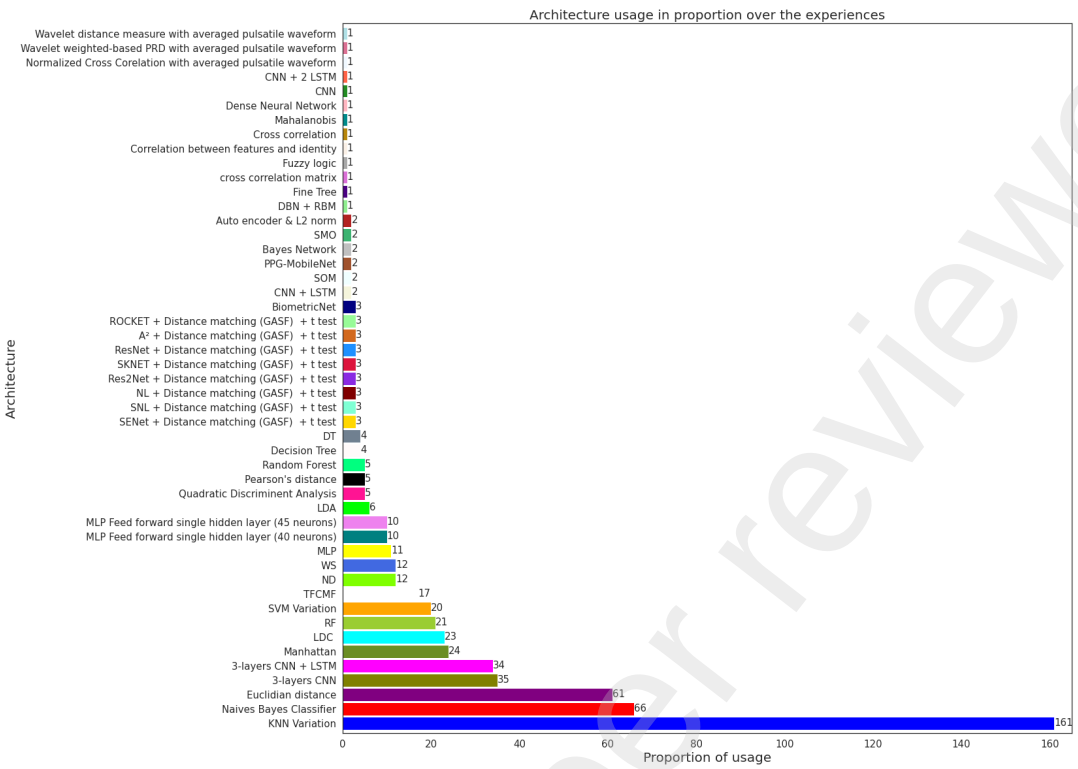


Figure 30. Algorithm architecture usage over the experiences

In all the studied papers, we identified four main classification methods: statistical, machine learning, template matching, and deep learning. We will briefly explain the main models used for each method. Statistical models include Bayesian networks and Hidden Markov Models (HMMs). Machine learning models include Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs). Template matching methods involve feature extraction and matching, such as Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). Deep learning models include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks. These methods have been used in various combinations to classify biometric data, such as fingerprints, faces, and voices.

Machine learning has been widely used in classification, as shown by Figure 28. Its popularity can be observed throughout the years, as depicted by Figure 29.

Figure 30 displays the various architectures employed by the community, with a missing value rate of 0.81%. We group in one category all the KNN and SVM variation to improve readability. All the variation of KNN and SVM are detailed in the Figure 31. We observed that 80 tested architectures were reported in the papers, with some architectures being tested multiple times, such as Naive Bayes Classifier and Euclidean distance in template matching. Figure 31 shows that the most commonly used architecture is KNN, accounting for 26.89% of all architectures used. Naive Bayes and CNN-based architectures are also popular.

8.1 Statistical

Statistical classification methods, primarily based on cross-correlation, were initially used by research teams to develop PPG-based biometric authentication systems. Some teams also employed direct statistical analysis techniques such as LDA for creating classification systems. The first study on PPG authentication utilized fuzzy logic with a Gaussian function [31]. They computed a score between an enrollment template and a given signal using the Gaussian function parameters (μ and σ). The Gaussian function was calculated for each pulse to maximize the overlap of the PPG signal's maximum area.

[84] computed the correlation of each extracted feature for each subject. They found that the selected features were highly correlated within each subject but not correlated with others. However, they did not attempt to create a full authentication system and did not provide any performance metrics. Later, [63] used Euclidean distance between FFT features extracted from PPG signals. They demonstrated that this distance was significantly higher between two pulses from different subjects compared to those from the same subject. Again, no performance metrics were provided.

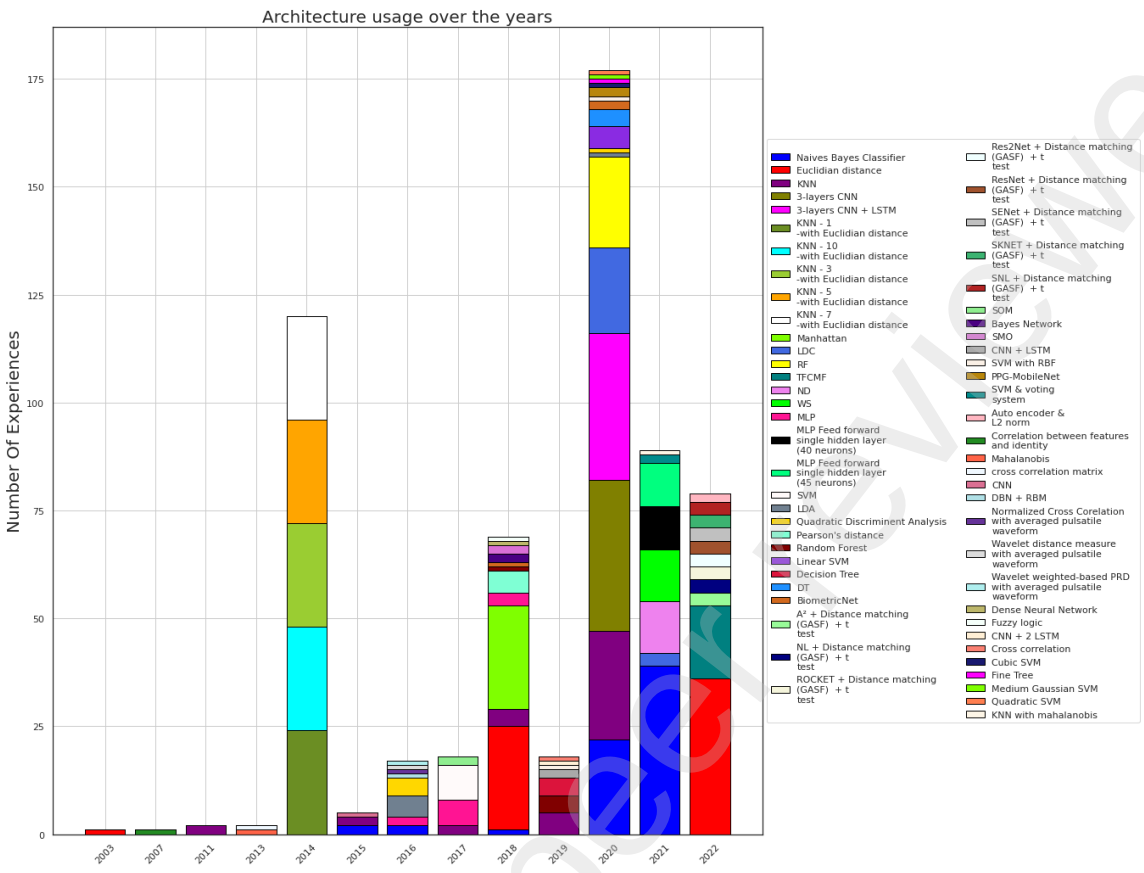


Figure 31. Algorithm architecture usage over the years

8.2 Template matching

The second most common method employed by the community for PPG-based biometric authentication is template matching. This technique, which is often used in other biometric systems, involves creating a template that is stored in the system and comparing it to each input signal. A matching score is computed for each input signal using distance metrics such as Euclidean or Manhattan distances. When classification occurs with these distance metrics, there is only one parameter to set: the threshold value. When a template is computed for a claimed user, the system measures the distance between the two templates. If this distance exceeds the threshold value, the authentication is rejected, but if it is below the threshold, the authentication is accepted. Distance metrics can be used with any kind of template and are the basis of machine learning algorithms such as k-Nearest Neighbors (KNN) [28].

The main advantage of template matching classification is that no machine learning algorithm needs to be trained. We simply need to create a template using specific features, store it in the system, and then compare it with the input signal's template.

8.3 Machine Learning

Machine learning algorithms are widely used in the selected studies. These algorithms need to be trained using a subset of the available data (the training set) and tested using another subset of data (the testing set). Sometimes, when hyperparameters need to be tuned, another subset of data is used for this purpose (the validation set). In general, around 4.3% of the available data is used for validation, 8.7% for testing, and the rest is used for training. For example, in the ImageNet Challenge 2014 [61], these proportions were used to split the data into training, validation, and testing sets.

These machine learning algorithms provide a function that is refined and corrected with data. They can generalize their function to new data, which is why they need to be tested using previously unseen data. The goal of each algorithm is to classify data into at least two classes. For example, the Support Vector Machine (SVM) algorithm [86] will split the data into multiple classes by maximizing the space between two classes using a combination of multiple linear functions. If the clustering function is not linear, it can be improved by using a "kernel trick" [86]. Other machine learning algorithms used in the selected studies include Random Forest [49], Decision Tree [58], Naive Bayes [48], and

Bayesian Network [26]. However, KNN and SVM are the most commonly used algorithms in PPG-based biometric authentication.

The performances of these machine learning algorithms are highly variable and depend on factors such as the quality of the signal, the training data set, and the testing dataset. As a result, one algorithm may perform better than another in one study, while a different algorithm may perform better in another study. For example, [47] shows that Random Forest achieves 99% accuracy, while KNN achieves 98% accuracy. However, [32] reports that KNN has an accuracy of 94.44% and Random forest has an accuracy of 90.39%. This is why it is essential to compare studies that use the same data and the same validation method.

8.4 Deep Learning

Deep learning algorithms are the second most commonly used in the selected studies. These algorithms are based on neural networks such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and other types of multi-layered neural networks. Many different kinds of deep learning algorithms can be combined because they are multi-layered. Deep learning algorithms can be used for all the described steps: noise reduction, feature extraction, selection, and classification. The most efficient algorithms in our selected studies use a combination of CNN and Long Short Term Memory (LSTM) [74].

LSTMs are very efficient at processing time-dependent signals and was initially developed for natural language processing. For example, [16] uses a two-layer CNN to extract features from the PPG signal, followed by an LSTM layer that feeds into two final neurons activated using the Soft-Max function. The final layer outputs the class of the signal, corresponding to the authenticated user, with a certain probability. They claim to have achieved 96% accuracy but do not provide any EER values. Similarly, [24] uses a similar architecture and claims to obtain the same results. However, they do not explain how they train the algorithm, which data is used for training, and which one is used for testing.

The major problem with learning methods is the impact of the test methodology. For example, if data used for training and testing are unbalanced, the algorithm can perfectly fit the data and achieve a very high accuracy rate, but the accuracy drops when fed with different data. This issue has been observed in various studies, such as [10], where it was demonstrated that algorithms like Bayes Network, Naive Bayes, and Multi-Layer Perceptron achieved 100% accuracy on unbalanced datasets. However, this is true only for a small subset of the data. They split data into gender and ages groups and achieved 100% accuracy only for people aged between 16 and 35. For other categories, the accuracy drop between 80% and 95%. If the same data are split based on gender, they obtain an accuracy rate between 80% and 90%. This appears to show a large variability in the algorithm

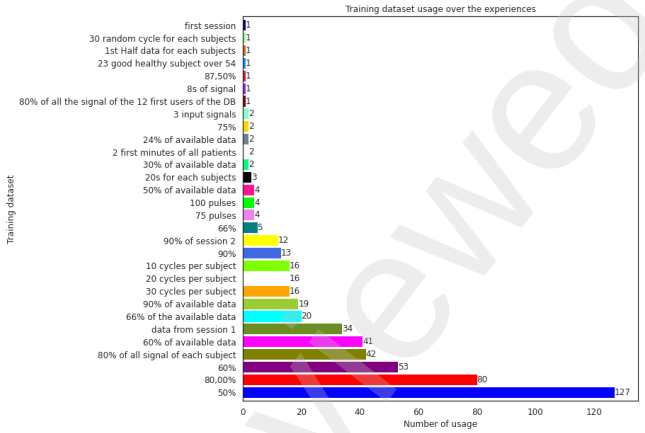


Figure 32. Training datasets usage over the experiences

performances and further metrics and test must be done to determine if it can be due to unbalanced data or if there is a problem based on age and gender that needs to be considered. A common testing methodology must be thus used to correctly benchmark all the algorithms.

9 Study comparison

In this section, we will compare the results of various studies related to learning methods. First, we will analyze the training and testing sets used in these studies. Then, we will analyze the validation methods employed. Finally, we will analyze the results in terms of accuracy, EER, lowest FMR, and lowest FNMR.

9.1 Training and validating sets

The training and testing sets are crucial to determine the validity of the results. To achieve a valid result, the dataset used for training should be large enough and representative of the population, while the testing set should be different from the training set and not used in the training process. In most of the studies analyzed, the training and testing sets comprised the same users, with the training phase acting as enrollment and the testing phase as authentication.

Figure 32 shows the usage of training datasets across various experiences. For this parameter, we observe 12.76% missing values. We observe significant heterogeneity in the creation of datasets, with most studies using a percentage of available data or a fixed size of signal for each subject. Very few studies split users into genuine and impostor categories. Figure 33 shows the evolution of dataset usage over time. Figure 35 and Figure 34 show the composition of the validating dataset. We can observe that most of the experience use at least 10% of the available data. It is obvious that there is no consensus on how to split training and validating datasets, making it difficult to compare studies due to the significant heterogeneity in dataset creation methods.

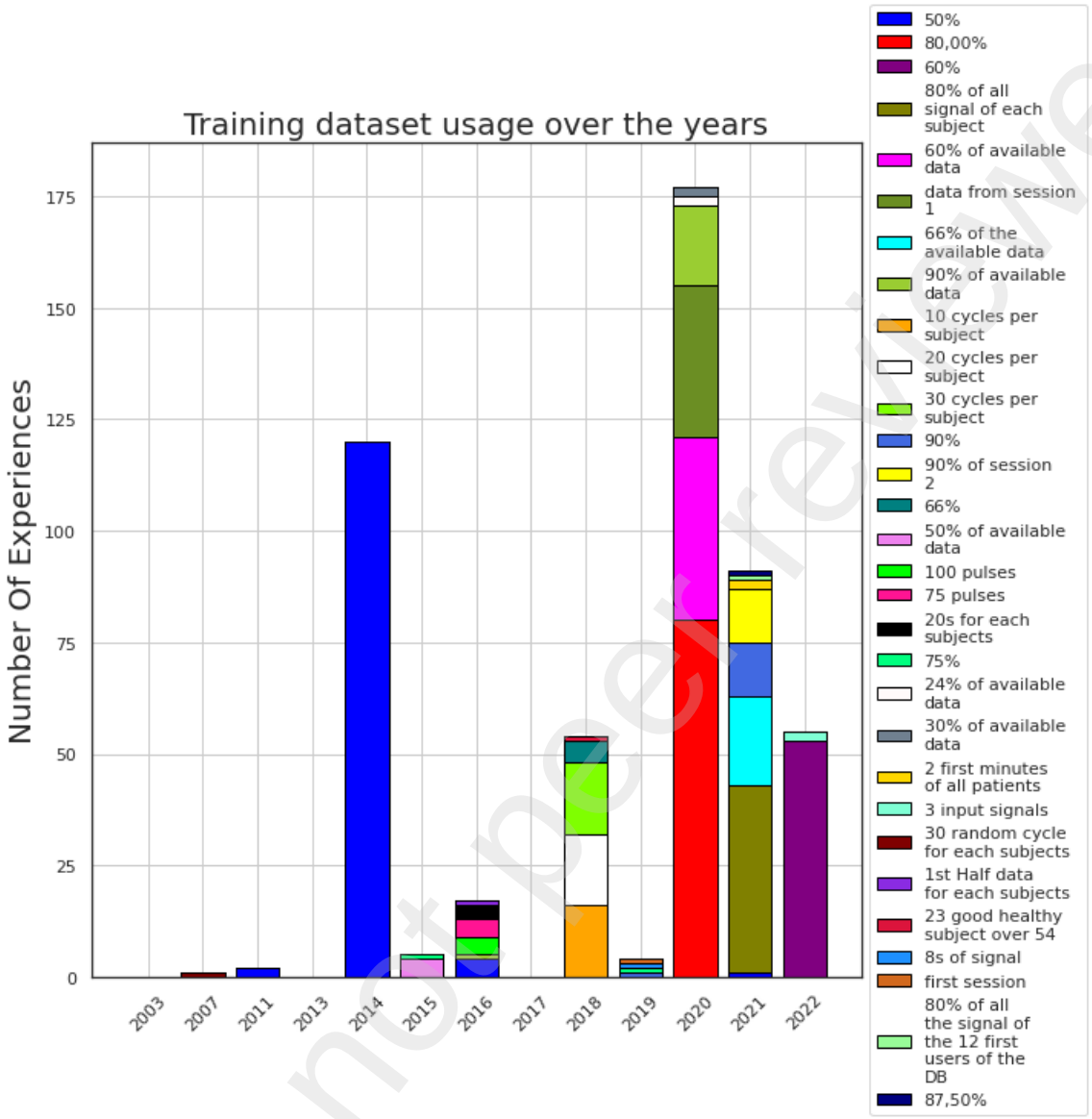


Figure 33. Training datasets usage over the years

9.2 Validation methods

The validation method is critical to ensure the validity of the results and avoid troubles such as over fitting. For this parameter we observe 78.78% of missing values. This shows that most of the community does not use a validation method and provide results that may contains bias and may not be valid. This is why it is of the utmost important to define a clear methods to reproduce and benchmarks all the provided methods.

Figure 36 shows the different validation techniques used across various studies. We observe five methods, including cross-fold validation, with variations in the number of folds used. The most commonly used technique is 10-fold cross-validation with L2 regularization. It is evident that there is significant heterogeneity in validation methods employed, making it difficult to compare studies and ensure valid results.

Figure 37 shows the usage of validation techniques over time. We observe that 10-fold cross-validation was used in 2015 by a few studies, and this technique dominated until 2020 when L2 regularization became more commonly employed. However, there is significant heterogeneity in validation methods employed across various studies, indicating a lack of methodological consistency within the community.

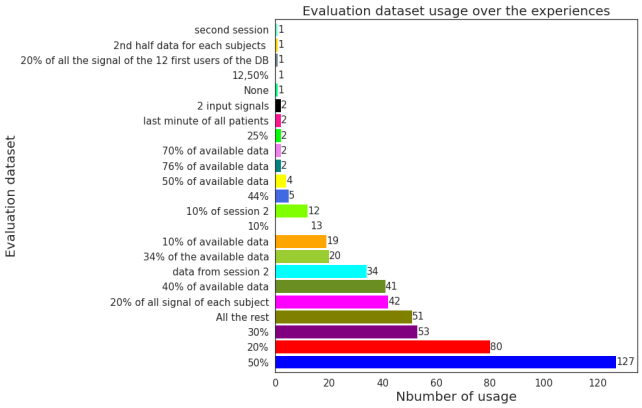


Figure 34. Validating datasets usage over the experiences

9.3 Performances metrics

To compare various studies and find the best architectures, we wanted to analyze multiple performance metrics, including accuracy and EER, which are commonly provided in study results. However, we observe significant missing values for these parameters, with 14.75% missing values for accuracy and 68.49% missing values for EER. Due to this lack of consistency in reported metrics, we will not be able to compare studies using FMR or FNMR. Instead, we will focus on comparing few studies that provide good global performance results over multiple experiences. Figure 38a shows the evolution of global performance over time, with most studies achieving accuracy rates between 80% and 100%. Figure 38b show the global EER over the years. It seems to reduce overall, although the extremes are still significant, showing an instability of the techniques employed.

While there is variability in performance across studies, with some showing accuracy rates as low as 20% or lower, it is clear that not all architectures are equal. To achieve a valid comparison, we need to compare studies using the same dataset. Additionally, while accuracy is commonly used to evaluate biometric systems, ROC or DET curves provide more insightful information and should be employed for comparisons. However, despite this being recognized in various papers, only a few provide the data needed to draw these curves, making it difficult to compare different works effectively. A good metric derived from ROC curves is the area under the curve (AUC), which could potentially serve as a more meaningful comparison measure.

While only three studies provide AUC values, representing 13 experiences, their datasets are different, making it difficult to compare and contrast them effectively. [9] provides some mean AUC values for some of its experiences, but does not provide raw data or standard deviations. [51] provides multiple ROC curves and the raw mean AUC values for all of the experiences and at every stage (validation, develop and test), as well as standard deviations for each experience. We can observe significant variability in AUC performances across studies. [59] provides ROC curves and AUC values for two of their experiences, but it seems they have reproduced some parts of previous studies by [16], [81], and [31]. It also appears that they have made changes to preprocessing and feature extraction techniques. Overall, the methodology employed in these papers is good and should serve as a basis for future studies. Two other papers[17] and [50] provide the DET curves which is similar to the AUC curve, but with no exploitable values.

Other metrics, such as precision, recall, specificity, and f-measure, could also be informative and should be provided for each experience. [42] is the only study to provide these measures for all subjects, while [43] provides mean accuracy, specificity, and sensitivity values for its experiences, but not for all subjects. It is evident that there is significant heterogeneity in the metrics employed across various studies, making it difficult to compare and contrast them effectively.

[39] presents accuracy, precision, and recall values for cluster identification but not for each individual or the mean scores for all subjects. [5] provides only accuracy, specificity, and error rate values for one experience, and only for a subset of subjects (14 out of 57). The values are biased and not exploitable. [33] provides some sensitivity metrics for single-session experiences only. It is evident that there is significant heterogeneity in the metrics employed across various studies, making it difficult to compare and contrast them effectively.

Finally, [40], [41], [81] , [35], [34], [83] draw some ROC curves for some or all of their experiences but never provide any exploitable AUC values.

9.4 Experiences comparison

Here we compare studies that use similar datasets or provide a comparison on multiple algorithms or methods. We compared works by Sancho et al. [64], Yadav et al. [81], and Yang et al. [82]. They compared multiple algorithms and methods using multiple online databases, including the Canopbase IEEE TBME dataset, which provides only one signal record for each subject and is therefore suitable for testing short-term scenarios.

Figure 39 shows the accuracy of each algorithm in studies that use the Capnobase IEEE dataset. We only included studies that provided the accuracy and architecture used, resulting in 112 exploitable studies. Table 7 indicates that most studies report accuracies between 90% and 100%. The Naive Bayes Classifier has the lowest mean performance and two outliers, while KNN provides good accuracy with a mean of 98.54% and low variability (std=1.85). The Euclidean matching method shows a mean accuracy of 94.4% and variability (std=2.07). The architecture that appears to provide the best performance with the most stability is the 3-Layer CNN, with a mean accuracy of 99.25% and low variability (std=0.89), although only six studies used this architecture.

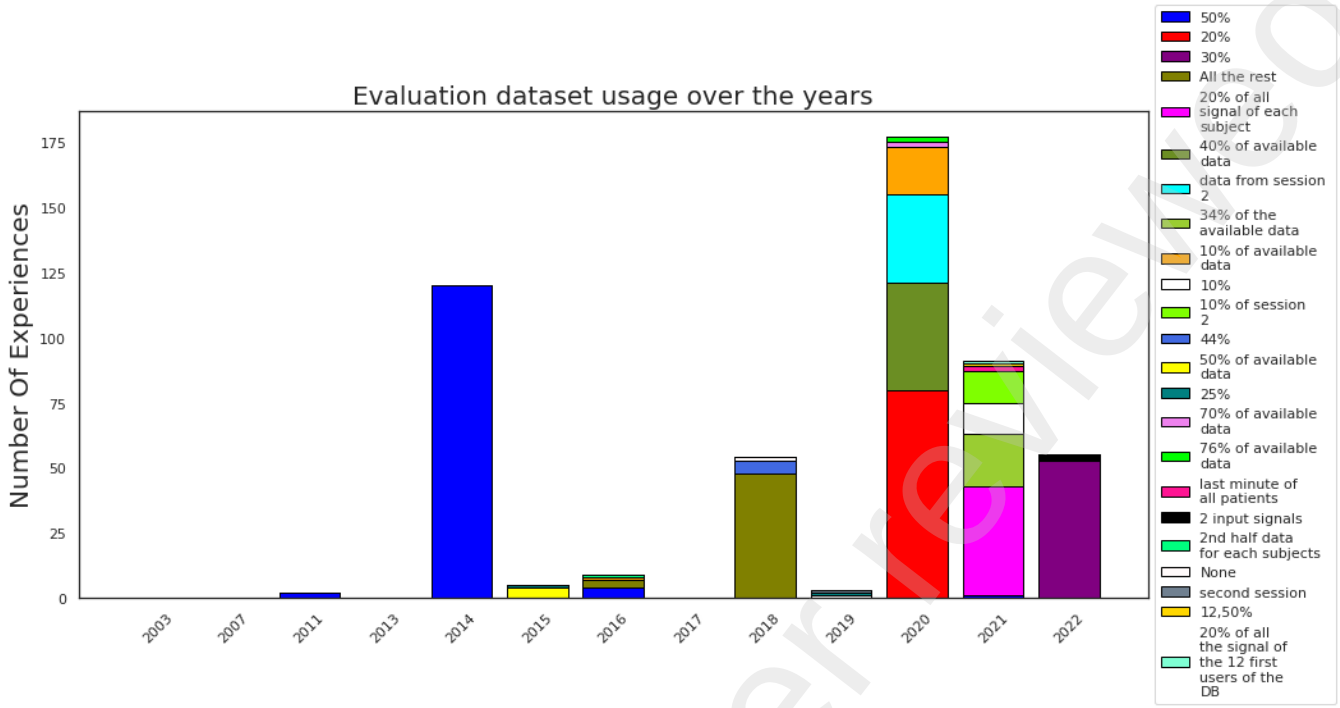


Figure 35. Validating datasets usage over the years

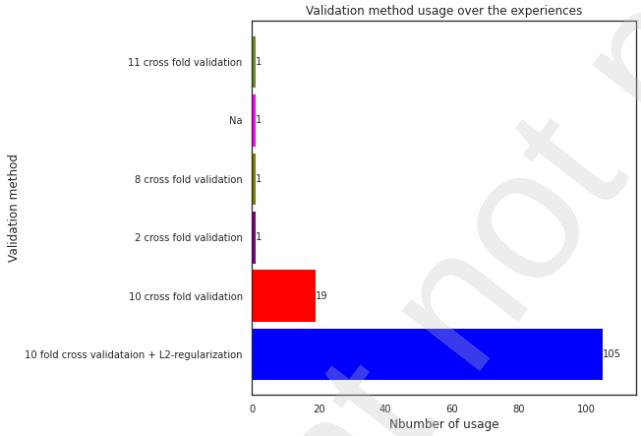


Figure 36. Validation usage over the experiences

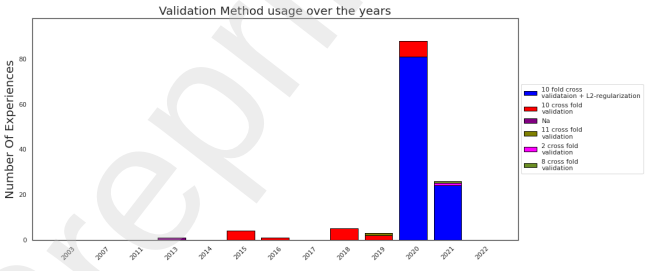
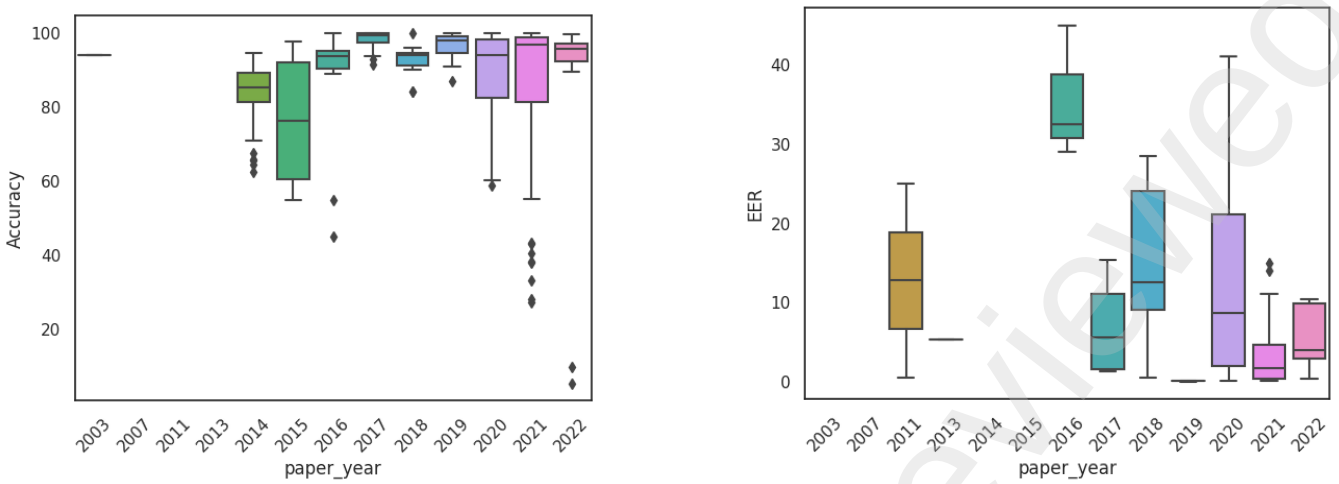


Figure 37. Validation methods usage over the years

The first study we analyzed is the one by Sancho et al. [64], where they compared multiple feature extractors using classification with Manhattan and Euclidean distances. They also investigated the differences between single-session (short-term) and multi-session recordings (long-term), computing their EER by training and testing each dataset separately and then calculating a mean EER. They mainly used 30 cycles for enrollment and testing, but found that using fewer cycles was not significant enough to improve accuracy, as they gained only approximately 1% in EER. However, the EER with long-term sessions was significantly higher, ranging from 8% to 24%, indicating that PPG signals can vary greatly between recording sessions and that their methods cannot be generalized.

Another relevant study is the one by Yadav et al. [81], where they mainly compared feature selection algorithms using a CWT transformation to extract features and a Pearson's distance [15] matching method for recognition. They used the Canopbase dataset and trained their model on the first 45 seconds of each user's signal. The testing data consisted of random segments of varying duration between 6 and 7 seconds. They found that DLDA was the best feature selection algorithm, as it had the lowest EER rate of all (0.46%). They used this technique with other datasets and achieved EER rates between 2% and 3%, except for the exercise sessions in the Biosec Database, which had an EER rate of 5%. This study shows a good comparison between feature selection methods but does not provide an accuracy



(a) Accuracy evolution over the years

(b) EER evolution over the years

Figure 38. Evolution of the accuracy and EER over the years

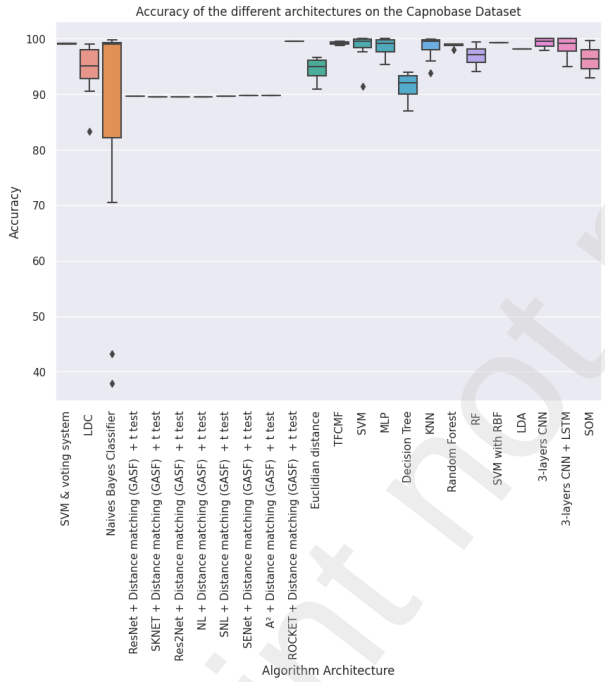


Figure 39. Accuracy of the different architectures using the Capnobase dataset

rate. From this study, we can conclude that DLDA works well with CWT extraction and Pearson's metrics, but using different feature extraction and classification methods may yield other effective combinations.

In their study, Yang et al. [82] used three databases and applied the same methods to each dataset, providing an accuracy rate. The three datasets were BIDMC, MIMIC-II, and Canopbase. To extract features, they developed a new algorithm using a three-layer model that produces a sparse softMax vector. They used 80% of available data for training and 20% for testing, then tested KNN, Random Forest, Linear Discriminant Classifier, and Naive Bayes as classifiers. For the Canopbase dataset, they obtained accuracy rates of 97.59% with Random Forest and 99.92% with KNN. These accuracy rates were consistent across all three datasets. They also tested their methods with one-layer and two-layer feature extraction models but found that the three-layer model was best. From this study, KNN appears to be slightly more efficient than the other algorithms, although no tests were conducted with genuine and impostor data to calculate an EER. This study provides a good comparison of some machine learning algorithms, but it cannot be generalized and applied to every algorithm. Moreover, feature extraction played a significant role in their method, though this approach can be effective when used with other feature extraction techniques as well.

9.5 Time stability: One session VS Two sessions

Kavsaoglu et al. [42] were among the first teams to use two separate sessions for training and testing, but they did not provide any information on the time between the two sessions. This leads us to conclude that their studies cannot be used to test long-term stability. The first work to highlight the importance of long-term stability was conducted by Hwang et al. [35] in 2020. They dedicated one study to evaluate the impact of using a single session for enrollment and another session for testing. They used the Biosec1 and Biosec 2 datasets, which provide two sessions recorded 14 days apart in the same conditions (Biosec1) or in different states (Biosec 2). They conducted the same experiments as

Accuracy								
Architecture	**count**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
3-layers CNN	6.0	99.25	0.89	97.90	98.70	99.50	100.00	100.00
3-layers CNN + LSTM	6.0	98.46	2.03	94.90	97.72	99.20	100.00	100.00
A^2 + Distance matching (GASF) + t test	1.0	89.80	NaN	89.80	89.80	89.80	89.80	89.80
Decision Tree	4.0	91.25	3.09	87.00	90.00	92.00	93.25	94.00
Euclidian distance	12.0	94.40	2.07	90.85	93.34	94.96	96.15	96.64
KNN	13.0	98.549	1.85	93.76	98.00	99.54	99.79	99.95
LDA	1.0	98.11	NaN	98.11	98.11	98.11	98.11	98.11
LDC	7.0	94.11	5.57	83.24	92.75	95.05	97.97	99.04
MLP	6.0	98.47	1.87	95.31	97.65	99.20	99.80	100.00
NL + Distance matching (GASF) + t test	1.0	89.50	NaN	89.50	89.50	89.50	89.50	89.50
Naives Bayes Classifier	20.0	88.81	18.72	37.86	82.11	98.98	99.28	99.81
RF	7.0	96.90	1.87	94.05	95.72	97.17	98.13	99.40
ROCKET + Distance matching (GASF) + t test	1.0	99.50	NaN	99.50	99.50	99.50	99.50	99.50
Random Forest	4.0	98.75	0.50	98.00	98.75	99.00	99.00	99.00
Res2Net + Distance matching (GASF) + t test	1.0	89.50	NaN	89.50	89.50	89.50	89.50	89.50
ResNet + Distance matching (GASF) + t test	1.0	89.60	NaN	89.60	89.60	89.60	89.60	89.60
SENet + Distance matching (GASF) + t test	1.0	89.70	NaN	89.70	89.70	89.70	89.70	89.70
SKNET + Distance matching (GASF) + t test	1.0	89.50	NaN	89.50	89.50	89.50	89.50	89.50
SNL + Distance matching (GASF) + t test	1.0	89.60	NaN	89.60	89.60	89.60	89.60	89.60
SOM	2.0	96.30	4.73	92.96	94.63	96.30	97.97	99.65
SVM	8.0	98.30	2.89	91.46	98.32	99.47	99.91	100.00
SVM & voting system	2.0	99.09	0.09	99.03	99.062	99.09	99.12	99.16
SVM with RBF	1.0	99.30	NaN	99.30	99.30	99.30	99.30	99.30
TFCMF	5.0	99.23	0.30	98.79	99.06	99.31	99.43	99.56

Table 7. Accuracy of the different algorithm architectures using the Capnobase Dataset

for the single-session scenario and observed a significant drop in performance. In all their studies, they observed a decrease in accuracy of approximately 30% on average and an increase in EER of up to 41%. In the single-session scenario, the accuracy ranged from 91% to 100% and the EER ranged from 0.1% to 10%. In the double-session scenario, the accuracy ranged from 58% to 81% and the EER ranged from 18% to 41%.

The second work on this topic was conducted by the same team in 2021 [34], in which they attempted to improve stability using a Generative Adversarial Network (GAN) technique. They did not provide the accuracy for the single-session scenario, only the EER, which ranged from 0.1% to 15%. For the double-session scenario, they did not provide an average EER but only the accuracy, which ranged from 77% to 88%. This shows better stability than in their previous paper, but there is still a significant gap between the two scenarios, and in both papers, we can observe that performance varies significantly between datasets. Overall, we concluded that no single algorithm outperforms all others on all datasets, and different architectures are more effective for each dataset.

These two papers show the need to investigate further the time stability of the PPG biometric recognition. They allow us to partially answer *RQ 1.2): the biometric authentication using PPG are still unstable in long term scenario*. However, the performances in the first study were good enough to encourage further investigation on this topic. This case must be included in a dedicated benchmark of the algorithms. We also need a larger array of datasets to study this phenomenon. These datasets would need to be recorded on consecutive days and hours, and, when possible, for a duration of 24 hours.

10 Future works

We saw through our analysis that many issues occur in all studies about human authentication through PPG. Some biases in the data set, learning, testing, noise suppression, along with others, do remain. Hence, we want to provide tracks for the community in order to increase the quality of future studies. We propose solutions for the constitution of the datasets along with a testing protocol to measure performance that would result in less biases. We then propose

a benchmark method that will be implemented in future works.

10.1 Database or federated learning

As stated in Section 5, most PPG-based biometric authentication studies are conducted in controlled environments where subjects are asked to sit and relax, which may not be realistic in real-world situations such as opening a door or going for a run. Additionally, the PPG signal can vary significantly from one record to another, as shown by [64]. To address these issues, we need to create a large dataset with multiple records taken in different and various conditions, ideally over the course of a full day during at least two sessions. The BioSec data set provided by the University of Toronto is a good start, but the community needs to provide more datasets like this one, including those with different materials. Since every individual is different and multiple PPG sensors coexist with multiple sampling rates and different frequencies, algorithms need to account for these factors when analyzing the data. Therefore, we need a more heterogeneous dataset.

To achieve a robust PPG-based biometric authentication system, we need as many data as possible. Creating a big database where each team can add small amounts of data is one solution to this problem. It is difficult for research teams to gather more than 20 volunteers, which leads to variability in the total number of patients. By adding gathered data to a public database, we can reduce this variability and help compare algorithms. This will also help determine how an algorithm scales with the number of patients. After all, one key point of an authentication system is its ability to be used by the highest number of people.

To address privacy concerns surrounding biomedical data, federated learning [72] can be used as a new paradigm. With this method, the learning phase of a deep learning algorithm can be distributed through multiple centers, allowing each center to train the algorithm with its data without the need to publish it. This approach can help protect the privacy of sensitive biometric data while still enabling research and development in PPG-based biometric authentication systems.

10.2 Benchmark

One big results of our study is the the need to determine a common methodology to evaluate the experiences and the multiples architectures. Most of the studies use different quantity of data to train and test their models and few of them used methods to prevent over-fitting such as L2 Regularization or 10 cross fold validation. Thus, we need to define a full benchmark method which provides fixed methods and metrics for all experiences. The methods should fix or test multiples parameters :

- Training dataset
- Testing dataset
- Validation dataset
- Enroll process and times
- Identification process
- Authentication process
- Single case scenario
- Long time stability
- Validation method
- Continuous authentication

Then multiple performances metrics representing the security of the system, its usability and stability should be computed :

- Accuracy (global and detailed for each subject)
- EER (global and detailed for each subject)
- ROC curve and AUC
- Number of subject that can not use the system (FTE)
- Memory performances of the system
- Number of signals rejected for poor quality (FTA)
- Enroll time

In our future works, we will propose one benchmark method providing these metrics. Additionally, we will split the dataset to enroll and test it as splitting can influence many of said metrics. Then, we will apply this benchmark to a maximum of different architectures in order to find the prime ones.

11 Conclusion

In conclusion, we gathered 44 studies with the same goal: creating an efficient way of authenticating people through PPG records. We extracted around 600 experiences made during the past twenty years. These works provide tracks to explore this topic, however, many methodological biases remain, thus leading to the impossibility to compare most of the available works. We identify the four main phases in the development of an algorithm able to recognize a person with its PPG signal. For each phase, we define objective criteria, but the heterogeneity of the gathered studies has lead to the impossibility to clearly define which method is the best, or the advantages and disadvantages of each part. Ultimately, we were able to compare and contrast some studies presented in Section 9 and to answer some of our crucial research questions.

11.1 RQ 1.1.

The performances in short term scenarios are favorable for most of the tested architectures and can be exploited.

11.2 RQ 1.2.

The performances of PPG-based biometric authentication systems in long-term scenarios are less certain than in short-term scenarios, with a drop of performance around 20%. While this is not too much and the systems are still better than random choices, it is not sufficient or relevant enough for real-world use. Therefore, further research on this topic is needed.

11.3 RQ 1.3.

In general, most PPG-based biometric authentication datasets contain less than 50 subjects. Only the Biosec2, VORTAL, and VITAL datasets provide more than 100 subjects. However, few tests have been conducted with the full Biosec2 dataset, and no studies have been done with the full VORTAL or VITAL datasets. The performances with the whole Biosec2 dataset were conclusive enough to be used in the real world, but further studies are still needed to confirm this. Additionally, we need to build larger datasets with at least 1000 and 10,000 users to fully scale up and confirm the performance of these systems.

11.4 RQ 1.4.

Very few studies have been conducted to test this hypothesis. Only eight experiences using the DEAP dataset provide records with different emotional states, and only 11 use the TROIKA dataset, which provides signals recorded during physical exercise and at rest. The results are slightly lower (between 95% and 96% accuracy) than others, but this may not be significant due to the relatively small number of experiences. This shows that PPG-based biometric authentication may be robust to physiological changes, but further research is needed to confirm this.

11.5 RQ 2.1 and 2.2

The validation methods and metrics provided in the studied experiences do not allow us to answer these two research questions. None of the selected studies compute a Failed To Enroll metric or the mean number of tries that an user needs to be authenticated. These metrics should be included in future studies on PPG-based biometric authentication to provide more accurate and comprehensive evaluations of the performance of these systems.

11.6 RQ 3.1

Considering one common dataset, 24 classification architectures have been tested. If we consider the entire architecture (feature extraction, selection, etc.), we obtain 112 architectures. If we consider all the experiences with all the datasets, 315 different architectural combinations have been tested. However, these tests should be re-run with the correct datasets and validation methods to provide more accurate and reliable results.

11.7 RQ 3.2

Considering all the datasets and all the possible elements or architectures (noise reduction, feature extraction and selection, segmentation etc.) we observed: 21 different methods for segmentation, 19 for normalization, 34 for noise reduction, 56 features types, 45 features length, 74 feature selection methods and 67 different classification architectures algorithms. With only these parameters, and considering the usage of only one methods for each category we can define 169 495 774 560 different pipeline architectures. This gives us a coverage of 1.85×10^{-7}% which is very low. Moreover, many methods can also be added for each parameter and multiples parameter of one category can be used. This shows that most of the experiences have not been done and most of the work need to be done.

11.8 RQ 3.3

Some classification architectures are more efficient than others, the neural network based architectures seem the better ones. We still need to test this hypothesis with more experiences to ensure that this is the right answer. For other pieces of architectures (features extraction, selection, noise reduction etc.), we need to conduct more experiences to be able to answer. Moreover, each piece of architecture influences the final score and some pieces may work synergistically for this problem, whilst others don't. We need further investigation.

11.9 RQ 3.4

Some classification architectures have been more used than others. The segmentation in single cycles is very popular, the Butterworth filter was the most used to reduce noise. In the normalization usage, it is the zero-mean normalization that is the most popular. However, this popularity is not as sharply outlined as for the Butterworth filter or the single cycle segmentation. For the extracted features, the fiducial features were very popular at first, but progressively many other kinds of features were tested. The frequencies extracted with FFT seem a little bit more popular than the others. For the selection process, it seems that the most popular method is to select all the extracted features. Even with all that said, this popularity is very low as it only represents 10% of the experiences though it is the most reused one over the papers. As for the feature size, we saw no real popularity. For the classification algorithms architectures, the Deep-learning architectures using CNN seems the most popular, followed by machine learning algorithms like KNN and SVM. The popularity of an element can show a certain form of consensus of the community; specific experiences should be done to determine if the popularity of these elements is due to simplicity of the implementation or if they are truly better. For example, the single cycle segmentation may not be as efficient as 10-cycles segmentation.

In the end, we were able to show the evolution of the usage over the years in the community. We observe the increase usage of publicly available datasets over the years which provide the same basis for every one. However the validation methods still lacks. Thus we need to define one unique method and benchmark all the tested experiences. This benchmark will have to show metrics to represent the time stability of the system, the security level, the ergonomic level and the usability level (using the Failure to Enroll problem). Moreover the lack of open source code does not allow

the community to reproduce the experiences. This is why we need to provide one unique platform where each team can upload its code and compute all the relevant associated metrics.

In our future works, we will implement some of the proposed algorithms in this literature review and benchmark them with the proposed method. We will also provide one platform where every team can test their algorithms: this will enable us to dispense finer answers to our research questions.

References

[1] 2005. Décret n°2005-1726 du 30 décembre 2005 relatif aux passeports. https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000268015/.

[2] Mohammed Abo-Zahhad, Sabah M Ahmed, and Sherif N Abbas. 2014. Biometric authentication based on PCG and ECG signals: present status and future directions. *Signal, Image and Video Processing* 8, 4 (2014), 739–751.

[3] D Agrò, R Canicattì, A Tomasino, A Giordano, G Adamo, A Parisi, R Pernice, S Stivala, C Giaconia, AC Busacca, et al. 2014. PPG embedded system for blood pressure monitoring. In *2014 AEIT Annual Conference-From Research to Industry: The Need for a More Effective Technology Transfer (AEIT)*. IEEE, 1–6.

[4] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. 1974. Discrete cosine transform. *IEEE transactions on Computers* 100, 1 (1974), 90–93.

[5] Aya Al Sidani, Ali Cherry, Houssein Hajj-Hassan, and Mohamad Hajj-Hassan. 2019. Comparison between K-Nearest Neighbor and Support Vector Machine Algorithms for PPG Biometric Identification. In *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*. IEEE, 1–4.

[6] Aya Al-Sidani, Bilal Ibrahim, Ali Cherry, and Mohamad Hajj-Hassan. 2018. Biometrie identification using photoplethysmography signal. In *2018 Third International Conference on Electrical and Biomedical Engineering, Clean Energy and Green Computing (EBECEGC)*. IEEE, 12–15.

[7] Mohammed Aledhari, Rehma Razzak, Basheer Qolomany, Ala Al-Fuqaha, and Fahad Saeed. 2022. Biomedical IoT: Enabling Technologies, Architectural Elements, Challenges, and Future Directions. *IEEE Access* 10 (2022), 31306–31339.

[8] John Allen. 2007. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement* 28, 3 (2007), R1.

[9] Turky N Alotaiby, Fatima Aljabarti, Gaseb Alotibi, and Saleh A Alshebeili. 2020. A Nonfiducial PPG-Based Subject Authentication Approach Using the Statistical Features of DWT-Based Filtered Signals. *Journal of Sensors* 2020 (2020).

[10] Siti Nurfarah Ain Mohd Azam, Khairul Azami Sidek, and Ahmad Fadzil Ismail. 2018. Photoplethysmogram Based Biometric Identification Incorporating Different Age and Gender Group. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10, 1-5 (2018), 101–108.

[11] Sangeeta Bagha and Laxmi Shaw. 2011. A real time analysis of PPG signal for measurement of SpO2 and pulse rate. *International journal of computer applications* 36, 11 (2011), 45–50.

[12] Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* 18, 1998 (1998), 1–8.

[13] Jean Baptiste Joseph baron de Fourier. 1822. *Théorie analytique de la chaleur*. Firmin Didot.

[14] Lucas Bastos, Bruno Cremonezi, Thais Tavares, Denis Rosário, Eduardo Cerqueira, and Aldri Santos. 2021. Smart Human Identification System Based on PPG and ECG Signals in Wearable Devices. In *2021 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 347–352.

[15] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.

[16] Dwaipayan Biswas, Luke Everson, Muqing Liu, Madhuri Panwar, Bram-Ernst Verhoef, Shrishail Patki, Chris H. Kim, Amit Acharyya, Chris Van Hoof, Mario Konijnenburg, and Nick Van Helleputte. 2019. CorNET: Deep Learning Framework for PPG-Based Heart Rate Estimation and Biometric Identification in Ambulant Environment. *IEEE Transactions on Biomedical Circuits and Systems* 13, 2 (2019), 282–291. https://doi.org/10.1109/TBCAS.2019.2892297

[17] Angelo Bonissi, Ruggero Donida Labati, Luca Perico, Roberto Sassi, Fabio Scotti, and Luca Sparagino. 2013. A preliminary study on continuous authentication methods for photoplethysmographic biometrics. In *2013 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*. IEEE, 28–33.

[18] Stephen Butterworth et al. 1930. On the theory of filter amplifiers. *Wireless Engineer* 7, 6 (1930), 536–541.

[19] Passport Canada. 2011. The ePasseport. https://web.archive.org/web/20110728085939/http://www.ppt.gc.ca/eppt/index.aspx?lang=eng.

[20] Samik Chakraborty and Saurabh Pal. 2016. Photoplethysmogram signal based biometric recognition using linear discriminant classifier. In *2016 2nd International Conference on Control, Instrumentation, Energy & Communication (CIEC)*. IEEE, 183–187.

[21] Tilendra Choudhary and M Sabarimalai Manikandan. 2016. Robust photoplethysmographic (PPG) based biometric authentication for wireless body area networks and m-health applications. In *2016 Twenty Second National Conference on Communication (NCC)*. IEEE, 1–6.

[22] Ruggero Donida Labati, Vincenzo Piuri, Francesco Rundo, Fabio Scotti, and Concetto Spampinato. 2021. Biometric recognition of PPG cardiac signals using transformed spectrogram images. In *International Conference on Pattern Recognition*. Springer, 244–257.

[23] Simon Eberz, Kasper B. Rasmussen, Vincent Lenders, and Ivan Martinovic. 2017. Evaluating Behavioral Biometrics for Continuous Authentication: Challenges and Metrics. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (Abu Dhabi, United Arab Emirates) *(ASIA CCS '17)*. Association for Computing Machinery, New York, NY, USA, 386–399. https://doi.org/10.1145/3052973.3053032

[24] Luke Everson, Dwaipayan Biswas, Madhuri Panwar, Dimitrios Rodopoulos, Amit Acharyya, Chris H Kim, Chris Van Hoof, Mario Konijnenburg, and Nick Van Helleputte. 2018. BiometricNet: Deep learning based biometric identification using wrist-worn PPG. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.

[25] Paul Faragó, Robert Groza, Liliana Ivanciu, and Sorin Hintea. 2019. A correlation-based biometric identification technique for ECG, PPG and EMG. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 716–719.

[26] Nir Friedman, Dan Geiger, and Moises Goldszmidt. 1997. Bayesian network classifiers. *Machine learning* 29, 2 (1997), 131–163.

[27] Mohammad Golparvar, Hossein Naddafnia, and Mahmood Saghaei. 2002. Evaluating the relationship between arterial blood pressure changes and indices of pulse oximetric plethysmography. *Anesthesia & Analgesia* 95, 6 (2002), 1686–1690.

[28] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep learning*. Vol. 1. MIT press Cambridge.

[29] John N Gowdy and Zekeriya Tufekci. 2000. Mel-scaled discrete wavelet coefficients for speech recognition. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, Vol. 3. IEEE, 1351–1354.

[30] YY Gu and YT Zhang. 2003. Photoplethysmographic authentication through fuzzy logic. In *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering, 2003.* IEEE, 136–137.

[31] YY Gu, Y Zhang, and YT Zhang. 2003. 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.* IEEE, 13–14.

[32] Shi-Jinn Horng, Xuan-Zi Hu, Bin Li, and Naixue Xiong. 2018. Personal Identification via Heartbeat Signal. In *2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP).* IEEE, 152–156.

[33] Dae Yon Hwang and Dimitrios Hatzinakos. 2019. PPG-based personalized verification system. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE).* IEEE, 1–4.

[34] Dae Yon Hwang, Bilal Taha, and Dimitrios Hatzinakos. 2021. PBGAN: Learning PPG representations from GAN for time-stable and unique verification system. *IEEE Transactions on Information Forensics and Security* 16 (2021), 5124–5137.

[35] Dae Yon Hwang, Bilal Taha, Da Saem Lee, and Dimitrios Hatzinakos. 2020. Evaluation of the Time Stability and Uniqueness in PPG-Based Biometric System. *IEEE Transactions on Information Forensics and Security* 16 (2020), 116–130.

[36] Tonislav Ivanov, Ayush Kumar, Denis Sharoukhov, Francis Ortega, and Matthew Putman. 2020. DeepDenoise: a deep learning model for noise reduction in low SNR imaging conditions. In *Applications of Machine Learning 2020*, Michael E. Zelinski, Tarek M. Taha, Jonathan Howe, Abdul A. S. Awwal, and Khan M. Iftekharuddin (Eds.), Vol. 11511. International Society for Optics and Photonics, SPIE, 20 – 28. https://doi.org/10.1117/12.2568986

[37] Nur Azua Liyana Jaafar, Khairul Azami Sidek, and Siti Nurfarah Ain Mohd Azam. 2015. Acceleration plethysmogram based biometric identification. In *2015 International Conference on BioSignal Analysis, Processing and Systems (ICBAPS).* IEEE, 16–21.

[38] Arne Jensen and Anders la Cour-Harbo. 2001. *Ripples in mathematics: the discrete wavelet transform.* Springer Science & Business Media.

[39] Vasu Jindal, Javad Birjandtalab, M Baran Pouyan, and Mehrdad Nourani. 2016. 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, 6401–6404.

[40] Nima Karimian, Zimu Guo, Mark Tehranipoor, and Domenic Forte. 2017. Human recognition from photoplethysmography (ppg) based on non-fiducial features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 4636–4640.

[41] Nima Karimian, Mark Tehranipoor, and Domenic Forte. 2017. Non-fiducial ppg-based authentication for healthcare application. In *2017 IEEE EMBS international conference on biomedical & health informatics (BHI).* IEEE, 429–432.

[42] A Reşit Kavsaoğlu, Kemal Polat, and M Recep Bozkurt. 2014. A novel feature ranking algorithm for biometric recognition with PPG signals. *Computers in biology and medicine* 49 (2014), 1–14.

[43] Muhammad Umar Khan, Sumair Aziz, Syed Zohaib Hassan Naqvi, Ahmed Zaib, and Aiman Maqsood. 2020. Pattern Analysis Towards Human Verification using Photoplethysmograph Signals. In *2020 International Conference on Emerging Trends in Smart Technologies (ICETST).* IEEE, 1–6.

[44] Ruggero Donida Labati, Vincenzo Piuri, Francesco Rundo, and Fabio Scotti. 2022. Photoplethysmographic biometrics: A comprehensive survey. *Pattern Recognition Letters* (2022).

[45] Anthony Lee and Younghyun Kim. 2015. Photoplethysmography as a form of biometric authentication. In *2015 IEEE SENSORS.* IEEE, 1–2.

[46] Eugene Lee, Annie Ho, Yi-Ting Wang, Cheng-Han Huang, and Chen-Yi Lee. 2020. Cross-Domain Adaptation for Biometric Identification Using Photoplethysmogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 1289–1293.

[47] Sun-Woo Lee, Duk-Kyun Woo, Yong-Ki Son, and Pyeong-Soo Mah. 2019. Wearable Bio-Signal (PPG)-Based Personal Authentication Method Using Random Forest and Period Setting Considering the Feature of PPG Signals. *JCP* 14, 4 (2019), 283–294.

[48] K Ming Leung. 2007. Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering* 2007 (2007), 123–156.

[49] Andy Liaw, Matthew Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[50] Chunying Liu, Jijiang Yu, Yuwen Huang, and Fuxian Huang. 2022. Time–frequency fusion learning for photoplethysmography biometric recognition. *IET Biometrics* 11, 3 (2022), 187–198.

[51] Jordi Luque, Guillem Cortes, Carlos Segura, Alexandre Maravilla, Javier Esteban, and Joan Fabregat. 2018. End-to-end photopleth Ysmography (PPG) based biometric authentication by using convolutional neural networks. In *2018 26th European Signal Processing Conference (EUSIPCO).* IEEE, 538–542.

[52] Soumik Mondal and Patrick Bours. 2017. A study on continuous authentication using a combination of keystroke and mouse biometrics. *Neurocomputing* 230 (2017), 1–22. https://doi.org/10.1016/j.neucom.2016.11.031

[53] Henri J Nussbaumer. 1981. The fast Fourier transform. In *Fast Fourier Transform and Convolution Algorithms.* Springer, 80–111.

[54] University of Toronto. 2011. The Biosec1 Dataset. https://www.comm.utoronto.ca/~biometrics/PPG_Dataset/.

[55] Jiapu Pan and Willis J Tompkins. 1985. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering* 3 (1985), 230–236.

[56] Dung Phan, Lee Yee Siong, Pubudu N Pathirana, and Aruna Seneviratne. 2015. Smartwatch: Performance evaluation for long-term heart rate monitoring. In *2015 International symposium on bioelectronics and bioinformatics (ISBB).* IEEE, 144–147.

[57] João Ribeiro Pinto, Jaime S Cardoso, and André Lourenço. 2018. Evolution, current challenges, and future possibilities in ECG biometrics. *IEEE Access* 6 (2018), 34746–34776.

[58] Anuja Priyam, GR Abhijeeta, Anju Rathee, and Saurabh Srivastava. 2013. Comparative analysis of decision tree classification algorithms. *International Journal of current engineering and technology* 3, 2 (2013), 334–337.

[59] Limeng Pu, Pedro J Chacon, Hsiao-Chun Wu, and Jin-Woo Choi. 2022. Novel Robust Photoplethysmogram-Based Authentication. *IEEE Sensors Journal* 22, 5 (2022), 4675–4686.

[60] Tim Ring. 2015. Spoofing: are the hackers beating biometrics? *Biometric Technology Today* 2015, 7 (2015), 5–9.

[61] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[62] NS Girish Rao Salanke, N Maheswari, and Andrews Samraj. 2013. An enhanced intrinsic biometric in identifying people by photopleythsmography signal. In *Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012).* Springer, 291–299.

[63] NS Girish Rao Salanke, N Maheswari, Andrews Samraj, and S Sadhasivam. 2013. Enhancement in the design of biometric identification system based on photoplethysmography data. In *2013 International Conference on Green High Performance Computing (ICGHPC).* IEEE, 1–6.

[64] Jorge Sancho, Álvaro Alesanco, and José García. 2018. Biometric authentication using the PPG: A long-term feasibility study. *Sensors* 18, 5 (2018), 1525.

[65] Abhijit Sarkar, A Lynn Abbott, and Zachary Doerzaph. 2016. Biometric authentication using photoplethysmography signals. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 1–7.

[66] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1997. Kernel principal component analysis. In *International conference on artificial neural networks*. Springer, 583–588.

[67] Muhammad Shahzad and Munindar P Singh. 2017. Continuous authentication and authorization for the internet of things. *IEEE Internet Computing* 21, 2 (2017), 86–90.

[68] Claude Elwood Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE* 37, 1 (1949), 10–21.

[69] Ali I Siam, Atef Abou Elazm, Nirmeen A El-Bahnasawy, Ghada M El Banby, Abd El-Samie, and E Fathi. 2021. PPG-based human identification using Mel-frequency cepstral coefficients and neural networks. *Multimedia Tools and Applications* 80, 17 (2021), 26001–26019.

[70] Khairul Azami Sidek, Nur Khaleda Naili Kamaruddin, and Ahmad Fadzil Ismail. 2018. The study of ppg and apg signals for biometric recognition. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10, 1-6 (2018), 17–20.

[71] Khairul Azami Sidek, Munieroh Osman, SNA Mohd Azam, and Nur Izzati Zainal. 2016. Development of an Acceleration Plethysmogram based Cardioid Graph Biometric Identification. *International Journal of Bio-Science and Bio-Technology* 8, 3 (2016), 9–20.

[72] Santiago Silva, Boris A. Gutman, Eduardo Romero, Paul M. Thompson, Andre Altmann, and Marco Lorenzi. 2019. Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 270–274. https://doi.org/10.1109/ISBI.2019.8759317

[73] Petros Spachos, Jiexin Gao, and Dimitrios Hatzinakos. 2011. Feasibility study of photoplethysmographic signals for biometric identification. In *2011 17th International Conference on Digital Signal Processing (DSP)*. IEEE, 1–5.

[74] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

[75] Issa Traore. 2011. *Continuous Authentication Using Biometrics: Data, Models, and Metrics: Data, Models, and Metrics*. Igi Global.

[76] Junia Valente, Matthew A Wynn, and Alvaro A Cardenas. 2019. Stealing, spying, and abusing: Consequences of attacks on internet of things devices. *IEEE Security & Privacy* 17, 5 (2019), 10–21.

[77] Daomiao Wang, Qihan Hu, and Cuiwei Yang. 2022. Biometric recognition based on scalable end-to-end convolutional neural network using photoplethysmography: A comparative study. *Computers in Biology and Medicine* (2022), 105654.

[78] Claes Wohlin. 2014. Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. http://doi.acm.org/10.1145/2601248.2601268. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (London, England, United Kingdom) *(EASE '14)*. ACM, New York, NY, USA, Article 38, 10 pages. https://doi.org/10.1145/2601248.2601268

[79] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[80] Jian Xiao, Fang Hu, Qiang Shao, and Sizhuo Li. 2019. A Low-Complexity Compressed Sensing Reconstruction Method for Heart Signal Biometric Recognition. *Sensors* 19, 23 (2019), 5330.

[81] Umang Yadav, Sherif N Abbas, and Dimitrios Hatzinakos. 2018. Evaluation of PPG biometrics for authentication in different states. In *2018 International Conference on Biometrics (ICB)*. IEEE, 277–282.

[82] Junfeng Yang, Yuwen Huang, Fuxian Huang, and Gongping Yang. 2020. Photoplethysmography Biometric Recognition Model Based on Sparse Softmax Vector and k-Nearest Neighbor. *Journal of Electrical and Computer Engineering* 2020 (2020).

[83] Junfeng Yang, Yuwen Huang, Ruili Zhang, Fuxian Huang, Qinggang Meng, and Shixin Feng. 2021. Study on ppg biometric recognition based on multifeature extraction and naive bayes classifier. *Scientific Programming* 2021 (2021).

[84] Jianchu Yao, Xiaodong Sun, and Yongbo Wan. 2007. 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, 4576–4579.

[85] Yalan Ye, Guocheng Xiong, Zhengyi Wan, Tongjie Pan, and Ziwei Huang. 2021. PPG-based biometric identification: Discovering and identifying a new user. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1145–1148.

[86] Alexander Zien, Gunnar Rätsch, Sebastian Mika, Bernhard Schölkopf, Thomas Lengauer, and K-R Müller. 2000. Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics* 16, 9 (2000), 799–807.