The history of biometric authentication and identification using photoplethysmography (PPG): a twenty years systematic literature review

Benjamin Vignau, Patrice Clemente, Pascal Berthomé

The history of biometric authentication and identification 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 persons 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 which gather all the papers talking about biometric authentication with PPG published between 2003 and late 2022.

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

ACM Reference Format:
Benjamin Vignau, Patrice Clemente, and Pascal Berthomé. 2023. The history of biometric authentication and identification using photoplethysmography (PPG): a twenty years systematic literature review. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 45 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

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 intimates devices, etc. [77] 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. Today passport also uses fingerprint authentication. For example, France started to deliver them in 2006 [1] and Canada in 2013 [20]. 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 [76]. The main problem with static authentication is the impossibility to remedy a hijacked session. Continuous Authentication aims to re-authenticate the user multiples 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) [24]. Recently the usage of IoT to enforce continuous authentication is studied [68]. 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 smarts wearable 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 theses signals, first measured for medical purpose, to recognize people and develop authentication systems more ergonomic and robust.

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, induce by heartbeat [28]. 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.

Heartbeats signal is a biological trait, it can be easily measured, and, such as voice, iris, or fingerprint it is 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 it. To measure ECG multiples electrodes need to be stuck on persons, 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 those works. 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 questions, 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 summary year by year all the collected papers, describing their methodology and their results.
- Section 5 explained 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 comparison of multiples studies that used the same dataset.
- 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 it's cost (few dollars for a PPG sensor), it's difficulty to counterfeit and the possibility to add this sensor inside wearable devices (watches, T-shirt etc). This lead to the ability to provide a new ergonomic, simple and non invasive form of continuous authentication. Finally this technology also provide medical data that can be exploited to provide a medical monitoring to users.

The technology description, it's 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 their are the only one to provide a survey on this problem. But, they study 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 provide full dataset of all the experiences realised for this topic.

2.2 Problem definition

First we need to define the differences between authentication and identification. The authentication is the action to prove the identity of someone. The user give the claimed identity with a proof and the system only check the proof. In an authentication system with PPG a user could claim an identity and give it's 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 have 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. Theses methods are a kind of 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 have 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 it's PPG signal.

From this definition, we understand that the process of identification and authentication needs an enrollment phase where the user give a first proof of it's identity. Then, there is verification phase where the system have to check if the second proof given by the user match 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 thief and impostors. Thus we need to have a good attention on the False positive match (where user A match the identity of User B) and False negative match (Where user A present a valid proof but is not recognized by the system). The False positive matches are a big problem in term of security because it let 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 researches questions splitted in three mains axis : the robustness of the system, the ergonomic of the system and the 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 user ?
- 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 theses question, 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 element to answer our question. 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 the paper to compare their works, 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

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 mains phases of this common methodology to structure our paper.

3.1 Papers collection

To make a good systematic literature review, we followed the guidelines provided by Wohlin [79]. 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 firsts 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 duplicates papers resulting in 136 different papers. For each of these papers, we analyzed, the title, the abstract, and if needed introduction and conclusion. We excluded all the papers which matched at least one of the criteria defined in Table 3.

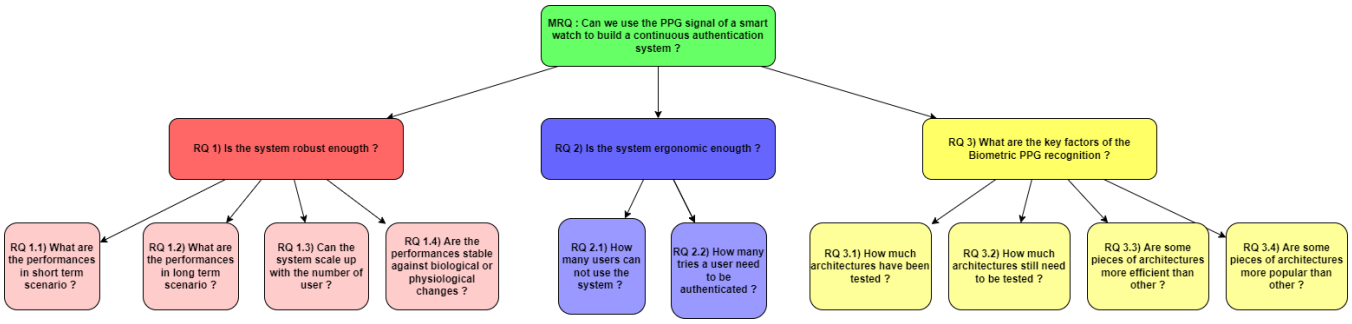


Figure 1. Research questions of this paper

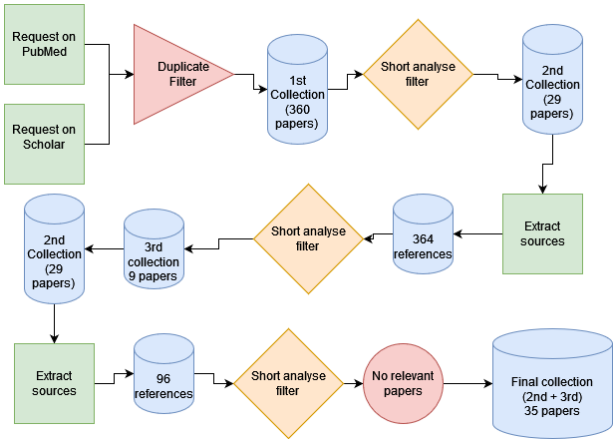


Figure 2. Methods for paper collection in 2021

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 paper in those 96 references was new or relevant. It results in no new reference 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 is used to authenticate patient change over the years. In the oldest papers, simple metric calculus was used to classify the subjects, like 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 criteria and its 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

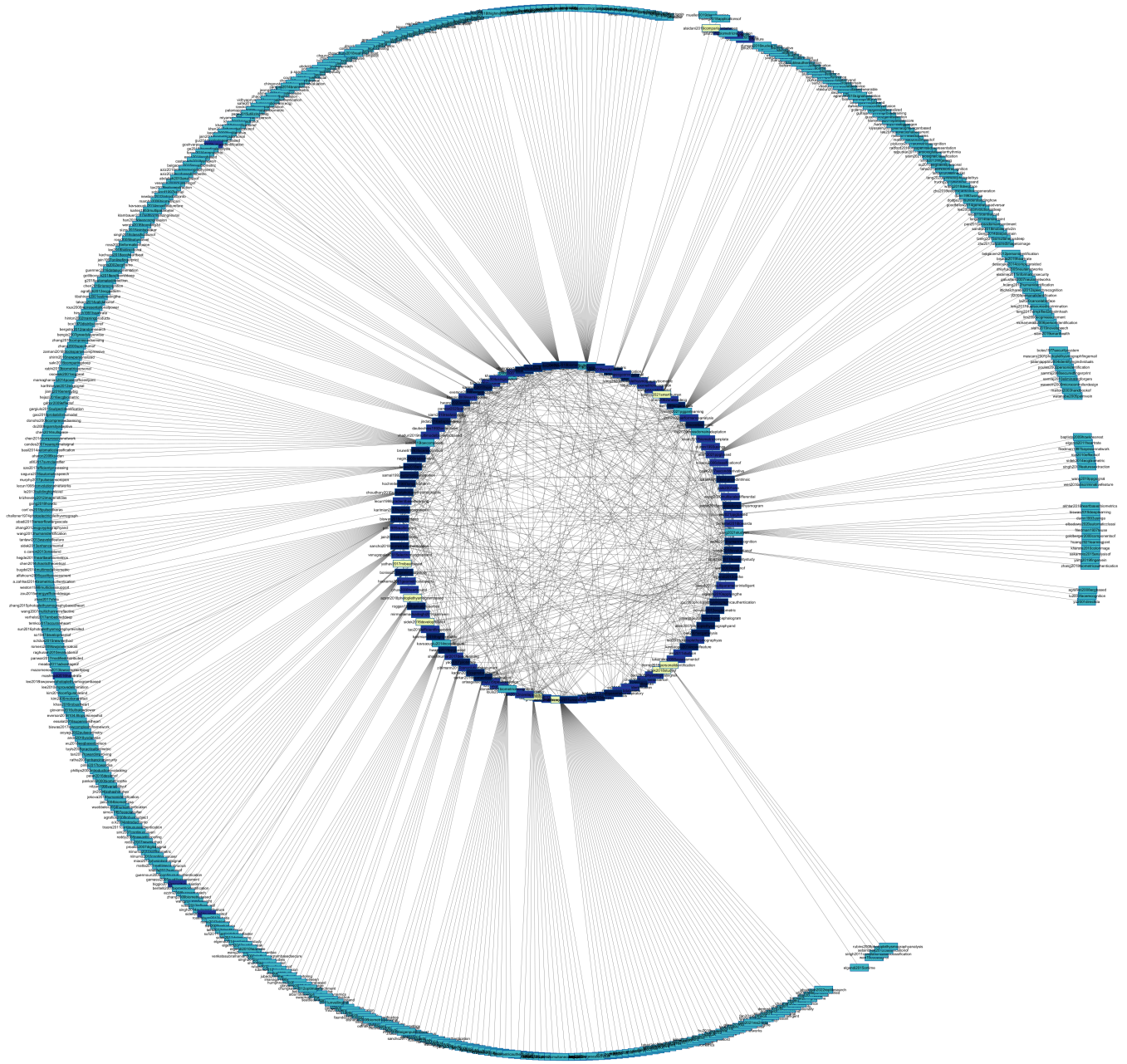
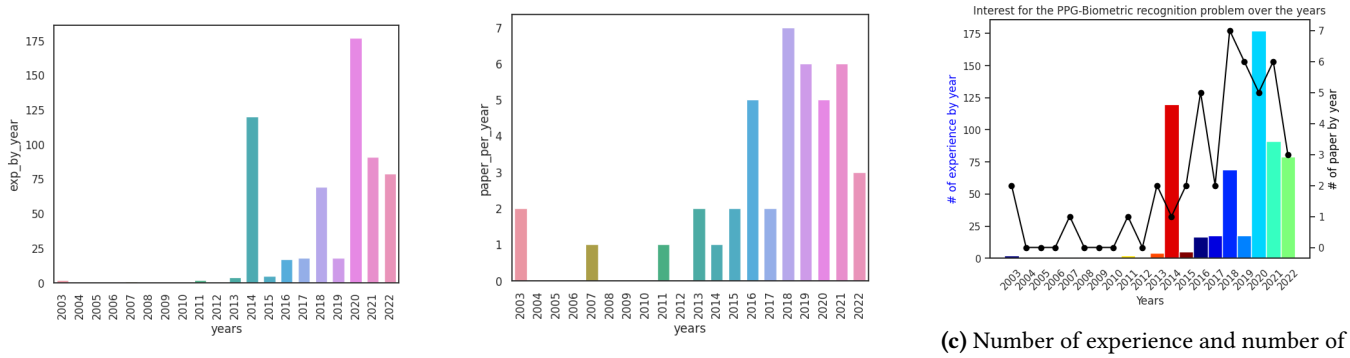


Figure 3. Representation of all the papers collected and their references

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 differences dataset (ex two publicly available database of PPG), we counted two experiences. Each time where the value of one criteria change, 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

The Figure 4 represents generals statistics of the number of paper and experiences done from 2003 to 2022. We can observe two main periods : before 2014 and after 2014. Figure 4a represents the distribution of the experience over the years. We can see very few experiences done before 2014. Less than 5 experiences by years where made during this period. Figure 4b represent the number of publication per year. Here again we can observe two main periods : before



(a) Number of experience by year

(b) Number of published paper per year

(c) Number of experience and number of published paper per year

Figure 4. Three representation of the evolution of the research interest for the PPG biometric recognition from 2003 to 2022

2013 and after. Before 2013 only three paper 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). Figure 4c combine the two previous figures to provide a better understanding. Here we can see that the number of publication 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 represent 19.9% of the total experiences conducted from 2003 to 2022.

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 the 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 work, they tested only true identity, they did not test impostors. These first work were good enough to start the researches on this topic. However 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 paper were published during 4 years, until the one made by Yao et al. in 2007 [86]. In this work, the authors extracted fiducial features from filtered PPG, and its two derivatives. They collect 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 conclude to the possibility to identify people with the PPG. However no identification or authentication metrics are provided.

4.1.3 2011. In 2011, Spachos et al. [74] published a study made on 29 subject taken in two public dataset : 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 pusle 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 achieve 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 [18, 63, 64], gathering 4 experiences. Salanke et al. published two papers [63, 64]. In the first one they splitted the signal in single pusle 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 parameter 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 test two feature selection methods. In the first one they only took the first 5 coefficients after SDD for each subjects. In the second one they took the q first coefficients where q change 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. At the opposite, Bossini et al [18] made a study on 44 subjects where 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 pulse and computed the correlation between the template and input matrix. They tested 6 different ways to fusion the data and obtain a matching score but only present the result for one : The maximum value of the correlation. Finally they provided multiples metrics on this system. They tested identification with genuine and impostors. They achieve a 5.29% EER wich is quite good.

In conclusion, between 2003 and 2013 few papers were published with few experimentation. Most of the papers provided simple studies, with poor metrics and poor methodology. Most of the dataset are not publicly available and most of the study 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, the research about PPG-biometric recognition increased a lot.

4.2.1 2014. In 2014 Kavsaoglu et al. [42] provided 120 different experimentation 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 less 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 subject. They tested multiples 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 time 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 better parameter combination for the KNN algorithm. They results show that the ranking process increased significantly the accuracy. However the optimal number of extracted feature change from one dataset to another. They achieve good accuracy, over 90%.

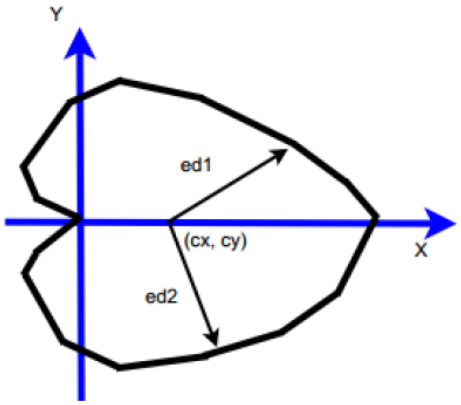
4.2.2 2015. In 2015 two papers were published, gathering 5 experiences [37, 45]. In the first one, the authors studied the impact of using only the APG (second derivative of the PPG) to authenticate people. To do so, they used the MIMIC dataset, split signal in single pulse and derive the signal two times. They extracted 5 fiducial points in the APG and used it in a classifier. To class the people, they used the Naives Bayes and KNN. They used the 10 cross fold validation methods to avoid over fitting 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 Naives 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 use a public dataset, their architecture must now be tested on bigger dataset.

The second one extracted 22 physical features from single pulse (different length and angle in the signal) and use a CNN to class 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.

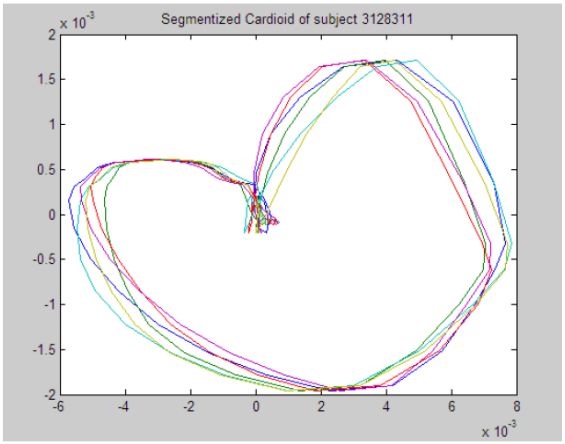
4.2.3 2016. In 2016 17 experiences where made for 5 published papers [21, 22, 39, 66, 72].

Sidek et al. [72] 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 frequency 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. These representation are show in Figure 5.

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



(a) Cardioid representation from [72]



(b) Cardioid representation of single pulse from one patient. From [72]

Figure 5. Two cardioid representation from [72]

Choudhary et al [22] used a public dataset MIT-BIH Polysomnographic Database and build a simple architecture. They splitted 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 good in comparison with the 5% of EER achieved by older experiences.

Chakraborty et al [21] made one experience, using a custom dataset composed of 3min signal 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 dataset is very 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 class the subjects. They achieved 96.1% of accuracy using a 10 cross fold validation method. The results are good and this architecture should be selected for experiences with higher number of people.

Sarkar et al [66] 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 kind 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 use the Gaussian Decomposition to generate features. They tested to extract 3 and 5 features. Then 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 where 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 features type and 3 algorithm. 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 experience 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 uses the same technique to select the best 10 features. Finlay they tested SVM, SOM and KNN algorithm. They achieve good accuracy scores with 99,84% in the maximum, using KNN and the DWT features. From their results, the difference of performances between the algorithm is very

low and may not be significant. However the usage of DWT features seems to really improve the performances compared to fiducial features.

In their second paper, the authors made 12 experiences, using the same dataset and same filtering method. However, in this paper, they added a Zero mean normalization of the signal before filtering. They tested 3 different kinds of features : Fiducial, Wavelet decomposition domain (DWT) and fiducial. 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 hited 100% accuracy. However, no details were provided about training and testing set, no validation methods was 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, 25, 32, 51, 65, 71, 82].

Sidek et al [71] 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 to extract 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 case, 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 multiples 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 the same for all algorithm, except for Random Forest and NB which were less efficient. All the others achieved 94% accuracy which is in the norm compared to others experiences.

Yadav et al [82] 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 much features they 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 good. 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 [25] made one experience on the TROIKA dataset. They did not provide any details about the pre processing methods. They simply build a CNN called BiometricNet and feeded it with the raw signal. They said to achieve 96% accuracy score but they did not provide any details on training, testing and validation methods.

Sancho et al [65] made 49 experiences.In this study, the authors gather 4 publicly available dataset 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 multiples 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, where EER went over 20% while in short term it stayed around 10%.

Most of the architecture provide similar results. We can say that the augmentation of cycles for the training improve the performances. However between the algorithms, the standard deviation are overlapping thus, we can not conclude to that one outperform another.

Luque et al [51] made one experience where they use 1s 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 [5, 16, 26, 33, 47, 81]

Xiao et al [81] made one experience on a custom dataset. They segmented the signal in single cycle, the a 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 achieve 91.31% of accuracy which is good and match other papers levels.

Lee et al [47] made 12 experiences using the Capnobase IEE TBME dataset. They tested multiples 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 it's two derivative and used a 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 very high compare to all other study and the lacking of details in the paper did not allow us to conclude to a good architecture.

Farago et al [26] 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 splitted the signal in 1 000 points sample, 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 dataset to train the algorithm and the rest to test it. They used the 10 cross fold validation and achieve 99.8% of accuracy which is very good. It was the first time that natural language algorithm 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 again the 10 cross fold validation. They achieved 99.8% showing the robustness of their architecture for long time stability.

Biswas et al [16] use the TROIKA dataset to made 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 extract the best features. Then the output of this CNN fed 2 LSTM which provide the classification. They achieve 96% of accuracy.

4.2.7 2020. In 2020, 177 experiences were made, across 5 papers [9, 35, 43, 46, 84].

Yang et al [84] made 80 experiences using 3 publicly available dataset : BIDMC, MIMIC II and Capnobase. They segmented the signal using a Sliding windows. 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 combination 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 good accuracy score, with most of the scores where 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.

The usage of the combination of multiples extractor seems to improve the performances for all the algorithm, expect for the KNN algorithm. It's the better algorithm from theses experiences, it's 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 over 5 on the BIDMC dataset and two times over 5 on the Capnobase dataset. The experiences provided by this paper were good but must be reproduced with a good 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 Naives Bayes. Their results showed a better performance from the quadratic SVM with 93.1% of accuracy. This was good and the result should be reproduced using a public dataset.

Lee et al [46] made 4 experiences using two dataset : 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 their is no difference in accuracy between the two tested model, 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 splitted in different frame lengths (1 , 3, 5, 7, 10 or 15s). To extract the desired feature, they created multiples vectors of statistical features extracted from the raw PPG signal, its first derivative and on the signal filtered with DWT. Then they made multiples experiences with the usage of one or multiples 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 multiples classifiers like KNN, SVM, RF etc. with only 15s vectors and the vector from DWT decomposition. They achieved 99.3% of accuracy with a 0.02% of EER which is quite good. 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, spaced 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 where the heart rate was too low or too high. They tested multiple features extraction : 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 over-fitting problem, they used 10 fold cross validation with L2-regularization. All the details for each layer were provided, which is much appreciable. About the experiences concerning the selection methods, the differences in average accuracy are very low and the standard deviation should have been computed. Now we can not conclude on the efficiency about the selection method. They also made experiences to compare the architecture and the selection method, depending on the database. With the two channels data (DTW + IN) the results are better compared to the one channel usage. Moreover in both case the selection feature method 2 provided better results. However in the two channel scenario it's the CNN architecture which performed best while in one channel experience it was the CNN + LSTM architecture which dominates. 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 author experiment 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, 23, 34, 70, 85, 87].

Donida et al [23] made two experiences using the Capnobase dataset. In this paper, the authors tried to identify users using a template matching methods 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 K-NN or an package of SVM. They hit 99.16% of accuracy. Remark: they use 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 use 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 multiples single filtered cycles. They assumed that this step returned a matrix, where each column meant a sample of the wave, and each line meant represents a wave. Then 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 was not good. 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 a good accuracy but, we can see 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 positive. This can not be used for real usage.

Yang et al [85] produced 42 experiences over the BIDMC, MIMIC II and Capnobase datasets. In this paper the authors studied the PPG biometric recognition based on multiples 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, etc. 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 multiples models : Linear discriminant classifier and Naives 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 subject at each test and use 80% of the available signals of each subject for training. They obtains interesting results, showing how the adding of frequency and wavelet domains features improves accuracy. However theses feature only were not good 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. In all case, the NBC show better performances against the LDC. Next they tested the Euclidean distance as classifier and obtains similar results for all kind of feature vectors and database (between 96% and 98% of accuracy). This paper is interesting and showed the effect of the different types of extracted features on the classification.

Ye et al [87] 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 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 [70] 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. 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 re train 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 multiples 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 good idea but if it worked, this could be used to attack the system by generating synthetic signal of the user. They used multiples databases such as TROIKA, PRRB and Biosec 1 and 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.
- In this paper, two scenarios are tested, one with single session and one with two-sessions. In the two session scenario, the signal for registration was taken on session 1 and the signal for testing was taken in session 2. In both cases a L2 regularization and 10-fold validation was used.
- 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.

Once the tests were done in one session scenario, they tried to improve the performances of their model in a two session scenario with GAN. To do that, they tested 3 GAN.. The GAN are only used for improving two sessions scenarios. They try GAN DCGAN , WGAN eand 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 version of PGGAN. We can see that the best algorithm depended on the database. In two session case, the PGBAN-DC out performed the other on Biosec3 while it's PGBAN-LS that out-performed 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, 78].

Pu et al. [59] made 2 experiences. They fusioned the Capnobase and Biosec 1 dataset 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). Finally they extracted a template and normalized it by meaning all pulse. 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 [78] 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 is 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 dataset, the ROCKET extraction feature gives the best result, but for the VITAL DB it's 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. They showed that the performances of all their algorithms can decrease a lot (it could be halved for some).

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 matrix U.V, one based on the features and the other on the sample. The V matrix should have represented 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 multiples figures are available on our Github repository : https://github.com/bvignau/PPG_SLR_dataset.

In Figure 6 we have represented the most common fully pipeline (from signal segmentation to classification). To draw our graph, we computed the graph representation of the experiences. In our representation, each step (eg signal segmentation in single cycle) is a node, weighted by the total usage in the experience dataset. Each connection is also weighted in the same way eg : if the transition 'Single Cycle Segmentation' to 'No Normalization' is observe two times in the experiences dataset, then the edges is weighted to 2. In Figure 6 we only represent the nodes with weight at 6 and more, and with a fully pipeline (from signal segmentation to classification).

This representation show 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 answers the major question: how the data are 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 dataset and did not share them, but they gave the parameters of the dataset. We will present in dedicated sub section, the data set 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 signal's acquisition. Finally we define a new metric to measure the contribution of each dataset to the community.

5.1 Data set usage

In this section we draw the evolution of the usage with Figure 8 which represents the number of experience that use each data set, and Figure 9 which group the number of experience by years and show how each dataset is used over the years. We have 20 different datasets categories. We have one category to represent each publicly available dataset, plus one which gather all the custom dataset. A custom

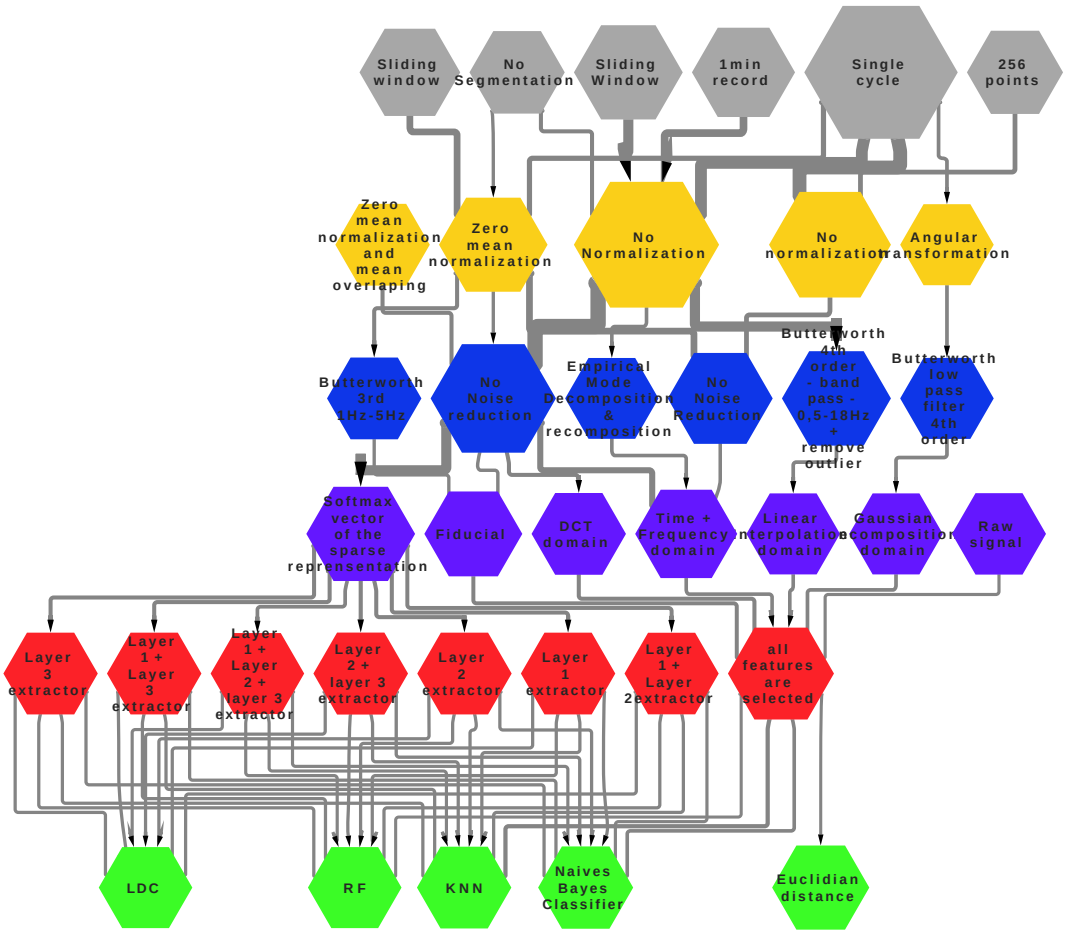


Figure 6. The representation of the most popular models pipelines

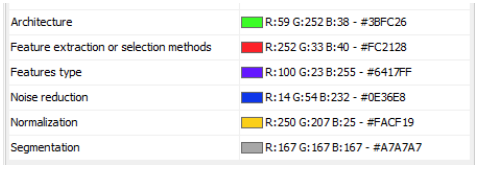


Figure 7. Color Legend for our graph

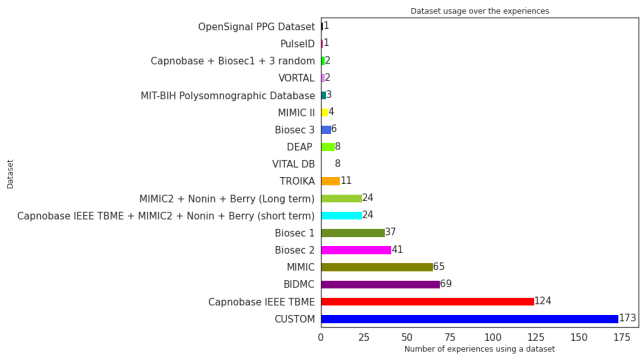


Figure 8. Data set usage over the experiences

dataset represent a dataset gathered and not published by the authors of one study. Thus, all the experiences made with a custom dataset can not be reproduced. The details of each custom dataset is given in Table 4. The characteristics of the public datasets are given in Table 5. The acquisition time is written as "time x number". Here the first part is the time of one acquisition session where researchers collect the PPG signals. The second part corresponds to the number of different sessions. For some dataset their is a third number, representing the number of recorded channels. The PPG signal can be measured using green light, red light and infrared light. Some sensors provide the three channels and some research teams recorded more than one channel. The time interval correspond to the duration between two recorded sessions.

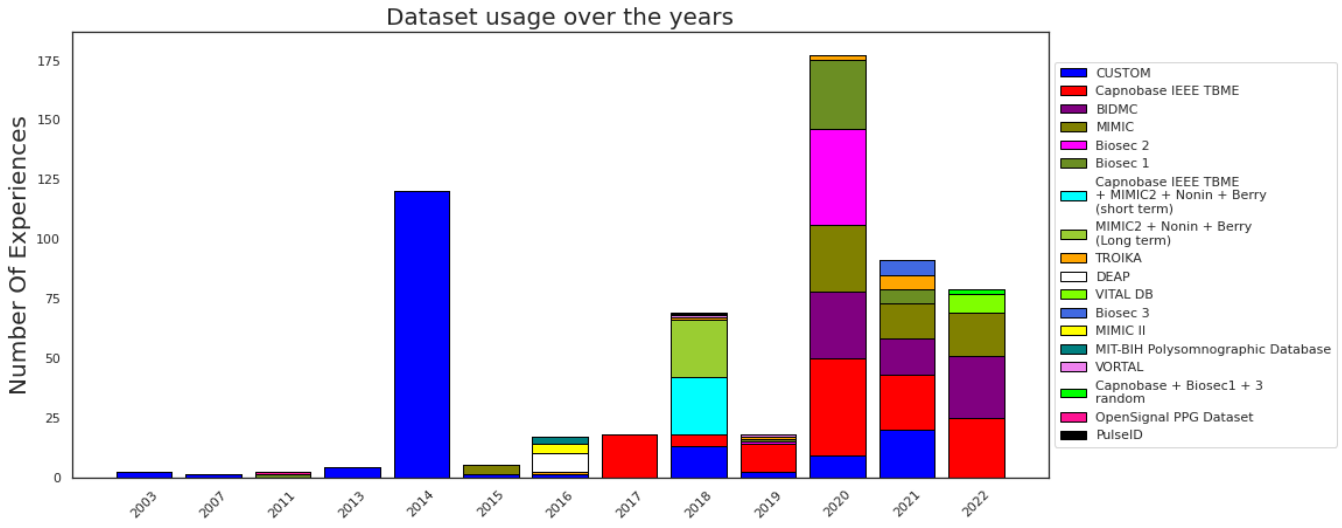


Figure 9. Data set usage over the years

The first interesting thing to point in Figure **??** is the massive usage of custom dataset usage over the total experiences. 28% of the experiments are made with a custom dataset. All of this experiences can not be reproduced. Then multiples publicly available dataset are used. However each experiences use a different subset of these dataset. In an ideal world, each experience must be done with all the available datasets. Also in future works, each paper should produce the same number of experience with each dataset. Some works started to do that, mainly using the Capnobase, BIDMC and MIMIC II dataset, like [14, 85].

With Figure 9, we can observe that the custom dataset where mainly used in 2014 and before. few usages persist across the experiences after but they are very low compare to all 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 dataset, making the experiences robust and easier to reproduce.

5.2 Sampling frequency

In this section we study the evolution of the sampling frequency over the experiences. The sampling frequency is very important because it plays a major role in the signal quality. The Shanon's sampling theorem said that the sampling frequency should be at least twice time higher than the highest frequency in the signal. Most of the studied papers in this literature review shows that the PPG signal range from 0.5 to 20 Hz. So we need to use at least 40Hz as sampling frequency. Using higher sampling frequency add noise in the signal because it will measure electromagnetic perturbations.

Figure 12 represent the distribution of the sampling frequency used by the datasets and the evolution over the years.

On Figure 12a 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 show a huge disparity in the dataset, which will influence the quality of the dataset and the final performances of each algorithm.

On Figure 12b we can observe that the sampling frequency does not show any particular law. Their is no clearly augmentation or reduction of the sampling frequency and the distribution over the years and datasets seems random.

We plot the heat map of the numerical values of our dataset in Figure 10 and the pair plot of the same variables in Figure 11. In this two figures, we can see that their is no clear relation between the sampling frequency and the performances metrics (Accuracy, EER etc.)

5.3 Acquisition time

The acquisition time is another fundamental parameter of a dataset. We represent it with two or three number (a x b x c). The first number correspond to the time length of the measured signal. Some recorded long signal of multiples minutes and other gather short signal of few heart beats. The second number represent the number of sessions, or number of signal available for each subjects. The final number, present only on three dataset correspond to the number of available channel. In deed, to measure the PPG signal, sensors can use three different light : red, green and infra-red. Most of datasets keeps only one channel but some keep all.

Increasing the acquisition time increase the total available data, which can help to build better algorithm. Having multiple session with a long time interval will help to build more realistic scenario and allow to study the time stability of algorithm.

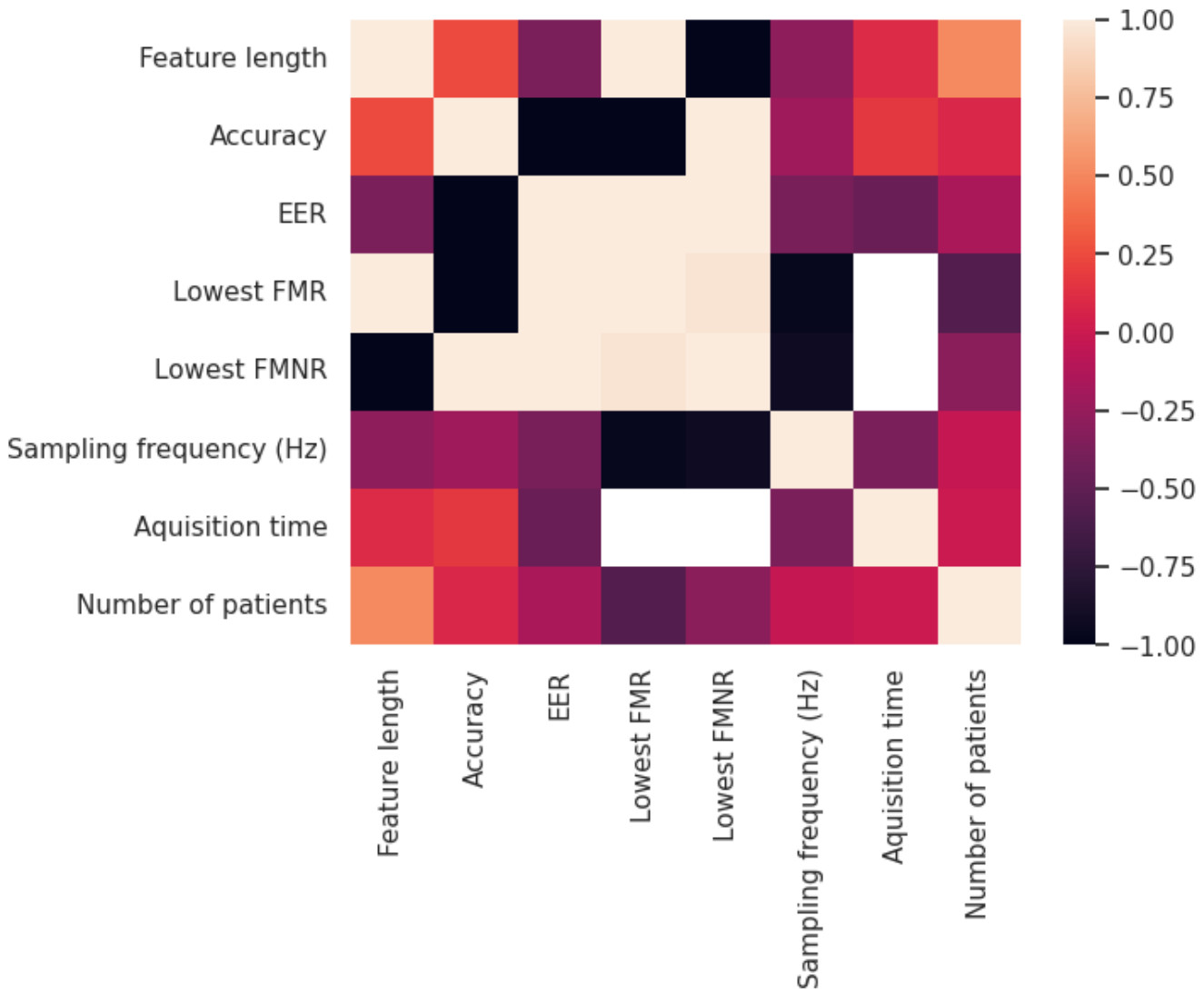


Figure 10. Heat map of the numerical variables of our dataset

In Figure 13 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 dataset does not provide this parameter. For other like the VITAL DB, the acquisition time is too heterogeneous and is different for each patient. For this dataset, the acquisition time range from few minutes to ten hours. On Figure 13a we can see that most of the dataset have it's own acquisition time thus leading to heterogeneous datasets.

This observation is confirmed by Figure 13b where we plot the sampling evolution over the time. We define the acquisition time in second by multiplying all the numbers given in definition for each dataset. Here we can observe that most of the dataset provide less than 500s of signal for each subjects. This show a big difficulty to gather long PPG signal over multiples sessions.

From the Figure 10 and Figure 11 we can not see any big correlation between the acquisition time and the performances metrics. One kind of experience that was never made is to train and tune an algorithm with one dataset and test it with another one. With this kind of experience we will be able to see the influence of each dataset on the algorithms.

5.4 General conditions

The general conditions of the signal acquisition are very important and can influence a lot the performances of an algorithm. Most of the datasets recorded the signal in a controlled environment, where subject where seating with the PPG sensors on the index. Another big parts of the dataset

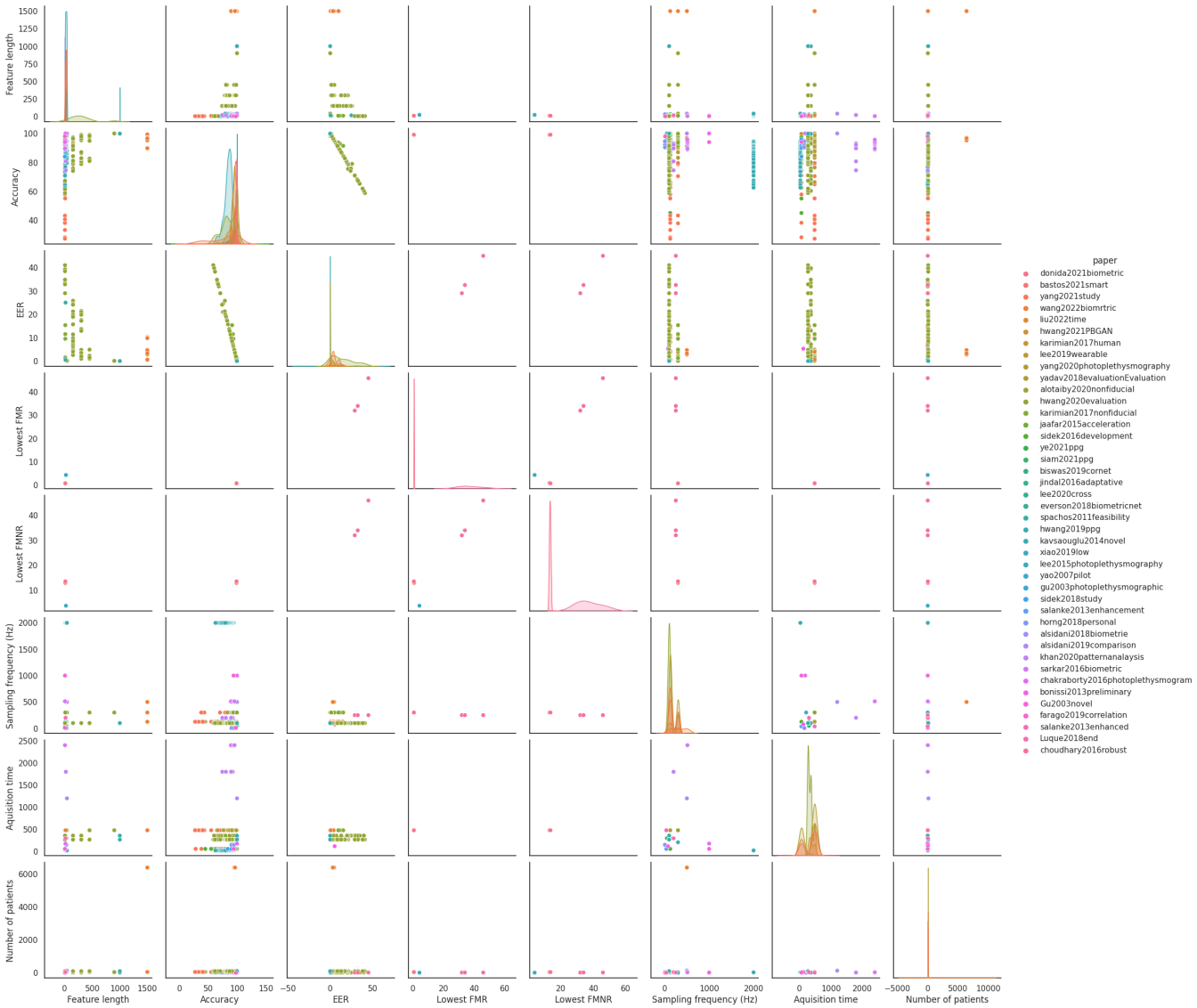


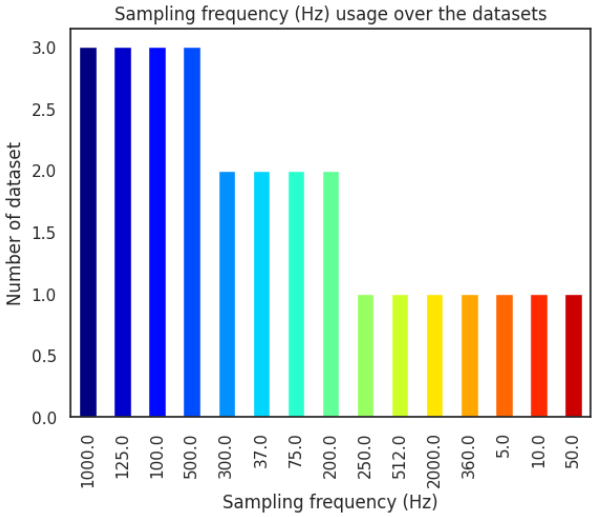
Figure 11. Pair plot of the numerical variables of our dataset

are collected for medical purpose. For example the MIMIC II dataset gather signals on patient in intensive care units. The VITAL DB collect signals on peoples This provide in general good quality signal but theses signals are far from real world signals. One of the major goal of PPG-Biometric recognition is to provide continuous authentication. To do that, a system must collect in real time and over long period, PPG data on subjects. The system must recognize the subject in all conditions : at rest, during sport, in different emotional states etc. So it's important to collect PPG signal in different condition in order to construct a robust system that can be used in real world scenario. The TROIKA dataset provides PPG signal acquired during an test effort on a treadmill which is very good for building a system robust to heart rate variation. The DEAP dataset was collected with patient watching different emotional videos, which is very good to build a system robust to emotional variations.

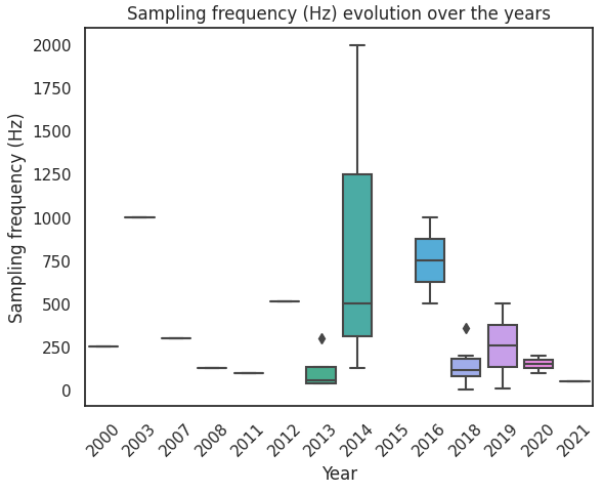
5.5 Datasets contribution to the community

To show how much a dataset contribute to the research topic, we define a new measure that we called 'data consumption'. This metrics is the product of the total acquired time by the number of subject by the number of experiences. In Figure 14, we first plot the tree-map of the total consumed data for each dataset. Figure 14a represent all the dataset contribution while only show the contribution of the public datasets.

We can observe in Figure 14a that the custom dataset provide very low contribution for the research topic. The total contribution of all the custom dataset is lower than the

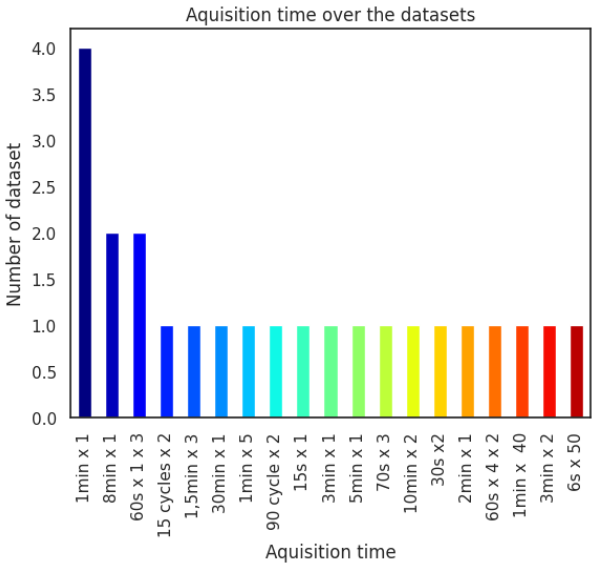


(a) Sampling frequency used by dataset

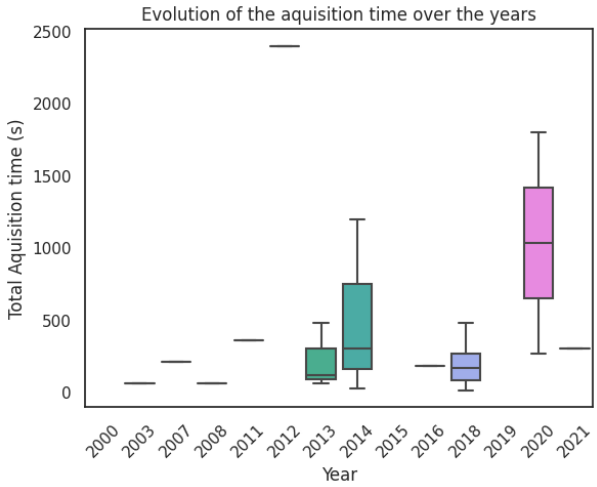


(b) Evolution of the sampling frequency over the years

Figure 12. Sampling frequency distribution and evolution



(a) Acquisition time over the datasets



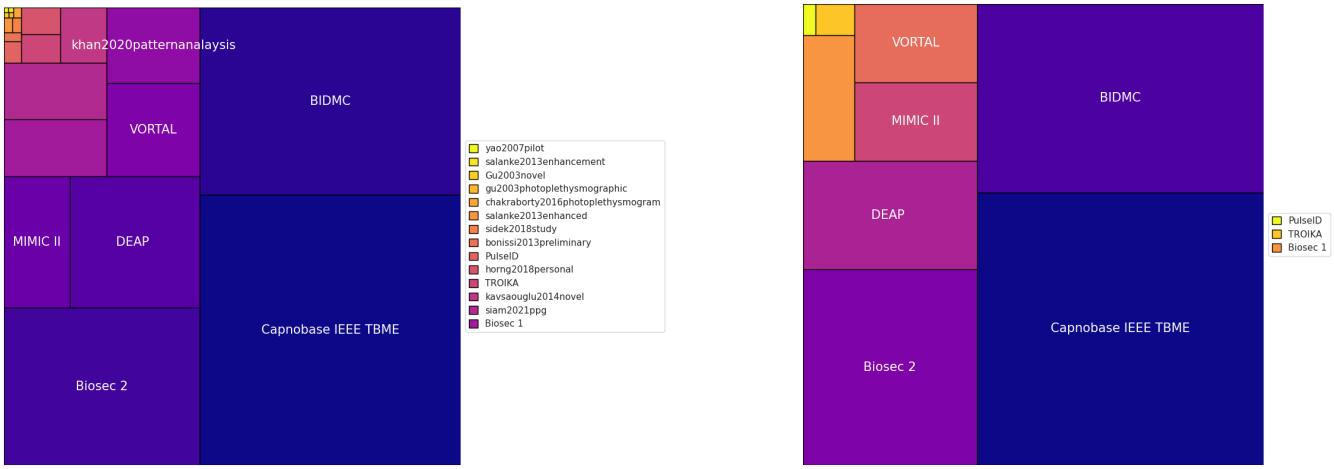
(b) Evolution of the acquisition time over the years

Figure 13. Time acquisition distribution and evolution

contribution of the BIDMC dataset. However some custom dataset contribute more than public dataset. For example the dataset of Kavsaoglu et al [42] or the one made by Khan et al [43] contribute more than the TROIKA data set. This is mainly due to higher number of experience and higher time acquisition.

In Figure 14b we focused on public dataset to show their relative contribution. We can observe that the Capnobase and BIDMC datasets contribute the most, and represent around 2/3 of all the contribution of the public datasets. They are the most used dataset as shown in Figure ??. Moreover the Capnobase provide one of the highest number of patient (42) with a good time record (8min). The BIDMC provide the same time record but with 53 patients.

In Figure 15, representing the evolution of the contribution of the dataset over the years, we can see that very few contribution was made before 2014. We can explained that by the poor number of publication and the poor number of experiences done during this period. Moreover, during this period, researchers mainly used custom datasets with few



(a) Datasets contribution

(b) Public datasets contribution

Figure 14. Data set usage over the experiences

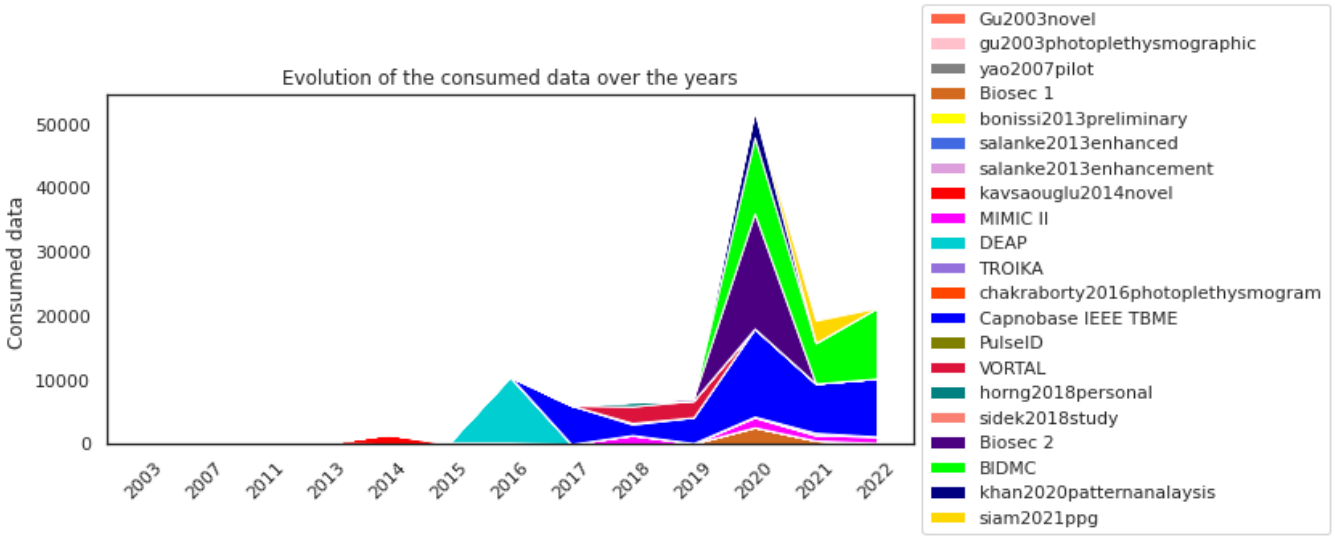


Figure 15. Datasets contribution over the years

number of subjects and little acquisition time. We observe a first good contribution in 2014 with the work of Kavsaouglu et al [42] and the 120 experiences it provides. The contribution of other custom dataset are very little.

Figure 16 is a focus on the contribution of public dataset. We can observe that the Capnobase dataset is the only one to contribute every year from 2016 to 2022 and is always one of the most contributing dataset each years. This show a high popularity of this dataset among the community. The MIMIC II dataset is also very popular but does not provide much contribution due to it's low record time (60s). The BIDMC dataset gain in popularity from 2019 and contribute a lot to community.

5.6 Proposed methodology to build future datasets

From all the observation 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 [78] who used the VITAL DB, only keep 100 subjects over the 6000 available. So we need to build a dataset with at least 1 000 peoples with all ages. The materials used to measure the signal should be a smart watch or stuff like that. People won't wore a Pulse Oxymeter to their right index

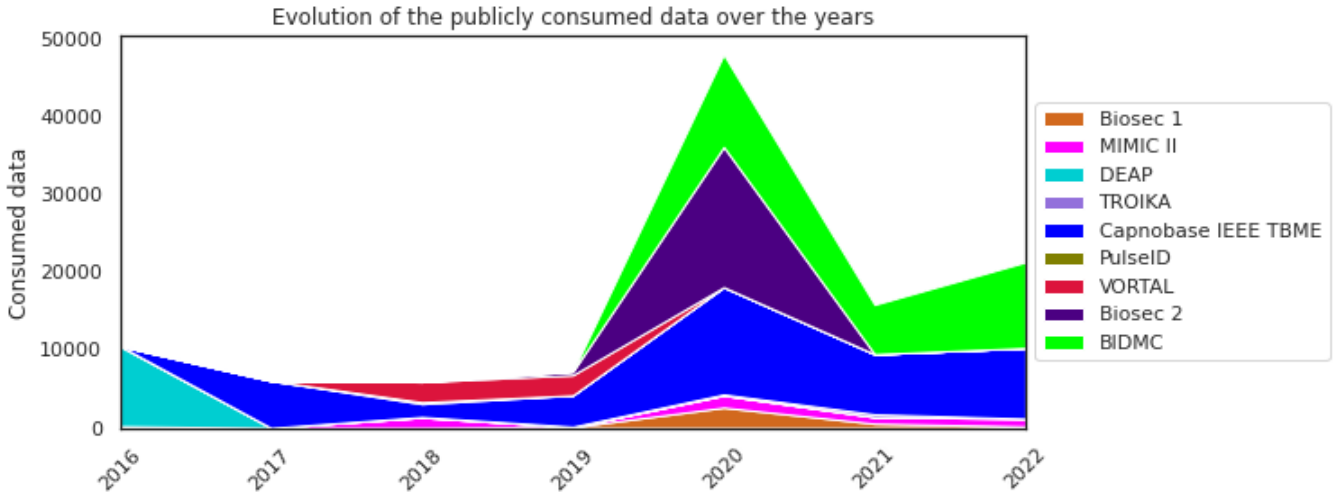


Figure 16. Public datasets contribution over the years

paper	Year	Sampling frequency (Hz)	Acquisition time	Number of patients	Conditions	Time interval
[70]	2021	50	6s x 50	35	rest, wrist	NA
[42]	2014	2000	15 cycles x 2 sessions	30	Finger PPG; in relaxation	NA
[81]	2019	500	NA	23	Finger PPG; in relaxation	NA
[45]	2015	NA	NA	10	Finger PPG	NA
[86]	2007	300	70s x 3	3	Finger PPG; in relaxation	NA
[30]	2003	1000	1min x 1	17	Finger PPG, in relaxation	NA
[71]	2018	NA	1min x1	10	NA	NA
[64]	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
[21]	2016	1000	3min x 1	15	Finger PPG, in relaxation	NA
[18]	2013	75	2min x 1	44	Finger PPG, in relaxation	NA
[31]	2003	1000	1min x1	17	Finger PPG, in relaxation	NA
[26]	2019	10	NA	5	NA	NA
[63]	2013	37	60s x 4 x 2	9	Relax & stress	NA
[65] Nonin	2018	75	60s x 1 x 3	24	NA	7 days
[65] 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

in continuous. However, may people wear smart watches which already measure the PPG to provide heart rate. So it's highly probable that the continuous authentication systems using PPG will use smart watches. Finally, the signal should be acquired during multiple days, with a dedicated protocol done at least one each measured day. This protocol should be : rest the subject on a silent and dark room. Then subject should make and effort test on treadmill or bike or whatever. Then it should rest again and meditate (or do whatever it wants to reduce heart beat at the minimum). Finally it should make some activities that will induce multiples different emotional states. This can be done by watching different emotional videos, playing to a defined video game etc. The goal here is to measure the change in PPG signal induce by the different emotional states. After that, the subject should go back to it's daily life. This will be interesting to collect such dataset on three consecutive days, two or three times, with at least one month between each three-days sessions.

Constituting such a dataset is very difficult. This is why it can be interesting to create a collaboration between university or labs to create it. Such a dataset will benefits the community of PPG-Biometric recognition but also the medical community and the Human-Machine Interface community. The PPG represent 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. At first, the DEAP dataset was collected to achieve this goal.

6 Pre-processing

In this section, we will explain the main parts of signal preprocessing. We explain the methods used to reduce noise in the signal, segment, and normalize heartbeats. This step is critical to improve performances of the PPG-biometric recognition algorithms.

6.1 Noise reduction

As we stated in our introduction, PPG signals are recorded by the amount of light that is absorbed or reflected by blood vessels. The signal measures the variation in blood volume. So they are subject to many kinds of noise: motion artifact, electromagnetic perturbations, etc. Moreover, the PPG signals frequencies are in the range 0.5-5 Hz [11]. However, as we shown in Figure 12 most studies recorded their PPG signals with a sampling rate above 100 Hz. This gives more details on the signal but adds some noise. The sampling theorem proved by Shannon [69] impose a sampling frequency at least twice more important as the most elevated frequency of the signal we want to represent. Thus, the sampling frequency for the PPG signal must be at least 10Hz and a sampling frequency over 300Hz may be too much and add too much noise. This is why all teams had to filter the noise to clean the PPG signal.

Figure 17 represent the different noise reductions techniques used by each experiences. For this parameter, we observe around 25% of missing values. Then we can observe 33 different methods, over the 44 papers, this show that each paper used a different method to reduce the noise of the PPG signal. This show a big heterogeneity in this parameter and no consensus.

Figure 18 show the different usage over the years. Here again we can observe the huge heterogeneity of the methods. However, we can observe that many studies used a Butterworth filter [19], but each teams test different parameter (different order and cut off frequency). In this figure we color all the methods using a Butterworth filter with a different green color (green, olive, forestgreen etc). We can observe that the usage of Butterworth filter started mainly in 2016 and was the most used technique until 2021. They use low pass, high pass, and band-pass filters. However, each team selects different kinds of filters, different ranges. The most common filtering range is pass-band between 0.5-15 Hz. We can observe that, in rare cases, some teams use Gaussian filters [22] or Discrete Wavelet Transform [9] to lower the noise in the signal. Only the team of [42] use finite impulse response filter (FIR) and [64] keep only the most important coefficient in Fast Fourier Transform (FFT). They keep the 8 first coefficients for each pulse and recreate the signal with the inverse function and the chosen coefficients.

However, we can see that some studies do not provide enough details, the filter frequency is missing, and some times the filtering process is not precised ([30] [74], [18], [63], [37], [**?**], [66], [6], [10], [81], [47], [5]). Because all theses methods are parts of bigger experiments, with different data, and different classification methods we cannot determine the impacts of each filtering methods.

The community needs to provide a clear comparison between each method. To do this, multiples filtering methods have to be tested first on the same data set, then on heterogeneous data set, with a fixed algorithm and fixed feature extraction methods. This will help the community to determine the pro and cons of each filtering method. Each team uses a filtering method without explaining why they choose it to compare to the other. The filtering phase is critical because it removes some data in the signal (ideal noise) and influences all the next phases.

Finally, some other techniques may be tested to reduce noise, such as deep learning [36]. A filtering framework must be created to help the community to develop generalized and robust algorithms to authenticate people.

6.2 Segmentation and normalization

Most studies uses two more techniques in their pre-processing phases : signal segmentation and normalization.

First, we can see 18 different methods to segment the signal that will be used to classify the subjects. Virtually it exists an infinity of methods because we have an infinity of

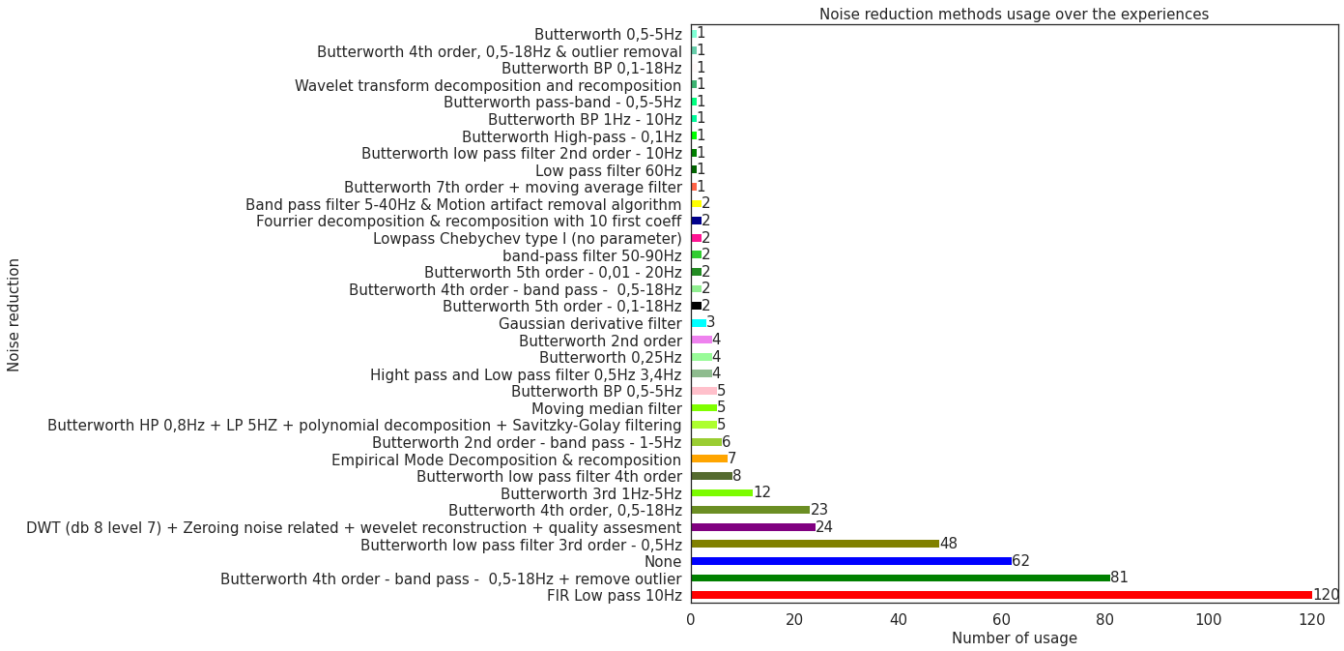


Figure 17. Noise reduction usage over the years

possibility to segment a continuous signal. However a PPG signal is a cyclic signal so intuitively we can assume that most of the information is contained in one cycle. Generally they split PPG signals in single beat or cycle. To do this a modified version of the Pan-Thomkins algorithm [55] is mostly used. It detect the systolic peak, allowing researchers to split signal from one peak to another. The other technique used to detect systolic peak is to compute the first derivative of the signal and finding the 0 crossing points.

Figure 19 show the different techniques used by the community to segment their signal. We present the data in raw values and in proportion. For this parameter we observe 2.65% of missing values, showing a good usage and description of this stage inside the community. We can observe that around 50% of the experiences were made with a single cycle segmentation. This show a consensus on this criteria. However, some studies on the optimisation of this parameter are needed.

Figure 20 show the segmentation technique usage over the years. We can see that the method of single cycle segmentation is the only one that is used all years, over multiples papers. This show again the beginning of the consensus on the usage of this technique.

The only 5 which used more cycles are [63], [82], [47], [33] and [51]. Finally, only [16] and [43] used all the signal, and [84] used a sliding windows decomposition. the 3 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 show that using more than 10 cycles was not improving the performances. However they did not test the most used techniques : single cycle.

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

Figure 21 draw the usage of the normalization techniques over the experiences. We can observe that in all the experience, the most used methods is to provide raw signal, with no normalization (40% of the provided experiences). Then the Zero Mean normalization is the second most employed method. We can observe 17 different methods but most of them are little used, only in one paper.

Figure 22 show 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. All other seems to have been tested only one time.

However, all the normalization techniques are known in statistics and machine learning (zero mean normalization, amplitude normalization, alignement etc.) except for one : cardioid normalization [**?**]. An example of this normalization is shown in Figure **??** and Figure 5b. This 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,

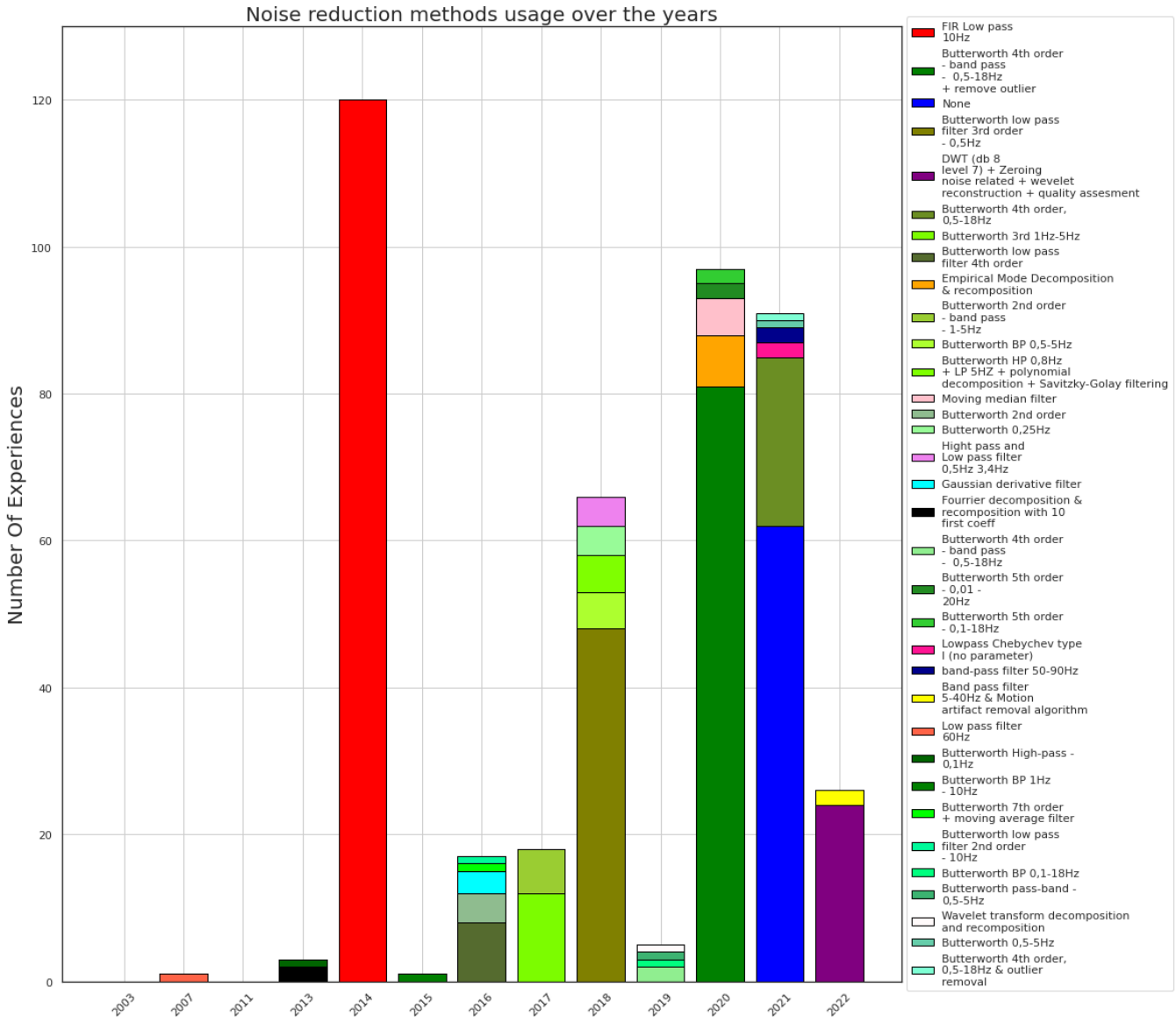


Figure 18. Noise reduction usage over the years

and we need to quantify the effect 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. No clear justification is given to do this phase. These phases induce computation usage and may not be useful, in the IoT world, where computational power and energy are limited, it can be worth deleting a phase that has few or few impacts.

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 majors kinds of features: fiducial and not fiducial. The first one took 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 extract by numerous technique, 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. Thus most of the papers provide details about this stage : type of features, number of features, extraction or selection methods. We only have 0.16% of missing values for the feature types. Figure 23 draw the feature type usage over the experiences. Figure 24 draw the feature type usage over the years.

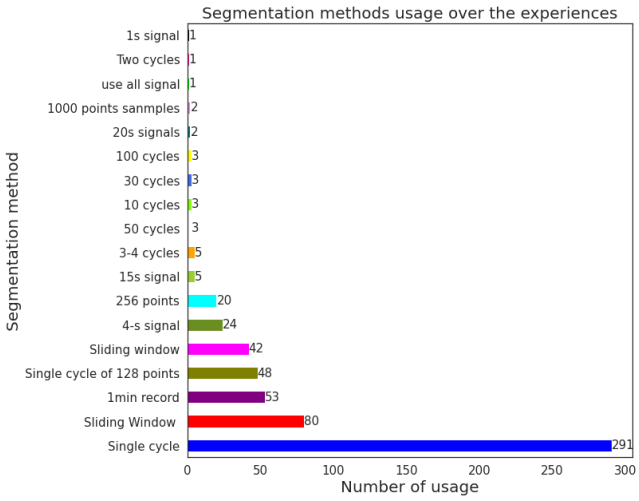


Figure 19. Segmentation usage over the years

The fiducial domain correspond to historical landmarks on the signal, taken from biology (like systolic peak). The time domain features correspond to statistics about the signal in time domain like min, max, kurtosis, skewness etc. Then the transformed domain correspond to all features extracted with a signal transformation like FFT, DWT, DCT etc. We observe in our dataset, 46 different methods over 46 different papers. Some features types were used by different papers, like the time domain, fiducial or statistical types. Other papers as hwang2020evaluation compare the efficiency of multiples type of features extraction like Interpolation in frequency, Zero Padding in Time, Dynamic Time Wrapping and all the possible combination of this features types. Moreover, some feature types were extracted on multiple version of the signal : raw signal, first and second derivative. For this parameter, there is no consensus at all. Before 2013 most of the papers used fiducial features. It was less used and replaced by other non fiducial features. After 2014 most of the collected papers used a different transformation on the signal to extract different features. Thus it's hard to compare the methods and tell the pros and cons of each method.

7.1 Fiducial features

Fiducial features are points, values, on the PPG signal or its derivative. They are also called "landmarks" [40]. These points and values are determined by the standards of physiology and statistics such as mean, standard deviation, systolic peak, etc. There are more than 40 different standard fiducial features, all in the time domain. The most used are given in Table ??. An example of fiducial points on a PPG signal and it's derivatives, from [42] is given in Figure 25 and 26

These points can be determined in the first and second derivatives of the PPG signals. Some studies use systolic, diastolic peak, etc in the three main versions of the signal (raw signal, 1st, and 2nd derivative) as features. The number

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

of extracted features is highly variable between studies, from 3 to 200. In their study [10] define a window of 200 points, centered on the systolic peak. All these points are combined in a feature vector of dimension 200, this why they have a high number of features. Otherwise, the maximum of fiducial features is around 40, when all the time differences of all fiducial points are taken as a feature. In this case, most of the features are combinations of standards points such as time differences. This can lead to repetition of some information and one study show that using fewer features can improve the accuracy of the system. For example, [42] extracted 40 fiducial features and compare the accuracy of their classification algorithm depending on the number of extracted features. They also provided a ranking algorithm between fiducial features. They show 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. Moreover the "best" features are dependent on the used algorithm, its configuration, and the signal used. For example, using KNN with k=3 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.

Fiducial features are time domain-specific points, values, and transformation. They were the first used and are very sensitive to noise and physiological change [41]. There is a limited number of interesting points and extract more fiducial features can reduce the accuracy of a system.

7.2 Non-fiducial features

Opposed to fiducial features, non-fiducial features are extracted with signal transformation [41] such as Fast Fourier Transform (FFT) [13] [53], Discrete Cosine Transform (DCT) [4], Discrete Wavelet Transform (DWT) [38], etc. They are not points in the time domain (eg. systolic peak) or statistical

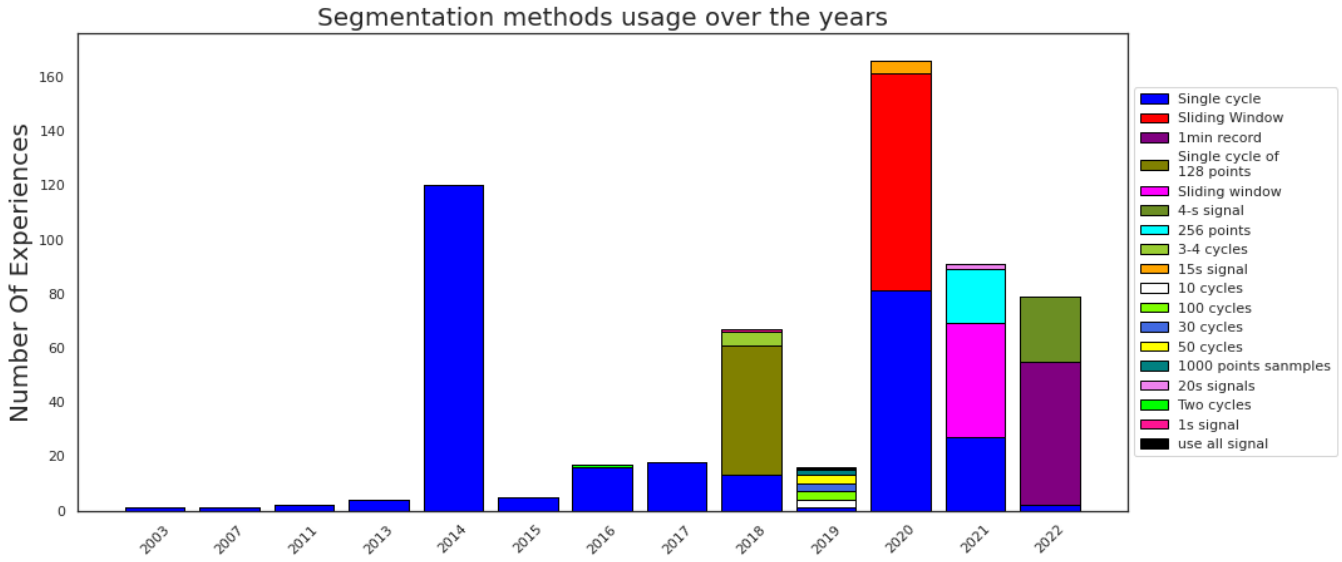


Figure 20. Segmentation usage over the years

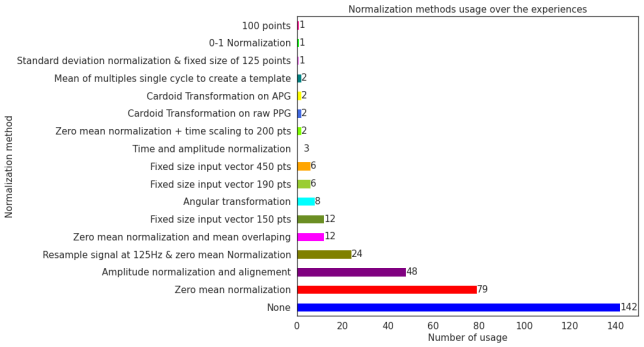


Figure 21. Normalization usage over the years

values (mean, standard deviation, etc) but features are given by the transformation. Each transform gives different types of information, different number of features and provides a different way of describing the signal. For example, FFT gives the frequency decomposition of the signal or the DWT give time-frequency domain features. They are less sensitive to noise compare to fiducial features.

They are many kinds of transformations, but the most used are DWT, DCT, and FFT. Others, such as Gaussian decomposition, Continuous Wavelet Transform (CWT) [61] are less used. Overall our studies, only [66] used the Gaussian decomposition, and only [82] used the CWT.

Because studies have different data, a different number of extracted features, and different algorithms it's hard to determine the best transformation to extract features. As for the signal filtering, many experiences need to be done to determine the impact of each transformation, the number of coefficients, and if the usage of multiple transformations can improve accuracy or not.

There is two other way of extract non-fiducial features: through classical dimensional reduction such as Linear discriminant analysis (LDA) [12] or principal component analysis (PCA) [80] or with a deep learning algorithm. For example, [74] used the LDA, [63] used a modified version of PCA, called KPCA [67] as feature extraction. For the deep learning feature extraction, teams used in general a Convolutional Neural Network (CNN) [29]. For example [25] dedicated the first layer of it's CNN to extract features, [84] and [46] used multi-layer CNN for features extraction. In this case, we cannot 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 [66] [21].

7.3 Features selection

Using many features increases the computational needs of a classification algorithm. Moreover, using highly correlated features does not add any information and they are not very useful for the classification algorithm. This is why dimensional reduction could improve the efficiency of the system, by reducing the computational needs while not decreasing the performances.

Common techniques to achieve dimensional reduction are Linear Discriminant Analysis (LDA) [12], Principal Component Analysis (PCA) [80] and their variations (DLDA [82] and KPCA [67]).

They can be used with all fiducial features and some non-fiducial features. For the latter ones, if they are extracted

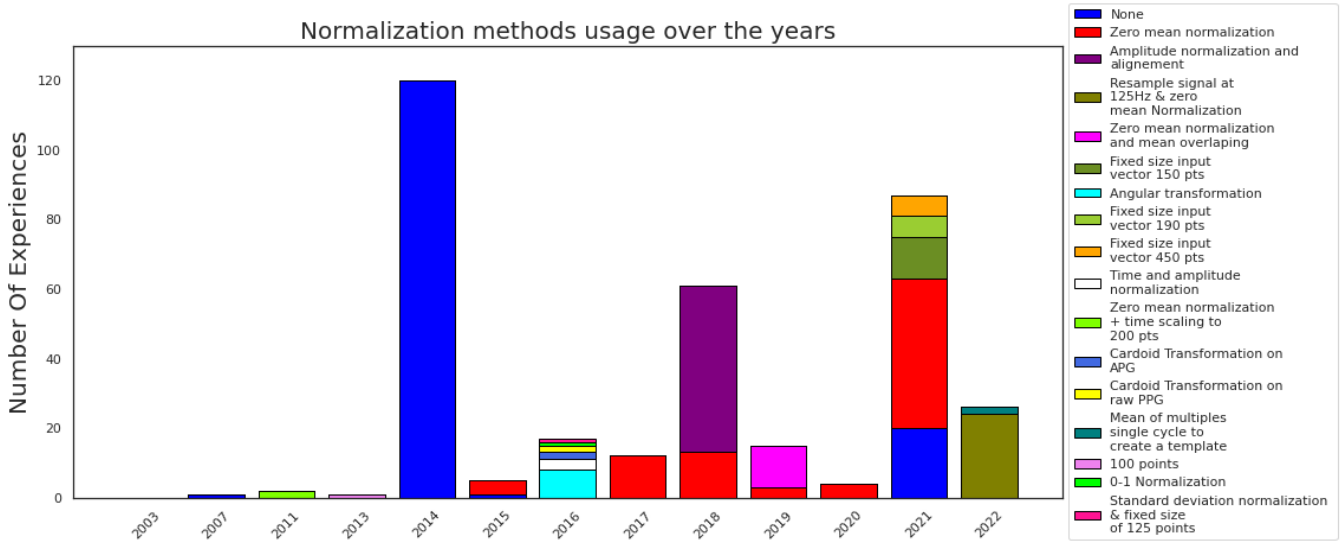


Figure 22. Normalization usage over the years

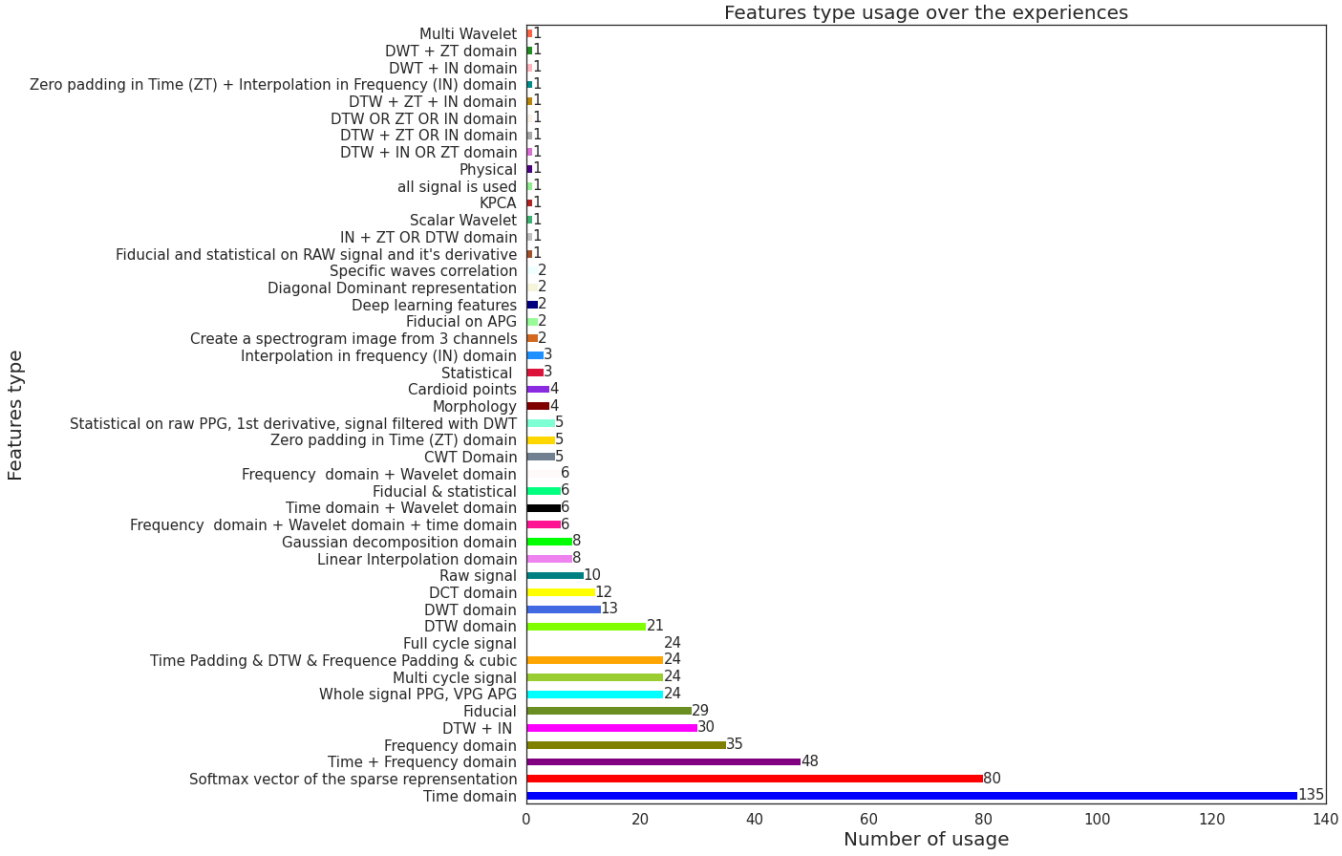


Figure 23. Feature type usage over the years

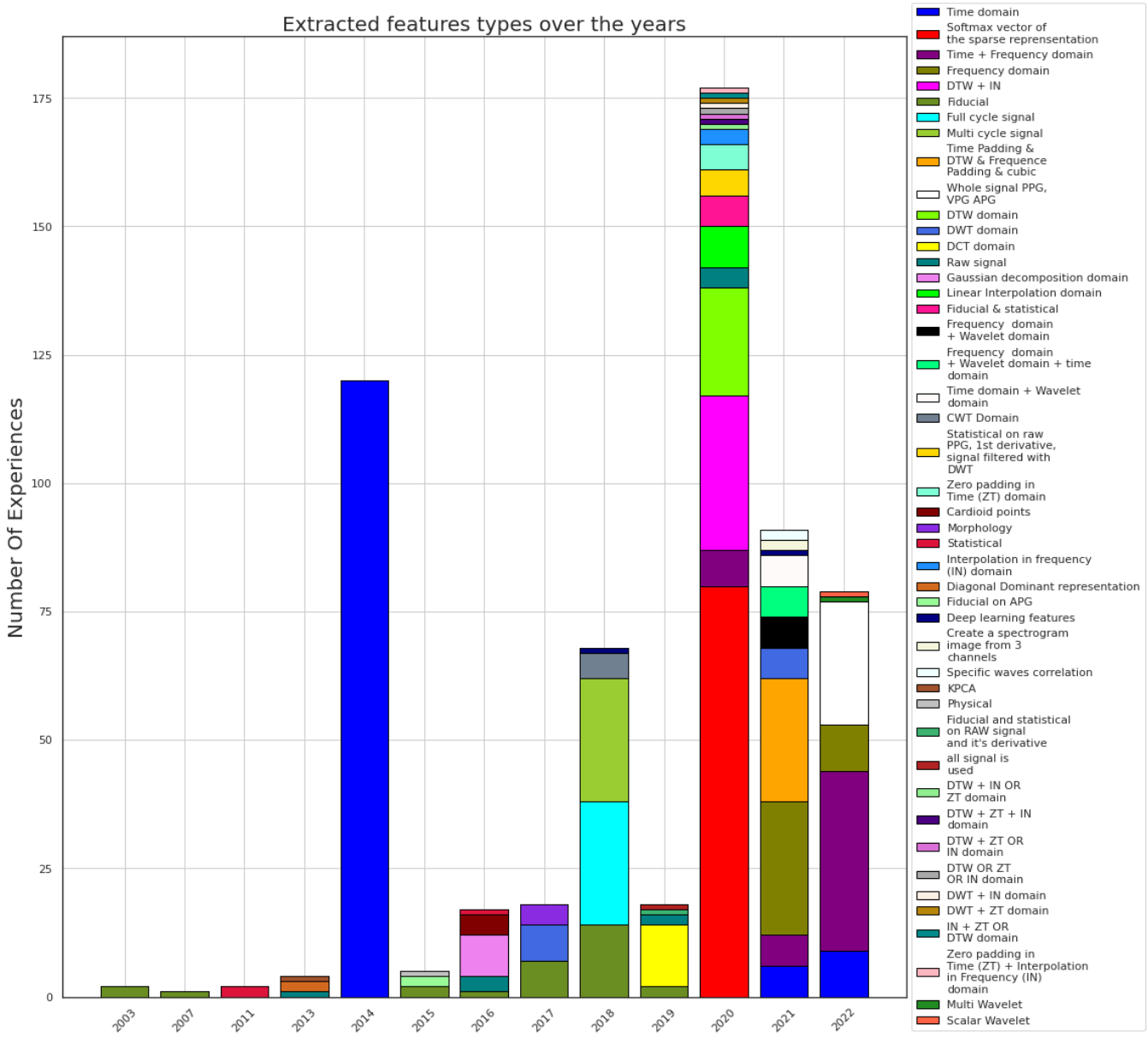


Figure 24. Feature type usage over the years

with traditional transformations such as FFT or DWT, PCA and LDA can be used to select the best features from those transformations. For example, [82] used DWT to extract time-frequency domain features and use DLDA to select the best features. This step is, however, generally unnecessary with representation learning since the features are automatically extracted and summarized. This is the case, for instance, with works that exploited convolutional neural networks (CNN) such as [16]. In this case, the number of "neurons" in the first layers and the number of outputs determine the selected features.

Figure 27 show the different techniques used over the experiences to select features. On this parameter we observed 1.49% of missing values, showing a good description of this stage in the literature. We can observe the usage of 78 different technique over 44 papers, showing a good exploration of multiples methods. Some papers test multiples methods or one method with different parameter. For example Sancho et al [65] use a KLT average on 10, 20 and 30 cycles. However most of the techniques were not tested in comparison with other, thus it's hard to determine if some are most efficient than other. Figure 28 show the usage of the multiple selection methods over the years. We observe that the only method

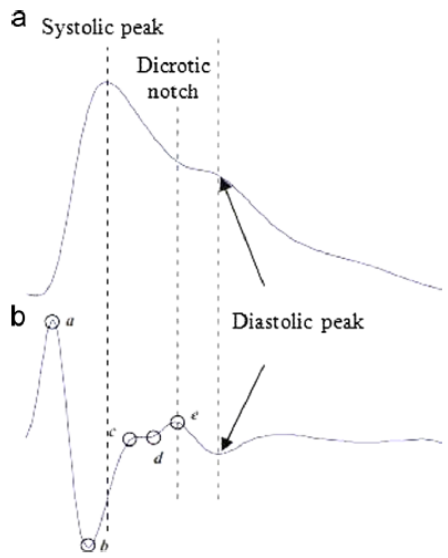


Figure 25. An example of fiducial points from PPG raw signal and APG. From [42]

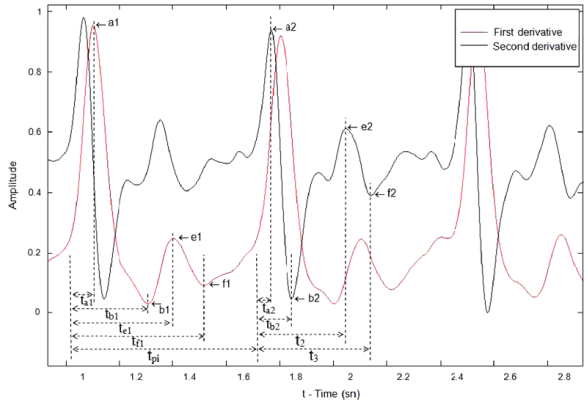


Figure 26. An example of fiducial points from PPG 1st and 2nd derivative. From [42]

used over multiple papers is the selection of all extracted features, so no use of technique to reduce dimensionality. This show no consensus on the community on this stage.

Kavsaoglu et al. [42] showed in their study that decreasing the number of features can significantly improve accuracy. It would be interesting to use both fiducial and nonfiducial features and apply some dimensional reduction techniques to find the best features to extract. Then it will be interesting to compare these extraction methods with an extraction done with Deep Neural Network, with the same data set and the same algorithms.

In conclusion, we can see that it is pretty difficult to compare studies and their methods. Moreover, some are lacking details, such as [37] who explains how works the used algorithms (KNN and Naive Bayes) but never explains which features are extracted and how. Finally, some are not using learning methods to make their systems. For example, [22] proposed a system where they create a signal template by combining several aligned pulses. Then when a user is presented to the system a PPG pulse is extracted and several distance metrics are used to compare its pulse to all templates. In this system, there is no need to extract and select features.

7.4 Features length

The final parameter on the features is the "feature length". This represent the information that will be used by the algorithm to recognize subjects. As we said earlier, the feature length will also play on the performances and computational needs. In our dataset we represent the feature length with one, two or three dimension vector (a x b x c) when possible. Figure 29 show the different usage of feature length over the experiences. For this parameter we observe 30% of missing values. Around 60 differents features length have been tested. As for the feature selection stage, some papers tested multiples feature length. For example, in 2014 Kavsaoglu et al [42] extracted 40 fiducial features and test to use 5, 10, 15, 20, 25, 30, 35 and 40 feature in the input vector. They show that in general, increasing the number of features used improve the performances until a limit. In their experiences we can see that in general, the limit is reached at 25 features. More features does not improve the performances. Some studies just used the whole signal, other one cycle, but do not provide the real size of the input vector. Some studies used one dimensional vector, for example when they just used 5 fiducial features.

Figure 30 show the feature length usage over the years. As for the feature selection, we observe that each paper use it's own parameter and no consensus exist. To compare the length used by the community, we convert all the feature length in one value. For the multi-dimensional vectors we just multiply each dimension to have the total number of features used. We only keep the numerical values of the dataset for this calcul.

Figure 31 show the evolution of the feature vector length over the years. We observe a global increase of the feature length over the years. Before 2018 most feature vectors have a length between 10 and 100. In 2018 the feature length increase a lot with a mean size around 10^5. Then the size decrease and stabilize around 1000 feaures.

8 Recognition algorithm

The last part of an authentication system is the classification algorithm. It's this piece of software that will recognize registered people and reject unknown people. Thereby two metrics are important to determine the performances of an

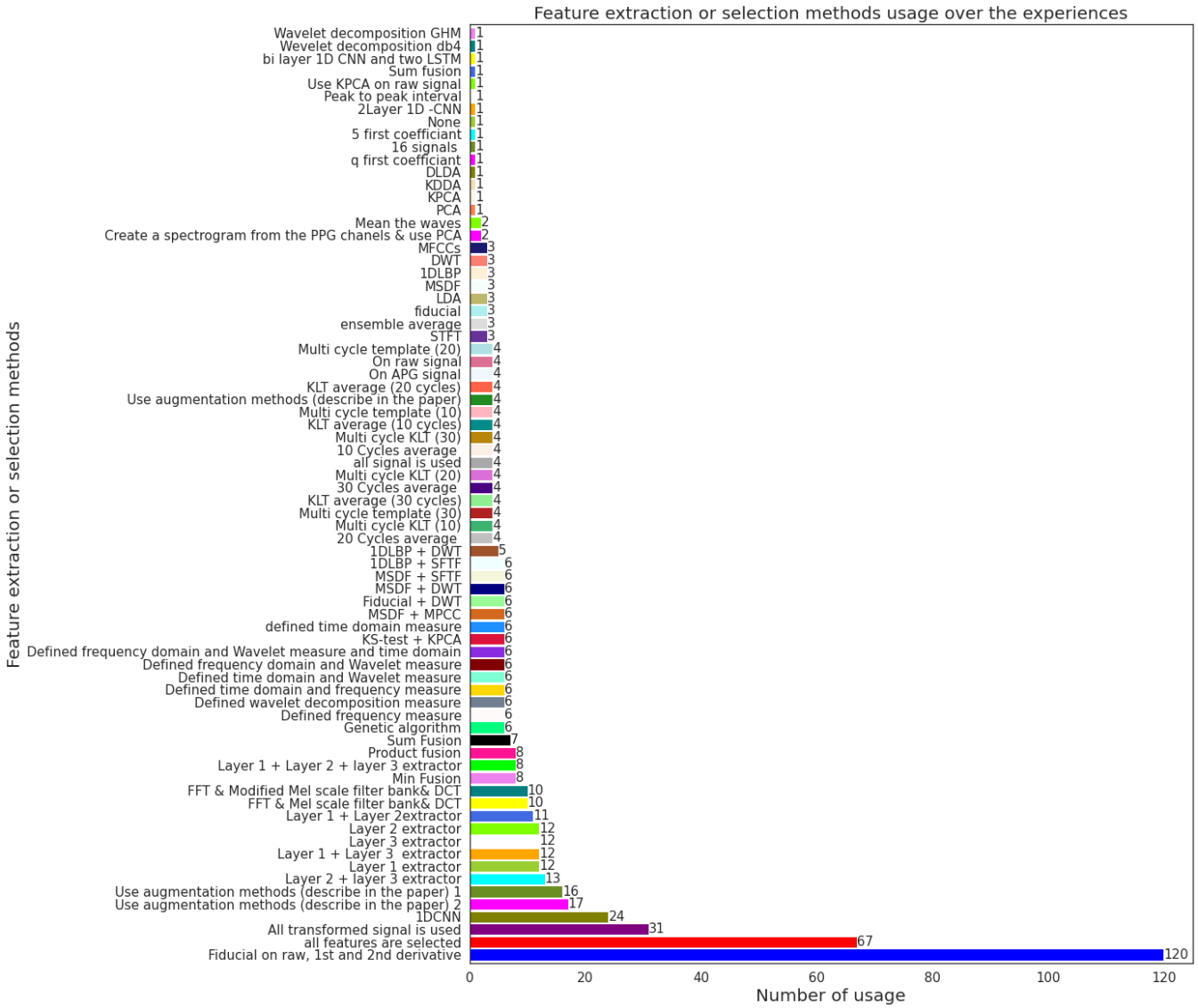


Figure 27. Methods to select or extract features over the experiences

algorithm: accuracy and equal error rate. The accuracy represents the ability to well identify a subject. It's the true positive rate, and we want to maximize it. Moreover, two metrics are used to measure the errors of 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 who is tested. This represents the probability of an intruder has been treated as a genuine user. The other metric, the FMNR represents the probability of a genuine user treated as an impostor. Because biometric authentication is a template matching problem, an algorithm gives a matching probability and we need to apply a threshold. The value of the threshold influences deeply the FNMR and the FMR. However there always exists a threshold value where FNMR and FMR are equal, this point is called Equal Error Rate (ERR). This point is traditionally used to measure the performance of a biometric authentication system [17]. In this study, we will use it in the same way.

In all the studied papers, we identify four main types of classifications methods: statistical, machine learning, template matching and deep learning. We will explain the main used models for each part.

Figure 32 show the repartition of the algoritms type over the experience. We observe that most of the experiences used machine learning for classification (around 60%). Figure 33 show the algorithm type usage over the years. We observe that the machine learning was tested in early stages and is the most used over the years. This show a big popularity for this kind of algorithms.

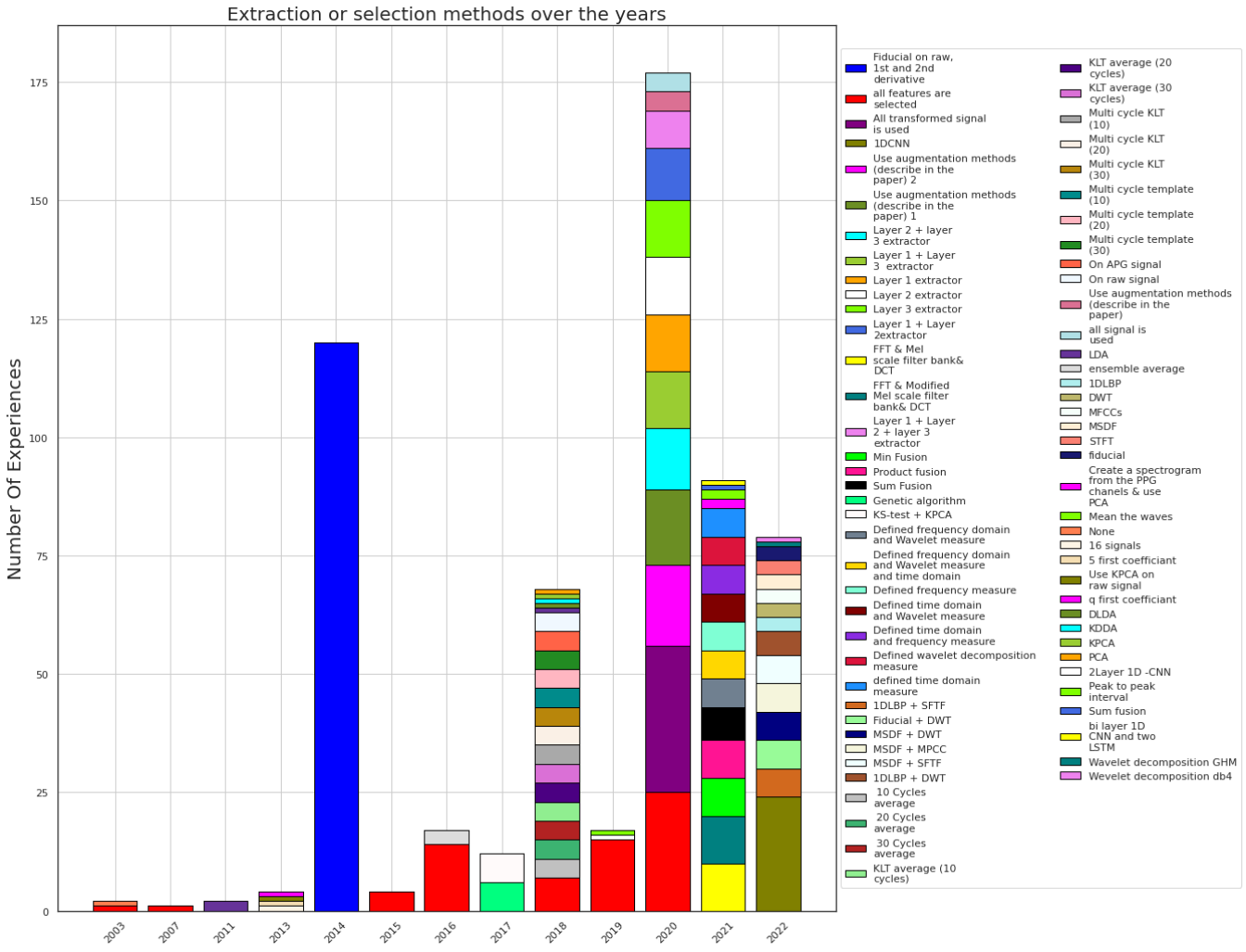


Figure 28. Methods to select or extract features over the years

Figure 34 show the different architectures of the algorithms used by the community. For this parameter we have 0.81% of missing value. When avalaibale, we provide the parameter of the architecture (K for a KNN, number of neurones etc.). We observe 80 tested architecture showing multiple tests per papers. However, some architectures have been tested by many papers, such as Naive Bayes Classifier or the euclidian distance in a template matching. Figure 35 show that the most used architecture is KNN. We observe that the different KNN architectures gather 26.89% of all the used architectures. Naive Bayes and architectures based on CNN are the two other most famous architectures.

8.1 Statistical

Statistical classification is the first method used by research teams to create PPG-based biometric authentication. These methods are mainly based on cross-correlation. Some teams also used directly statistical analysis such as LDA to make a classification system. The first study on PPG authentication used fuzzy logic with Gaussian function [31]. They compute a score between an enrollment template and a given signal. This score is based on the Gaussian function parameter (μ and σ). The Gaussian function is computed for each pulse, to overlap the maximum area of a PPG signal. [86] compute the correlation of each extracted feature for each subject. They show that the selected features were highly correlated for each subject and not correlated to others. However, they did not try to make a full authentication system and do not provide any performances. Later, [64] use Euclidean distance between FFT features extracted from PPG signals. They show that this distance was significantly higher between two pulses from two different subjects than for pulses from one subject. Here again, no metrics are provided.

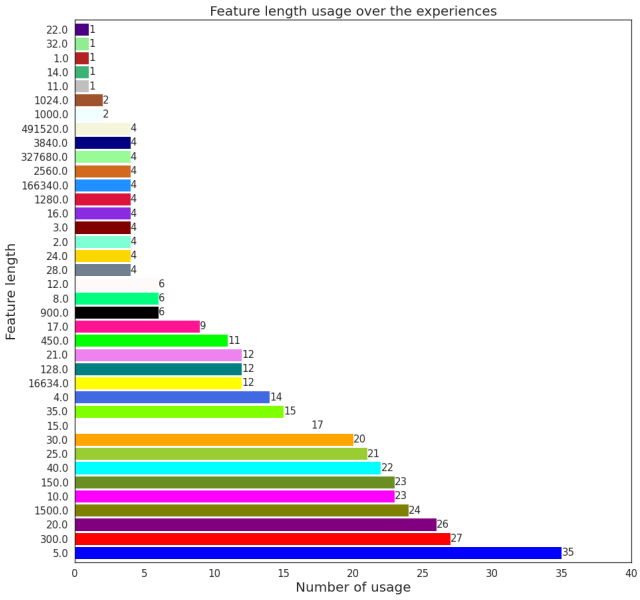


Figure 29. Different feature length usage over the experiences

8.2 Template matching

The second kind of methods used by the community is the template matching. This techniques are often used for other biometric systems (TODO CITE). It consits in the creation of a template that is stored in the system and that will be compared to each input. Each time a matching score is computed. When classification occurs with distance metrics such as Euclidean or Manhattan distances, there is only one parameter to set: it's the threshold value. When a template is computed for a claimed user, the system will measure a distance between the two templates, if this distance is over the threshold, the authentication is rejected, however, it's accepted. Distance metrics can be used with any kind of template and it's the fundamentals of a machine learning algorithm: k-Nearest Neighbors (KNN) [29].

The main advantage of template matching classification is the no-need to train an algorithm. We just need to create a template with features, store it, and then compare it with the tested template.

8.3 Machine Learning

Machine learning algorithms are the most used in the selected studies. They are algorithms that need to be trained with a subset of available data and test with another subset of data, they are called the training and testing set. Sometimes, when algorithms have hyperparameters, another subset is used to tune them. This subset is called the validation set. In the end, to make good training and testing, the majority of the available data is used to train the classifier, then a small subset is used to tune hyperparameters, and finally, another subset is used to test the algorithm, to determine its performance in unseen data. For example in the ImageNet Challenge 2014 [62], around 4.3% of available data are used for validation, 8.7% is used for testing and the rest is for training.

These algorithms provide a function and they refine and correct it with data. They can generalize their function to new data, this is why we have to test them with previously unseen data. The goal of each algorithm is to classify data in at least two classes. For example, the Support Vector Machine (SVM) algorithm [88] will split the data into multiples classes. To do this, we have to provide the algorithm data from each class we want it to class. This algorithm maximizes the space between two classes, with a composition of multiple linear functions. If the clustering function is not linear, it can be improved by using a "kernel trick" [88]. KNN is the extended version of the metric template comparison. We need to provide as many templates as possible and store them. Then when a user request authentication, a template is computed and its class will be the same as the k closest template. Other machine learning algorithms are used in selected studies but less than KNN and SVM : Random Forest [49], Decision Tree [58], Naive Bayes [48] and Bayes Network [27].

Their performances are highly variables and depend on the signal, training data set and testing data set. We can see, in one study than one algorithm may be better than another, and in another study than the second is better than the first. For example, [47] show that Random Forest have 99% accuracy and KNN 98%. However, [32] shows that KNN has 94.44% accuracy and Random forest has 90.39%. This is why it's very important to compare studies that use the same data and same validation method.

8.4 Deep Learning

Deep learning algorithms are the second most used in the collected studies. They are quite more difficult to develop and are based on neural networks such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), etc. Many kinds of algorithms can be mixed because they are multi-layer algorithms. However, they can be used for all the described steps: noise reduction, feature extraction, selection, and classification. The most efficient algorithms in our selected studies used a combination of CNN and Long Short Term Memory (LSTM) [75]. LSTM is very efficient at processing time-dependent signals. They were first used for natural language processing. Here, for example, [16] makes a framework to authenticate patients with PPG. In their study, they used a two-layer of one dimension CNN that extracts features. Then this CNN fed an LSTM which fed two final neurons, activated with the SoftMax function. The final layer output the class of the signal, corresponding to the authenticated user, with a certain probability. They claimed to achieve 96% of accuracy but they did not provide any EER values. Everson et al. [25] claim to obtain the same results

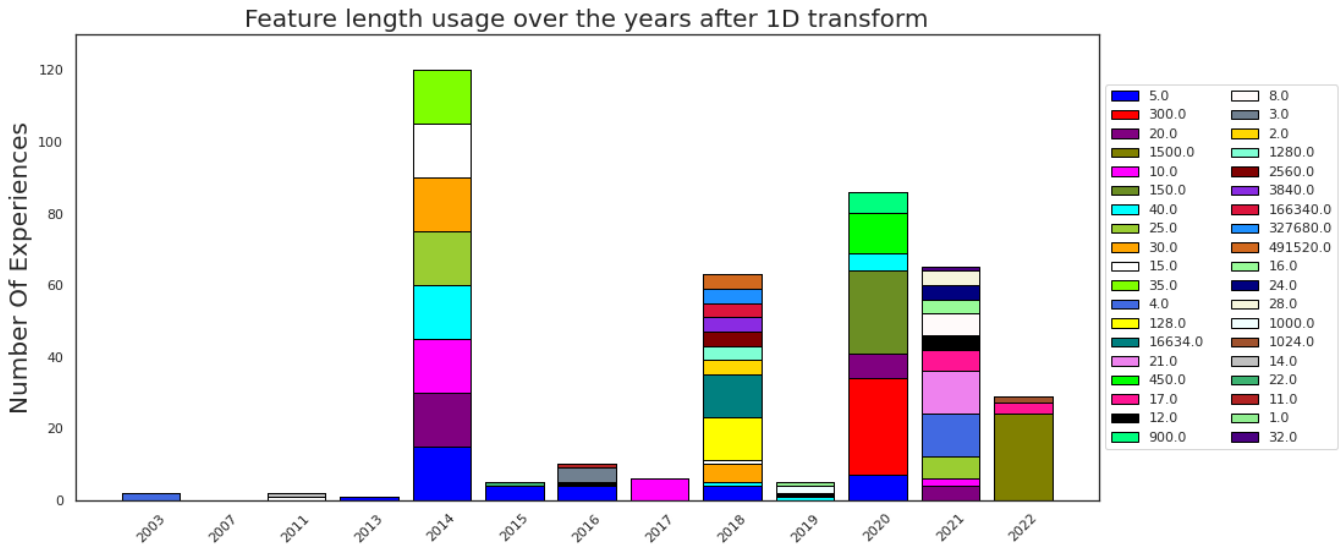
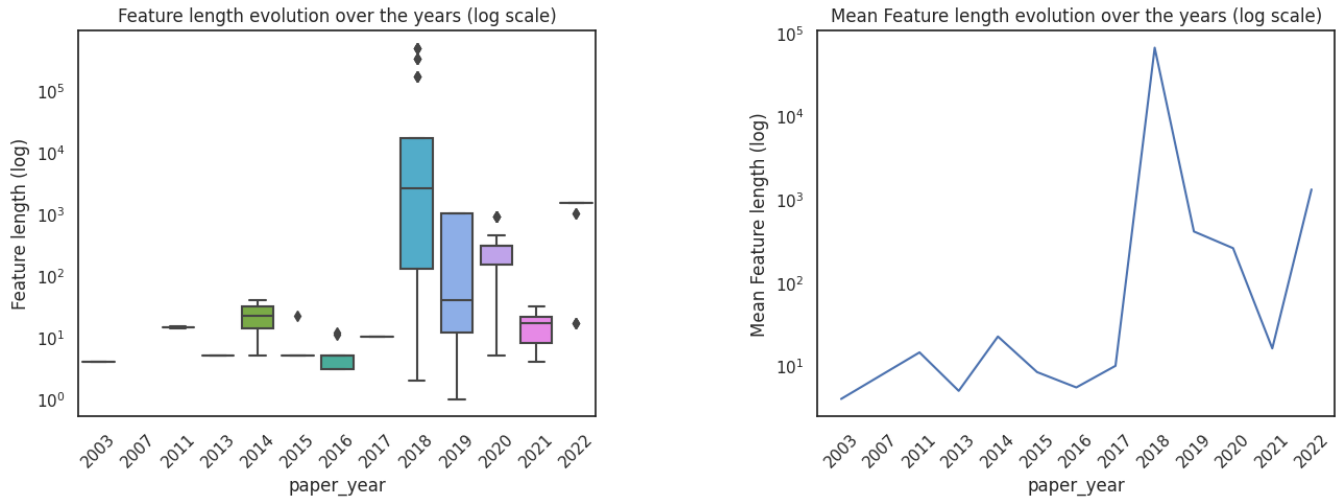


Figure 30. Feature length type usage over the years



(a) Feature length evolution over the years

(b) Mean Feature length evolution over the years

Figure 31. Representiation (box plot and mean evolution) of the feature length evolution over the years (log scale)

with similar architecture. 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. For example, [10] showed 100% accuracy for different algorithms (Bayes Network, Naive Bayes, and Multi-Layer Perceptron). However, this is true only for a small subset of the data. They split data into gender and ages groups. They achieve 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 show huge variability in the algorithm performances and further metrics and test must be done to determine if it's due to unbalanced data or if there is a problem based on age and gender that needs to be considered. Thus, a common testing methodology must be used to correctly benchmark all the algorithms.

9 Study comparison

In this section we will try to compare the results of the collected studies. First we will analyse the training and testing

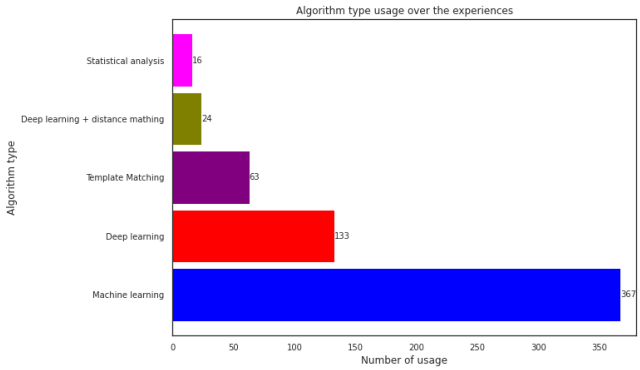


Figure 32. Algorithm type usage over the experiences

sets, then we will analyse the validation methods. Finlay we will analyse the results with the accuracy, EER, lowest FMR and lowest FNMR.

9.1 Training and testing sets

The training and testing sets are very important to determine the feasibility of the results. To have a valid result the experience should have a good and big enough training set and a good different testing set. The testing set must not be used in training. In general in the studied experiences, the training and testing set have the sames users. The training phase act as the enrollment and the testing phase as the authentication phase.

Figure 36 show the usage of the training datasets over the experiences. For this parameter we observe 12.76% of missing values. We observe a big heterogeneity in the creation of the datasets. Most of them take a percentage of the available data, other a fixed size of signal for each subjects. Very few experiences splits the users in genuine and impostors.

Figure 37 show the evolution of the usage over the time. We can see that each paper use it's own method. There is no consensus and it's hard to compare studies due to the huge heterogeneity of this parameter.

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 show 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's very important to define a clear methods to reproduce and benchmarks all the provided methods. Figure 40 show the different validation techniques usage. We observe 5 methods, based on cross fold validation. Each method is a variation with more or less cross fold. The most used technique is the 10 cross fold validation with L2 regularization.

Figure 41 show the validation technique usage over the years. We observe that the 10 cross fold validation was used in 2015 with few experiences. This was the main used technique until 2020 where the add of L2-regularization dominate the experiences. However, the usage of validation techniques is very heterogeneous over the years, showing the lack of methodology across the community.

9.3 Performances metrics

To compare the studies and find the best architectures, we wanted to compare multiple performances metrics : accuracy and EER because they are the most provided metrics in the studies. However we observe a huge proportion of missing values for this parameter. We observe 14.75% of missing values for the accuracy, 68.49% for the EER. In the end, we can not use the FMR and FNMR to compare the studies, and the huge lacking values for accuracy and EER is also a big problem for comparison. This is why we will only compare few studies which provide the good metrics over multiples experiences. Figure 42b show global goods performances over the years. We observe that most of the experiences range from 80% to 100% of accuracy. However the variability is huge with some experiences that show a drop under 20% of accuracy. This show that all the architecture are not good and some are better than other. To have a valid comparison, we need to compare study with the same dataset. Moreover the accuracy is not very good to compare biometric systems, the ROC or DET curves are better choices. However, in all the studied papers, only X draw such curves but none gives the data to draw it and allow the community to compare different works. A good metrics derive from the ROC curve is the Area Under the Curve (AUC).

Only three papers provide AUC values, representing 13 experiences, but their dataset are different. This is why they are hard to compare. [9] provide some mean AUC values for some of it's experiences. [51] provide multiples ROC curves and the raw mean AUC values for all of it's experience and at every stage (validation, develop and test). They also provide the standard deviation for each experience. We can observe a huge variability in the AUC performances. They did not provide any other metrics. [59] Provide ROC curves and AUC of two of their experiences. They seems to have reproduced some part of the previous experiences from [16], [82] and [31] but they seems to have change some part such as prepossessing and features extraction. Finnaly, the methodology of theses papers is good and must serve as basis for others.

Two other papers[18] and [50] provide the DET curves wihch is similar than the AUC curve, but with no exploitable values.

Other metrics can also be interesting such as the one provided by [42] : precision, recall, specificity and f-measure. All of theses measures should be provided for each experiences. However they are the only one to provide them for all subjects. [43] provide the mean accuracy, specificity and sensitivity for all of its experiences, but not for all subjects.

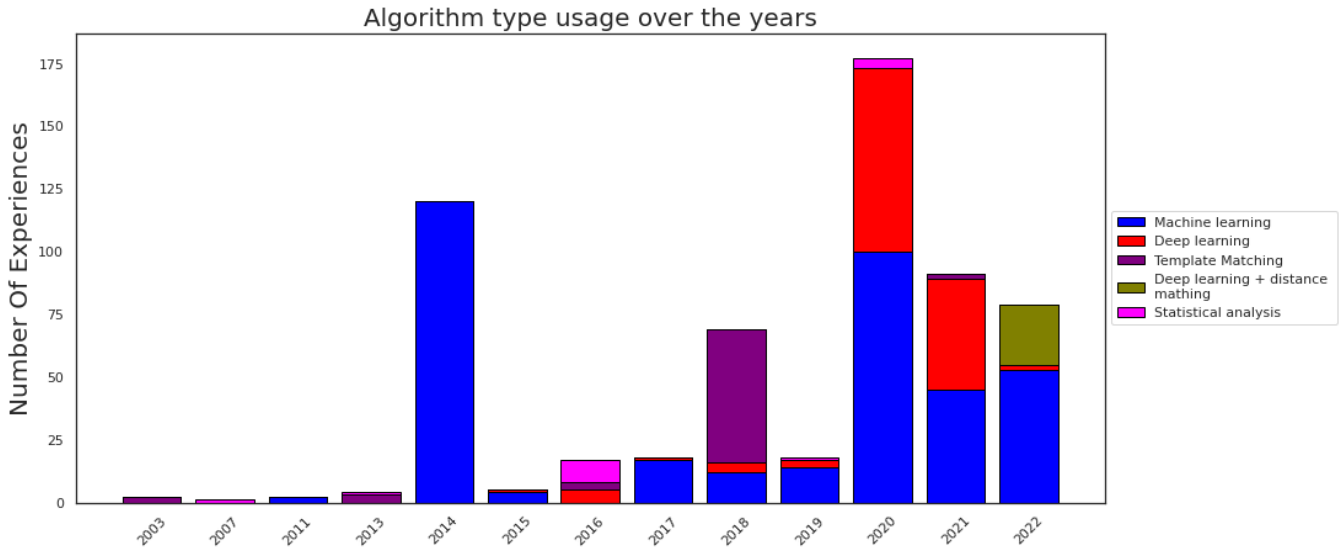


Figure 33. Algorithm type usage over the years

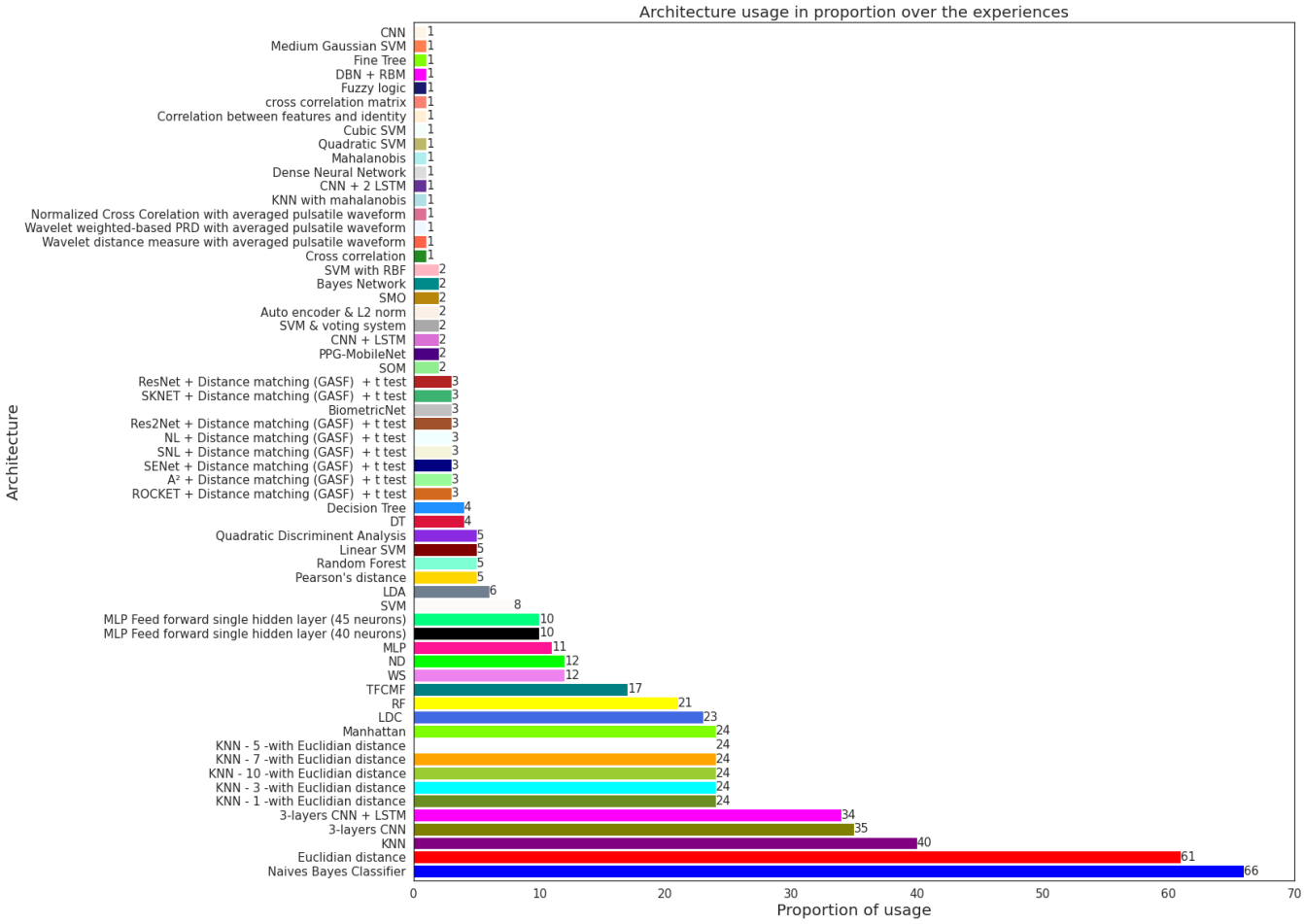


Figure 34. Algorithm architecture usage over the experiences

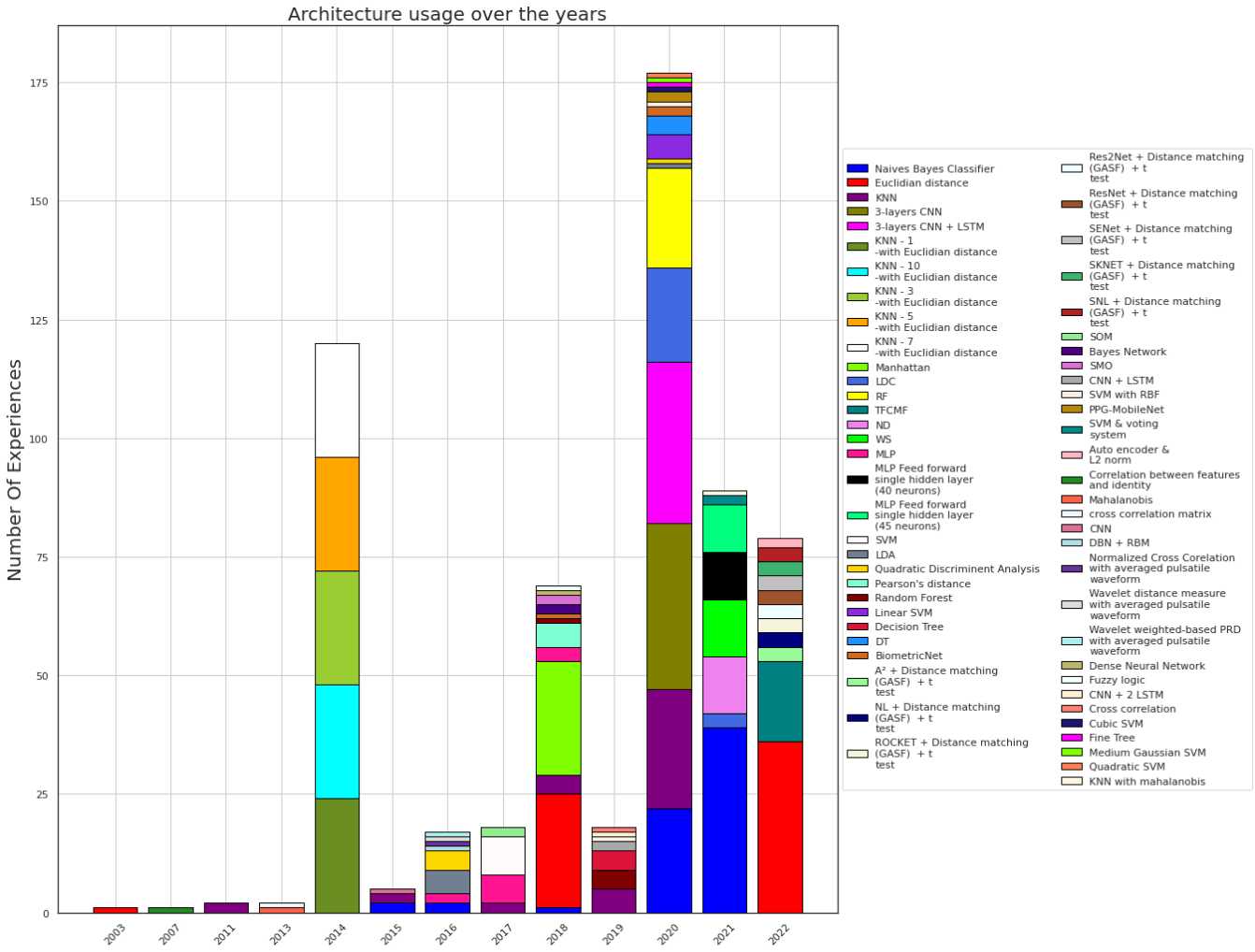


Figure 35. Algorithm architecture usage over the years

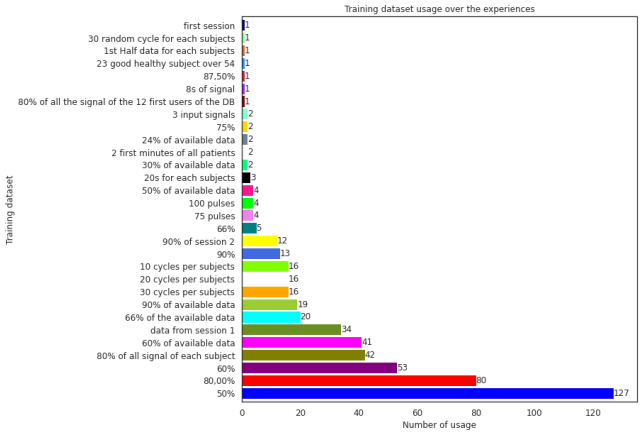


Figure 36. Training datasets usage over the experiences

[39] present accuracy, precision and recall for cluster identification but not for each individual or the mean scores for all subjects. [5] provide only accuracy, specificity and error rate for only one experience, and only for a subset of subjects (14 over 57). The values are biased and non exploitable.[33] provide some sensitivity metrics for the single session experiences only.

Finlay [40], [41], [82] , [35], [34], [85] 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 which use similar data sets or which provide a comparison on multiple algorithm or methods. We compared works made by [65] [82] and [84]. They compare multiples algorithms and methods, using multiples online databases. They have one data set in common which is the Canopbase IEEE TBME dataset. As stated earlier, this dataset provide only one signal record for each subject so it can be used only to test short term scenario use case.

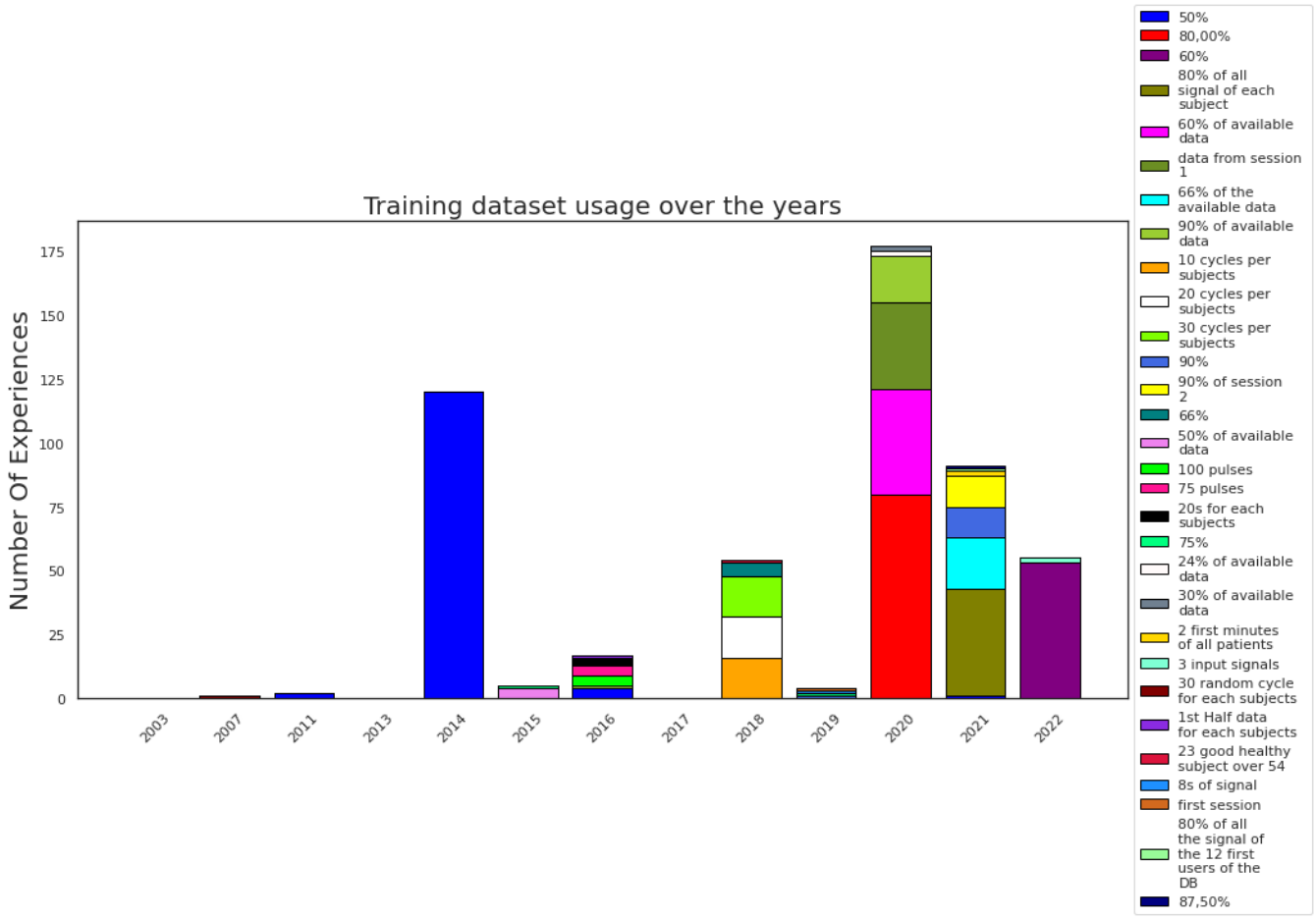


Figure 37. Training datasets usage over the years

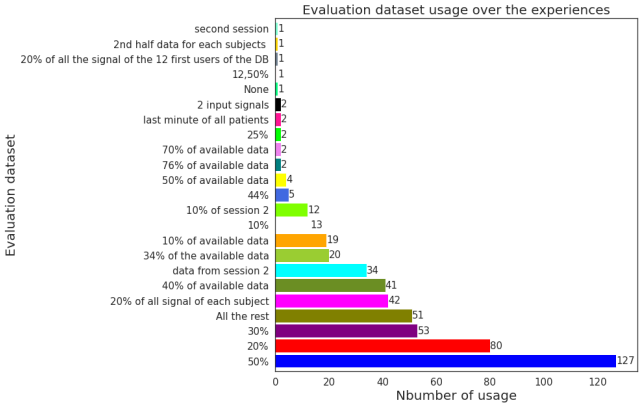


Figure 38. Validating datasets usage over the experiences

Figure 43 show the accuracy of each algorithm in experiences that use the Capnobase IEEE dataset. We only keep the experience that provide the accuracy and the architecture used. Thus lead to 112 exploitable experiences. On On Tab 7 we observe that most of the experience provide an accuracy between 90% and 100%. However we observe very few experiences for each architecture, the three more tested architectures on this dataset are NBC (20), KNN (13) and Euclidean distance matching (12). The Naives Bayes Classifier show the lowest mean performance but has two outliers and the biggest variability. KNN provide good accuracy in mean (98.54%) and low variability with std=1.85. The Euclidean matching show 94.4% of mean accuracy and and std=2.07. The architecture that seems to provide the best performances with the best stability is the 3-Layer CNN with 99.25% of mean accuracy and std=0.89. However only 6 experiences were conducted with this architecture.

The first study we analyze is the one made by [65] where they compare multiple features extractors with classification using Manhattan and Euclidean distances. Plus they compare the difference between single session (short term) and multiple session recording (long term). To compute their EER, they train and test each data set separately and then made a mean EER. They mainly use 30 cycles for enrollment and testing. They try with fewer cycles but they show that using 30 was a good tread-off between training time and

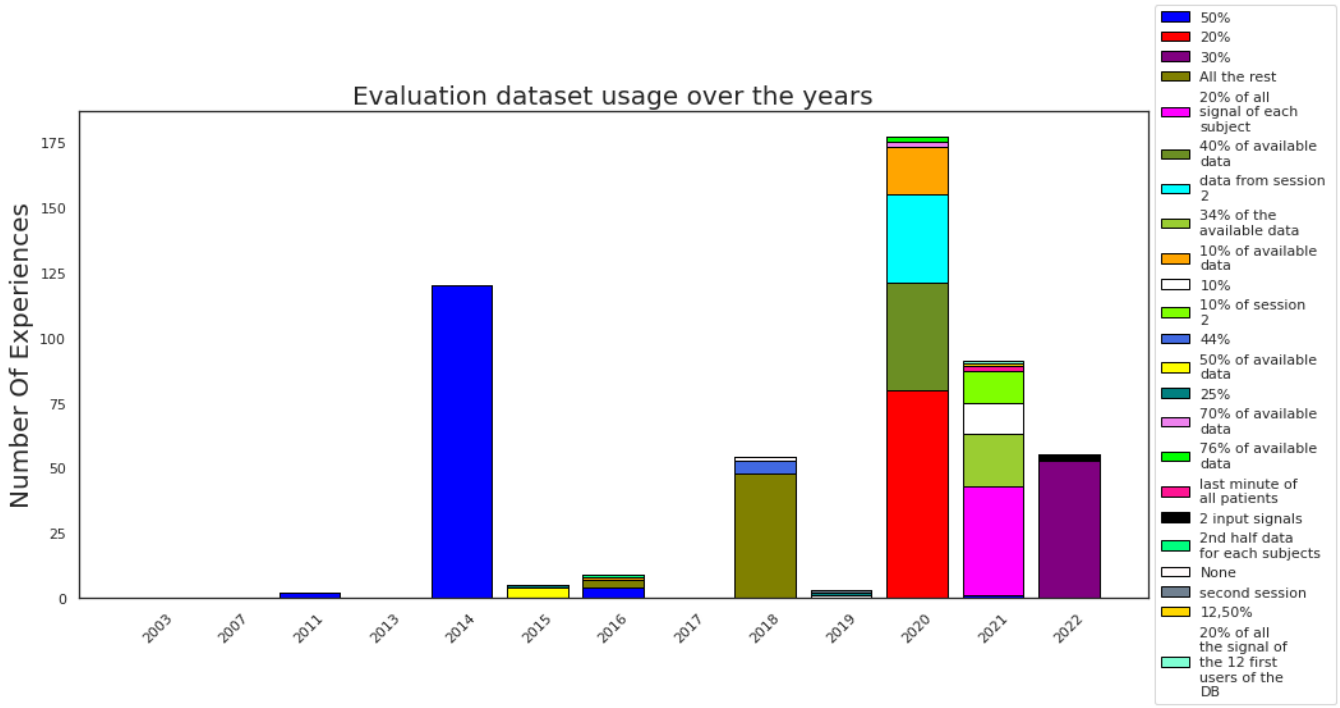


Figure 39. Validating datasets usage over the years

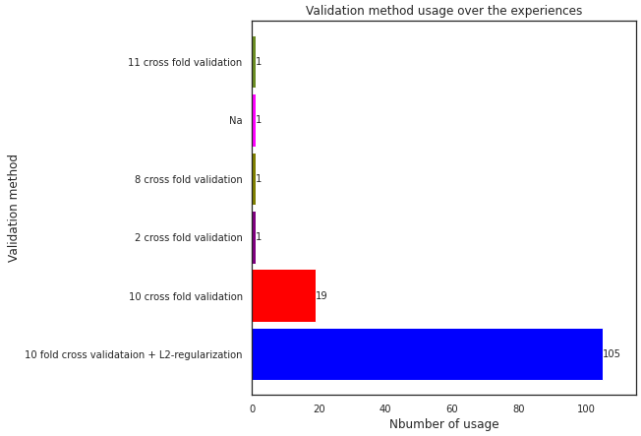


Figure 40. Validation usage over the experiences

accuracy. Moreover, they show that using multiples cycles to extract feature decrease EER. However the difference is not impressive, they gain approximately 1%. However, the EER with long-term sessions is quite higher, up from 8% to 24%. This show two things: PPG signal can variate a lot between two recording session, and their methods do not generalize well.

Another good study is the one made by [82], where they compare mostly the feature selection algorithm. With the Canopbase data set, they use a CWT transformation to extract features and use a Pearson's distance [15] matching methods to recognize users. To select the best features, they compare numerous techniques: DLDA, KDDA, KPCA, LDA, PCA, and LDA. The training data set where the 45 first seconds of each user's signal. Then the testing data was a random segment of the available signal, with duration included between 6 s and 7 s. They show that DLDA was the best feature selection algorithm because it had the lowest EER rate of all (0.46 %). So they use it with other data set. They achieve EER between 2 % and 3 % except with the exercise sessions of the Biosec Database which is 5 %. In this study, the accuracy rate is not calculated.nThis study shows a good comparison between feature selection methods but this is not enough. With this study, we can say that DLDA works fine with CWT extraction and Pearson's metrics, but if we use another feature extraction and another classification method, we could probably find another good combination.

In their work, Yang et al, [84] used three databases, apply the same methods to each dataset and provide an accuracy rate. The three datasets are BIDMC, MIMIC-II, and Canopbase. To extract features they develop a new algorithm using a three-layer model that produce a sparse softMax vector. Here 80 % of available data is used for training and 20 % for testing. Then they test k-NN, Random Forest, Linear Discriminant Classifier, and Naive Bayes as a classifier. For the Canopbase, they obtain an accuracy rate from 97.59 % with Random Forest and 99.92 % with k-NN. These accuracy rates are quite constant with other data set. They also test their

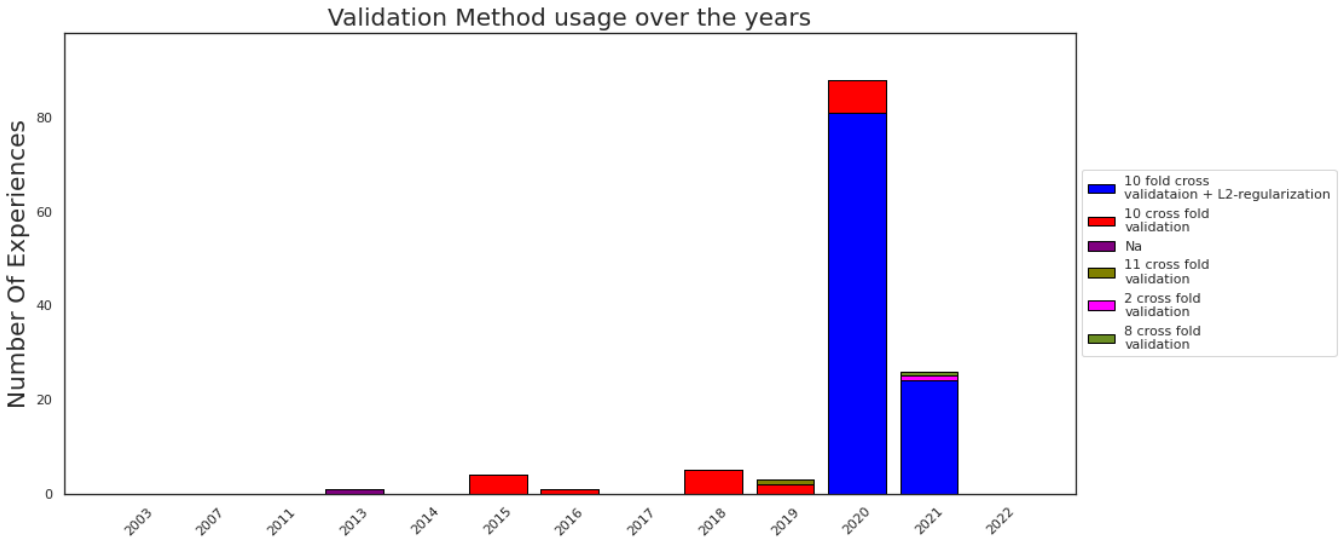
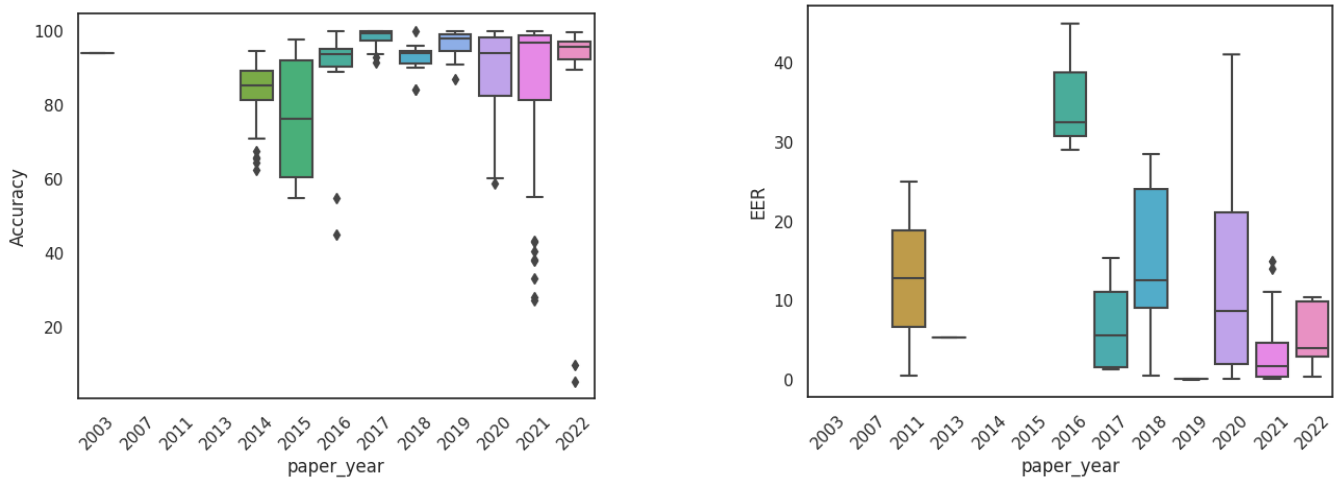


Figure 41. Validation methods usage over the years



(a) Accuracy evolution over the years

(b) EER evolution over the years

Figure 42. Evolution of the accuracy and EER over the years

methods with one layer and two layers for extracting features, but they show that the 3-layer model was the best. From their work, k-NN seems to be a little more efficient than the three others. However, no tests were made with genuine and impostors to measure an EER. This is a great comparison of some machine learning algorithms, however, it's not enough to generalize to all algorithms. Moreover, extraction features played a lot, and this method, in general, may be a good one, but others can exist.

9.5 Time stability : One session VS Two sessions

Kavsaoglu et al [42] were the first teams to use two separate sessions for training and testing. However they did not provide any precision on the time between the two sessions. So their experiences can not be used to test long time stability. The first work to spot the interest of long time stability was done in 2020 by Hwang et al. [35]. In their paper, they dedicated one experience to study the effect of using one session for enrolling and another one for testing. They used the Biosec1 and Biosec 2 datasets, where the one provides two sessions in the same conditions but recorded with 14 days of differences. The second one provide two sessions, recorded the same day but in two different state : relaxed for the first one and in exercise for the second. They conduct the same experience than the single session scenario and

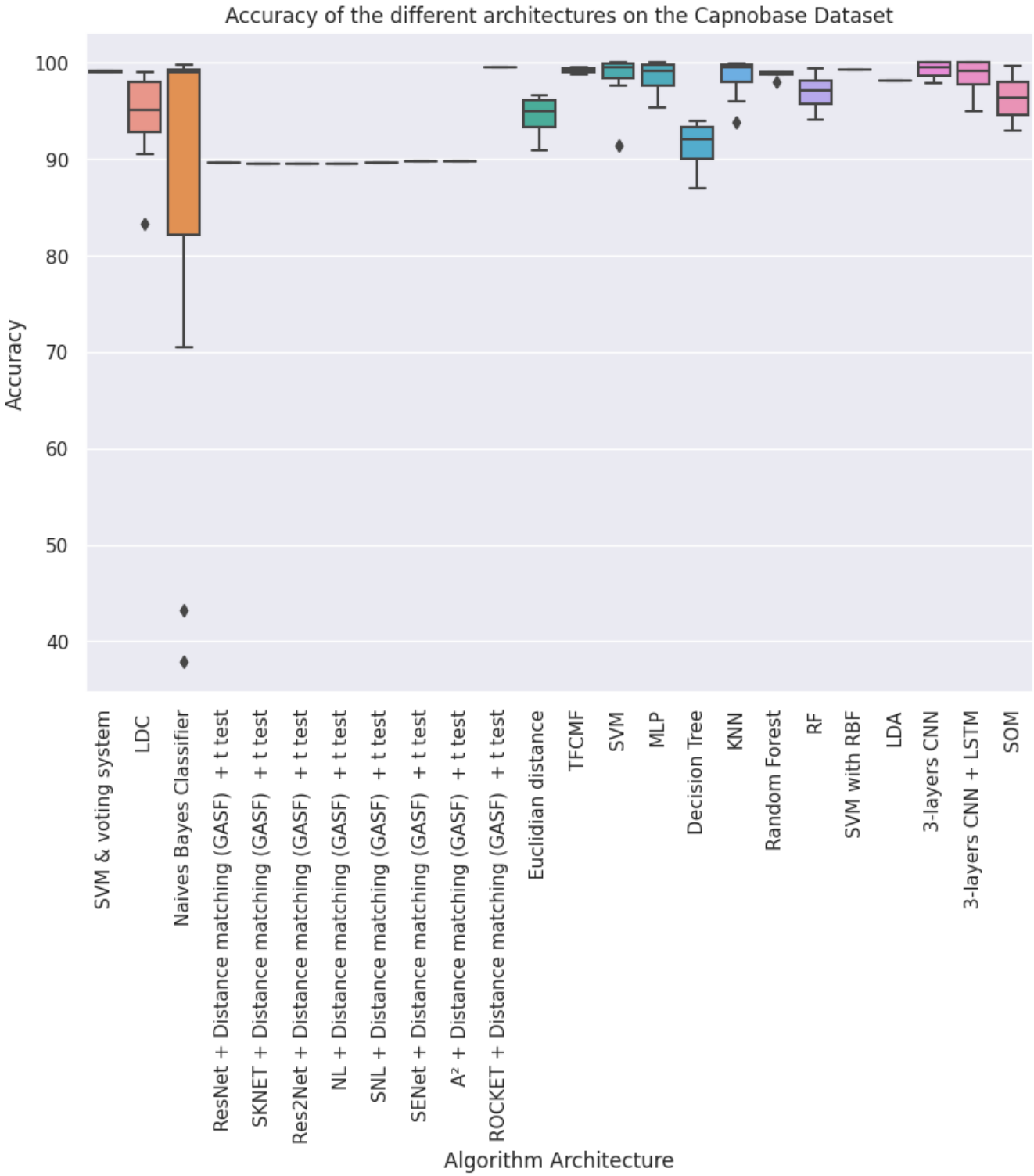


Figure 43. Accuracy of the different architectures using the Capnobase dataset

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

show a significant drop of the performances. In all their experiences, they show a drop of accuracy around 30% in mean and the EER increase up to 41%. In single session scenario, the accuracy ranges from 91% to 100% and the EER range from 0.1% to 10%. In the two session scenario, the accuracy range from 58% to 81% and the EER range from 18% to 41%. The second works on this topic was made in 2021 by the same team [34] where they try to increase the stability of the performances using a Generative Adversarial Network (GAN) technique. They did not provide the accuracy of the single session scenario but only the EER which range from 0.1% to 15%. For the two session scenario they did not provide the average EER but only the accuracy. They range from 77% to 88%. This show better stability than in their previous paper. However their is still a huge gap between the two scenarios. Moreover in both paper we can observe that the performances can change a lot for each dataset. In general we observe one different algorithm that outperform the other for each dataset. Their is no architecture witch outperform all the other with all the datasets.

These two papers show the need to investigate further the time stability of the PPG biometric recognition. They allow us to partially answer the *RQ 1.2): the biometric authentication using PPG are still unstable in long term scenario*. However the performances in the first study are good enough to encourage the investigation on this topic. This use case must be include in a dedicated benchmark of the algorithms. Moreover we need more dataset to study this phenomenon. We need datasets with records taken on multiples days, during multiples hours and if possible during 24h.

10 Future works

We saw through our analysis that many issues occur in all studies on human authentication through PPG. Some bias in the data set, learning, testing, noise suppression, etc remain. Hence, we want to provide tracks for the community in order to increase the quality of studies. We propose solutions for the data set constitution, a testing protocol to measure performance with less bias. Then we propose a benchmark method that will be implemented in future work.

10.1 Database or federated learning

As we stated at the end of Section 5, most of the data are acquired in a controlled environment. Most of them ask the patient to sit and relax which is unrealistic in real-world situations. For example to open a door, if someone takes a run and wants to go back home, we can not ask him to sit, wait to cool down to open the door. Moreover, [65] shows that the PPG signal can vary a lot from one record to another. This is why we need to create a big data set, with as many patients as possible, with multiple records, taken in multiples conditions. The best will be to record the patient's PPG over one full day, during at least two sessions. The BioSec data set [83] made by the University of Toronto started to gather data with this approach. The community needs to provide more data set like this one, with different materials. In real life, multiples PPG sensors coexist, with multiple sampling rates, different frequencies are used to measure PPG, populations are all different, and algorithms have to work for all. This is why we need a more heterogeneous dataset.

Moreover, we can merge some datasets and provide a bigger set that can be a reference to compare algorithms. The main goal is to achieve a robust authentication system. To do that we need as much data as possible. This why it is interesting to create a big database where each team can add a small amount of data. It's hard for a research team to gather more than 20 volunteers and this is why we observe a huge variability in the total number of patients. Adding gathered data to one public database can be a solution to achieve less variability in experience and would help to compare algorithms. It will also help to determine how an algorithm scale with the number of patients. Such authentication systems won't be employed with a reduced number of subjects. One key point of an authentication system is its ability to be used by the highest number.

However, biometric data are really sensitive, and privacy concerns make it harder for research teams to publish them. To address this problem of privacy of biomedical data, a new paradigm can help: federated learning [73]. With these new methods, we can distribute the learning phase of a deep learning algorithm through multiple centers. Then each center train the algorithm with its data without the need to publish them.

10.2 benchmark

One big result 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 model 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 provide 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 multiples performances metrics representing the security of the system, it's 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 futur works, we will propose one benchmark methods that provide at least all of this metrics. Moreover, we will fix the dataset for enroll and test because this splitting can influence a lot the metrics. Then we will apply this benchmark to the maximum of different architecture in order to find the bests.

11 Conclusion

In conclusion, we gather 44 studies with the same goal: creating an efficient way of authenticating people through PPG records. We extract around 600 experiences made during the twenty past 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 objectives criteria but the heterogeneity of the gathered studies leads to the impossibility to clearly define which method is the best, or the advantages and disadvantages of each part. Finlay we were able to compare some studies and we are able to answer some of our research questions :

11.1 RQ 1.1.

The performances in short term scenario are quite good for most of the tested architecture and can be exploited.

11.2 RQ 1.2.

The performances in long term scenario are less good than in short time scenario. The drop of performance is around 20% which is not too much and the systems are still better than random choice. However they are not good enough to

be used in real world use case. Thus further research on this topic are needed.

11.3 RQ 1.3.

In general, the dataset are composed with less than 50 subjects. Only the Biosec 2, VORTAL and VITAL provide 100 and more subjects. However few tests have been done with the full Biosec2 dataset and non where conducted with the full VORTAL or VITAL dataset. The performances with the whole Biosec2 dataset are good enough to be used in real world but, further studies are needed to confirm that. Moreover, we need to build clean dataset with 1 000 and 10 000 users at least to be able to confirm a full scaling up.

11.4 RQ 1.4.

Very few experiences have been conducted to test this hypothesis. Only 8 experiences using the DEAP dataset providing records with different emotional state and only 11 woth the TROIKA that provide signal recorded in physical exercise and at rest. The results are a little bit lower (between 95% and 96% of accuracy at best) than other but this may not be significant due to the few number of experiences. This show that a biometric authentication based on PPG may be robust to physiological change but further research are needed.

11.5 RQ 2.1 and 2.2

The validation methods and the metrics provided in the studied experiences don't allow us to answer to these two research questions. In deed, none of the selected studies compute a Failed To Enroll metrics or the mean number of tries that an user need to be authenticated. Theses metrics should be include in the next studies on this topic.

11.6 RQ 3.1

Considering one common dataset, 24 classification architecture have been tested. If we consider the whole architecture (feature extraction, selection etc.) we obtain 112 architectures. If we consider all the experiences with all dataset, 315 different architectural combination have been tested. However the test must be replay with correct dataset and validation methods.

11.7 RQ 3.2

Considering all the dataset and all the possible elements or architecture (noise reduction, feature extraction and selection, segmentation etc.) we observe : 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 theses parameters we can define 169 495 774 560 different pipeline architectures. This give us a coverage of $1.85.10^{-7}$% which is very low. Moreover, many methods can also be added for each parameter. This show 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 architecture are more efficient than other, the neural network based architectures seems the better ones. However we need to test this hypothesis with more experience to ensure this answer. For the 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 influence the final score and some pieces may works in synergy for this problem while other don't. We need further investigation .

11.9 RQ 3.4

Some classification architecture have been more used than other. The segmentation in single cycle is very popular, the Butterworth filter were the most used to reduce noise. In the normalization usage, it's the zero-mean normalization that is the most popular. However this popularity is not clear as for the Butterworth filter or the single cycle segmentation. For the extracted features, the fiducial features where very popular at first, but progressively many other kinds of features where tested. The frequencies extracted with FFT seems a little bit more popular than the others. For the selection process, it's seems that the most popular is to select all the extracted features. However this popularity is very low because it's only represent 10% of the experiences but it's the most reused over the papers. There is no popularity for the feature size. 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 for this element. Specific experiences should be done to determine if the popularity of theses elements is due to easiness of implementation or if they are truly better than other. 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 dataset over the years which provide the same base for every one. But the validation methods still lacks and we need to define one unique methods and benchmark all the tested experience. 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 it's 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 all teams can test their algorithms. This will allow us to provide better answers for 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] Jorge Blasco, Thomas M Chen, Juan Tapiador, and Pedro Peris-Lopez. 2016. A survey of wearable biometric recognition systems. *ACM Computing Surveys (CSUR)* 49, 3 (2016), 1–35.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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

[25] 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.

[26] 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.

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

[28] 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.

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

[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] Olivier Rioul and Pierre Duhamel. 1992. Fast algorithms for discrete and continuous wavelet transforms. *IEEE transactions on information theory* 38, 2 (1992), 569–586.

[62] 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

[63] 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.

[64] 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.

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

[66] 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.

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

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

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

[70] 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.

[71] 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.

[72] 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.

[73] 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

[74] 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.

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

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

[77] 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.

[78] 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.

[79] 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

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

[81] 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.

[82] 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.

[83] 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)*. 277–282. https://doi.org/10.1109/ICB2018.2018.00049

[84] 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).

[85] 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).

[86] 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.

[87] 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.

[88] 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.