DOI: 10.1016/j.imavis.2020.103980 | Image and Vision Computing - Volume 102, October 2020

# Demographic classification through pupil analysis

Virginio Cantoni[a], Lucia Cascone[b], Michele Nappi[b], Marco Porta[a]

[a] Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Via A. Ferrata 5, 27100 Pavia, Italy
[b] Dipartimento di Informatica, Università di Salerno, Via Giovanni Paolo II 132, 84084 Fisciano (Salerno), Italy

**Abstract**
An area of biometrics that has recently attracted much attention is gender and age classification. Its applications can be found not only in the fields of security and surveillance, but also in the context of marketing and demographic information gathering. In addition, extracting this information from a biometric sample can help to decrease the time to identify the exact individual. In this paper, we exploit pupil size as a discriminating feature for the estimation of gender and age. Data obtained from the free observation of face images have been used to train two classifiers (Adaboost and SVM), considering both the best results produced by each classifier and their fusion through weighted means. With experiments involving more than 100 participants, we have found that pupil size can provide significant results, better than those achievable using data on fixations and gaze paths. Pupil Diameter Mean (PDM) has proved to be the best discriminating feature for both gender and age. To the best of our knowledge, there are no other studies trying to perform such a classification using pupil size only.

**Keywords**: Biometrics, Pupil analysis, Gaze analysis

## 1. Introduction

Biometric traits can be *physiological* (such as fingerprints and iris) or *behavioral* (like gait and eye movements) [25]. Distinctive behaviors are normally developed by people depending on their cultural, social, and psychological experiences [34]. Behavioral biometrics is practically always *soft biometrics* [15], which means that it needs to be used together with other recognition methods because, alone, it cannot achieve very high recognition rates. Gaze analysis is typically soft biometrics.

Eye movements can be basically classified as *saccades* and *fixations*. Saccades are extremely fast (< 100 ms) shifts between fixations, which are longer periods (normally with durations between 100 and 600 ms) of relative stability of the eye.
The connection between gaze behavior and "emotional states" has been considered in different contexts, from e-learning [3][31] to subtitle reading in movies [29]. In particular, attention is interrelated with gaze behavior, as stated by the Eye-Mind Hypothesis [17] — what people look at reflects what they think or attend to. For example, few blinks are associated with those tasks that require greater attention and concentration. The inhibition of blinks is necessary to minimize the loss of information caused by the interruption of visual perception. Therefore, in conditions that require considerable attention investment, the number of blinks is widely decreased. Instead, a large number of fixations suggests a difficulty in the interpretation of information, while a longer fixation duration can have different interpretations: more demanding cognitive processing or greater interest. The saccadic velocity, on the other hand, is a reliable indicator of fatigue.

Also, we can distinguish between *overt* and *covert* vision, with this second "unconscious" modality being directly connected with the psychological and cognitive processes of a person [14]. Overt attention can be directly observed in the alignment of oculomotor resources with the interested target. Instead, in covert attention there is an inhibition of the saccadic response, so the attentional change is not evident in the form of eye movements.

Eye data are obtained through *eye trackers* [7], i.e., devices able to detect the user's gaze position (typically, but not necessarily, on a screen). The recent availability of cheap eye trackers has opened the door to new application contexts of these tools, which in the past were only exploited as assistive technologies or in the psychology and marketing fields. An eye tracker records the user's gaze position (X and Y coordinates) a certain number of times per second (e.g., 50) and indirectly detects fixations as sequences of gaze samples concentrated in specific spots. Besides gaze coordinates, an eye tracker can often also measure the user's pupil size.

Pupil size can be considered both a physical and a behavioral biometrics, as its variation is influenced by emotional and psychological factors. The pupil is the circular opening located at the center of the iris. Its size is variable and, in adults, is generally between 2 and 8 mm. Positive emotions or excitement situations such as stress and fear, induce pupil dilation, while past thoughts and negative inner feelings are associated with a reduction in size. A fully dilated pupil (*mydriasis*) has normally a size between 4 and 8 mm, while a constricted pupil (*miosis*) is typically between 2 and 4 mm (Figure 1). Obtaining information about the pupil is not invasive, because there is no need for contact, and so it is considered a user-friendly biometrics.

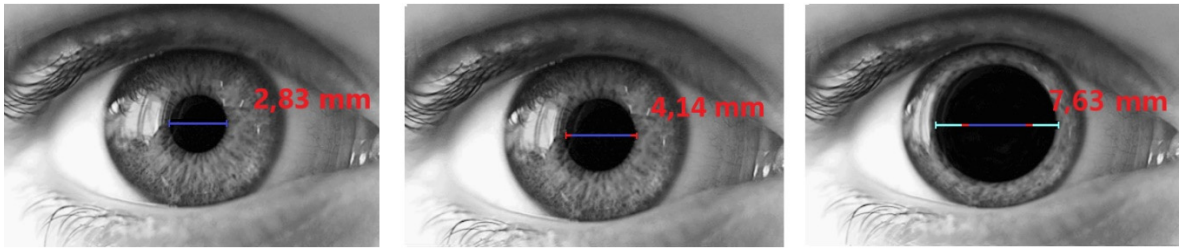



**Fig. 1.** Pupil constriction (miosis) and dilation (mydriasis).

In this paper, we present a study in which pupil size is used to identify the user's gender and age range. Specifically, we compare pupil data with those obtained from a previous investigation [9] in which gender and age were extracted from gaze behavior. The results attained show that pupil size is better than gaze data in recognizing gender and age category. To the best of our knowledge, this is the first work that tries to operate such classifications through data extracted from pupil only.

The paper is structured as follows. Section 2 shortly presents some previous works related to the use of pupil data for biometric purposes. Section 3 describes in detail the implemented method, exploring SVM and AdaBoost classifiers. Section 4 describes the data acquisition and preparation processes. Section 5 illustrates the implemented experimental protocols. Section 6 shows the obtained results. Section 7, lastly, summarizes and discusses the achieved results.

## 2. Background

The use of eye tracking data for biometric purposes dates back to 2004, when Kasprowski and Ober [18] proposed a first research on the subject. Several works have appeared since then. In the context of "traditional" biometrics (i.e., to implement identification and verification tasks), the existing studies can be roughly classified into four categories, specifically fixations and scanpaths, eye speed, oculomotor features, and pupil size.

Regarding fixations and scanpaths, reading scans [13], face stimuli [2], and several fixation and saccadic features [10] have been exploited. As for eye speed, basic velocity and acceleration have been used [1], as well as normalized velocity histograms over the travelled eye angles [20], saccadic movements [16][33], and combined features [36]. Oculomotor characteristics have been mainly analyzed through complex oculomotor plant models [18][22][23].

Studies of pupil as a biometric indicator are few, although some investigations encourage its use as a soft biometrics for both verification and identification tasks [1][26]. The way pupil size evolves over time can provide valuable information to identify an individual or to verify a declared identity. This biometric trait not only allows the use of non-intrusive acquisition systems, but is also characterized by a strong aptitude to be combined with other biometric features to improve the overall selection power of the system.

Pupil size is influenced by physical factors such as light, chemical components such as certain medications and the abuse of alcohol or drugs, and health states connected with some pathological conditions or dysfunctions. Changes in pupil diameter can also provide important information on cerebral activity, since such involuntary variations are a real-time indicator of the activity of the nervous system: emotional factors such as fear or surprise, and cognitive aspects such as thought, learning processes, execution of mental calculations, etc., can be accessed. Pupil size is considered a good indicator of attention level [24], fatigue [37], emotional states [30], and other factors as well.

Pupil diameter is an individual measure, as it varies from person to person, but it also changes over time: young people have larger pupils than the elderly and children. Indeed, pupil size increases from childhood to the adolescent phase, where it reaches its maximum, and then gradually decreases in subsequent years [11].

Pupillary changes have also been studied to investigate the reaction of people to images of objects that they considered unpleasant, neutral, or pleasant. A larger pupil diameter in pleasant and neutral images has been recorded for women compared to the values of men, but smaller than those found for unpleasant images [32]. Further investigations have shown that, also for neutral auditory stimuli, pupillary responses are significantly greater in female subjects [28].

From an anatomical point of view, an emmetropic eye is a non-pathological, normal eye, and it has been demonstrated [35] that pupil diameters are greater in emmetropic women than men.

Ortiz et al. [27] proposed a linear regression model to evaluate pupil dilation as a function of age. With age, not only pupil diameter decreases, but also the pupillary response to a stimulus loses speed [19] and the dilation elicited by an emotionally unpleasant image reduces [12].

In order to improve the accuracy achieved through pupil size alone and overcome the limits of a single biometric system, pupil size could be also fused with information extracted from other biometrics (e.g., features of the face or periocular area, such as blinks, saccadic movements, fixations, etc.).

## 3. The Method

The proposed method (Figure 2) is based on the most popular and efficient learning models for addressing classification problems: Support Vector Machines (SVM) [6] and AdaBoost. These two techniques for learning both having received a significant attention in the last years and many successful applications have been described in the literature [38][39]. After, the best results

produced by each classifier have been fused through weighted means where the weights were the average accuracies reached by classifiers and the gender/age accuracies for each classification class.

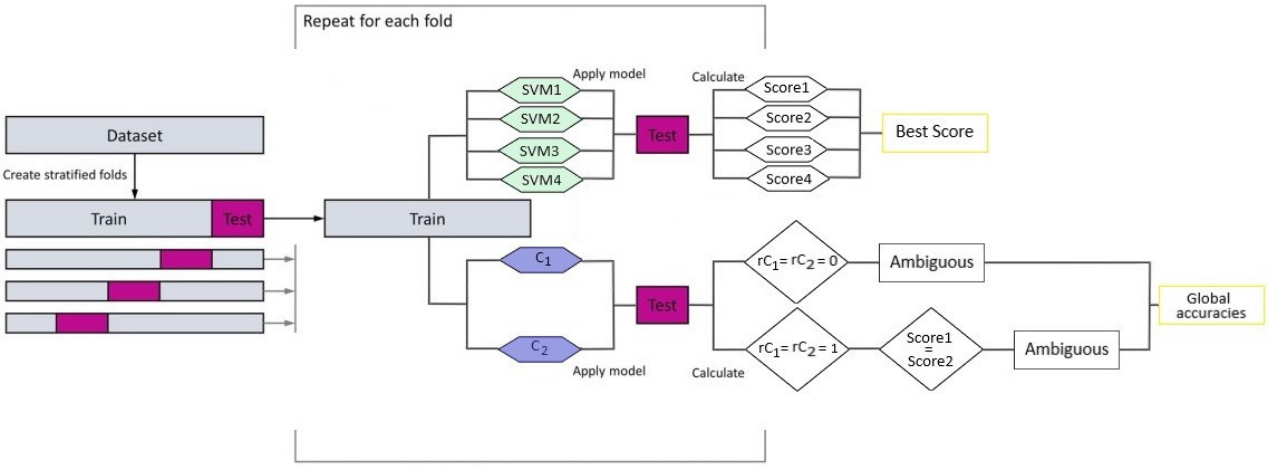



**Fig.2.** Proposed method: using 4-fold cross validation each training partition has been provided to classifiers. See 3.1 and 3.2 sections for more details.

### 3.1 SVM classifier

In Machine Learning, the classification problem can be addressed using supervised learning algorithms. One of the most common methods of supervised classification is the SVM algorithm. SVM can be used to solve many practical problems about classification, regression and other learning tasks. It has many applications in text categorization, sentiment analysis, diagnosis services, etc. SVM was originally formulated for binary problems, i.e. the classification of two classes. It is precisely on this type of problems that its performance is better, and in fact the extension to multiclass problems is not easy. SVM is particularly good at drawing decision boundaries on a small dataset.

Given a labeled training set of N data points $\{y_k, x_k\}_{k=1}^{N}$, where $x_k \in \mathbb{R}^t$ and $y_k \in \{-1,1\}$, and -1 and 1 indicate the class to which the point belongs, the basic idea behind SVM is to find an optimal hyperplane that distinctly classifies the data points maximizing the margin between the classes' closest points (Figure 3). The larger the distance of data points from the hyper-plane, the better the separation of the analyzed classes is. The form of the classifier is the following:

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i \, y_i K(x_i, x) + b\right) \qquad (1)$$

where $\alpha_i$ are positive real constants, $b \in \mathbb{R}$ is a bias, and $K(x_i, x)$ is the kernel function. The kernel can be any of the following:

- *linear* if $K(x_i, x) = x_i^T x$
- *rbf* if $K(x_i, x) = \exp\{-\left|\left|x - x_i\right|\right|_2^2/\sigma^2\}$
- *polynomial of degree d* if $K(x_i, x) = (x_i^T x + 1)^d$

where $\beta, \sigma, \theta$ are constants.

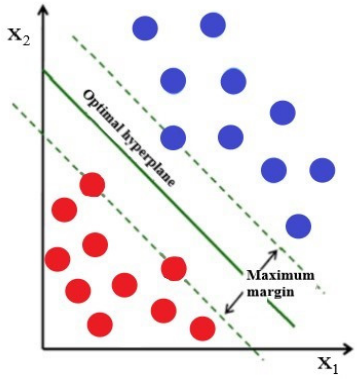



**Fig.3.** SVM classifier.

For our study, using K-fold cross validation [21], an SVM classifier has been trained through training sets balanced according to class subject of analysis (gender or age). Feature vectors in the test sets have been provided to the classifier to obtain single results, then averaged to obtain global accuracies.
Four different models with different parameters (among which the kernel) have been trained (Figure 2), and the best results have then been selected. SVM performance is highly dependent on parameter setting and kernel selection. The choice of SVM parameters and kernel functions obviously affects learning and generalization performance. We have implemented four different strategies:

1. Linear kernel function. Regularization parameter (C) set to 1. No limits on iterations.
2. Similar to SVM with parameter kernel= 'linear', but implemented in terms of 'liblinear' rather than 'libsvm'. Since there are only two classes, only one model is trained. C set to 1. The maximum number of iterations to be run is set to 10,000.
3. Rbf kernel function. Kernel coefficient (*gamma*) set to 0.7. C=1. No limits on iterations.
4. Polynomial kernel function. Degree of the polynomial kernel function set to 3. Gamma = 1 / (*number-of features* * *matrix-variance*). C=1. No limits on iterations.

### 3.2 AdaBoost classifier

AdaBoost, or Adaptive Boosting, is an iterative ensemble boosting classifier. AdaBoost is an efficient learning algorithm to build accurate classifiers. It is one of most widely used boosting algorithms in different fields such as diagnostic, dynamic gesture recognition, security of mobiles intelligent terminals [40], and so on. The basic concept of AdaBoost is to combine multiple "weak classifiers" into a single "strong classifier". Let $\{y_k, x_k\}_{k=1}^{N}$ a labeled training set of N data points where $x_k \in \mathbb{R}^t$ and $y_k \in \{-1,1\}$ and $\{D_1, \dots, D_t\}$ a set of weak classifiers where $D_j(x_i) \in \{-1,1\}$, AdaBoost assigns equal weights to all the training points at first. From the training set and the distribution of the weights at the round $j$ the algorithm generates $D_j$. Then, it uses the training examples to test $D_j$, and the weights of the incorrectly classified examples will be increased. So, an updated weight distribution is obtained.
This process iterate until the complete training data fits without any error or until reached to the specified maximum number of estimators. Therefore, the last result is a combination of the predictions from all classifiers through a weighted majority vote:

$$D(x) = sign\left(\sum_{i=1}^{t} \alpha_i D_i(x)\right) \quad (2)$$

where $\alpha_i \in \mathbb{R}$.

In our experiment, using K-fold cross validation, each training partition has been provided, as an input, to 100 weak classifiers forming a strong classifier [5][8]. The base estimator is a decision tree classifier with the maximum depth of the tree set to 1. Testers in the partitions have been subdivided into balanced groups according to their class (gender or age). Two separate classifiers, $C_1$ and $C_2$, have been trained for the two sub-classes ('male' and 'female' for gender and 'under 30' and 'over 30' for age). Vectors from the test sets have been presented to both $C_1$ and $C_2$, obtaining two values, $rC_1$ and $rC_2$, from each, namely a *response* (a Boolean output that is true if the predicted class is correct) and a *score* (a real number reflecting the classifier's accept/reject confidence of the sample). Those samples for which both $C_1$ and $C_2$ have returned the same score have been considered "ambiguous"— with an associated label only in the "true" case, however. Ambiguous responses have not been included in the computation of the global classification accuracies (Figure 2).
K-fold cross validation results have been averaged to obtain global accuracies.

## 4. Data acquisition and preparation

Our study involves two-class classification problems, i.e., the distinction between 'male' and 'female' for gender and 'under 30' and 'over 30' for age. The database used for the experiments [4] includes a total of 112 volunteer testers. Since the dataset is unbalanced, because it includes more males than females and more testers under 30 than testers over 30, for both gender and age we have trained models with equal numbers of samples for each class, determined by the number of females and 'under 30' testers.

The dataset derives from three different gaze recording sessions, each carried out at a time distance from the previous or next one between five and nine days. Each tester was asked to look at a sequence of 16 photos of human faces, which were labeled with a numerical value from 1 to 16. The photos included eight female and eight male faces.

Gaze data were acquired through the Tobii 1750 eye tracker. Pupil size for both left and right eyes were recorded 50 times a second, the eye tracker's sampling rate being 50 Hz. Fixations, derived from sequences of gaze samplings, were characterized by their screen X and Y coordinates, start times, and durations.
In a preprocessing phase, poor eye data (explicitly marked by the eye tracker) were excluded from the analysis.
Images were shown for 10 seconds each, in random order. Moreover, before the first image (for five seconds) and between consecutive images (for three seconds), a black cross at the center of a white background was displayed, so that the tester's visual exploration path always started from the same position.

For gaze behavior analysis, we have used the GANT method [2]. Basically, GANT (Gaze ANalysis Technique) partitions the fixations of each tester using a 6x7 grid. A different node is associated to each cell, along with weights. Three features are employed, namely *density* (number of fixations in

a cell), *duration* (sum of fixation durations in a cell), and *weighted arcs* (undirected links between pairs of nodes whose weights are given by the number of times the tester's gaze moved from one cell/node to the other).
Matrices with elements corresponding to nodes are used to represent densities and durations, while an adjacency matrix is exploited for weighted arcs. The comparison between testers then occurs by calculating the Euclidean distances between pairs of matrices.

For pupil size, we have derived the following features:

a) *Pupil Diameter Mean* (PDM): for each subject, it is the mean pupil diameter for each image;
b) *Pupil Diameter Ratio* (PDR): after calculating the mean pupil diameter for each subject over all the images, the percentages of maximum dilation and constriction with respect to this value have been computed;
c) *Pupil Diameter Order* (PDO): for each subject, the images have been sorted according to their associated average pupil size. A vector has then been created, still for each subject, in which indexes refer to the images and values indicate their position in the ordered sequence.

## 5. Experimental protocols

For our analysis, we have considered only the data from the first round of acquisitions, which involved 111 testers (one was excluded from the original 112 because of poor gaze data). Participants included 72 males and 39 females with ages ranging from 17 to 80 years (Table 1).

Table 1: Ages of testers

| Age range | Number of testers |
|---|---|
| a (17–18) | 11 |
| b (21–30) | 57 |
| c (31–40) | 9 |
| d (41–50) | 16 |
| e (51–60) | 8 |
| f (61–70) | 9 |
| g (71–80) | 1 |

However, as said, some testers repeated the experiment two more times after some days. We have then used these newly acquired data with the previously trained classifiers to predict their classes. Subsequently, the predicted values have been combined through two different weighted means where weights were the mean accuracies achieved by the two classifiers and the gender/age accuracies for each classifier class.

Using K-fold cross validation as a data selection strategy (subdivision into training and test sets), we have exploited AdaBoost and SVM. The protocol of our method is the same as that proposed in [9]. In this way, we guarantee an adequate comparison between the results obtained.

## 5.2 Gender

With the purpose of verifying whether the gender of the observed faces can affect the classification accuracy of testers, three experiments have been performed: a global analysis for both the male and female faces used as stimuli and two further evaluations for male and female faces separately. Since the original dataset includes 39 female and 72 male testers, we have considered two different experiments (*Exp1* and *Exp2*) and corresponding sub-sets with 39 male testers in each, as shown in Table 2 (which also specifies the involved testers' ID ranges).

Table 2: Experiments for gender classification

| Experiment | Stimuli (faces) | # Female testers | # Male testers |
|---|---|---|---|
| Exp1 | 8 females, 8 males | 39 | 39 (ID 002 – ID 059) |
| Exp2 | 8 females, 8 males | 39 | 39 (ID 054 – ID 112) |
| Exp1_F | 8 females | 39 | 39 (ID 002 – ID 059) |
| Exp2_F | 8 females | 39 | 39 (ID 054 – ID 112) |
| Exp1_M | 8 males | 39 | 39 (ID 002 – ID 059) |
| Exp2_M | 8 males | 39 | 39 (ID 054 – ID 112) |

## 5.3 Age

Two kinds of experiments have been carried out. In the first (indicated with *Exp1* and *Exp2* in Table 3), the 'Over 30' class is composed of age ranges *c*, *d*, and *e* in Table 1 (i.e., ages from 30 to 50, including 33 testers). In the second kind of experiments (indicated with *Exp1_10years* and *Exp2_10years* in Table 3), the 'Over 30' class is composed of age ranges *d*, *e*, *f,* and *g* in Table 1 (i.e., ages from 40 to 80, including 34 testers). The purpose of this subdivision is to verify whether an increased age gap among testers can improve the classification accuracy.

Table 3: Experiments for age classification

| Experiment | # Over 30 testers | # Male testers |
|---|---|---|
| Exp1 | 33 | 33 (ID 002 – ID 059) |
| Exp2 | 33 | 33 (ID 060 – ID 109) |
| Exp1_10years | 34 | 39 (ID 002 – ID 060) |
| Exp2_10_years | 34 | 39 (ID 060 – ID 110) |

## 5.4 Metrics

In the statistical analysis of binary classification the most used evaluation measures are recall, precision, and F1 score. The precision is defined as:

$$P = \frac{T_p}{T_p + F_p} \tag{3}$$

where $T_p$ is the number of true positive and $F_p$ is the number of false positives. The recall is defined as:

$$R = \frac{T_p}{T_p + F_n} \tag{4}$$

where $T_p$ is the number of true positive and $F_n$ the number of false negatives. The F1 score provides a single score that combines precision and recall values:

$$F1 = \frac{2(P * R)}{P + R} \tag{5}$$

where $P$ is precision and $R$ is recall.

## 6. Results and discussions

We conducted five groups of experiments to show the performance of pupil as a promising biometric trait for gender and age classifications. The experimental protocols were the same as those used in the work described in [9]. Thus, we split the dataset appropriately. We then used four-fold cross validation with two different classifiers, Adaboost and SVM. At the end, we considered gender and age classification separately, comparing the results of our pupil-based analysis to those of the previous study presented in [9]. In our analysis, we considered both best results from the two classifiers and results from their combination (data fusion) based on the two mentioned weighted means (Figure 4). The best scores of the two classifiers have been combined through two different weighted means. In the first, the weights are the average accuracies reached by each classifier. Instead, in the second weighted means, the weights are the gender/age accuracies for each classification class.

We also considered other classifiers:

a) Random Forest
b) Gaussian Process
c) KNN
d) Decision Tree
e) Quadratic Classifiers
f) Gaussian Naive Bayes

However, their average results compared to those obtained by Adaboost and SVM were not particularly significant. Thus, we decided not to include them in the paper and use the same classifiers presented in [9].

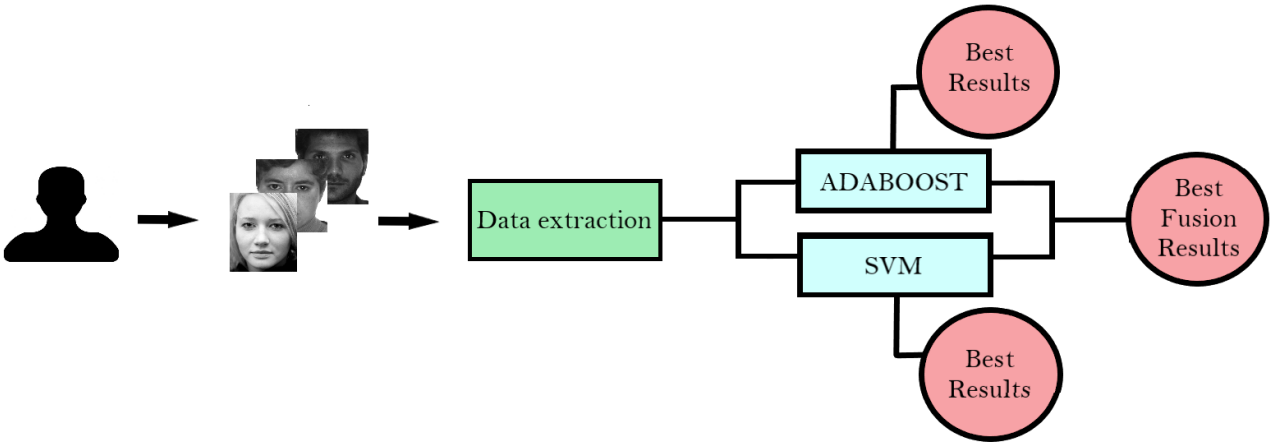


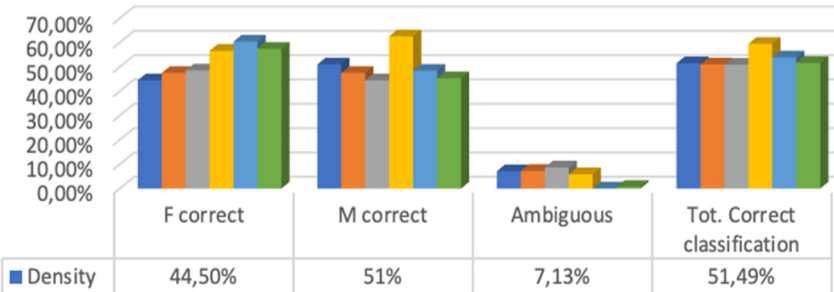

**Fig. 4.** Workflow of experiments and analysis.

### 6.1 Gender Results

We conducted three different kinds of experiments for gender classification:
1. *Exp1* and *Exp2*, where we considered all faces as stimuli;
2. *Exp1_F* and *Exp2_F*, where we examined only female faces as stimuli;
3. *Exp1_M* and *Exp2_M*, where we utilized only male faces as stimuli.

The distribution of testers in the various experiments is shown in Table 2.

In Fig. 5, the maximum values obtained with AdaBoost are shown. *F correct* and *M correct* identify, respectively, the correct classifications for female and male participants. Those situations in which the scores obtained for females and males are the same are indicated as "ambiguous". These ambiguous samples have not been considered in the calculation of the percentages of total correct classifications.

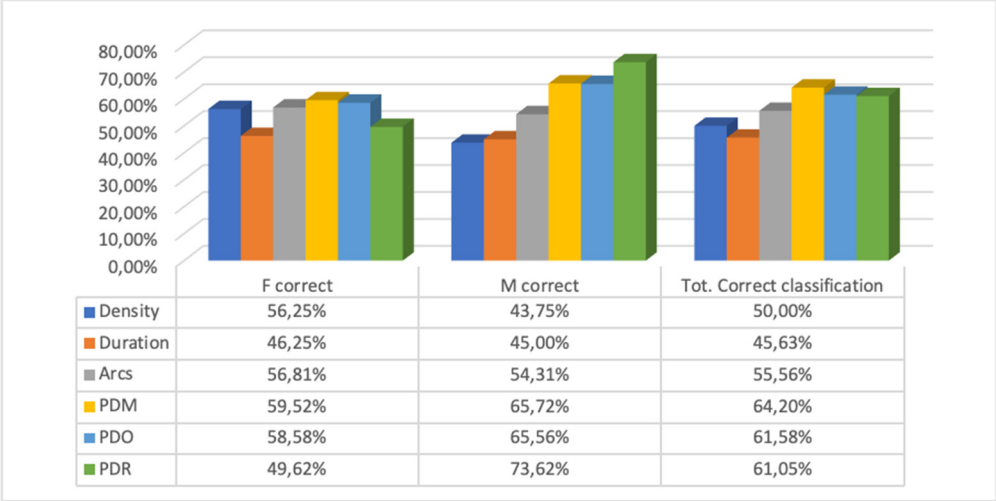



| | F correct | M correct | Ambiguous | Tot. Correct classification |
|---|---|---|---|---|
| Density | 44,50% | 51% | 7,13% | 51,49% |
| Duration | 47,50% | 47,50% | 7,25% | 50,99% |
| Arcs | 48,54% | 44,45% | 8,66% | 50,85% |
| PDM | 56,61% | 62,55% | 5,85% | 59,47% |
| PDO | 60,45% | 48,43% | 0% | 53,92% |
| PDR | 57,43% | 45,34% | 0,65% | 51,64% |

**Fig. 5.** Gender classification accuracies with AdaBoost, average scores from *Exp1* and *Exp2*.

In Fig. 6, the results of gender classification obtained through the SVM classifier are shown.

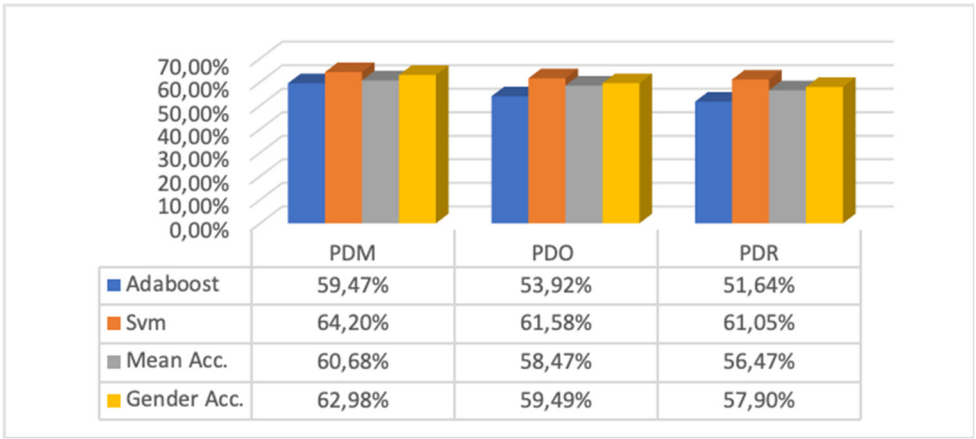
From these results, it is evident that PDM is the best discriminating index for both classifiers. For SVM, the highest accuracies have been achieved with a linear kernel for PDR and with a *gamma* value of 0.7 for PDM and PDO.
With SVM, the performances of PDM and PDR are decidedly better in the recognition of males than females, while with AdaBoost the differences are relatively significant for female recognition with PDO and PDR.




| | F correct | M correct | Tot. Correct classification |
|---|---|---|---|
| Density | 56,25% | 43,75% | 50,00% |
| Duration | 46,25% | 45,00% | 45,63% |
| Arcs | 56,81% | 54,31% | 55,56% |
| PDM | 59,52% | 65,72% | 64,20% |
| PDO | 58,58% | 65,56% | 61,58% |
| PDR | 49,62% | 73,62% | 61,05% |

**Fig. 6.** Gender classification accuracies with SVM, average scores from *Exp1* and *Exp2*.

Fig. 7 shows a comparison between the results for PDM, PDO, and PDR obtained with the two classifiers and with their combination according to the two kinds of weighted means (mean accuracy of both classifiers and gender accuracy for each classifier class). As can be seen, the gender-weighted accuracies for all three indicators are higher than the mean AdaBoost-SVM values, although they are lower than the accuracies of SVM.




| | PDM | PDO | PDR |
|---|---|---|---|
| Adaboost | 59,47% | 53,92% | 51,64% |
| Svm | 64,20% | 61,58% | 61,05% |
| Mean Acc. | 60,68% | 58,47% | 56,47% |
| Gender Acc. | 62,98% | 59,49% | 57,90% |

**Fig. 7.** Gender classification. Comparison between PDM, PDO, and PDR for the two classifiers and for their combination using two weighted means (mean accuracy and gender accuracy).

Table 4 also shows the values of precision, recall, and F1 score.

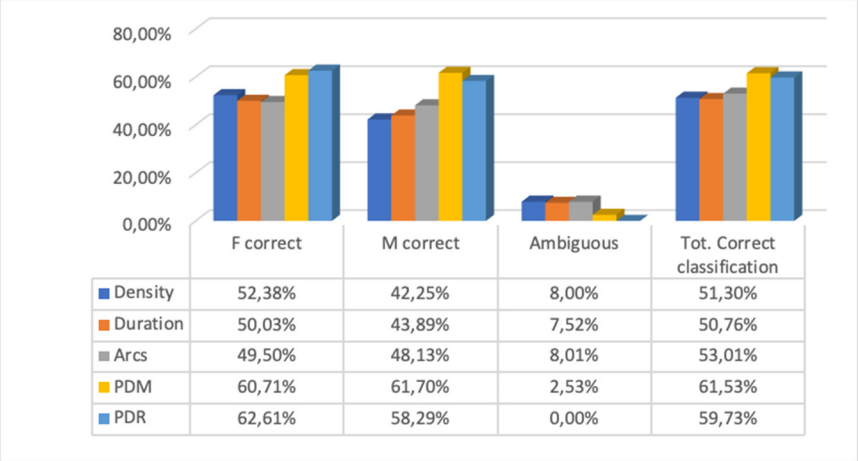
Table 4: Precision, recall and F1 score for *Exp1* and *Exp2*

| | Precision | Recall | F1 score |
|---|---|---|---|
| PDM (SVM) | 68,37% | 61,20% | 60,22% |
| PDO (SVM) | 62,97% | 61,58% | 61,12% |
| PDR (SVM) | 73,16% | 61,05% | 55,89% |
| PDM (Adaboost) | 60,34% | 59,47% | 59,21% |
| PDO (Adaboost) | 58,70% | 56,37% | 56,85% |
| PDR(Adaboost) | 53,39% | 51,64% | 50,49% |

Figs. 8 and 9 show the results achieved with, respectively, AdaBoost and SVM, considering only female faces as stimuli.

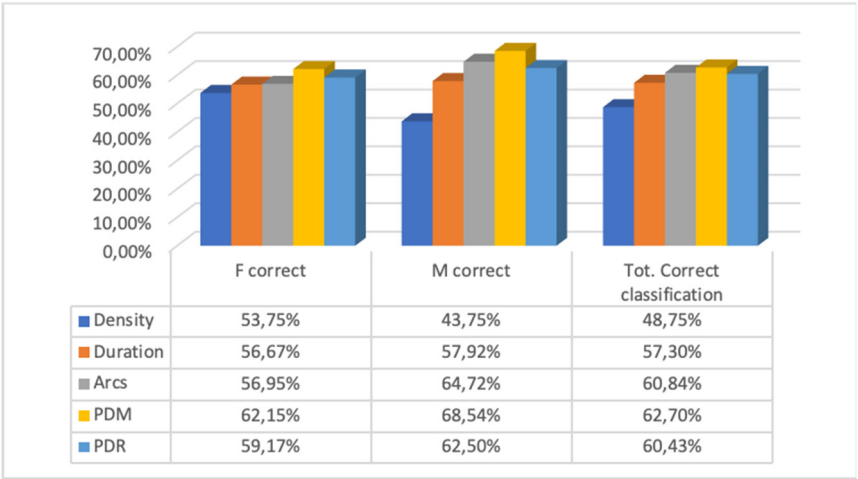
The higher discriminatory power of pupil data (PDM and PDR) compared to gaze data (density, duration, and arcs) is evident in this case.
The best results with SVM have been achieved with a linear kernel for both PDM and PDR, with a maximum of 10,000 iterations for PDM.

Fig. 10 shows the comparison between the results for PDM and PDR obtained with the two classifiers and with their combination according to the two kinds of weighted means. As can be seen, with PDR and the gender accuracy weighted mean the result is markedly higher than those provided by each single classifier.




| | F correct | M correct | Ambiguous | Tot. Correct classification |
|---|---|---|---|---|
| Density | 52,38% | 42,25% | 8,00% | 51,30% |
| Duration | 50,03% | 43,89% | 7,52% | 50,76% |
| Arcs | 49,50% | 48,13% | 8,01% | 53,01% |
| PDM | 60,71% | 61,70% | 2,53% | 61,53% |
| PDR | 62,61% | 58,29% | 0,00% | 59,73% |

**Fig. 8.** Observation of female faces only. Gender classification accuracies with AdaBoost, average scores from *Exp1_F* and *Exp2_F*.




| | F correct | M correct | Tot. Correct classification |
|---|---|---|---|
| Density | 53,75% | 43,75% | 48,75% |
| Duration | 56,67% | 57,92% | 57,30% |
| Arcs | 56,95% | 64,72% | 60,84% |
| PDM | 62,15% | 68,54% | 62,70% |
| PDR | 59,17% | 62,50% | 60,43% |

**Fig. 9.** Observation of female faces only. Gender classification accuracies with SVM, average scores from *Exp1_F* and *Exp2_F*.

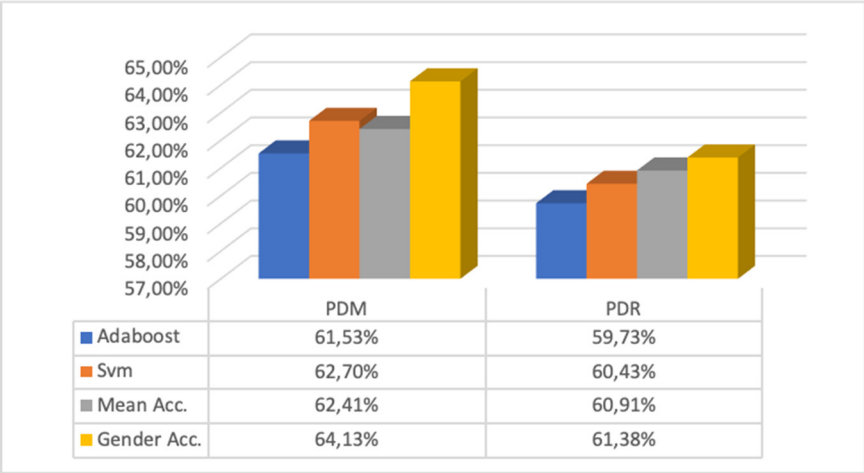



**Fig. 10.** Observation of female faces only. Comparison between PDM and PDR for the two classifiers and for their combination using two weighted means (mean accuracy and gender accuracy).

Table 5 shows the values of precision, recall, and F1 score.

Table 5: Precision, recall and F1 score for *Exp1_F* and *Exp2_F*

| | Precision | Recall | F1 score |
|---|---|---|---|
| PDM (SVM) | 70,40% | 62,70% | 61,49% |
| PDR (SVM) | 65,10% | 60,43% | 54,77% |
| PDM (Adaboost) | 63,27% | 61,53% | 60,93% |
| PDR(Adaboost) | 61,37% | 59,74% | 58,68% |

Figs. 11 and 12 show the results achieved with, respectively, AdaBoost and SVM, considering only male faces as stimuli. With AdaBoost, PDR achieves the highest value of accuracy, while the *arcs* gaze feature is better than PDM. On the contrary, with SVM (*gamma* = 0.7) both pupil indicators achieve accuracies higher than 60%, although comparable with those of arcs.

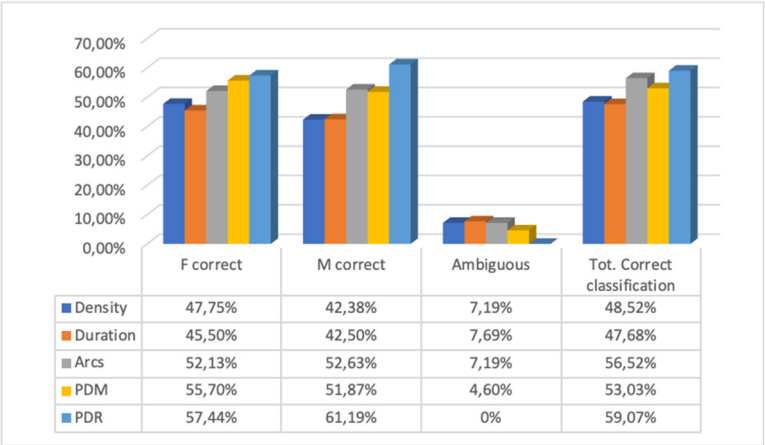
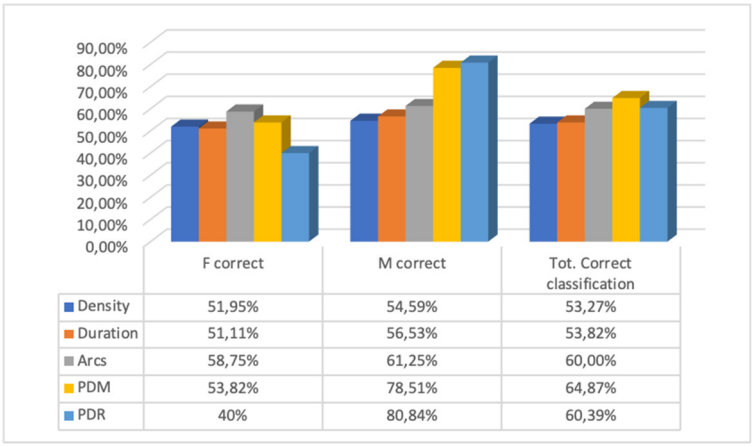


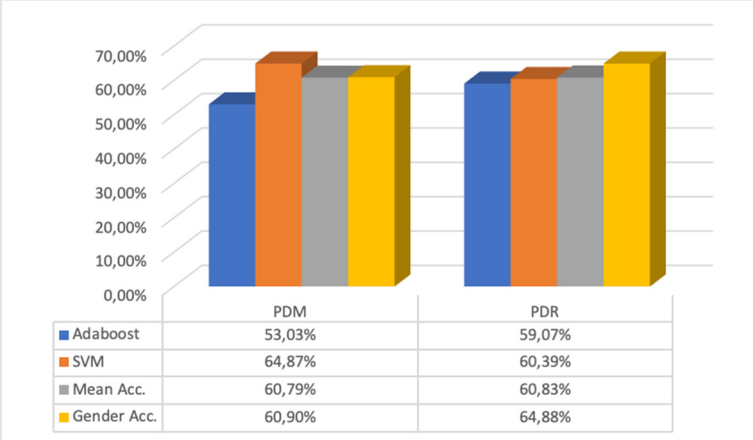

**Fig. 11.** Observation of male faces only. Gender classification accuracies with Adaboost, average scores from *Exp1_M* and *Exp2_M*.




| | F correct | M correct | Tot. Correct classification |
|---|---|---|---|
| Density | 51,95% | 54,59% | 53,27% |
| Duration | 51,11% | 56,53% | 53,82% |
| Arcs | 58,75% | 61,25% | 60,00% |
| PDM | 53,82% | 78,51% | 64,87% |
| PDR | 40% | 80,84% | 60,39% |

**Fig. 12.** Observation of male faces only. Gender classification accuracies with SVM, average scores from *Exp1_M* and *Exp2_M*.

Fig. 13 shows the comparison between the results for PDM and PDR obtained with the two classifiers and with their combination according to the two kinds of weighted means.




| | PDM | PDR |
|---|---|---|
| Adaboost | 53,03% | 59,07% |
| SVM | 64,87% | 60,39% |
| Mean Acc. | 60,79% | 60,83% |
| Gender Acc. | 60,90% | 64,88% |

**Fig. 13.** Observation of male faces only. Comparison between PDM and PDR for the two classifiers and for their combination using two weighted means (mean accuracy and gender accuracy).

With the gender accuracy weighted mean, both PDM and PDR achieve values over 60%.

Table 6 shows the values of precision, recall, and F1 score for PDM and PDR.

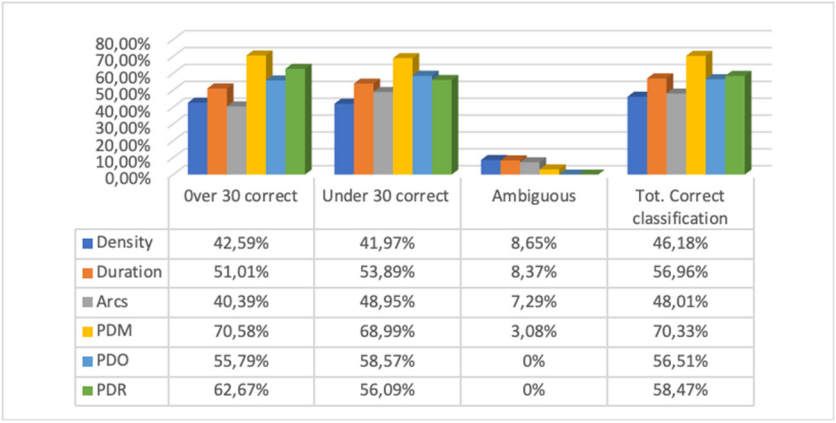
Table 6: Precision, recall and F1 score for *Exp1_M* and *Exp2_M*

| | Precision | Recall | F1 score |
|---|---|---|---|
| PDM (SVM) | 72,45% | 64,86% | 63,08% |
| PDR (SVM) | 71,90% | 60,39% | 55,27% |
| PDM (Adaboost) | 53,54% | 53,02% | 52,87% |
| PDR (Adaboost) | 61,17% | 59,08% | 58,77% |

### 6.2 Age Results

We conducted two kinds of experiments for age classification:
1. *Exp1* and *Exp2*, where the 'Over 30' class was composed of ages from 30 to 50;
2. *Exp1_10years* and *Exp2_10years*, where the 'Over 30' class was composed of ages from 40 to 80.

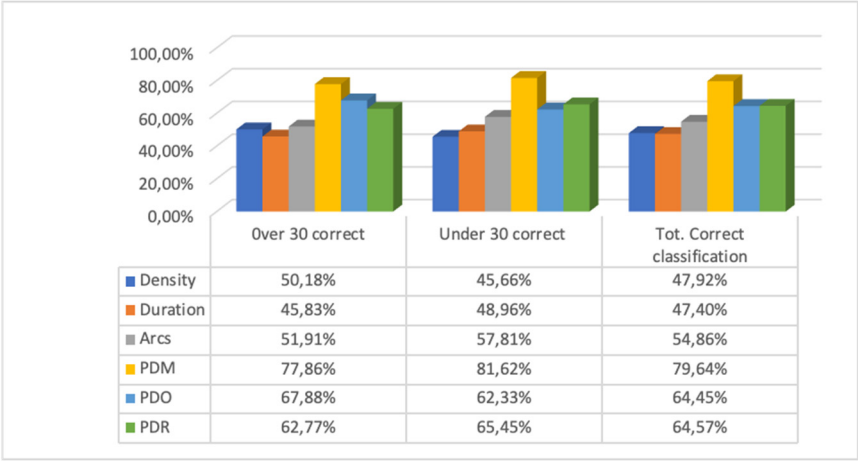
Considering, in Table 1, the ranges *c*, *d*, and *e* for ages over 30, Fig. 14 shows the maximum values obtained with AdaBoost.




| | Over 30 correct | Under 30 correct | Ambiguous | Tot. Correct classification |
|---|---|---|---|---|
| Density | 42,59% | 41,97% | 8,65% | 46,18% |
| Duration | 51,01% | 53,89% | 8,37% | 56,96% |
| Arcs | 40,39% | 48,95% | 7,29% | 48,01% |
| PDM | 70,58% | 68,99% | 3,08% | 70,33% |
| PDO | 55,79% | 58,57% | 0% | 56,51% |
| PDR | 62,67% | 56,09% | 0% | 58,47% |

**Fig. 14.** Age classification accuracies with AdaBoost, average scores from *Exp1* and *Exp2*.

In Fig. 15, the results of age classification obtained through the SVM classifier are shown.

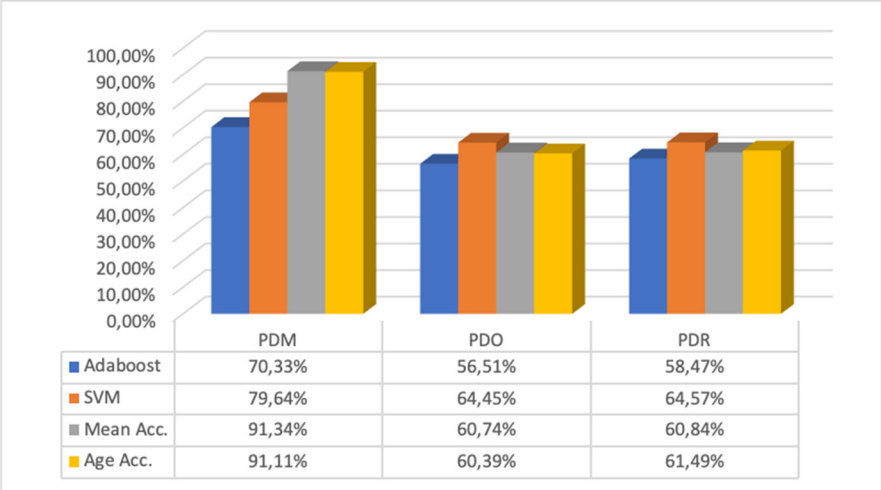



| | Over 30 correct | Under 30 correct | Tot. Correct classification |
|---|---|---|---|
| Density | 50,18% | 45,66% | 47,92% |
| Duration | 45,83% | 48,96% | 47,40% |
| Arcs | 51,91% | 57,81% | 54,86% |
| PDM | 77,86% | 81,62% | 79,64% |
| PDO | 67,88% | 62,33% | 64,45% |
| PDR | 62,77% | 65,45% | 64,57% |

**Fig. 15.** Age classification accuracies with SVM, average scores from *Exp1* and *Exp2*.

Pupil results obtained for both classifiers are clearly higher than those achieved with gaze features, and PDM is the index providing the best performance. The parameters with which SVM has attained the highest values are the following: for PDM, the linear kernel with at most 10,000 iterations; for PDO, the linear kernel; for PDR, a *gamma* value of 0.7.

Fig. 16 shows the comparison between the results for PDM, PDO, and PDR obtained with the two classifiers and with their combination according to the two kinds of weighted means. PDM achieves very good results with both mean and age accuracy, markedly higher than those obtainable through the two classifiers singularly.




| | PDM | PDO | PDR |
|---|---|---|---|
| Adaboost | 70,33% | 56,51% | 58,47% |
| SVM | 79,64% | 64,45% | 64,57% |
| Mean Acc. | 91,34% | 60,74% | 60,84% |
| Age Acc. | 91,11% | 60,39% | 61,49% |

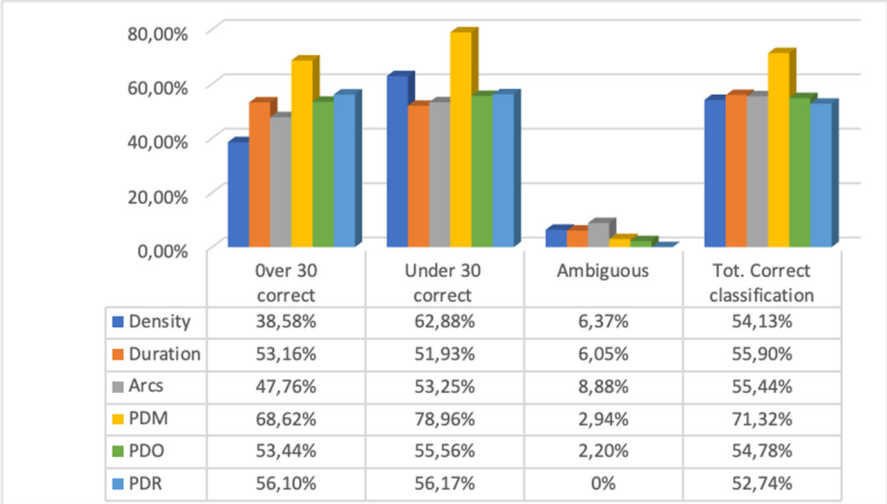
**Fig. 16.** Age classification (ranges *c*, *d*, and *e* in Table 1 for ages over 30). Comparison between PDM, PDO, and PDR for the two classifiers and for their combination using two weighted means (mean accuracy and age accuracy).

Table 7 shows the values of precision, recall and F1 score for PDM, PDR, PDO.

Table 7: Precision, recall and F1 score for *Exp1* and *Exp2*

| | Precision | Recall | F1 score |
|---|---|---|---|
| PDM (SVM) | 81,02% | 79,64% | 79,55% |
| PDR (SVM) | 66,83% | 64,57% | 62,23% |
| PDO (SVM) | 69,76% | 64,43% | 62,98% |
| PDM (Adaboost) | 72,98% | 70,33% | 70,27% |
| PDR(Adaboost) | 61,76% | 58,96% | 58,38% |
| PDO (Adaboost) | 58,19% | 56,52% | 56,12% |

Considering, in Table 1, the ranges *d*, *e*, *f*, and *g* for ages over 30, Fig. 17, shows the maximum values obtained with AdaBoost.




| | Over 30 correct | Under 30 correct | Ambiguous | Tot. Correct classification |
|---|---|---|---|---|
| Density | 38,58% | 62,88% | 6,37% | 54,13% |
| Duration | 53,16% | 51,93% | 6,05% | 55,90% |
| Arcs | 47,76% | 53,25% | 8,88% | 55,44% |
| PDM | 68,62% | 78,96% | 2,94% | 71,32% |
| PDO | 53,44% | 55,56% | 2,20% | 54,78% |
| PDR | 56,10% | 56,17% | 0% | 52,74% |

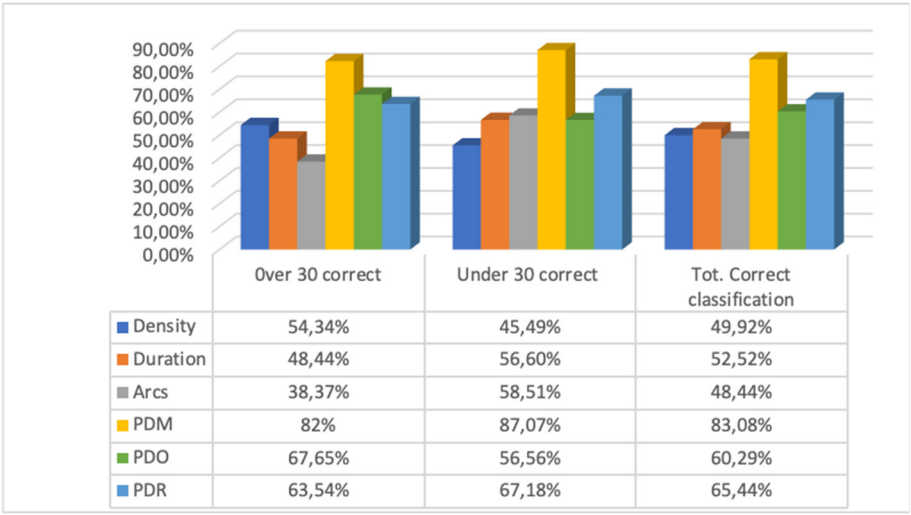
**Fig. 17.** Age classification accuracies with AdaBoost, average scores from *Exp1_10years* and *Exp2_10years*.

In Fig. 18 the results of age classification obtained through the SVM classifier are shown.

For PDO and PDR, AdaBoost does not provide better results than those obtained from gaze parameters, while for PDM the performance is much better. SVM, instead, always produces higher accuracies, almost achieving 83% (with the following parameters: for PDM and PDR, linear kernel and 10,000 iterations at most; for PDO, linear kernel).

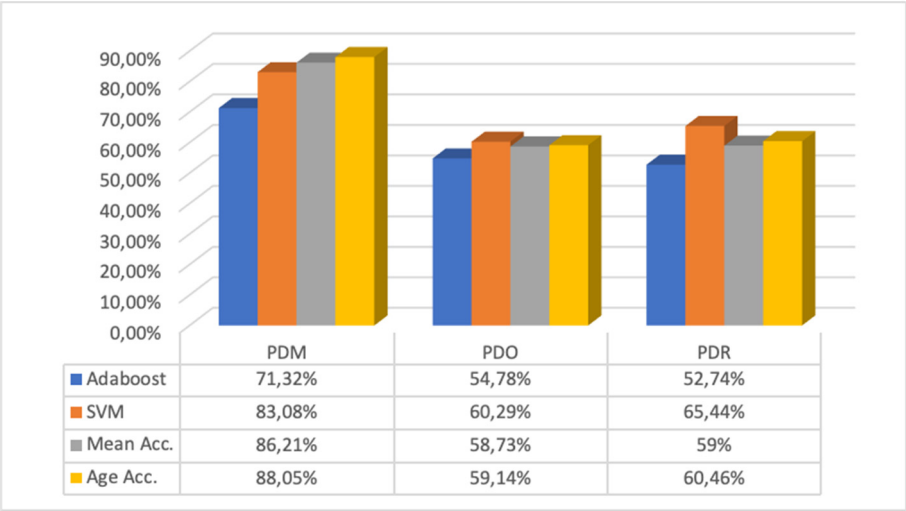
Fig. 19 shows the comparison between the results for PDM, PDO, and PDR obtained with the two classifiers and with their combination according to the two kinds of weighted means.

The combination of results definitely shows that PDM is the best age indicator, achieving an accuracy higher than 80%.




| | Over 30 correct | Under 30 correct | Tot. Correct classification |
|---|---|---|---|
| Density | 54,34% | 45,49% | 49,92% |
| Duration | 48,44% | 56,60% | 52,52% |
| Arcs | 38,37% | 58,51% | 48,44% |
| PDM | 82% | 87,07% | 83,08% |
| PDO | 67,65% | 56,56% | 60,29% |
| PDR | 63,54% | 67,18% | 65,44% |

**Fig. 18.** Age classification accuracies with SVM, average scores from *Exp1_10years* and *Exp2_10years*.




| | PDM | PDO | PDR |
|---|---|---|---|
| Adaboost | 71,32% | 54,78% | 52,74% |
| SVM | 83,08% | 60,29% | 65,44% |
| Mean Acc. | 86,21% | 58,73% | 59% |
| Age Acc. | 88,05% | 59,14% | 60,46% |

**Fig. 19.** Age classification (ranges *d*, *e*, *f*, and *g* in Table 1 for ages over 30). Comparison between PDM, PDO, and PDR for the two classifiers and for their combination using two weighted means (mean accuracy and age accuracy).

Table 8 shows the values of precision, recall, and F1 score for PDM, PDO, and PDR for the two classifiers.

Table 8: Precision, recall and F1 score for *Exp1_10years* and *Exp2_10years*

| | Precision | Recall | F1 score |
|---|---|---|---|
| PDM (SVM) | 85,61% | 83,08% | 83,36% |
| PDR (SVM) | 69,40% | 65,44% | 62,96% |
| PDO (SVM) | 65,33% | 60,29% | 58,30% |
| PDM (Adaboost) | 78,73% | 71,32% | 71,08% |
| PDR(Adaboost) | 58,91% | 57,35% | 57,16% |
| PDO (Adaboost) | 55,80% | 54,78% | 53,58% |

Almost all our results show that PDM is the feature that exhibits the best overall performance for both classifications. From the distribution of these values (Fig. 20), it is also evident why highly better results have been obtained in age classification.

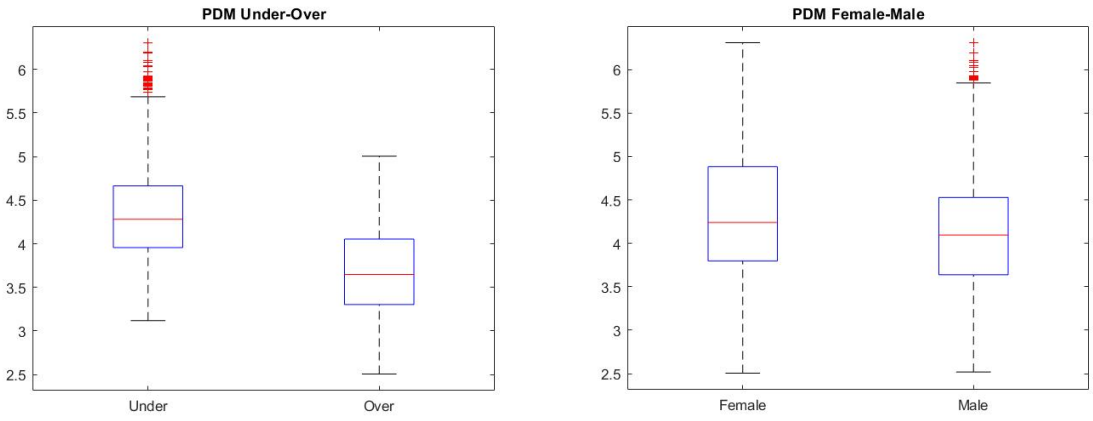



**Fig. 20.** Boxplots of PDM values. Comparison between the classes Under-Over and Female-Male.

## 7. Conclusions

In this paper we have considered pupil size as a discriminating factor for gender and age assessment. In particular, we have compared pupil indicators to the results obtained from a previous study [9] in which fixation data were analyzed through the GANT method [2].

Overall, the outcomes of our experiments show that pupil size is a good estimator of both gender and age, in most cases exhibiting better performances than the gaze-only counterpart. Two classifiers have been exploited, namely AdaBoost and SVM, trained and tested with four-fold cross validation. As pupil features, we have employed PDM (Pupil Diameter Mean), PDR (Pupil Diameter Ratio), and PDO (Pupil Diameter Order).

For gender, the best accuracy (64.20%) has been obtained with the SVM classifier and the PDM pupil feature. Still with SVM, a good result (61.58%) has also been achieved with PDO, clearly higher than the best outcome provided by the density feature with AdaBoost (51.49%). The combination of the classifiers using weighted means (with weights given by the mean accuracies achieved by the two classifiers and by the gender or age accuracies for each classifier class) does not provide any advantage over the use of the single classifiers.

When considering only female faces as stimuli, PDM and PDR obtain better accuracies with both classifiers compared to gaze data (density and duration, in particular). With the weighted combination of classifiers, PDM achieves an accuracy of 64.13%, significantly better than the best performance obtained with a single classifier (62.70% with SVM and PDM). On the contrary, with only male faces as stimuli, only SVM with PDM and PDR achieves an accuracy (64.87% and 60.39%) higher than the best result of gaze features (60% with arcs). No significant improvement is obtained with the weighted combination of classifiers. Although these results show some differences between the observation of female and male faces, we think they are not enough to claim that one of the two gender stimuli is preferable for the identification of the gender of the observer.

As regards age, considering the case in which testers older than 30 were 60 at most, pupil features exhibit a clearly better performance than the gaze approach, and PDM is the best index (70.33% with AdaBoost, against 56.96% of the duration feature; 79,64% with SVM, against 54.86% of arcs). When combining the two classifiers, PDM provides very good results with both mean (91.34%) and age (91.11%) accuracy - much higher than the 79,64% achieved with SVM.
Also considering the case in which testers older than 30 were up to 80, PDM turns out to be the best indicator, with accuracies of 71.32% for AdaBoost (against 55.9% of the duration gaze feature) and of 83.0.8% for SVM (against 52.52% of the duration gaze feature). The combination of the two classifiers confirms the better performance of PDM, with age accuracy weights being the best solution (88.05%). The wider age range of the second case (testers up to 80 years old) seems to provide an improvement of the discriminating power of pupil size in recognizing subjects younger and older than 30.

In conclusion, from the outcomes of our study we can say that pupil size is good at discerning gender and very good at distinguishing people younger and older than 30. In particular, the PDM (Pupil Diameter Mean) is the feature exhibiting the best overall performance.

In future investigations, we will further explore the potential of pupil size for gender and age detection, by considering other visual subjects as stimuli, identifying possible new pupil features, and investigating the use of other Machine Learning techniques.

## References


[1] R. Bednarik, T. Kinnunen, A. Mihaila, P. Fränti, Eye-movements as a biometric, in 2005 Scandinavian Conference on Image Analysis (SCIA), Joensuu, Finland, June 19-22, 2005, pp. 780-789.

[2] V. Cantoni, C. Galdi, M. Nappi, M. Porta, D. Riccio, GANT, Gaze analysis technique for human identification, Pattern Recognit. 48 (April (4)) (2015) 1027–1038.

[3] V. Cantoni, C. Jimenez Perez, M. Porta, S. Ricotti, Exploiting eye tracking in advanced E-learning systems, in 13th International Conference on Computer Systems and Technologies (CompSysTech 2012), Rousse, Bulgaria, 2012, pp. 376–383.

[4] V. Cantoni, C. Galdi, M. Nappi, M. Porta, D. Riccio, L. De Maio, R. Distasi, Gaze recording - Free observation of human faces, Mendeley Data, v2, 2020 (DOI: http://dx.doi.org/10.17632/3kn4jdd4kf.2).

[5] V. Cantoni, C. Galdi, M. Nappi, M. Porta, H. Wechsler, Gender and age categorization using gaze analysis, in 10th International Conference on Signal Image Technology & Internet Systems (SITIS 2014), Marrakech, Morocco, November 23-27, 2014.

[6] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[7] A.T. Duchowski, Eye Tracking Methodology –Theory and Practice, 2nd ed., Springer-Verlag, London, 2007.

[8] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.

[9] C. Galdi, H. Wechsler, V. Cantoni, M. Porta, M. Nappi, Towards Demographic Categorization using Gaze Analysis, Pattern Recognition Letters, Vol. 82, Part 2, 2016, 226-231.

[10] A. George, A. Routray, A score level fusion method for eye movement biometrics, Pattern Recognition Letters, Vol. 82, Part 2, 2016, pp. 207-215.

[11] M. Guillon, K. Dumbleton, P. Theodoratos, M. Gobbe, C. B. Wooley, K. Moody, The effects of age, refractive status, and luminance on pupil size, Optometry and Vision Science, vol. 93, no. 9, p. 1093, 2016.

[12] D. Hämmerer, A. Hopkins, M. J. Betts, A. Maaβ, R. J. Dolan, E. Düzel, Emotional arousal and recognition memory are differentially reflected in pupil diameter responses during emotional memory for negative events in younger and older adults, Neurobiology of aging, vol. 58, pp. 129-139, 2017.

[13] C. Holland, O. V. Komogortsev, Biometric identification via eye movement scanpaths in reading, in 2011 International Joint Conference on Biometrics (IJCB'11), Washington, DC, USA, October 11-13, 2011, pp. 1-8.

[14] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vis. Res. 40 (2000), 1489–1506.

[15] K. Jain, S.C. Dass, K. Nandakumar, Soft biometric traits for personal recognition systems, in International Conference on Biometric Authentication, LNCS, vol. 3072, 2004, pp. 731–738.

[16] M. Juhola, Y. Zhang, J. Rasku, Biometric verification of a subject through eye movements, Comput. Biol. Med., Vol. 43, No. 1, 2013, pp. 42-50.

[17] M.A. Just, P.A. Carpenter, Eye fixations and cognitive processes, Cogn. Psychol. 8 (1976) 441–480.

[18] P. Kasprowski, J. Ober, Eye movements in biometrics, in 2004 International Workshop on Biometric Authentication (BioAW), Prague, Czech Republic, May 15, 2004, pp. 248-258.

[19] S. Kasthurirangan, A. Glasser, Age related changes in the characteristics of the near pupil response, Vision Research, vol. 46, no. 8-9, pp. 1393-1403, 2006.

[20] T. Kinnunen, F. Sedlak, R. Bednarik, Towards task-independent person authentication using eye movement signals, in 2010 Symposium on Eye-Tracking Research & Applications (ETRA), Austin, Texas, USA, March 22-24, 2010, pp. 187-190.

[21] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI, vol. 14, 1995, pp. 1137–1145.

[22] O. V. Komogortsev, A. Karpov, C. D. Holland, H. P. Proença, Multimodal ocular biometrics approach: a feasibility study, in 5th IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), Washington DC, USA, September 23-26, 2012, pp. 209-216.

[23] O. V. Komogortsev, A. Karpov, L. R. Price, C. Aragon, Biometric authentication via oculomotor plant characteristics, in 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, March 29 - April 1, 2012, pp. 413-420.

[24] S. Mathôt, A. Siebold, M. Donk, F. Vitu. Large pupils predict goal-driven eye movements, Journal of Experimental Psychology General Vol.144, 2015, pp. 513-521. DOI: 10.1037/a0039168 2015;144.

[25] S. Mitra, B. Wen, M. Gofman, Overview of Biometric Authentication, in Biometrics in a data driven World: Trends, Technologies, and Challenges, 2017, CRC Press.

[26] N. Nugrahaningsih, M. Porta, Pupil size as a biometric trait, in International Workshop on Biometric Authentication. Springer, 2014, pp. 222-233.

[27] E. Ortiz, K. W. Bowyer, P. J. Flynn, A linear regression analysis of the effects of age related pupil dilation change in iris biometrics, in 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Sep. 2013, pp. 1-6.

[28] T. Partala, V. Surakka, Pupil size variation as an indication of affective processing, International Journal of Human-Computer Studies, Vol.59, 2003, pp. 185-198. DOI: 10.1016/S1071-5819(03)00017-X.

[29] E. Perego, F. Del Missier, M. Porta, M. Mosconi, The cognitive effectiveness of subtitle processing, Media Psychol. 13 (3) (2010) 243–272.

[30] G. Poerio, P. Totterdell, E. Miles, Mind-wandering and negative mood: Does one thing really lead to another?. Consciousness and Cognition, Vol.22, 2013, pp. 1412-1421. DOI: 10.1016/j.concog.2013.09.012.

[31] M. Porta, S. Ricotti, C. Jimenez Perez, Emotional E-learning through eye tracking, in 2012 IEEE International Conference on Collaborative Learning & New Pedagogic Approaches in Engineering Education (EDUCON 2012), Marrakesh, Morocco, 2012, pp. 1–6.

[32] Q.-X. Qu, F. Guo, Can eye movements be effectively measured to assess product design?: Gender differences should be considered, International Journal of Industrial Ergonomics, vol. 72, pp. 281-289, 2019.

[33] I. Rigas, G. Economou, S. Fotopoulos, Human eye movements as a trait for biometrical identification, in 5th IAPR International Conference on Biometrics (ICB), New Delhi, India, March 29 - April 1, 2012, pp. 217-222.

[34] K. Saeed, New Directions in Behavioral Biometrics, CRC Press, 2016.

[35] J. A. Sanchis-Gimeno, D. Sanchez-Zuriaga, F. Martinez-Soriano, White-to-white corneal diameter, pupil diameter, central corneal thickness and thinnest corneal thickness values of emmetropic subjects, Surgical and Radiologic Anatomy, vol. 34, no. 2, pp. 167-170, 2012.

[36] N. Srivastava, U. Agrawal, S. K. Roy, U. S. Tiwary, Human identification using linear multiclass SVM and eye movement biometrics, in 8th International Conference on Contemporary Computing (IC3), Noida, India, August 20-22, 2015, pp. 365-369.

[37] Y. Wang, G. Naylor, S. E. Kramer, A. A. Zekveld, D.Wendt, B. Ohlenforst, T. Lunner, Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task, Ear and Hearing, vol. 39, no. 3, pp. 573-582, 2018.

[38] C.Y. Sai, N. Mokhtar, H. Arof, P. Cumming, M. Iwahashi, Automated classification and removal of EEG artifacts with SVM and wavelet-ICA, IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 3, pp. 664-670, 2018.

[39] K. Randhawa, C. K. Loo, M. Seera, C.P. Lim, A.K. Nandi, Credit Card Fraud Detection Using Adaboost and Majority Voting, IEEE Access, vol. 6, pp. 14277-14284, 2018.

[40] F. Wang, D. Jiang, H. Wen, H. Song, Adaboost-based security level classification of mobile intelligent terminals, Journal of Supercomputing, vol. 75, pp. 7460-7478, 2019.