Ensemble-based systems in medical image processing

Andras Hajdu
Faculty of Informatics, University of Debrecen

SSIP 2011, Szeged, Hungary

IMAGE PROCESSING GROUP OF DEBRECEN
http://ipgd.inf.unideb.hu

Ensemble based systems

Ensemble learning is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem.

When to use ensembles?

- Not sufficient predictive performance
- Too much data
- Too few data
- Too complex data
- Multiple information sources

Not sufficient predictive performance

- Different algorithms have different predictive performances in different contexts
- Sometimes they do not have enough generalization capabilities to classify unknown instances using their learned model

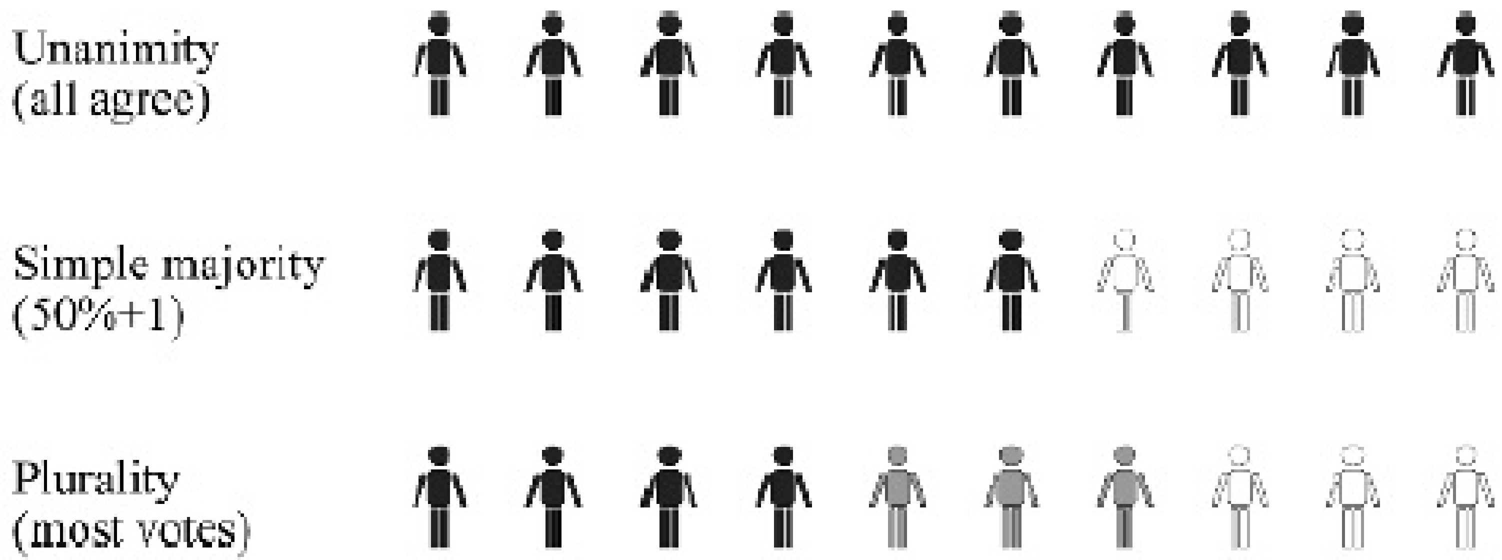
Solution

- Combining class labels provided by the individual predictors
- Combining real values provided by the individual predictors
- Other combinations methods

Combining class labels

- Non-learning based (majority voting, borda count)
- Learning-based (weighted majority voting, Behavioral Knowledge Space (BKS), Wernecke method)

Majority voting



Weighted majority voting

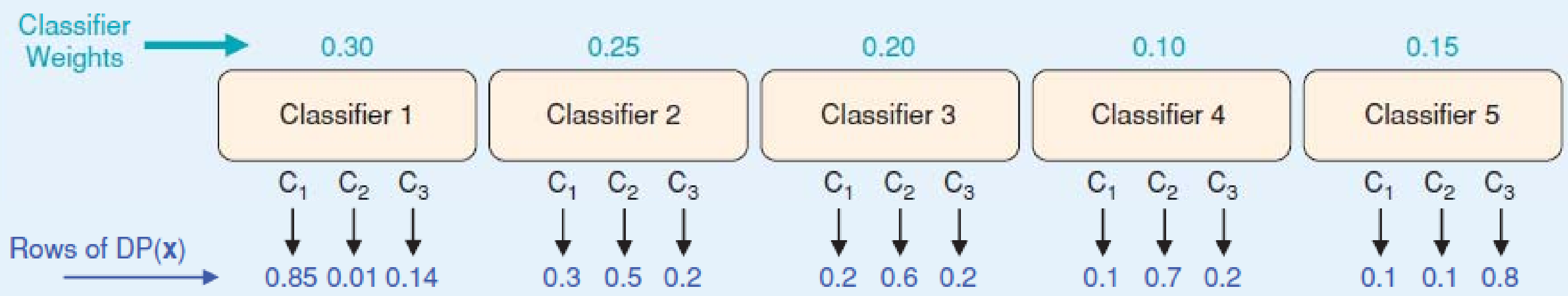
- We assign a weight to each algorithm based on its performance on a dataset
- The better the performance the larger weight assigned
- Usually, the following formula is used (p_t is the performance, w_t is the weight assigned to the predictor t):

$$w_t \propto \log \frac{p_t}{1 - p_t}$$

Other methods

- Behavioral Knowledge Space (BKS): stores the predictive outcomes for each voting combination during training.
- Wernecke method: extends BKS by introducing confidence intervals
- Borda Count: rank of the class membership probabilities

Combining real values



Product Rule:	**C_1: 0.0179,**	C_2: 0.00021,	C_3: 0.0009
Sum Rule:	C_1: 1.55,	**C_2: 1.91,**	C_3: 1.54
Mean Rule:	C_1: 0.310,	**C_2: 0.382,**	C_3: 0.308
Max Rule:	**C_1: 0.85,**	C_2: 0.7,	C_3: 0.8
Median Rule:	C_1: 0.2,	**C_2: 0.5,**	C_3: 0.2
Minimum Rule:	C_1: 0.1,	C_2: 0.01,	**C_3: 0.14**
Weighted Average	**C_1: 0.395,**	C_2: 0.333,	C_3: 0.272
Majority Voting:	C_1: 1,	**C_2: 3,**	C_3: 1 Vote
Weighted Majority Voting:	C_1: 0.3,	**C_2: 0.55,**	C_3: 0.15 Votes
Borda Count:	C_1: 5	**C_2: 6,**	C_3: 4 Votes

Figure 15. Example on various combination rules.

Too much data

- If we want to learn on too much data, we need to split the data into disjoint parts
- We train an algorithm on each part
- Finally, we combine the outcomes of the algorithms

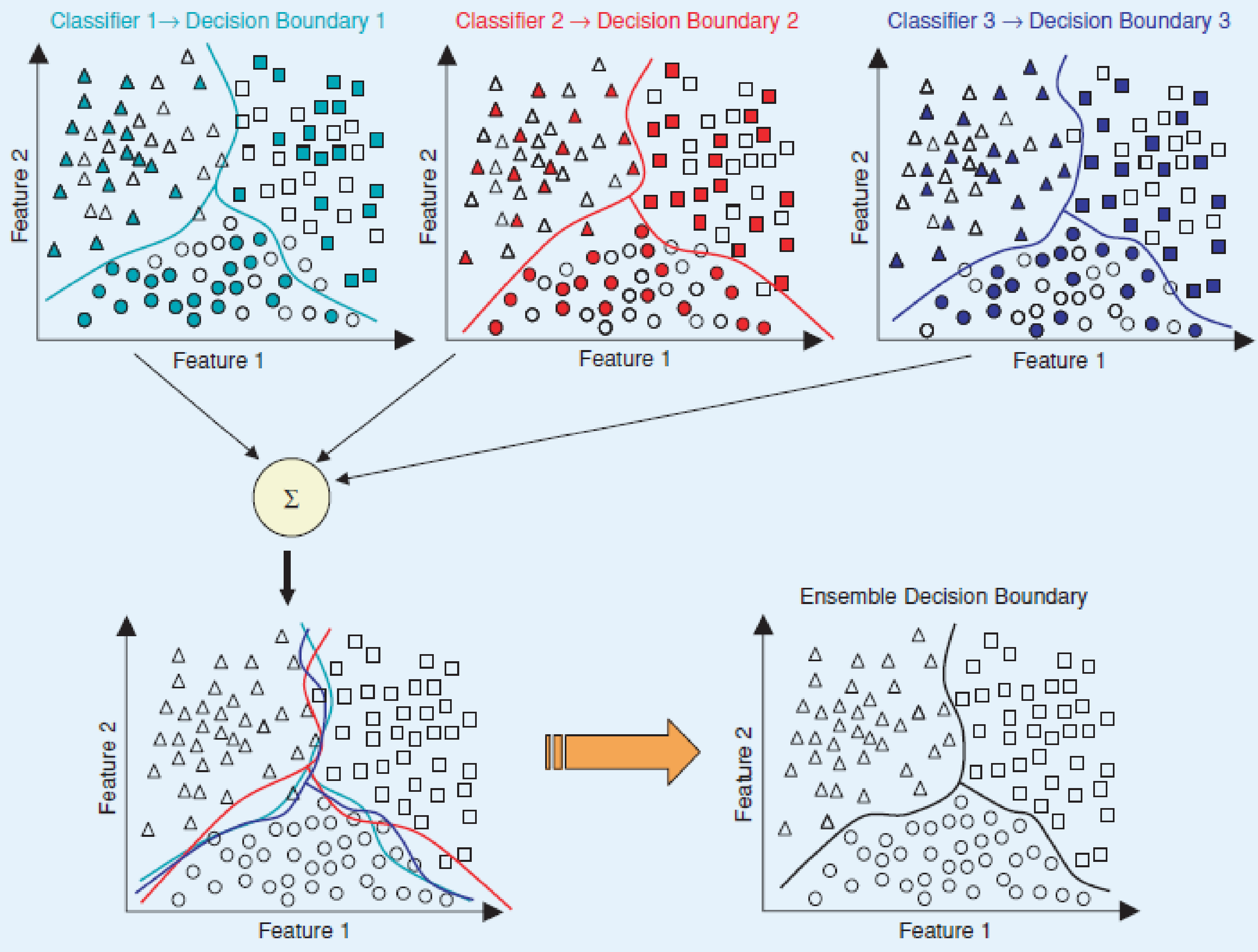


Figure 3. Combining classifiers that are trained on different subsets of the training data.

Too few data

- If we want to learn on too few data, we need to split the data into random, possibly overlapping parts
- We train an algorithm on each parts
- Finally, we combine the outcomes of the algorithms

Bagging

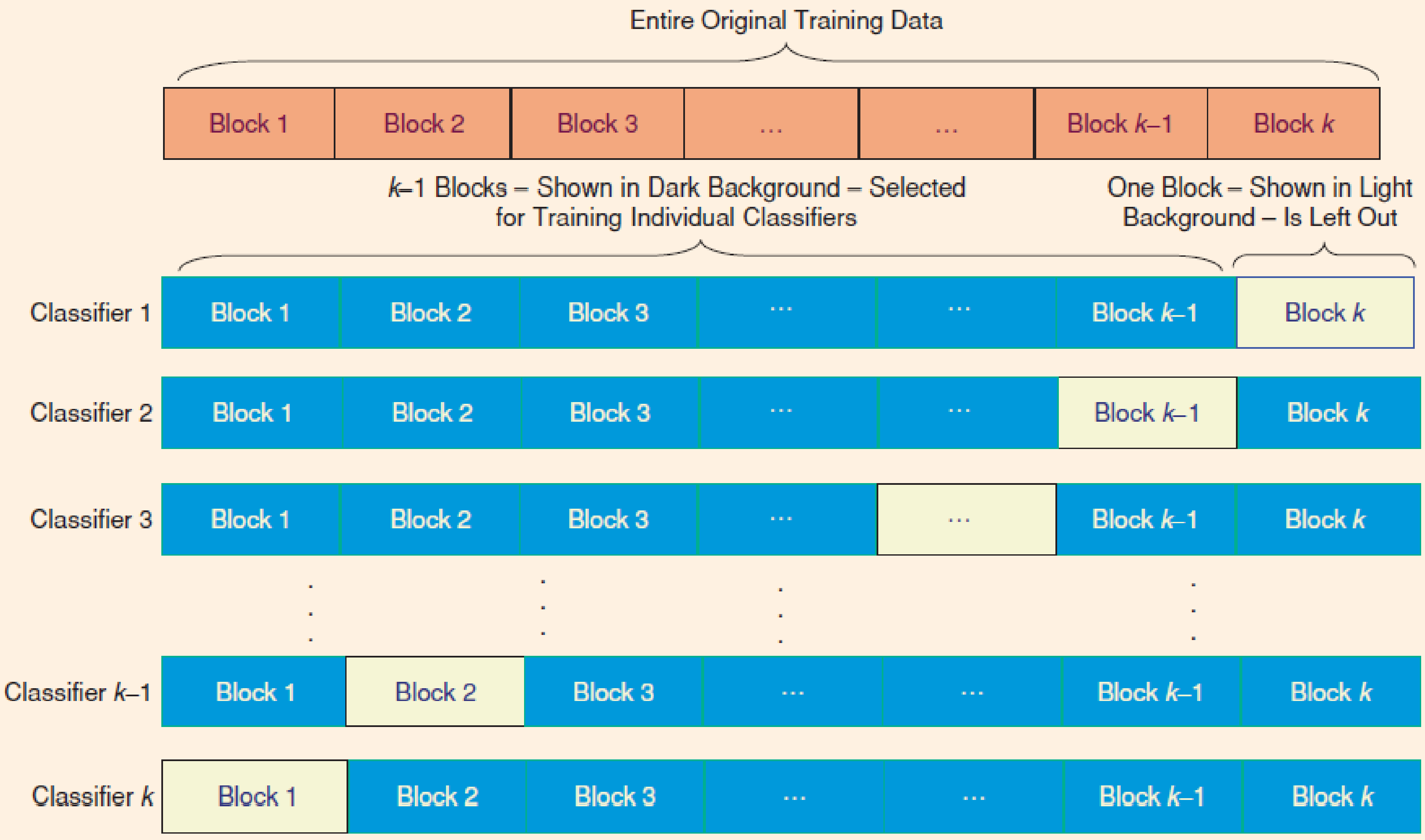


Figure 4. *k*-fold data splitting for generating different, but overlapping, training datasets.

Too complex data

- We use a „divide-and-conquer”-based solution strategy
- We use a voting among the algorithms trained for the different subproblems

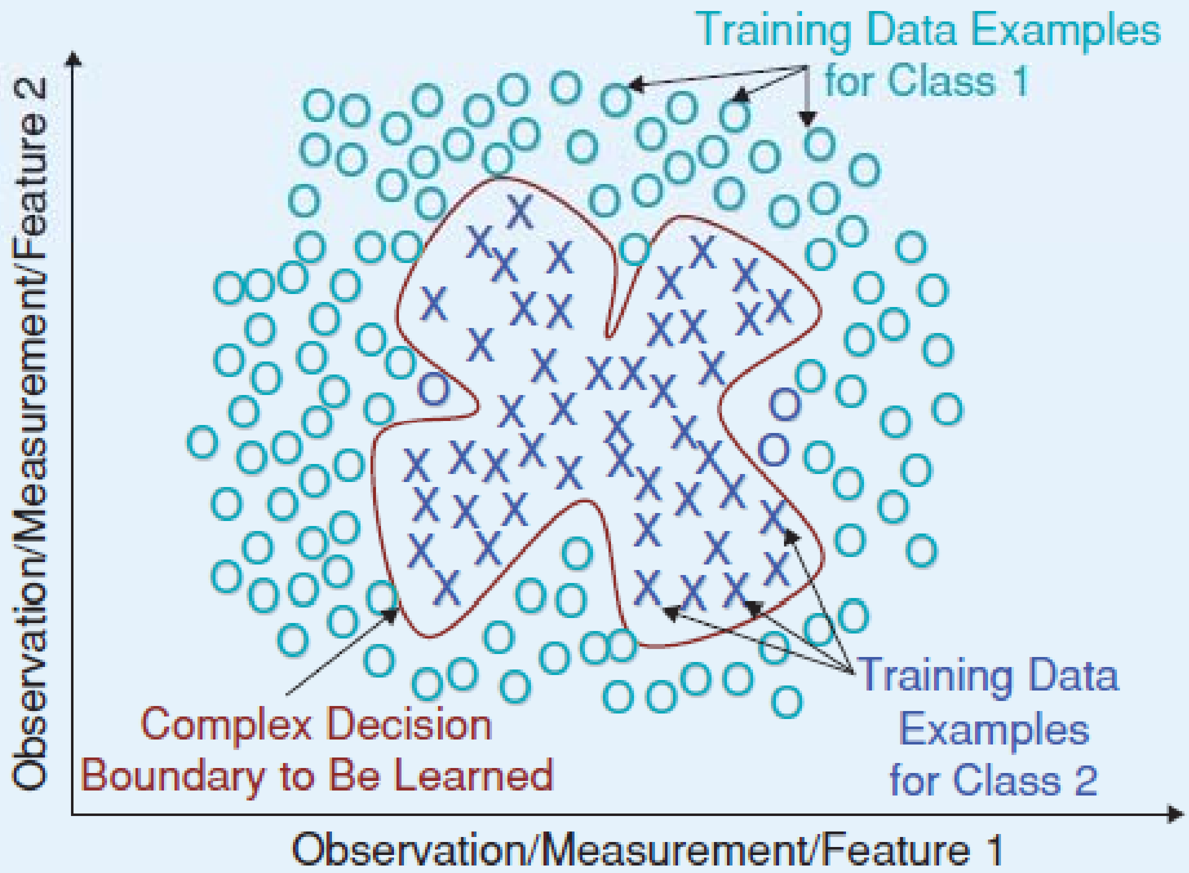


Figure 1. Complex decision boundary that cannot be learned by linear or circular classifiers.

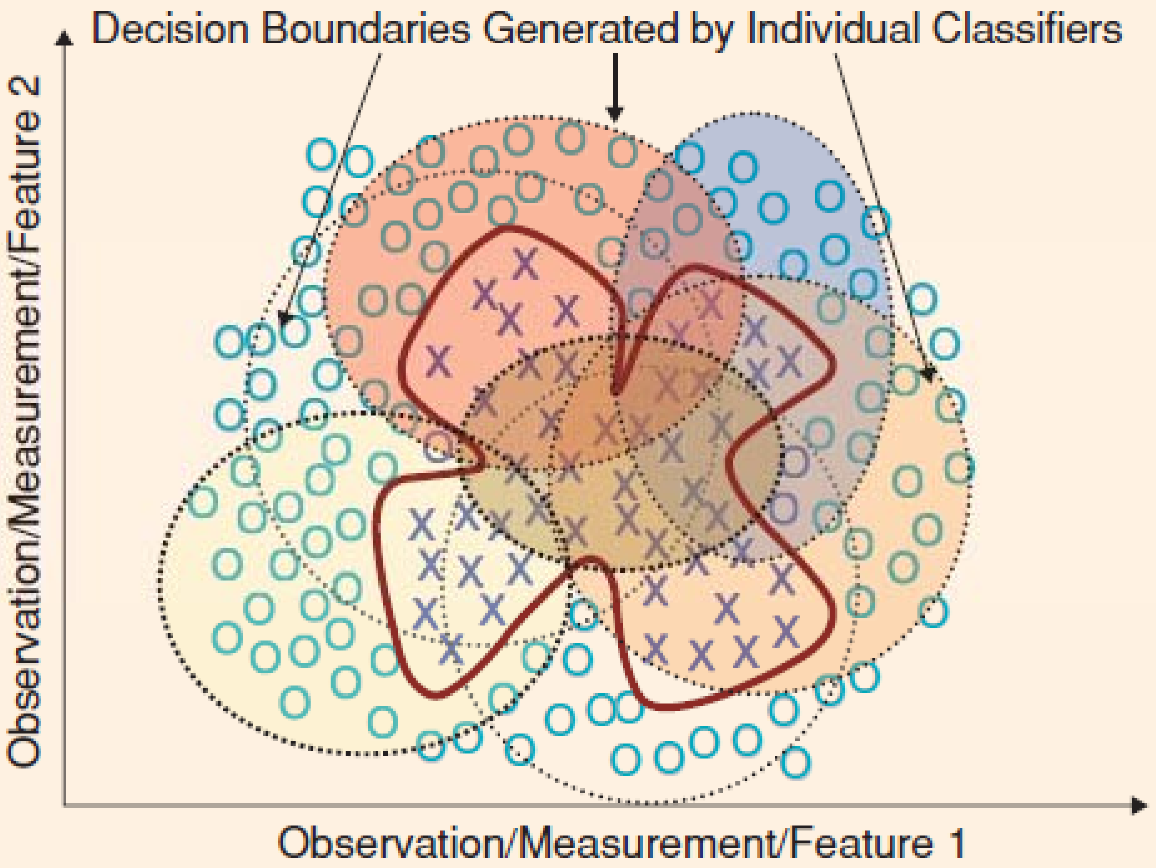


Figure 2. Ensemble of classifiers spanning the decision space.

Theoretical bounds of majority voting

- Does an ensemble-based system always performes better than an individual approach?
- The worst case scenario for 9 algorithms, each having 60% accuracy, is 28% accuracy!
- Weighted majority voting is proven to be better than majority voting when each participant have at least 50% accuracy.

How to choose participants?

- Diversity measures
- There are no evidence of a link between diversity and accuracy, but a good place to start investigating.
- The best case scenario is when the proportion of the correct votes equals the majority.

Diversity measures

	h_j is correct	h_j is incorrect
h_i is correct	a	b
h_i is incorrect	c	d

$$\rho_{i,j} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}, \quad 0 \leq \rho \leq 1.$$

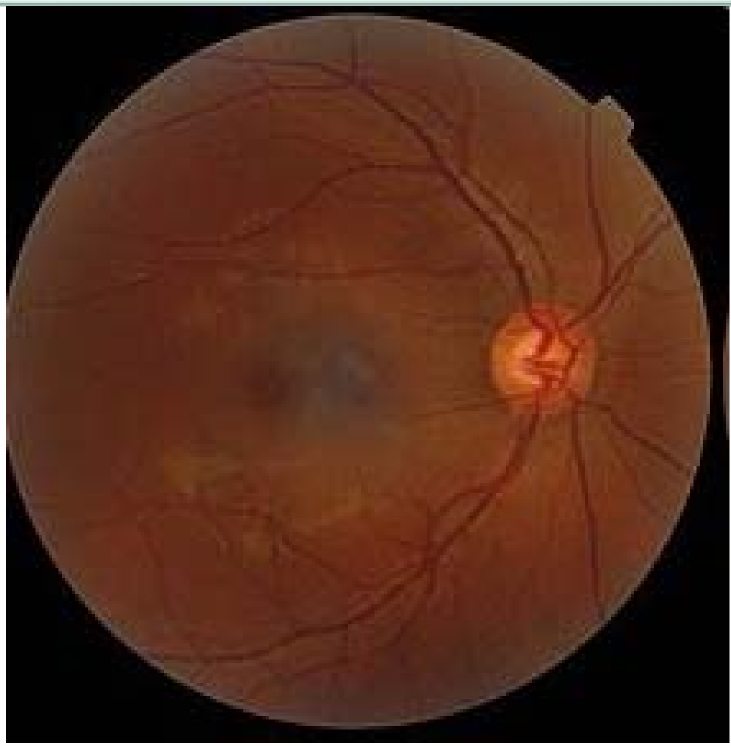
$$Q_{i,j} = (ad - bc)/(ad + bc)$$

$$D_{i,j} = b + c,$$
$$DF_{i,j} = d.$$

Clinical example – detection of the optic disc

- an important prerequisite for automatic screening of retina images: the accurate localization of the main anatomical features in the image, notably the optic disc (OD) and the macula.

Basic problem

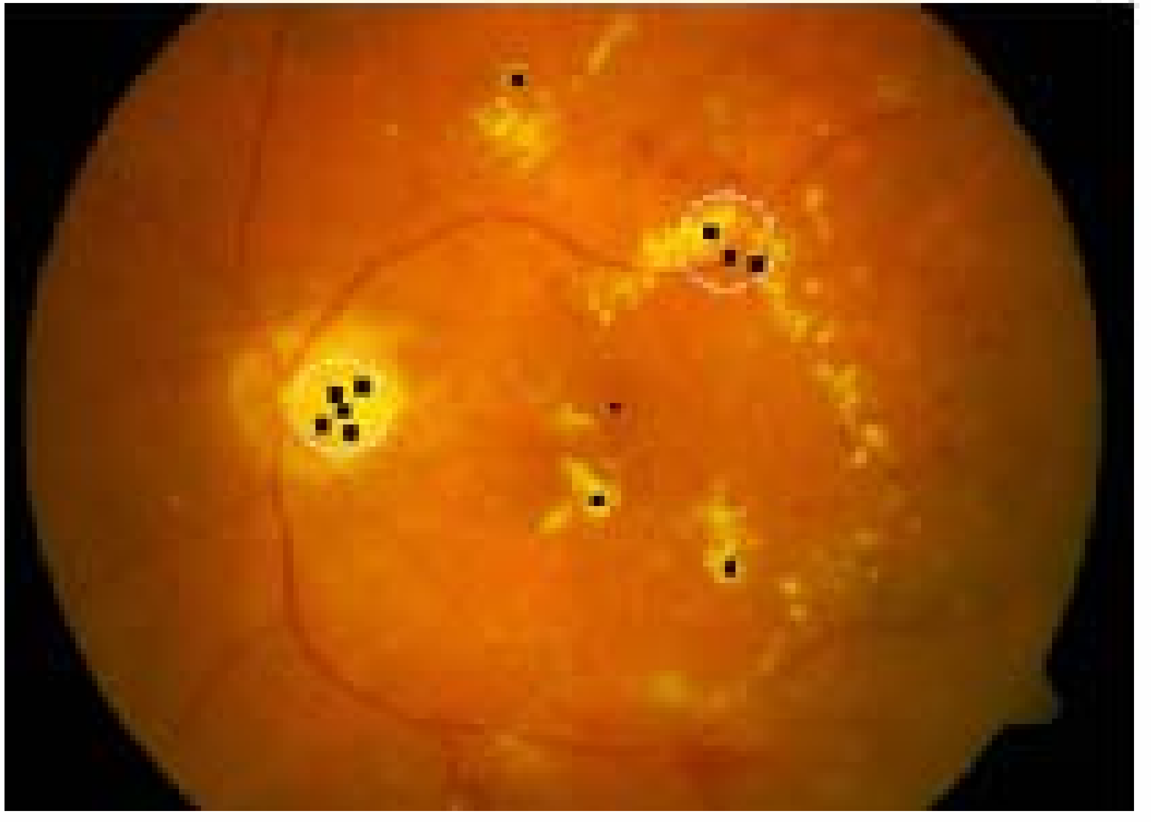


- optic disc - bright region with circular shape
- macula - oval-shaped highly pigmented spot
- fovea - responsible for the sharpest vision

Basic problem

- all of the OD algorithms return with the OD center as a single pixel

Basic problem



- the circle with maximal number of candidates is chosen for the optic disc

Basic problem

- to make a good decision even in the case when the bad candidates have majority

Majority voting

- Let $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ be a set (also called ensemble) of classifiers.
- $\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$ be a set of class labels.
- $D_i: \boldsymbol{R}^m \rightarrow \Omega \quad (i=1,..,n)$
- The majority voting method of combining classifier decisions is to assign the class label ω_i to $\boldsymbol{x}$ that is supported by the majority of the classifiers D_i.

Majority voting

- Let L be odd, $\Omega = \{\omega_1, \omega_2\}$ and all classifiers have the same classification accuracy p. The majority vote method with independent classifier decisions gives an overall correct classification accuracy calculated by the binomial formula:

$$P_{\mathrm{maj}} = \sum_{k=0}^{n/2} \binom{n}{k} p^{n-k} (1-p)^{k}$$

- When the classifiers are independent and $p>0.5$, this method is guaranteed to give a higher accuracy than individual classifiers.

Accuracy of correct classification

The majority voting method

	n = 3	*n*= 5	*n* = 7	*n*= 9
p = 0.6	0.6480	0.6826	0.7102	0.7334
p = 0.7	0.7840	0.8369	0.8740	0.9012
p = 0.8	0.8960	0.9421	0.9667	0.9804
p = 0.9	0.9720	0.9914	0.9973	0.9991

Spatial voting (optic disc geometry)

	n = 3	*n* = 5	*n* = 7	*n* = 9
p = 0.6	0.8208	0.8390	0.8895	0.9247
p = 0.7	0.9163	0.9373	0.9658	0.9823
p = 0.8	0.9728	0.9850	0.9942	0.9980
p = 0.9	0.9963	0.9988	0.9997	0.9999

Pattern of success

The „*pattern of success*" is a distribution of the L classifier outputs for **D** such that:

- The probability of any combination of

$[n/2]+1$ correct and $[n/2]$ incorrect votes is α.

- The probability of all L votes being incorrect

is γ.

- The probability of all other combinations is

zero.

- „Best” case: 1111000, „worst” case: 1110000.

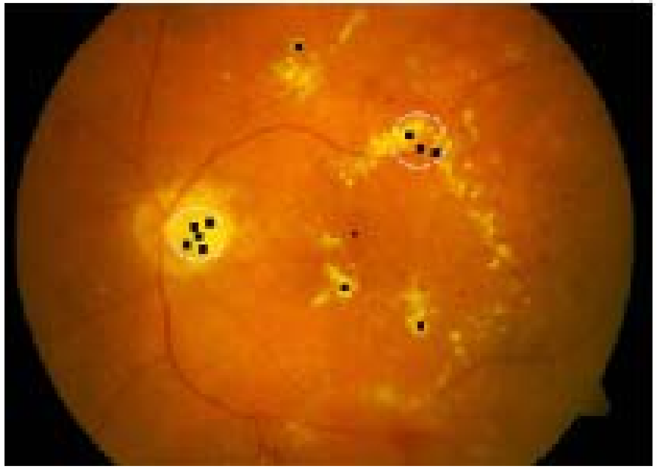
Pattern of success

The pattern of success and failure:

- useful information in clinical systems
- characterize the expected value of the system error and the boundary of the system accuracy:
  [minimum accuracy, maximum accuracy]

Spatial voting

- In such scenarios (algorithms vote by coordinates) it may happen that less number of „good” votes defeat larger number of „bad” votes.
- Model: p_{nk} the probability for good decision (n algorithms, k are correct)
- E.g. 1100000 still may be correct, $p_{7,2}$



Basic concepts

- $\eta = (\eta_1, \ldots, \eta_n)$: *n*-dimensional random variable
- the coordinates η_i of η are independent

  $$P(\eta_i = 1) = p;\ \ P(\eta_i = 0) = 1 - p\ \ (i = 1, \ldots, n)$$

  where $0 \le p \le 1$. (*n* algorithms)
- execute the experiment t times independently
- the outcomes in a table of size $n \times t$ (*j*-th column: the realization in the *j*-th experiment) (*t* objects)

Basic concepts

the random variables $\mu_1, \ldots, \mu_t$:

- if in the *j*-th column there are k ones then

  $$P(\mu_j = 1) = p_{nk}, \;\; P(\mu_j = 0) = 1 - p_{nk} \;\; (j = 1, \ldots, t);$$

  where the p_{nk}-s $(k = 0, 1, \ldots, n)$ are given numbers with

  $$0 \leq p_{n0} \leq \cdots \leq p_{nn} \leq 1.$$

- the μ_j-s are independent.

Basic concepts

Finally, put

$$\xi = |\{j : \mu_j = 1\}|$$

is the number of "good" decisions. Observe that all the individual decisions η_i $(i = 1,\ldots, n)$ are of binomial distribution with parameters (t, p). Then ξ is also of binomial distribution with the appropriate parameters.

Basic results

- *For any* $j = 1, \ldots, t$ *we have*

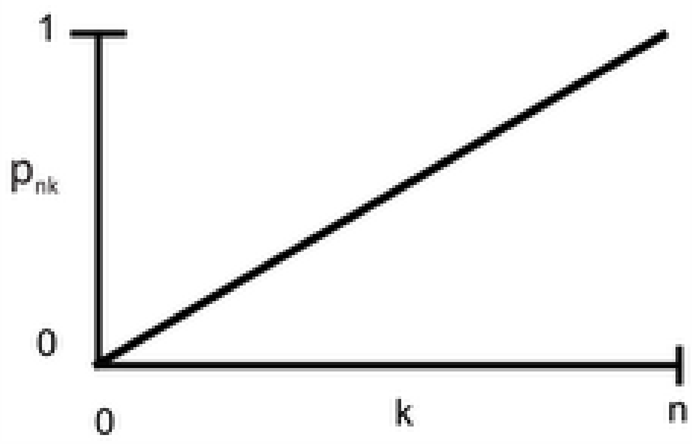
$$P(\mu_j = 1) = \sum_{k=0}^{n} p_{nk} \binom{n}{k} p^k (1-p)^{n-k}$$

- *Let* $q = P(\mu_j = 1)$. *The random variable* ξ *is of binomial distribution with parameters* (t, q).
- majority voting is "better" than the individual decisions, if $q \geq p$.

Basic results

- *Let* $p_{nk} = k/n$ $(k = 0, 1, \ldots, n)$. *Then we have*

  $q = p$ *and* $E\xi = tp$.



- *If we have* $p_{nk} \geq k/n$ *(*$k = 0, 1, \ldots, n$*), then*

  $q \geq p$ *and* $E\xi \geq tp$.

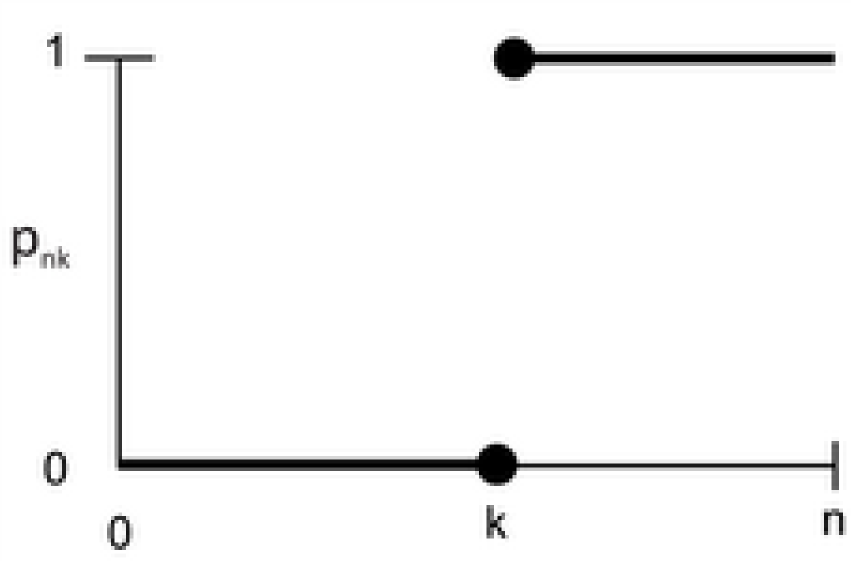
Special case (simple majority)

Suppose that n is odd, $p \geq 1/2$ *and*

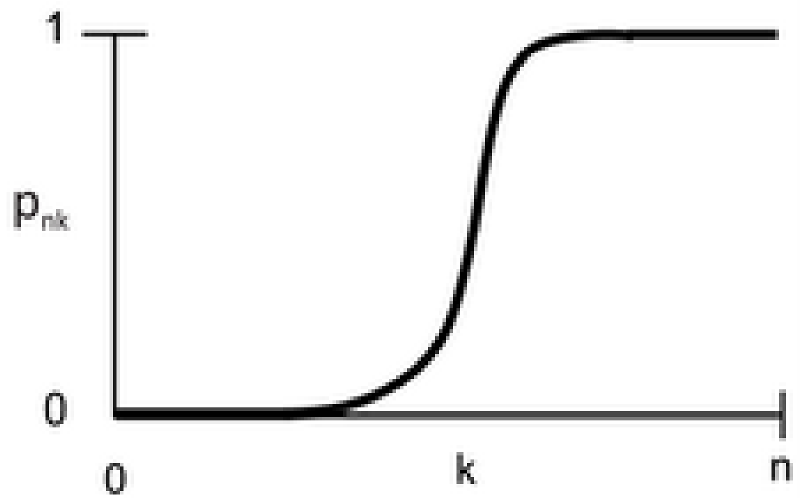
$$p_{nk} = 1, \text{ if } k > n/2$$

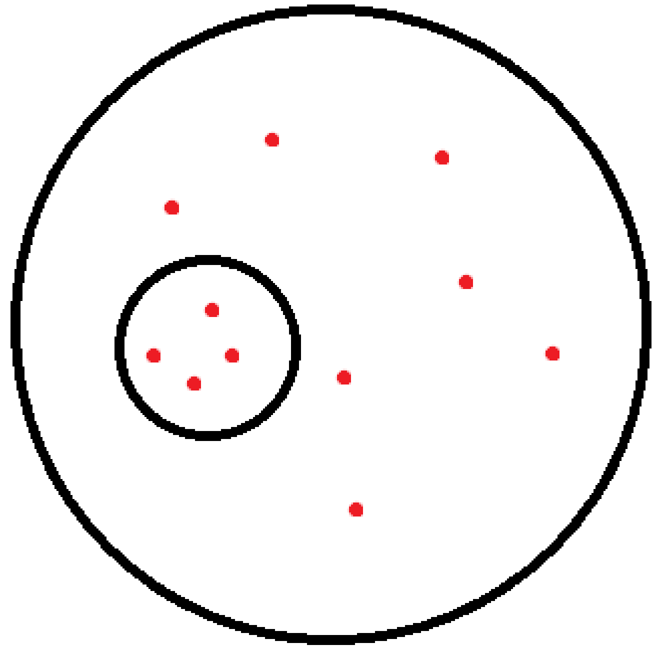
$$p_{nk} = 0, \text{ otherwise}$$

$(k = 0, 1, \ldots, n)$. *Then* $q \geq p$, *and consequently* $E\xi \geq tp$.



Optic Disc geometry





- increase exponentially in *k* for a given *n*.
- the probability that the diameter of a point set is not less than a given constant decreases exponentially (number of points to infinity)
- this diameter: the radius of the OD

Accuracy of correct classification

The majority voting method

	n = 3	*n*= 5	*n* = 7	*n*= 9
p = 0.6	0.6480	0.6826	0.7102	0.7334
p = 0.7	0.7840	0.8369	0.8740	0.9012
p = 0.8	0.8960	0.9421	0.9667	0.9804
p = 0.9	0.9720	0.9914	0.9973	0.9991

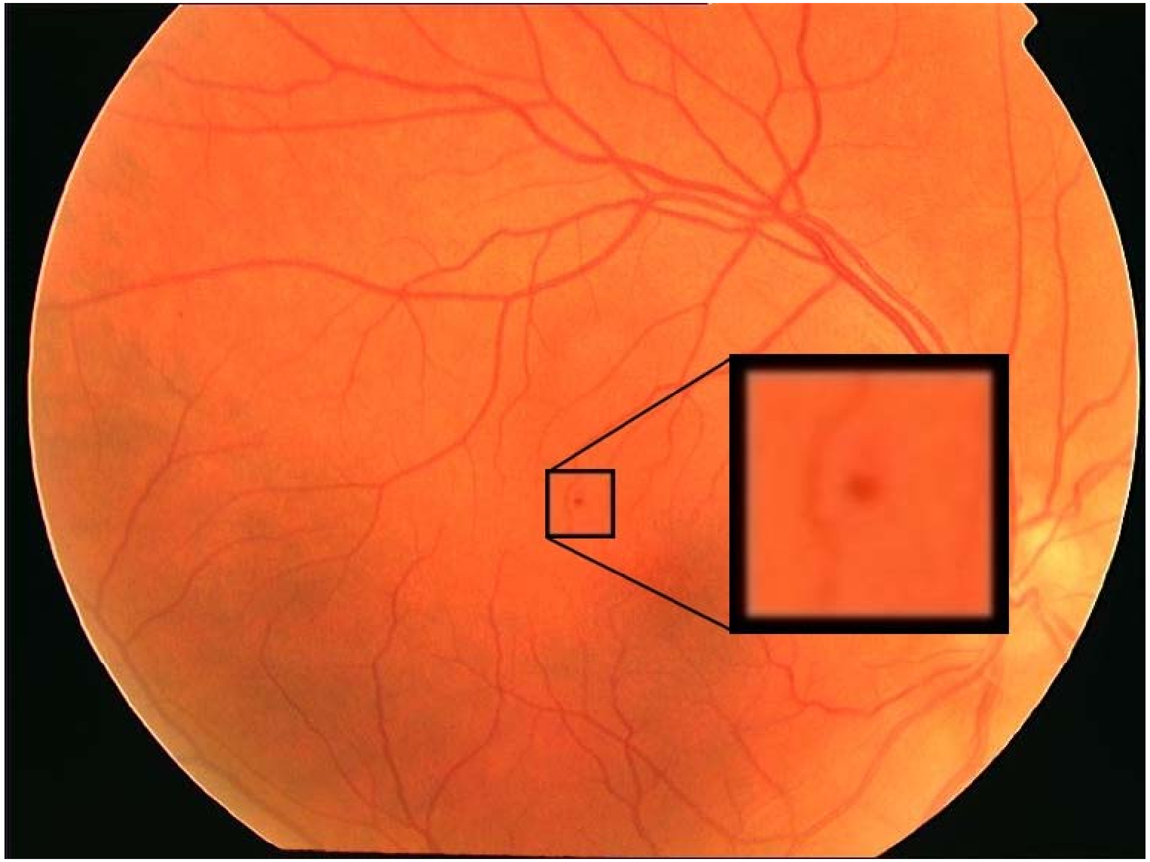
Spatial voting (optic disc geometry)

	n = 3	*n* = 5	*n* = 7	*n* = 9
p = 0.6	0.8208	0.8390	0.8895	0.9247
p = 0.7	0.9163	0.9373	0.9658	0.9823
p = 0.8	0.9728	0.9850	0.9942	0.9980
p = 0.9	0.9963	0.9988	0.9997	0.9999

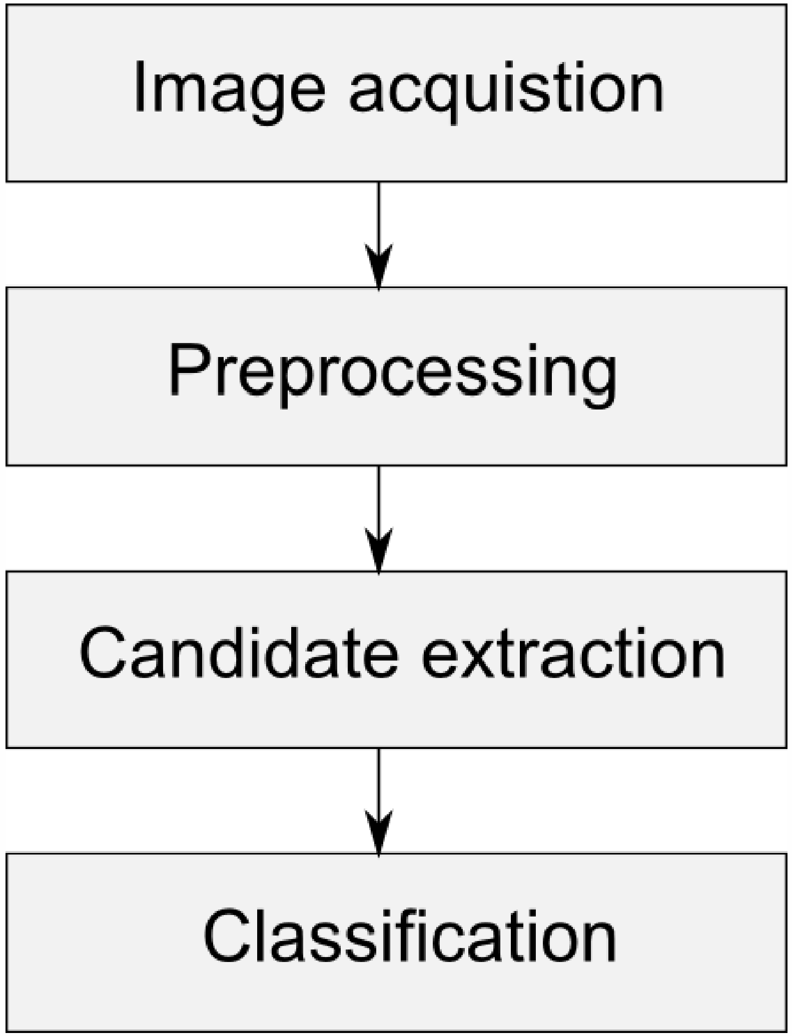
2nd example – microaneurysm detection

- Diabetic Retinopathy (DR)
- Early treatment
- Microaneurysm detection
- Hard to maintain reliability

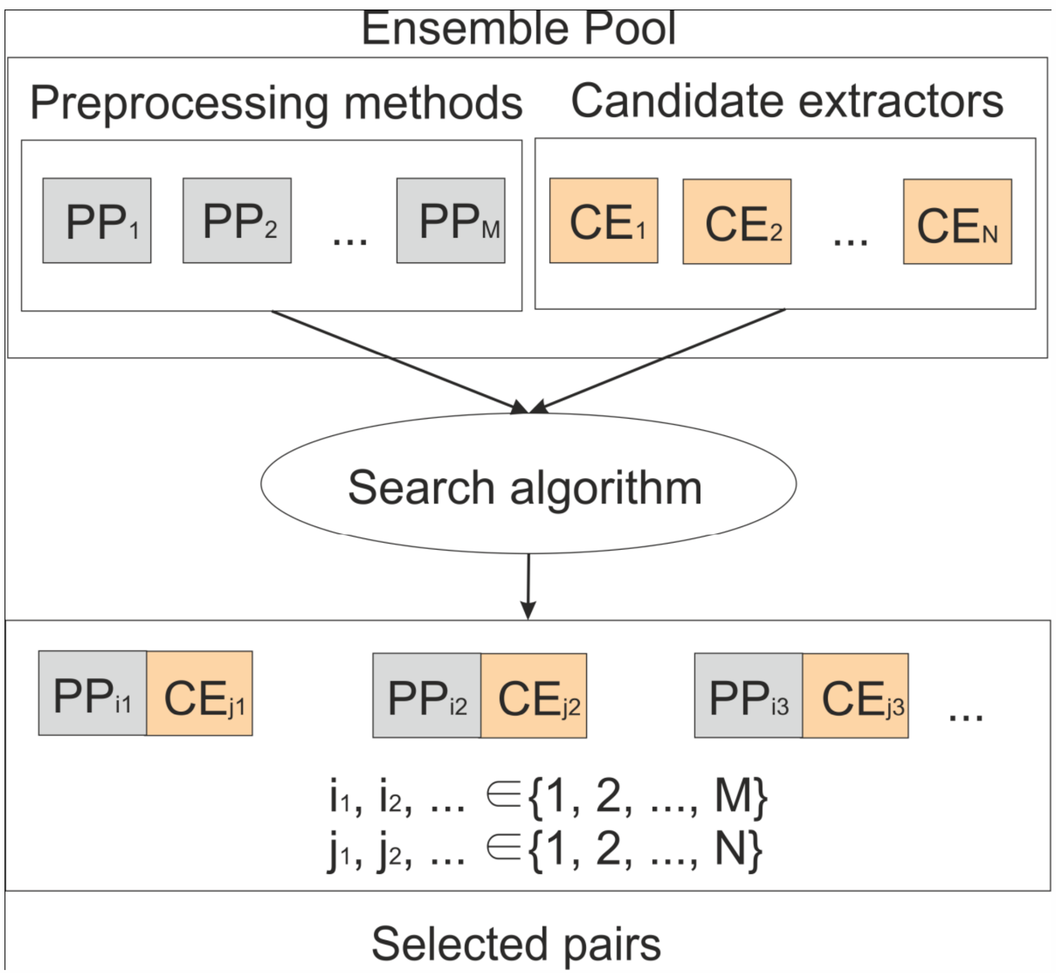
Automatic screening of DR



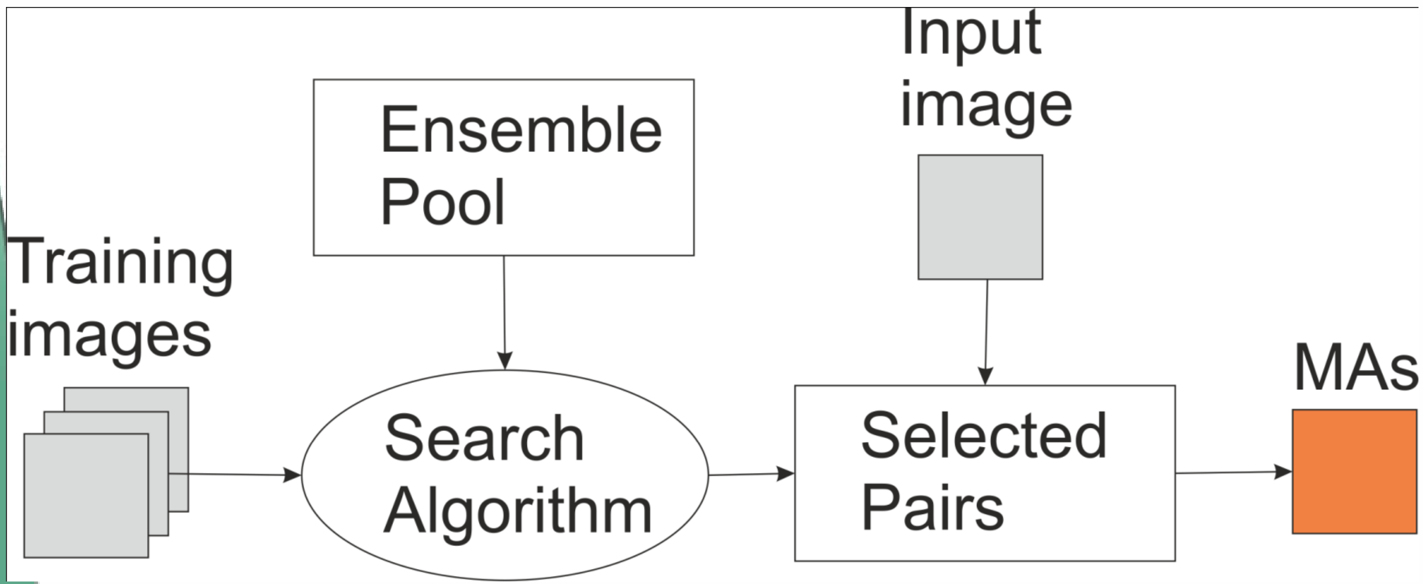
Usual steps of microaneurysm detection



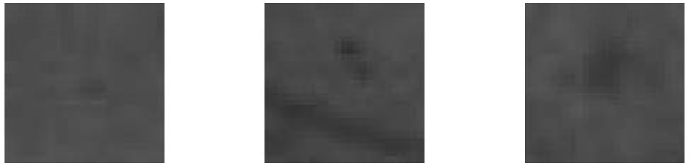
Steps of the proposed detector



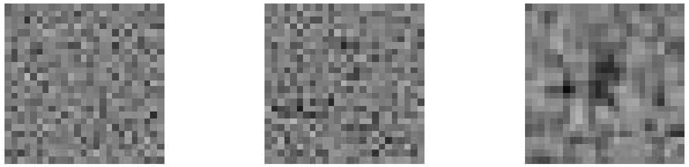
Steps of the proposed detector



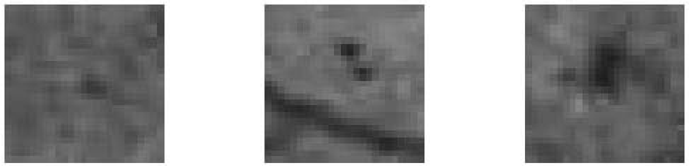
Preprocessing



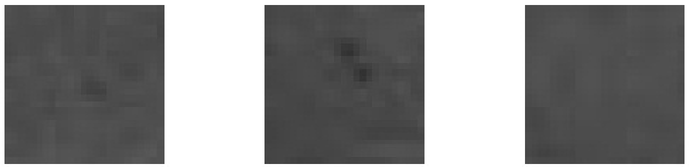
(a) Original



(b) Walter-Klein constrast enhancement

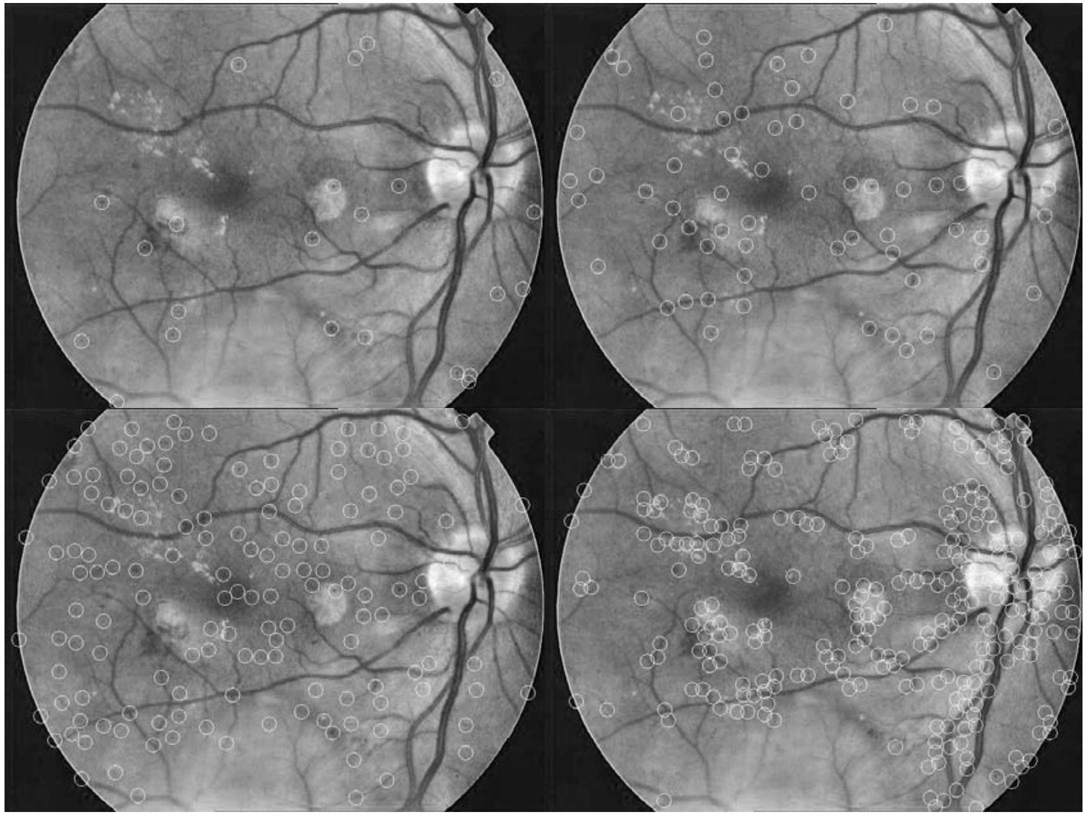


(c) CLAHE



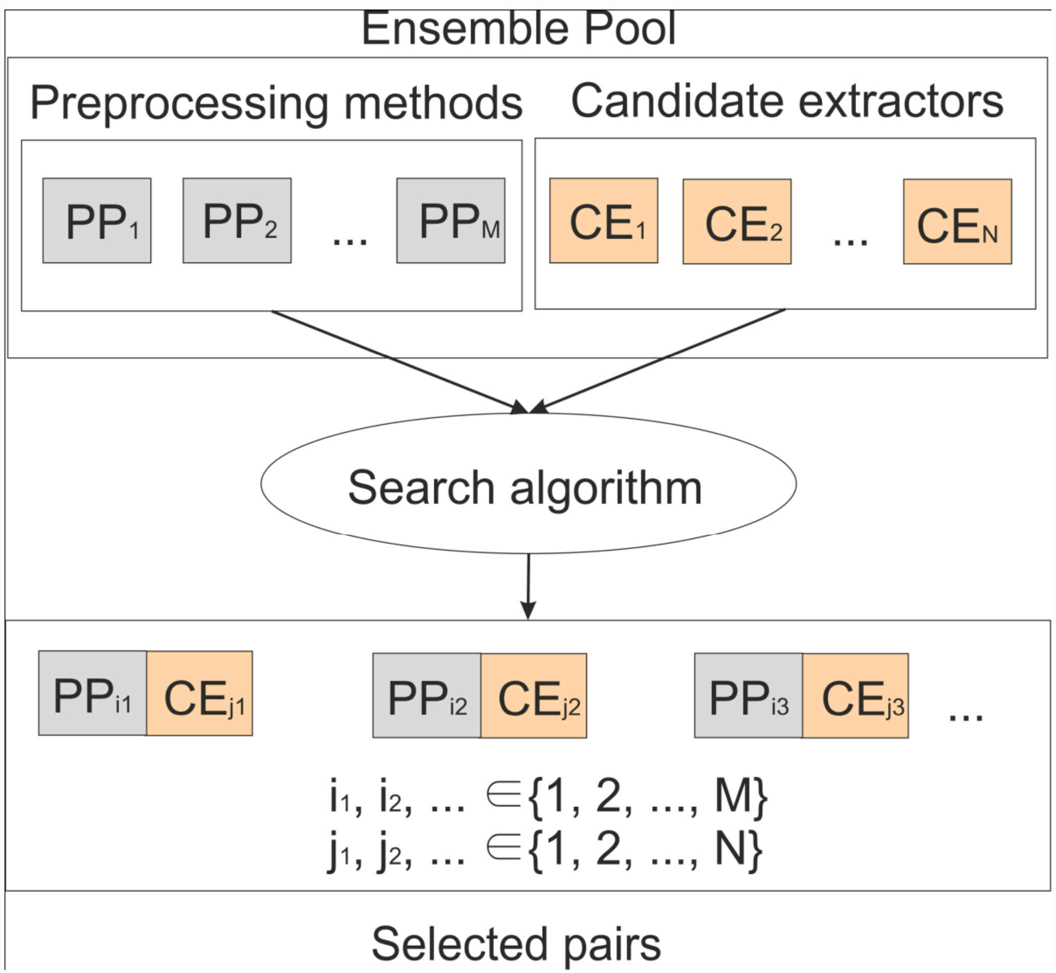
(d) Vessel removal and extrapolation

Candidate extractors



(a) Lazar (b) Walter
(c) Spencer (d) Hough

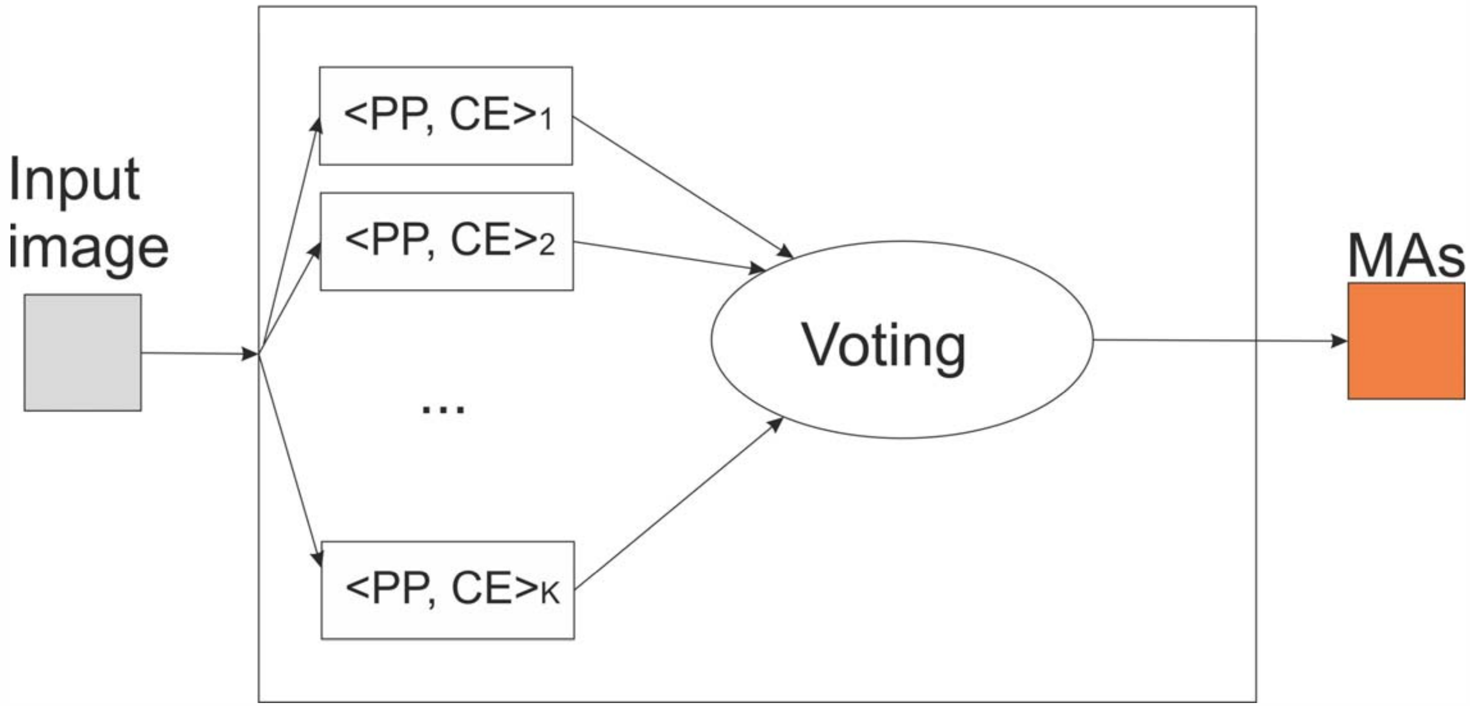
Ensemble creation



Searching

- We use a simulated annealing based algorithm
- We evaluate the possible ensembles using the Competiton Performance Metric (CPM): the average sensitivity at 7 fixed average false positive rates is calculated
- The ensemble with the highest CPM is selected

Voting



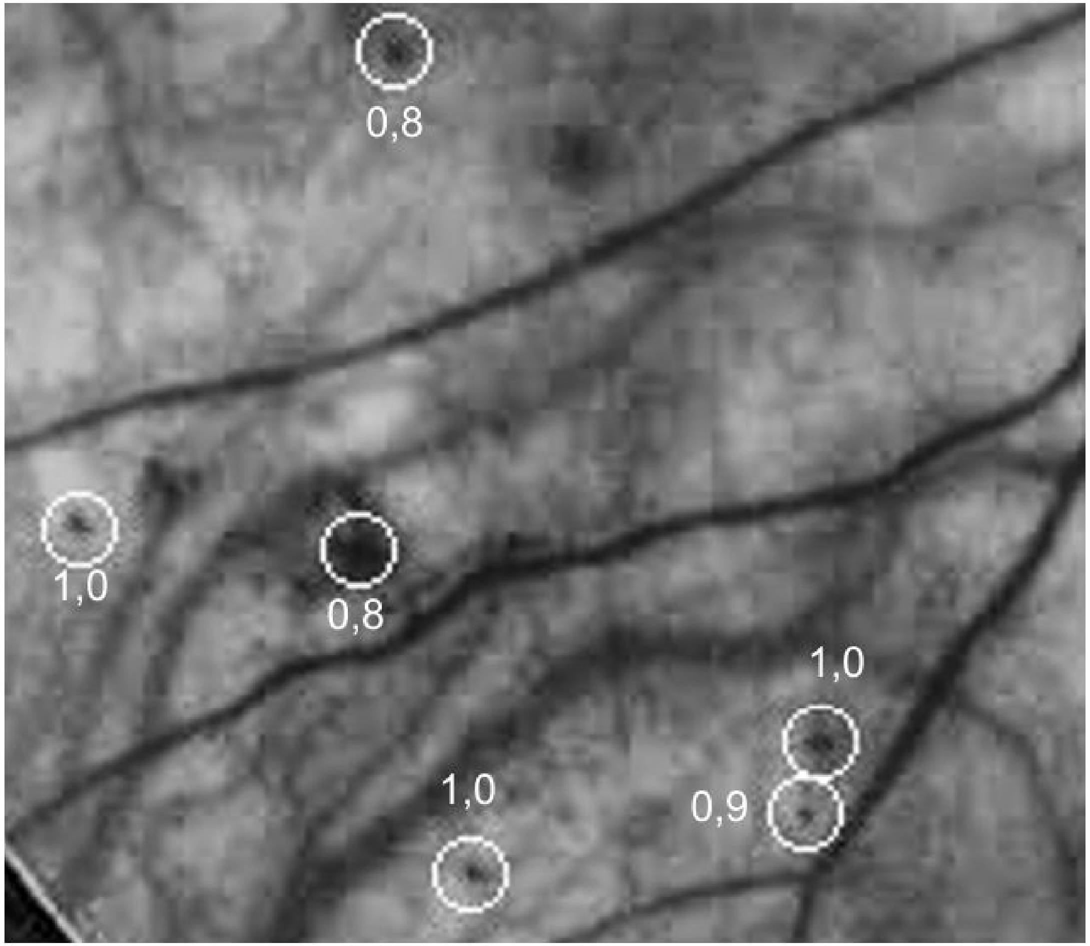
Voting scheme

•For each candidate, we count the number of pairs, for which the same candidate is present.

• We assign a confidence value C between 0 and 1 to each MA candidate c using the following formula:

$$C(c) = \frac{\text{the number of pairs where } c \text{ is present}}{\text{the number of pairs in the ensemble}}.$$

Result of voting



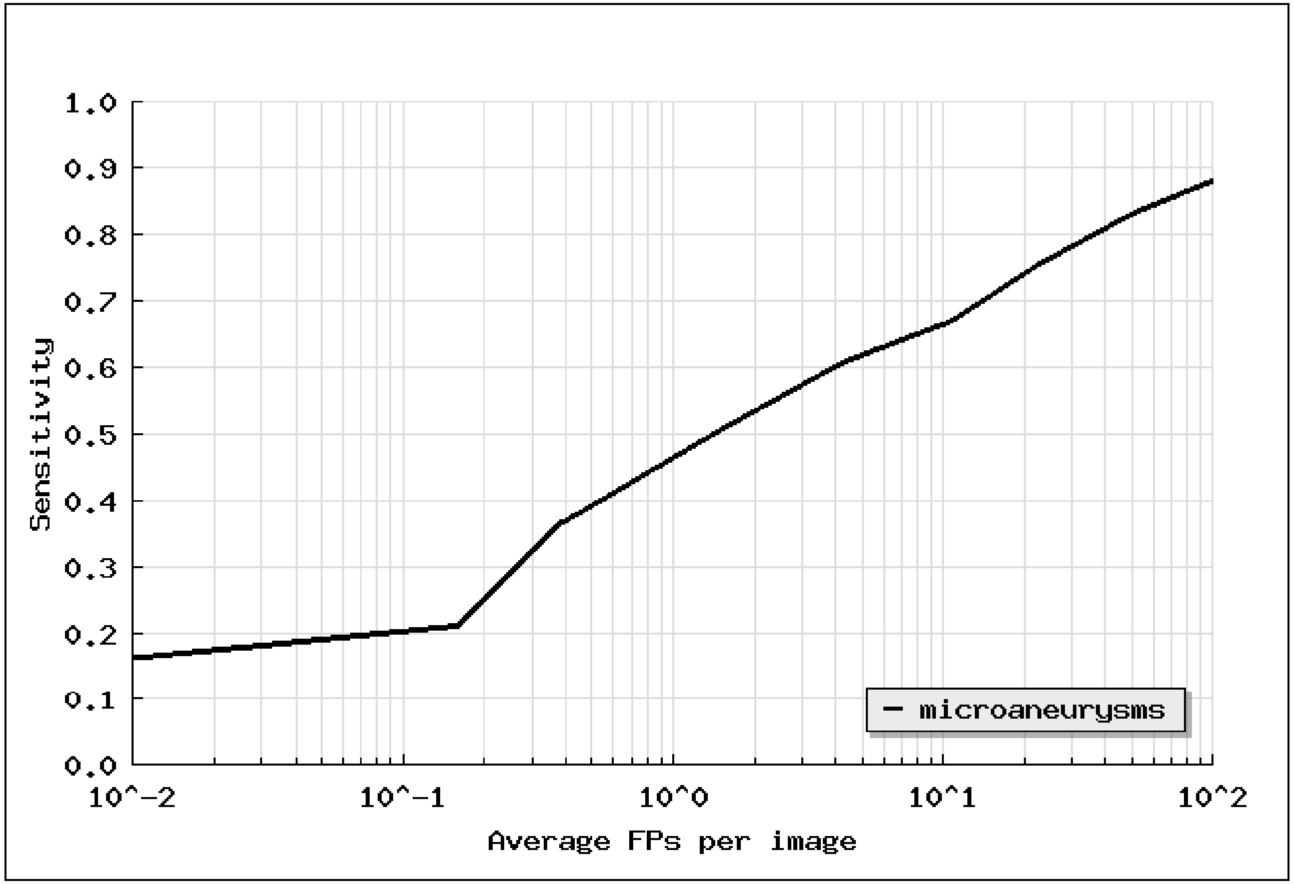
Results

- Retinopathy Online Challenge
- Independent evaluation of MA detectors
- 50 randomly selected image
- Detectors are compared using CPM

Pairs included in the ensemble

Preprocessing \ Candidate extractor	Walter	Spencer	Hough	Lazar	Zhang
Walter-Klein					•
CLAHE	•			•	
Vessel Removal				•	•
Illumination equalization				•	
No preprocessing	•			•	•

FROC curve


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Sensitivity
10^-2
10^-1
10^0
10^1
10^2
Average FPs per image
microaneurysms

CPM values

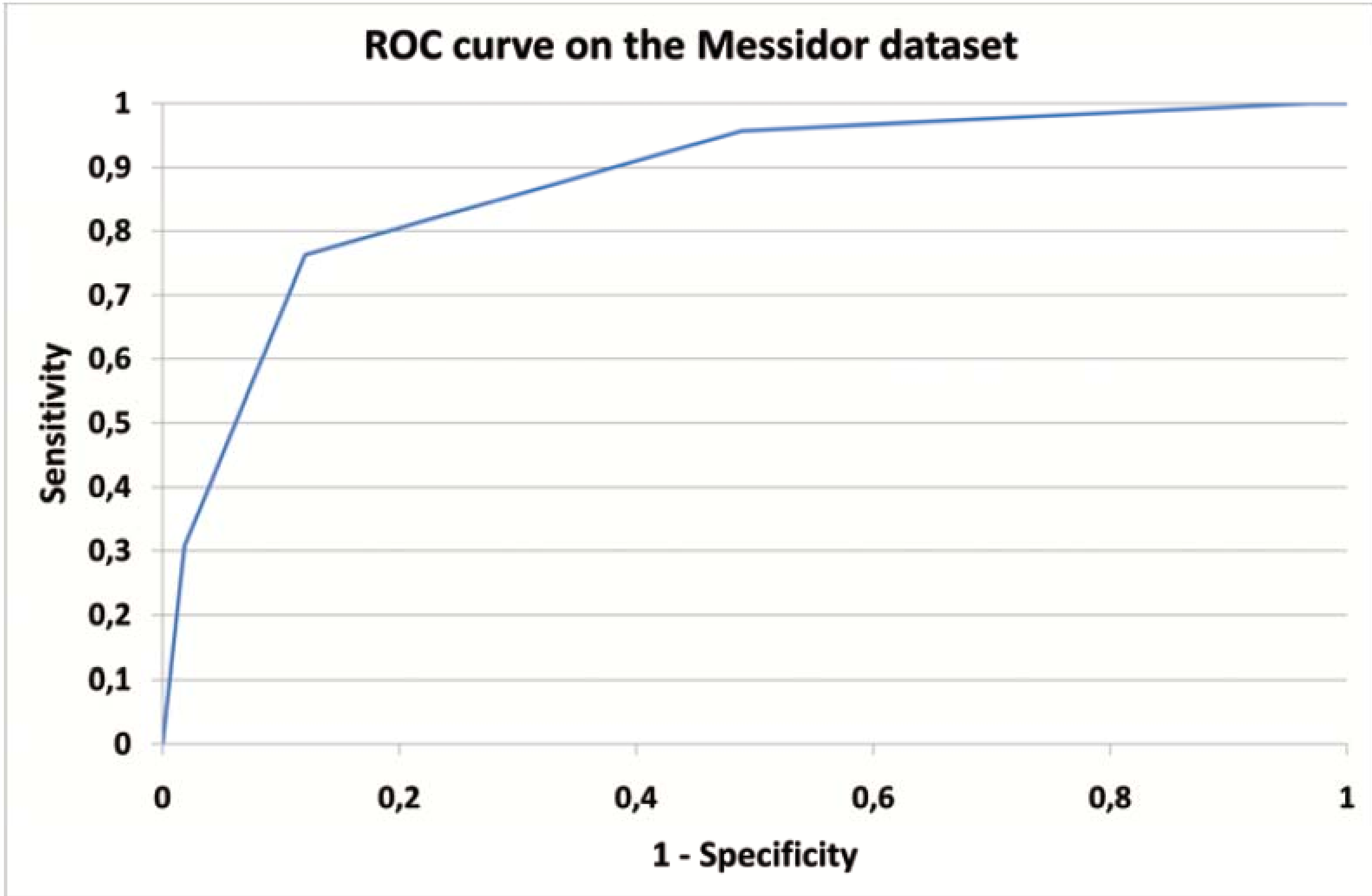
	1/8	1/4	1/2	1	2	4	8	avg.
DRSCREEN	**0.173**	**0.275**	**0.380**	**0.444**	**0.526**	**0.599**	**0.643**	**0.434**
Niemeijer et al.	0.243	0.297	0.336	0.397	0.454	0.498	0.542	0.395
LaTIM	0.166	0.230	0.318	0.385	0.434	0.534	0.598	0.381
OKmedical	0.198	0.265	0.315	0.356	0.394	0.466	0.501	0.357
Lazar et al.	**0.169**	**0.248**	**0.274**	**0.367**	**0.385**	**0.499**	**0.542**	**0.355**
GIB	0.190	0.216	0.254	0.300	0.364	0.411	0.519	0.322
Fujita	0.181	0.224	0.259	0.289	0.347	0.402	0.466	0.310
IRIA	0.041	0.160	0.192	0.242	0.321	0.397	0.493	0.264
ISMV	0.134	0.146	0.202	0.249	0.286	0.345	0.430	0.256
Waikato	0.055	0.111	0.184	0.213	0.251	0.300	0.329	0.206

Grading based on the presence of MAs

Measure \ Threshold	0.4	0.5	0.6	0.7	0.8	0.9	1.0
SEN	1	1	1	0.99	0.96	0.76	0.31
SPE	0	0.01	0.03	0.14	0.51	0.88	0.98
ACC	0.53	0.54	0.55	0.59	0.75	0.82	0.62

Class \ Threshold	0.4	0.5	0.6	0.7	0.8	0.9	1.0
R0	0.00	0.01	0.03	0.14	0.51	0.88	0.98
R1	1.00	1.00	1.00	0.97	0.92	0.60	0.18
R2	1.00	1.00	1.00	1.00	0.96	0.72	0.29
R3	1.00	1.00	1.00	1.00	0.98	0.92	0.42

Grading based on the presence of MAs



Final decision

- Several other features can be calculated besides MAs:
  - AM/FM
  - Prefiltering
  - MA detection
  - Exudate detection
  - Distance of the fovea and the optic disc
  - Compacteness of the ROI
  - Normalizing factor: diamater of the ROI

Results of the final decision

	ALL	FORWARD	BACKWARD
majority	99%/67%/81%	100%/0%/45%	98%/71%/83%
weighted majority	98%/67%/80%	100%/0%/45%	100%/0/%45%
avg	94%/79%/85%	**91%/83%/86%**	94%/77%/85%
mul	**94%/80%/86%**	**91%/86%/86%**	93%/78%/85%
max	60%/91%/77%	**93%/80%/86%**	64%/92%/71%
min	100%/52%/73%	86%/84%/85%	100%/54%/74%

Thank you

Thanks for your attention.