Artificial Intelligence in Medicine 60 (2014) 103–112

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Improving predictive models of glaucoma severity by incorporating quality indicators

Lucia Sacchi [a,*], Allan Tucker [b], Steve Counsell [b], David Garway-Heath [c], Stephen Swift [b]

[a] Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 1, 27100 Pavia, Italy
[b] Department of Information Systems and Computing, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
[c] NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, 162 City Road, London EC1V 2PD, UK

ARTICLE INFO

Article history:
Received 25 February 2013
Received in revised form 2 December 2013
Accepted 4 December 2013

Keywords:
Predictive modelling
Reliability indicators
Visual field testing
Glaucoma severity prediction

ABSTRACT

Objective: In this paper we present an evaluation of the role of reliability indicators in glaucoma severity prediction. In particular, we investigate whether it is possible to extract useful information from tests that would be normally discarded because they are considered unreliable.
Methods: We set up a predictive modelling framework to predict glaucoma severity from visual field (VF) tests sensitivities in different reliability scenarios. Three quality indicators were considered in this study: false positives rate, false negatives rate and fixation losses. Glaucoma severity was evaluated by considering a 3-levels version of the Advanced Glaucoma Intervention Study scoring metric. A bootstrapping and class balancing technique was designed to overcome problems related to small sample size and unbalanced classes. As a classification model we selected Naïve Bayes. We also evaluated Bayesian networks to understand the relationships between the different anatomical sectors on the VF map.
Results: The methods were tested on a data set of 28,778 VF tests collected at Moorfields Eye Hospital between 1986 and 2010. Applying Friedman test followed by the post hoc Tukey's honestly significant difference test, we observed that the classifiers trained on any kind of test, regardless of its reliability, showed comparable performance with respect to the classifier trained only considering totally reliable tests (p-value > 0.01). Moreover, we showed that different quality indicators gave different effects on prediction results. Training classifiers using tests that exceeded the fixation losses threshold did not have a deteriorating impact on classification results (p-value > 0.01). On the contrary, using only tests that fail to comply with the constraint on false negatives significantly decreased the accuracy of the results (p-value < 0.01). Meaningful patterns related to glaucoma evolution were also extracted.
Conclusions: Results showed that classification modelling is not negatively affected by the inclusion of less reliable tests in the training process. This means that less reliable tests do not subtract useful information from a model trained using only completely reliable data. Future work will be devoted to exploring new quantitative thresholds to ensure high quality testing and low re-test rates. This could assist doctors in tuning patient follow-up and therapeutic plans, possibly slowing down disease progression.

1. Introduction

Quality control during medical testing is an important issue in healthcare. Introducing unreliable, low-quality and noisy information into the care process can ultimately lead to incorrect diagnoses and therapy plans. High quality test delivery is important, but it should also take into account patient comfort and compliance, balancing the benefits of high quality measurements with the costs and harm potentially caused by over-repeating a test [1–3]. Checking reliability in a test can increase examination duration causing patient fatigue and affecting test outcome. Finding a trade-off between high quality care and patients' comfort during the testing procedure is still an open question in the medical community.

One of the most frequently applied techniques for ensuring high quality testing is to discard those tests that are considered unreliable. That said, it is often difficult to define the criteria for stating which tests are really trustworthy. These criteria are usually defined by the manufacturers of the testing devices and are conservative and general in nature. In this paper, we investigate whether it is possible to extract useful information from tests that may be discarded because they are considered unreliable. As a matter of fact, during clinical practice a qualitative evaluation of unreliable tests is often carried out to identify patterns that might

* Corresponding author. Tel.: +39 0382 985981.
E-mail address: lucia.sacchi@unipv.it (L. Sacchi).

0933-3657/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2013.12.002

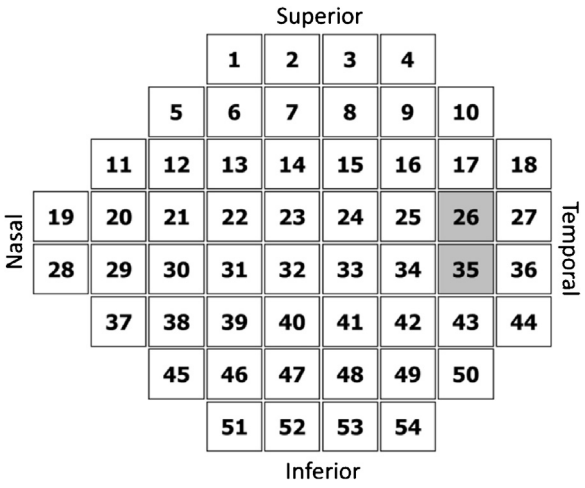

Fig. 1. Map of the visual field test locations for the right eye under the Humphrey Field Analyzer™ 24-2 program (the grey squares indicate the location of the blind spot).

suggest some useful information to the physician. This process, however, is not translated into a standard quantitative procedure and it is strongly dependent on the individual medical expert. The main objective of the present study is to evaluate the possibility of reducing the number of test repetitions while keeping the quality of the results and the patients' conditions at a high level. To this end, we consider an application of visual field testing (VF), a primary technique for detecting functional abnormalities related to glaucoma.

During a VF test, a subject is shown spots of light of varying brightness at several locations on the VF map. The number of locations on the map depends on the specific test program adopted in the clinical practice. VF tests are currently delivered through computerised automated perimetry. In this paper, we consider tests taken using second generation Humphrey Field Analyzers™ under the 24-2 program (HFA, Carl Zeiss Meditec, Inc., Dublin, CA). This protocol tests a total of 54 locations. At each location, the intensity of the stimulus that the subject is able to just see can be directly related to the sensitivity of retina to light. Each test location can be identified through use of a fovea-centred coordinates system or, more conveniently, a unique identification number, as shown for the right eye in Fig. 1.

VF testing is a good example of a testing paradigm where the reliability issue plays a crucial role. The process of defining retinal sensitivity needs to be performed using a small number of stimuli with the ultimate goal of minimising test duration. Algorithms devoted to the evaluation of patient reliability are included in the majority of testing programs. Besides sensitivity values, a VF test also collects global indices aimed at evaluating a patient's condition and the reliability of the test. Among these indicators, those intended to monitor reliability are the false positives rate (FP), the false negatives rate (FN) and fixation losses count (FL). In full-threshold automated perimetry, which is the technique we will consider in this study, an estimate of false positive and false negative rates is usually derived from a catch-trial strategy [4]. According to this technique, responses given during a pause in stimulus presentation are recorded as FP. On the other hand, when no response is given to a stimulus substantially brighter than the established threshold at a VF location, an FN response is recorded. FLs are determined by the count of positives responses to random stimuli at the location for the physiological blind spot [5–7].

Presently, tests results are not automatically adjusted for reliability and it is still up to the clinician to interpret an unreliable test. The challenge of determining which tests should be discarded is still open. Manufacturers of VF analysers have proposed certain thresholds on the reliability indices which, if exceeded, can be associated with a significant alteration of the visual field test results due to an unreliable patient. Once these thresholds are exceeded, a clinician is advised to discard the test and repeat it.

VF test reliability assessment has been addressed by many studies in the literature. In their early work, Katz and Sommer [8] described how manufacturer-defined reliability parameters are distributed into a clinic-based population. The effect of missed catch trials in normal subjects was studied in the work by Cascairo et al., where a population of normal controls was analysed when asked to purposely miss an increasing percentage of questions [9]. Bengtsson [5] considered reliability of VF test results expressed as threshold reproducibility, showing that it can be reasonably predicted by the amount of field loss with only a small contribution from reliability indices. Work by the same group [10] analysed the meaning of FN responses in a VF test, concluding that they were more related to the damage of the eye than to the patient's reliability characteristics. FP responses were addressed in both [6] and in [11]. In the first of these studies, the authors evaluated the accuracy of the catch-trial method for estimating the FP rate, proposing that tests showing less than 20% of FP should not be discarded and should be considered reliable instead. In the second study, two different test programs, full-thresholds and the Swedish interactive threshold algorithms (SITA standard) [12] were compared on their capability for estimating FP error frequency, with results showing that SITA suffered from FP underestimation on normal patients.

We address the problem of understanding the impact that test reliability indices have on the task of determining future disease severity. Many studies have investigated the appropriateness of reliability cut-offs for determining whether those thresholds reflect sensitivity alterations due to patient unreliability. To the best of our knowledge, the issue of how reliability indicators affect the results of the problem of predicting future disease severity has not yet been tackled.

We particularly focus on the following questions: is the information contained in unreliable tests totally impractical for determining the future evolution of the disease? How do different reliability indicators impact on prediction results? Are these contributions significantly different?

To answer these questions we setup a traditional predictive modelling framework [13,14], where we try to forecast future disease severity on the basis of visual field threshold sensitivities during a sample visit. In the literature, the problem of forecasting glaucoma evolution has been tackled by many authors [15]. It has been shown that machine learning classifiers (MLCs) out-perform traditional statistical analysis packages in glaucoma diagnosis [16–18]. Also, progression detection was shown to benefit from predictive modelling [19–21]. In these works, VF sensitivities have been used either to distinguish between glaucomatous and normal patients or to predict future disease progression based on specific criteria. In this paper we will consider disease severity, and we attempt to predict it in different reliability scenarios. To account for disease severity, we focus on scoring metrics used to assess visual loss.

A number of clinical trials have proposed scoring strategies to assess the severity of visual field loss for a specific test and to define disease progression during follow-up [22–24]. Advantages and limitations of such tests have been highlighted in several studies, mainly focusing on progression characterisation [25–28]. Even though effective visual field scoring and accurate progression definitions are highly desirable, the choice of standardised criteria is still very controversial in the clinical community. We selected the Advanced Glaucoma Intervention Study (AGIS) scoring system to assess VF test results. This choice comes from the current

practice adopted by our clinical partners at Moorfields Eye Hospital, together with the relevance of the criteria to the study objectives and the available data.

The rest of this paper is organised as follows: in Section 2 we state the predictive problem and we address and describe the methodologies used to tackle it. In Section 3, we outline the results we have obtained comparing different reliability scenarios in terms of classification performances. These results are then discussed in Section 4, where advantages and limitations of the study are highlighted. Finally, Section 5 provides concluding remarks.

2. Methods

One of the most important issues in glaucoma research is the attempt to model the temporal evolution of the disease. To this end, in collaboration with our clinical partners, we have collected a large database that stores approximately 54 thousand VF tests taken in a period from 1986 to the end of 2010. To the best of our knowledge, this is one of the largest databases considered in the literature for studies aimed at elucidating and forecasting glaucoma evolution.

2.1. Predictive task formalisation

We attempt to forecast field loss severity at the 'following' visit on the basis of VF data at the 'current' time point in different groups of patients classified on the basis of reliability indices. We address this issue as a standard classification problem [29], with the class variable being the AGIS score [22]. AGIS scores visual field defects based on the number and depth of adjacent depressed test locations in the visual field. Depressed locations are assessed with respect to the age-corrected normal thresholds available in the Humphrey perimeter. Depending on the sectors of the visual field, AGIS introduces different critical values for defining depression. AGIS is built as an additive score where the different areas of the visual field (nasal, upper hemi-field and lower hemi-field) contribute and sum to form the final metric value. The AGIS score ranges for 0–20, where a value of 0 stands for no defect while 20 stands for end-stage glaucoma.

Since our database is made up of time series of VF tests collected during patient follow-ups, to perform our classification task we selected a pair of visits from the time series to extract the variables and the outcome for the classification problem. Considering the nature of our study (a retrospective survey on patients who start being monitored at any stage of the disease), we chose to randomly sample one of the visits of the follow-up and the following visit for prediction. To this end, only patients with 2 or more visits were considered and, in the case one patient had only two tests, those two were selected. In the case one of the eyes of the patient was tested more than the other, that eye was selected for sampling. Otherwise, also they eye is randomly selected.

Once a visit is selected (sample visit at a generic time t), the variables related to that visit are extracted. Moreover, the AGIS score at the following visit (time $t+1$) is computed. As regards AGIS, we chose to use a 3-levels version of the score. The motivation underlying this choice is twofold: on the one hand, it allows a reduction in the number of values of the class variable and on the other it mimics the diagnostic process clinicians are currently using to rate a patient's status. AGIS values were grouped on the basis of clinical knowledge according to the qualitative categorisation presented in the AGIS paper [22]:

1. Mild defect: AGIS in the range [0,5]
2. Moderate defect: AGIS in the range [6,11]
3. Severe defect: AGIS in the range [12,20].

Table 1
Reliability scenarios considered in the predictive problem. Only reliable cases at $t+1$ were used.

Scenario	Reliability at time t
1	Reliable
2	Unreliable on FP
3	Unreliable on FN
4	Unreliable on FL
5	Any test

To better elucidate the real role of the quality indicators in forecasting visual loss severity we implemented a strategy that takes into account reliability indicators at time t and at time $t+1$. On the basis of the literature and the advice of the clinical expert, we considered false positives, false negatives and fixation losses as quality indicators. The reliability thresholds were set according to the manufacturer of the Humphrey Field Analyzer. As we consider the AGIS score as a means of evaluating disease severity, and since this score is specifically defined for tests delivered through full threshold strategy, we considered the reliability thresholds accordingly. In more detail, VF tests where FP or FN errors exceeded the 33% or where FL exceeded the 20% were considered unreliable [30]. Similarly, a test was considered totally reliable if all the constraints on the FP rate, FN rate and FL rate were met. We identified five possible scenarios, as listed in Table 1. These scenarios are built by considering several test reliability conditions at time t and totally reliable tests at time $t+1$. The restriction of taking into account only reliable tests at $t+1$ is made to avoid the introduction of noise on the outcome. Since the class variable is computed at time $t+1$, considering unreliable tests at that visit could lead to consideration of noisy class values that would negatively affect the training of the model.

In case 1, classification models are trained and tested on VFs which are completely reliable at time t. In case 2, we consider tests that fail to satisfy the reliability criteria on FP at time t. Case 3 investigates a situation where VFs are unreliable due to a high FN rate at time t. Scenario 4 considers a case where the visit at time t shows a failure due to fixation losses. Finally, case 5 considers a situation where we use any kind of test at time t, regardless of reliability indicators. As mentioned, for all five scenarios we require the tests to be completely reliable at time $t+1$.

The variables we considered in the analyses are the following:

1. VF sensitivities averaged according to groups of points related by their relationship to nerve fibre bundle (Fig. 2). To group the variables, we followed the well-known correspondence between

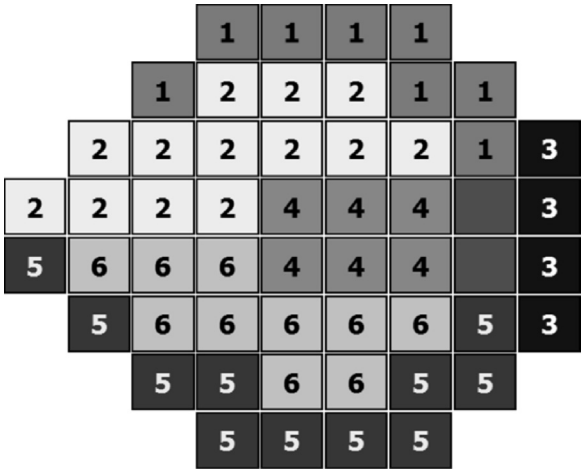

Fig. 2. Visual field sectors allocation to the nerve fibre bundles [31].

VF test locations on the retina and the nerve fibre bundle disposition [31]. According to this strategy, a total of 6 variables were obtained. Under the advice of the clinician, we used both variables at time t and variables at time $t+1$ as features.
2. AGIS score (3-levels version) at time $t+1$.

To perform the analysis presented so far, the data set was processed to extract only interesting pairs of visits (e.g., in Scenario 1 pairs of reliable tests). For this reason, even starting from a large set of patients, it is possible to be left with a small number of cases for some specific scenario. Reducing the size of the data set raises the important issues of sample size and class balancing. To overcome these potential problems, we have defined the bootstrap and class balancing procedure described in the following section.

2.2. Bootstrap and class balancing

To address the problems of small sample size and unbalanced classes, we implemented a strategy that couples bootstrap sampling with class balancing [32–34]. The methodology involves both the creation of the bootstrap samples to train the classifiers and of the test sets on which to evaluate the performance of the classification models. For each scenario, a balanced sampled data set is created according to the following pseudocode:

```
Input:
n_boot = number of bootstrap samples
C = number of classes in the problem
n_mC = number of examples in the minority class
D = complete data set (matrix where the number of rows
equals the total number of examples and the number of
columns equals the total number of variables)
Output: n_boot balanced data sets
Separate D into C subsets D_1, D_2, ..., D_C, each one
containing the examples of one class D = D_1 U D_2 U ... U D_C
for n = 1 to n_boot
    for i=1 to C
      B_i = {sample from D_i with replacement a number of
examples equal to the number of samples in D_i}
      If |B_i| > n_mC
        B_i = {take only the first n_mC examples from B_i}
      else
        num_test = number of examples NOT sampled from D_i to
make up B_i
      end
    end
    create a balanced training set B = B_1 UB_2 ... UB_C
    for i=1 to C
      Tt_i = {examples of D_i not sampled to create B_i}
      if |B_i| > n_mC
         Tt_i = {extract num_test randomly selected elements
from Tt_i}
      end
    end
    create a balanced test set T = Tt_1 U Tt_2 ... UTt_C
end
```

Once the *n_boot* training and test sets have been created, we build the classification models, obtaining performance indicators for each bootstrap sample we create.

To evaluate and compare the classification models we considered classification accuracy (CA) and the area under the ROC curve (AUC). The classifier we chose to exploit was Naïve Bayes. This choice was motivated by the characteristics of the algorithm, usually accurate despite the strong assumption of independency of the features [35]. This has been shown in many studies, especially related to clinical applications [36].

Since the models will be evaluated on several bootstrapped data sets and then compared on different scenarios, we needed to select suitable statistical instruments to correctly interpret the results. To perform multiple comparisons across scenarios, we used the Friedman test. When multiple scenarios are available, this test establishes whether there is a significant difference between those scenarios against the null hypothesis that they are all the same. To be able to state which pairs of scenarios are different we use the post hoc Tukey's honestly significant difference (HSD) test for multiple comparisons [37].

Table 2
Reliability indicators distribution in a sample of 1000 randomly extracted visits.

Indicator	Average % of patients (std)
FP	1.6 (0.1)
FN	8.3 (0.2)
FL	22.8 (0.3)

3. Results

3.1. Data set description

The global database we created contained 54,665 tests, including tests delivered with both full-threshold and SITA programs. As the AGIS score is specifically defined for full threshold tests, we considered only these tests in the analysis. A total of 28,778 tests met our requirements. These tests were related to 2318 patients, of which 2307 were tested on both eyes. The average number of follow-up visits was 6.2 (median: 3 visits, range: [1,42]). Overall, our data set contained 0.7% of tests with an FP rate greater than 33%, 4.5% with a FN rate greater than 33% and 13% of tests showing a FL rate outside the reliability boundaries (20%).

Following the methodology described in Section 2, for each patient we randomly extracted one visit and then consider that visit and the following to build the classification models. To understand how FP, FN and FL rates are distributed in the data set we will be analysing, we randomly performed the visit sampling 1000 times, obtaining the results shown in Table 2. In this table we report, for each reliability indicator, the percentage of patients that show a test outside the reliability thresholds, averaged over all the samplings; standard deviation (std) values are shown in brackets.

3.2. Glaucoma severity prediction models: selected results

After a first explorative analysis of the 5 proposed scenarios, the number of tests that were unreliable due to a high FP rate was small when compared with the total number of available tests in our study. In the whole data set made up of full threshold tests, only 200 were unreliable due to false positives. Moreover, once the 1000 balanced bootstrap samples had been created, the median number of FP faults was 12 per training set.

The behaviour observed in our data is confirmed by several studies in the literature that show that, among the three reliability indicators, false positives are the least variable and occur less frequently than fixation losses or false negatives [5,8,11].

For these reasons, we decided not to pursue the analysis related to those unreliable tests due to FP at time t (Scenario 2 in Table 1).

Fig. 3 shows the boxplots related to the classification accuracies obtained when using Naïve Bayes on 1000 bootstrap samples for all remaining scenarios.

From an initial qualitative analysis, we note that performance on tests that were unreliable due to a FN fault were poorer than those obtained by predicting on totally reliable tests (or on tests unreliable for other causes). As interestingly, the performances of the classifier appear comparable when we use totally reliable tests or any data at time t.

As shown by the boxplots in Fig. 3, we obtained average accuracy values greater than 80% in all cases, excluding only the one related to Scenario 3 (Unreliable FN → Good). These values are in line with the accuracy values published in the literature when MLCs were applied to glaucoma diagnosis prediction [17].

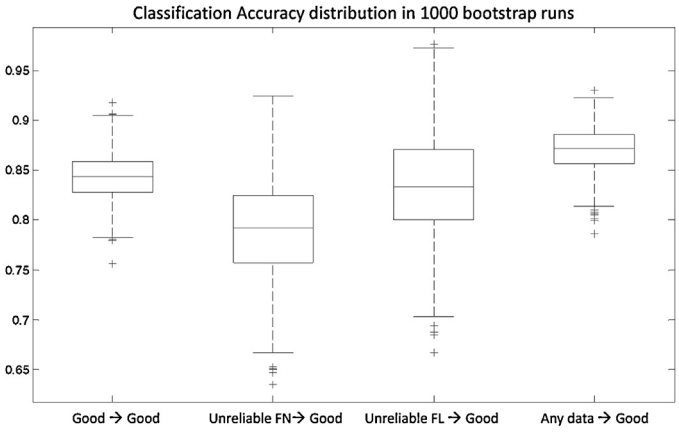

Fig. 3. Distribution of the classification accuracies obtained using a Naïve Bayesian classifier over 1000 bootstrapped datasets in the reliability scenarios detailed in Table 1.

To better evaluate the clinical relevance of these accuracies, together with the clinical expert we defined the baseline scenario to which the performance of our predictive models should be compared. In this scenario, we simply predict glaucoma severity at $t+1$ as the severity recorded at time t. For every scenario we used a Wilcoxon signed-rank test to compare the accuracies obtained by our model to the baseline accuracies over the 1000 bootstrap samples. We obtained significant results for all the scenarios with p-values beyond 0.01; our models out-performing the baseline classifier.

Observing the results presented in Fig. 3, we note that the variances of the accuracies are different across the four scenarios. This is confirmed by looking at Table 3 that reports, for each scenario, the values of the average classification accuracies and standard deviation, together with the number of examples contained in each of the bootstrapped training sets. Higher standard deviations indicate more variability in the classification results across the 1000 bootstrap samples. In our case, this happens for the scenarios involving the use of unreliable tests due to FN and FL.

Considering Fig. 3 and Table 3, we can observe that classification models trained on the largest data sets (i.e. the ones for the Good → Good and the Any → Good scenarios) show results with less variability and thus more robust. This is in line with the literature, where it has been shown that classification performance is correlated to data set size [38].

Starting from this observation, to be able to fairly quantify the differences among the four scenarios without advantaging those that can rely on larger training sets, we performed an additional analysis by slightly modifying the bootstrap algorithm presented in Section 2.2 with the aim of creating across-scenarios balanced data sets. Using this strategy, we have been able to obtain balanced data sets within a scenario and equal-size data sets across the different scenarios. This can be easily done by limiting the number of examples of each class in each data set to the number of examples of the less represented class across all scenarios.

Table 3
Data sets sizes and average classification performance in the different reliability scenarios.

Scenario	Mean classification accuracy (std)	Number of examples in each training set
Good → Good	0.8433 (0.0239)	543
Unreliable FN → Good	0.7897 (0.0490)	165
Unreliable FL → Good	0.8343 (0.0498)	138
Any data → Good	0.8709 (0.0213)	591

Fig. 4 shows the boxplots of the CAs obtained when using Naïve Bayes on the data sets balanced across the four scenarios. To cope with the reduced number of examples in the data sets, to perform these analyses we chose to increase the number of bootstrap samples to 10,000.

Comparing Fig. 4 to Fig. 3, it is possible to note that the performances of the classifier remain still comparable when we use totally reliable tests or any data at time t. Moreover, the result related to poor classification performances when training the model only on unreliable tests due to FN at time t is confirmed.

To compare classification results across the different scenarios, we applied the Friedman test and the post hoc Tukey's HSD test to these new data sets. The Friedman test resulted in a p-value <0.01, thus suggesting that there is a difference between the results obtained in the different scenarios. We proceeded by applying Tukey's test. As expected from Fig. 4, classification accuracies obtained using totally reliable tests and any type of test at time t are not significantly different (p-value >0.01). Results obtained using totally reliable tests and tests unreliable due to FL at time t are not significantly different (p-value >0.01). Predicting VF loss severity on unreliable tests due to FN at time t gave the worst results when compared to any other scenario (all the p-values result <0.01).

The same results in terms of statistical significance were obtained when examining AUCs across the different scenarios.

To describe the connections between variables at the same time point and shifting from one visit to the next, we used Bayesian networks (BNs). A BN was built for every case study considering the following variables: the class, the 6 VF sectors at time t and the 6 VF sectors at time $t+1$. Figs. 5–8 show the resulting networks coupled to a VF map representing the relationships between sectors at time t and time $t+1$. Variables are named according to the sector on the VF map and the time point they are related to (e.g. VF_1_t is the visual field sensitivity averaged according to sector 1 on the map considered at time t).

The networks that we show are summary networks derived by merging the results of 1000 runs of the bootstrap strategy. Networks were merged according to a strategy aimed at taking into account the number of times a link was preserved in the 1000 runs of the learning process. According to this strategy, a weight was given to each edge, representing the *confidence* of the connection it identified. The confidence of an edge is defined as follows [39]:

$$Confidence = \frac{1}{m}\sum_{i=1}^{m} e_i$$

where m is the number of bootstrap samples (1000 in this case) and e_i equals 1 if e is an edge in the ith network and 0 otherwise. In the presented graphs, we chose to plot only edges with a confidence >0.5.

In the figures, the networks are coupled with two VF maps, one for time t and the other for time $t+1$. An edge connects the VF areas linked in the BN. With this representation we aim to give a more intuitive picture of VF sector relationships from an anatomical viewpoint and to thus help the clinical interpretation of results.

4. Discussion

The aspect of checking patient's reliability during a VF test is very important since it should result in a more accurate threshold estimation. Nevertheless, it is also a time-consuming procedure that can increase test duration, thus fatiguing the patient [40,41]. Moreover, if a test turns out to be unreliable it needs to be repeated, resulting in additional testing time and possibly more noise. For these reasons, it is important to study how the quality of a test actually impacts the capability of assessing the damage related to VF loss. Finding a way to understand which tests need to be

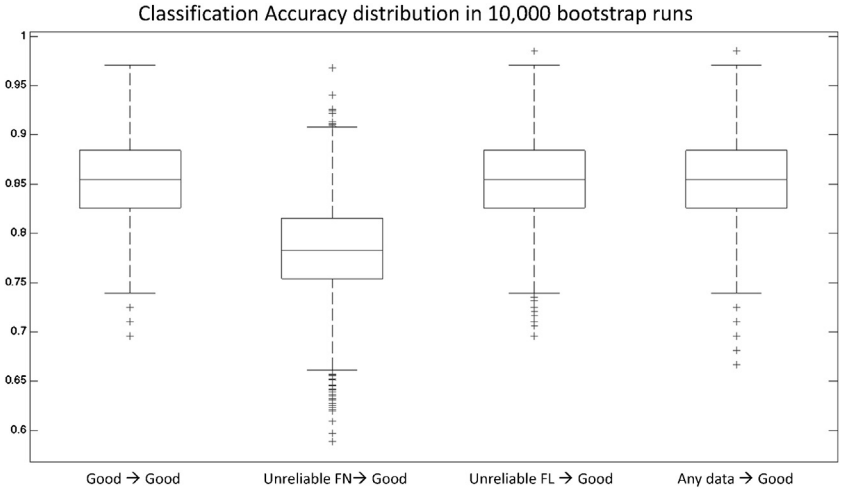

Fig. 4. Distribution of the classification accuracies obtained using a Naïve Bayesian classifier over 10,000 bootstrapped datasets balanced across the reliability scenarios detailed in Table 1.

discarded and which can still be helpful to elucidate the disease evolution is a crucial aspect of glaucoma research. The questions that we addressed in this paper are: did the information underlying an unreliable test make it useless to understand and predict glaucoma severity? Was there any difference among the quality indicators in the way they impact VF loss acuity prediction?

To answer these questions, we trained a Naïve Bayesian classifier in different clinical scenarios to evaluate the global meaning of

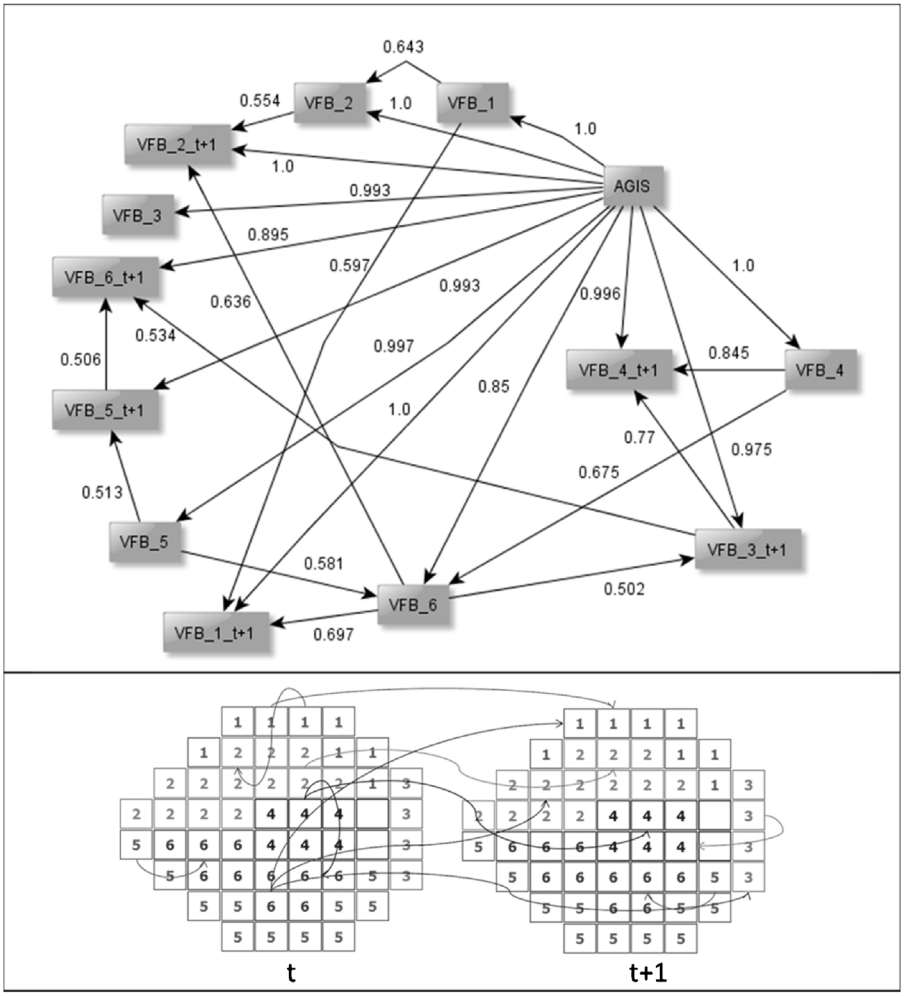

Fig. 5. Bayesian network and visual field maps related to reliability Scenario 1 (reliable data at time t and reliable data at time $t+1$).

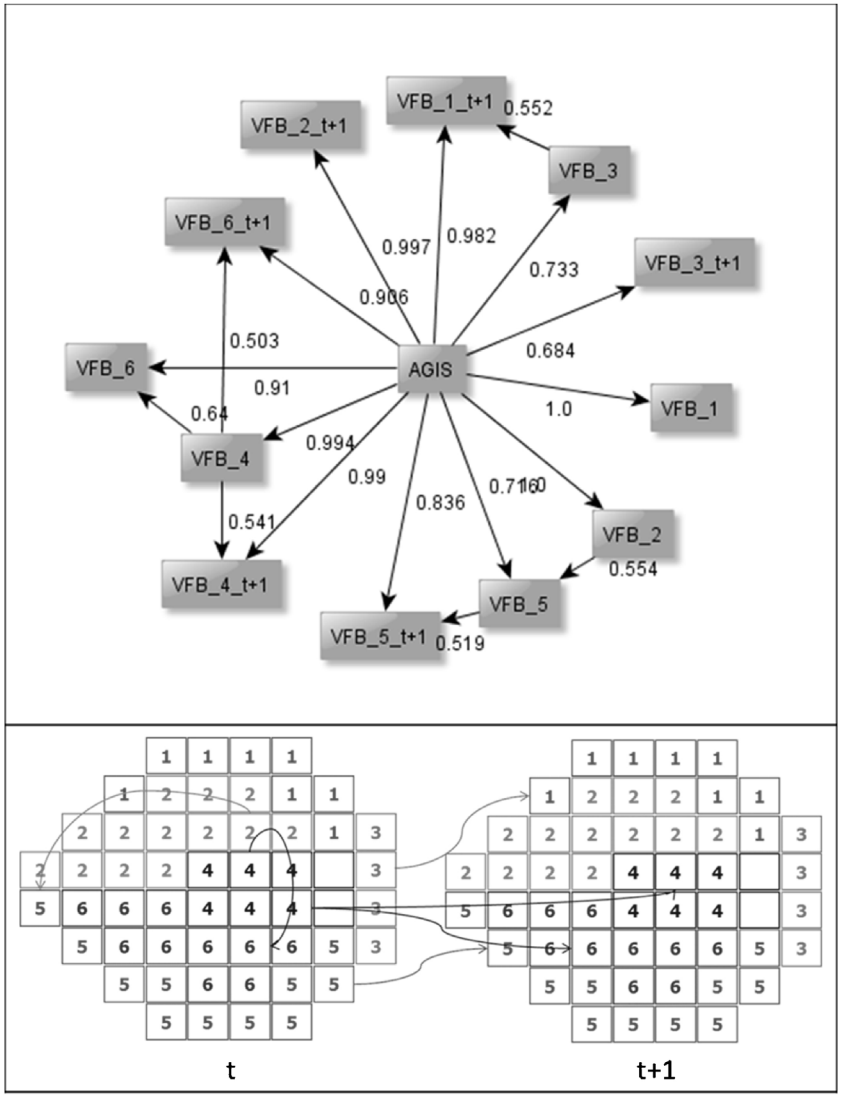

Fig. 6. Bayesian network and visual field maps related to reliability Scenario 3 (data unreliable do to FN at time t and reliable data at time $t+1$).

quality indicators and the effect that each of them had on prediction results.

A noteworthy result is that in Figs. 3 and 4. From these graphs we noticed that the classifier trained on any kind of test, regardless of its reliability, has performance comparable to the classifier trained only considering totally reliable tests at time t. This means that adding less reliable tests does not have a negative impact on the classification model obtained using only completely reliable data. These results are confirmed both by the CA and by the AUC performance indicators. This is an encouraging result, as it implies that discarding unreliable tests might not always be necessary and can possibly be avoided in cases when patients are particularly tired (e.g. aged or ill patients) or non-compliant. This would help in limiting patient's fatigue, which is often the cause for poor tests results.

As we were intuitively expecting, both AUC and CA showed high values when considering totally reliable tests in both consecutive visits. Interestingly, these values were higher but not significantly different from the ones obtained by using tests that failed for fixation losses at time t. This suggests that also including tests that do not respect criteria for fixation losses does not significantly affect the performance of the classifier. Fixation losses have been shown several times to be the most flexible reliability index since they depend on blind spot positioning. Some studies have showed that raising the FL Humphrey cut-off from 20% to 33% can significantly reduce the required re-test rates without compromising defect identification [42,43]. Coupling the literature conclusions with our results suggests that keeping FL unreliable tests in the analyses does not impact significantly on classification performance.

The worst classification results were obtained for Scenario 3, where variables related to tests exceeding the threshold on FN rate were used for prediction. In this case, using only tests which registered a high number of false negatives had a negative effect on classification results. Patients showing a high FN rate were patients that cannot see a supra-threshold stimulus given at a location where retinal sensitivity had been previously assessed. It is known from the literature that false negative responses correlate with patient status, increasing as glaucomatous VF loss increases [10,25,44]. This may result in visual field fluctuations that do not allow the patient to see spots of light (previously seen). Nevertheless, this has an impact on the quality of the recorded threshold values, originating from noisier and less reliable data. This direct effect on VF variables is probably what we detect in our classification results: poorer performances due to a higher level of noise.

The results from the analysis of the classification performances can be further supported by evaluating the BNs. These give us an

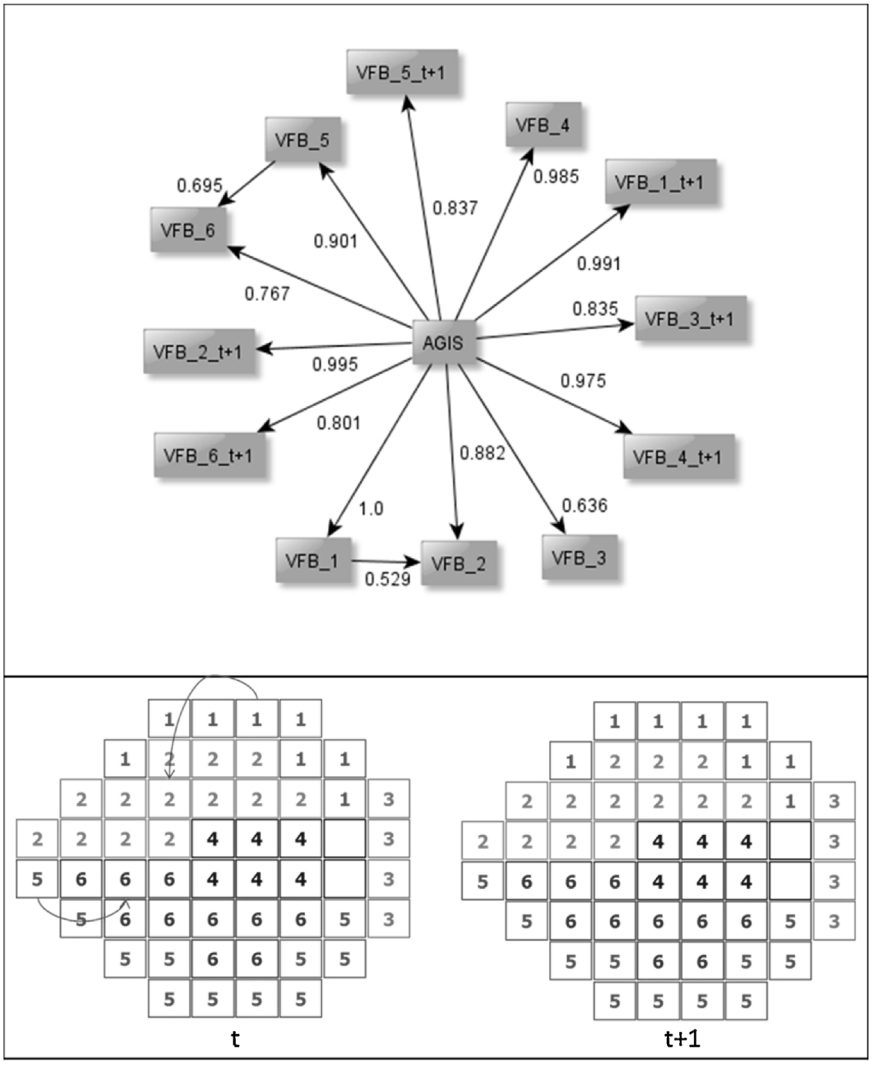

Fig. 7. Bayesian network and visual field maps related to reliability Scenario 4 (data unreliable do to FL at time t and reliable data at time $t+1$).

idea of how the variables (corresponding to different VF sectors) are linked to each other within the same visit and between two consecutive visits. As we see from Figs. 5 to 8, all the variables were found to be strongly connected to the class variable (AGIS), meaning that they were all useful for predicting the outcome. While this is more intuitive for variables at time $t+1$ (the AGIS score is related to the visual loss at the same visit), it is interesting that the AGIS score at the following visit can be predicted by considering the VF values at time t.

If we consider the links between the VF variables, we notice that the most connected network is the one extracted considering reliable tests at time t (13 connections between the variables, 10 of which link different visits). Both the networks extracted considering any data and unreliable tests due to FN at time t show 6 connections among the variables, while the less connected network is that related to the case of FL tests failing at time t (only 2 links between VF variables).

If we take a closer look at the strength of connections that characterise the networks, we can see that the network with, on average, the strongest links is the one corresponding to Scenario 5 (any data at time t), with an average weight of the links of 0.87. A high edge weight means that the link is strongly preserved across the 1000 bootstrap rounds. A high average edge weight in the network suggests that this is the most robust network and that its links can be trusted more than those in other networks. Interestingly, the network with weaker links turned out to be that extracted in the case of failed tests due to false negatives at time t (average links weight: 0.78). The fact that this network seems to be the less robust through the considered samples confirms results already obtained for classification performances, suggesting that using just FN unreliable tests to predict future disease severity is inadvisable. These tests are too noisy to be able to provide all the information needed to assess disease severity robustly.

To evaluate the links between the anatomical sectors of the visual field map, we decided to focus only on Scenarios 1 and 5. Considering the same time point, both scenarios extracted a link between sectors 1 and 2 and between sectors 5 and 6. It is interesting that the sectors involved in these pairs were anatomically adjacent to each other, suggesting that these areas were damaged concurrently. Moreover, the three sectors 2, 5 and 6 all include test locations belonging to the nasal part of the VF. The nasal step is a well-known glaucomatous visual field defect, which characterises the onset of the disease. This defect is often accompanied by more central defects [45], and this is confirmed in our results by the presence of the inner points belonging to regions 5 and 6 of the VF map.

The links between the variables at time t and variables at time $t+1$ are noteworthy, since they suggest a temporal pattern of

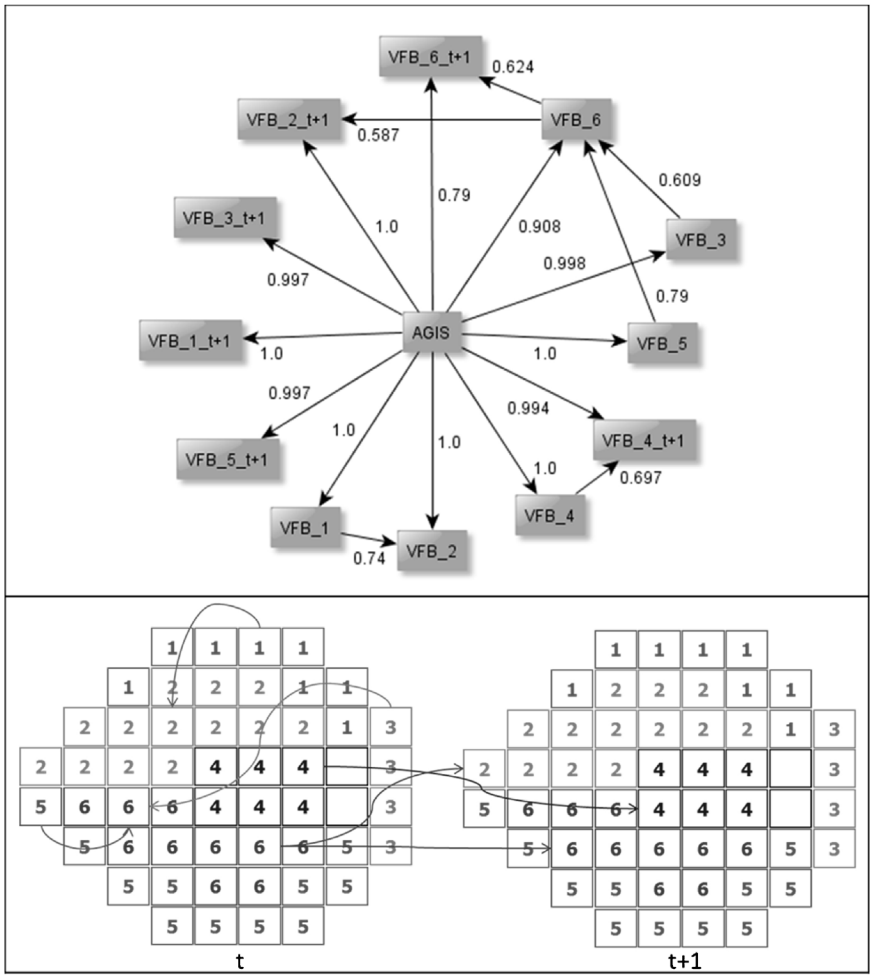

Fig. 8. Bayesian network and visual field maps related to reliability Scenario 5 (any data at time t and reliable data at time $t+1$).

evolution of the VF loss, describing how different areas of the visual field get depressed in time. This is important, as we are aware that visual defects follow specific patterns according to the anatomy of the retinal nerve fibre layer [46]. As derived using BNs, we can hypothesise that changes in a certain VF sector at one time can cause another sector at the following visit to change. Also from the literature, it is known that disease progression is most likely to occur in proximity to previous defects, as the affected part of the optic nerve head is weaker and more vulnerable to further changes [45]. The two scenarios that we consider agreed only on the link that connected VF6 at time t to VF2 at time $t+1$. This is a sign of a defect that, having developed in the lower hemi-field, crosses to the upper hemi-field. Some ring shaped defects are documented in the literature, where the damage starts affecting the upper or lower part of the visual field including a nasal step, which then moves towards the other hemi-field and spreads towards the centre [45].

5. Conclusions

In this study we presented an evaluation of the role of quality indicators in glaucoma visual field testing. We compared different reliability scenarios to understand the impact of introducing into the data analysis process tests that would normally be discarded as unreliable in clinical practice. We tackled the problem from a machine learning viewpoint, setting up a classification framework where we predicted glaucoma severity using AGIS categories as the outcome.

We obtained interesting results showing that classification modelling was not negatively affected by the inclusion of less reliable tests in the training process. Moreover, we showed that different quality indicators give different effects on prediction results. Training classifiers using tests that exceed the fixation losses threshold do not have a deteriorating effect on classification results. On the contrary, using only tests that fail to comply with the constraint on false negatives significantly decreases accuracy of the results.

The results of this study are encouraging and their translation into clinical practice is highly desirable as they could improve the current way tests are delivered. In order to complete this process though, some further investigations are required. Future work will be devoted to exploring new quantitative thresholds to ensure high quality testing and a low re-test rate. This could assist doctors in tuning patient follow-up and therapeutic plans, thus possibly slowing down disease progression.

Conflict of interest statement

Authors disclose that they do not have any known conflict of interest.

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. DFGH is also part funded by the International Glaucoma Association.

Acknowledgements

We would like to thank Dr Richard Russell for his precious help in data extraction. We are grateful to Francesca Mulas for her contribution on classifiers comparisons and. This project is funded by the EPSRC grant "Data Integrity and Intelligent Data Analysis Techniques Applied to a Glaucoma Progression Dataset" (EP/H019685/1). DFGH has part of his funding from the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology.

References

[1] Qaseem A, Alguire P, Dallas P, Feinberg LE, Fitzgerald FT, Horwitch C, et al. Appropriate use of screening and diagnostic tests to foster high-value, cost-conscious care. Ann Intern Med 2012;156(2):147–9.
[2] Institute of Medicine. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press; 2001.
[3] Vegting IL, van Beneden M, Kramer MH, Thijs A, Kostense PJ, Nanayakkara PW. How to save costs by reducing unnecessary testing: lean thinking in clinical practice. Eur J Intern Med 2012;23(1):70–5.
[4] Olsson J, Bengtsson B, Heijland A, Rootzén H. An improved method to estimate frequency of false positive answers in computerized perimetry. Acta Ophthalmol Scand 1997;75(2):181–3.
[5] Bengtsson B. Reliability of computerized perimetric threshold tests as assessed by reliability indices and threshold reproducibility in patients with suspect and manifest glaucoma. Acta Ophthalmol Scand 2000;78(5):519–22.
[6] Vingrys AJ, Demirel S. False-response monitoring during automated perimetry. Optom Vis Sci 1998;75(7):513–7.
[7] Heijland A, Krakau CE. An automatic perimeter. Acta Ophthalmol Suppl 1975;125:23–4.
[8] Katz J, Sommer A. Reliability indexes of automated perimetric tests. Arch Ophthalmol 1988;106(9):1252–4.
[9] Cascairo MA, Stewart WC, Sutherland SE. Influence of missed catch trials on the visual field in normal subjects. Graefe's Arch Clin Exp Ophthalmol 1991;229(5):437–41.
[10] Bengtsson B, Heijl A. False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? Invest Ophthalmol Vis Sci 2000;41(8):2201–4.
[11] Newkirk MR, Gardiner SK, Demirel S, Johnson CA. Assessment of false positives with the Humphrey Field Analyzer II perimeter with the SITA Algorithm. Invest Ophthalmol Vis Sci 2006;47(10):4632–7.
[12] Bengtsson B, Olsson J, Heijl A, Rootzén H. A new generation of algorithms for computerized threshold perimetry, SITA. Acta Ophthalmol Scand 1997;75(4):368–75.
[13] Bellazzi R, Ferrazzi F, Sacchi L. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdiscip Rev: Data Mining Knowl Discov 2011;1(5):416–30.
[14] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77:81–97.
[15] Bowd C, Goldbaum MH. Machine learning classifiers in glaucoma. Optom Vis Sci 2008;85(6):396–405.
[16] Chan K, Lee TW, Sample PA, Goldbaum MH, Weinreb RN, Sejnowski TJ. Comparison of machine learning and traditional classifiers in glaucoma diagnosis. IEEE Trans Biomed Eng 2002;49(9):963–74.
[17] Goldbaum MH, Sample PA, Chan K, Williams J, Lee TW, Blumenthal E, et al. Comparing machine learning classifiers for diagnosing glaucoma from standard automated perimetry. Invest Ophthalmol Vis Sci 2002;43(1):162–9.
[18] Boden C, Chan K, Sample PA, Hao J, Lee TW, Zangwill LM, et al. Assessing visual field clustering schemes using machine learning classifiers in standard perimetry. Invest Ophthalmol Vis Sci 2007;48(12):5582–90.
[19] Brigatti L, Nouri-Mahdavi K, Weitzman M, Caprioli J. Automatic detection of glaucomatous visual field progression with neural networks. Arch Ophthalmol 1997;115(6):725–8.
[20] Sample PA, Goldbaum MH, Chan K, Boden C, Lee TW, Vasile C, et al. Using machine learning classifiers to identify glaucomatous change earlier in standard visual fields. Invest Ophthalmol Vis Sci 2002;43(8):2660–5.
[21] Tucker A, Vinciotti V, Liu X, Garway-Heath D. A spatio-temporal Bayesian network classifier for understanding visual field deterioration. Artif Intell Med 2005;34(2):163–77.
[22] The Advanced Glaucoma Intervention Study Investigators. Advanced Glaucoma Intervention Study. 2. Visual field test scoring and reliability. Ophthalmology 1994;101(8):1445–55.
[23] Musch DC, Lichter PR, Guire KE, Standardi CL. The Collaborative Initial Glaucoma Treatment Study: study design, methods, and baseline characteristics of enrolled patients. Ophthalmology 1999;106(4):653–62.
[24] Leske MC, Heijl A, Hyman L, Bengtsson B. Early Manifest Glaucoma Trial: design and baseline data. Ophthalmology 1999;106(11):2144–53.
[25] Katz J, Congdon N, Friedman DS. Methodological variations in estimating apparent progressive visual field loss in clinical trials of glaucoma treatment. Arch Ophthalmol 1999;117(9):1137–42.
[26] Heijl A, Bengtsson B, Chauhan BC, Lieberman MF, Cunliffe I, Hyman L, et al. A comparison of visual field progression criteria of 3 major glaucoma trials in early manifest glaucoma trial patients. Ophthalmology 2008;115(9):1557–65.
[27] Vesti E, Johnson CA, Chauhan BC. Comparison of different methods for detecting glaucomatous visual field progression. Invest Ophthalmol Vis Sci 2003;44(9):3873–9.
[28] Katz J. Scoring systems for measuring progression of visual field loss in clinical trials of glaucoma treatment. Ophthalmology 1999;106(2):391–5.
[29] Hand DJ. Construction and assessment of classification rules. Chichester: Wiley; 1997.
[30] Heijl A, Lindgren G, Olsson J. Reliability parameters in computerized perimetry. Doc Ophthalmol Proc Ser 1987;49:593–660.
[31] Garway-Heath DF, Poinoosawmy D, Fitzke FW, Hitchings RA. Mapping the visual field to the optic disc in normal tension glaucoma eyes. Ophthalmology 2000;107(10):1809–15.
[32] Efron B, Tibshirani R. An introduction to the bootstrap. Chapman & Hall/CRC Monographs on Statistics and Applied Probability; 1993.
[33] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal 2002;6:429–49.
[34] Batista G, Prati R, Monard M. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explorations 2004:20–9.
[35] Hastie T, Tibshirani R, Friedman J. Elements of statistical learning: data mining, inference and prediction. 2nd ed. New York: Springer Series in Statistics; 2009.
[36] Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? Int Stat Rev 2001;69:385–98.
[37] Demsar J. Statistical comparisons of classifiers over multiple data set. J Mach Learn Res 2006;7:1–30.
[38] Sordo M, Zeng QT. On sample size and classification accuracy: a performance comparison. Lect Notes Comput Sci 2005;3745:193–201.
[39] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7(3-4):601–20.
[40] Heijl A, Drance SM. Changes in differential threshold in patients with glaucoma during prolonged perimetry. Br J Ophthalmol 1983;67:512–6.
[41] Hudson C, Wild JM, O'Neill EC. Fatigue effects during a single session of automated static threshold perimetry. Invest Ophthalmol Vis Sci 1994;35:268–80.
[42] Johnson CA, Keltner JL, Cello KE, Edwards M, Kass MA, Gordon MO, et al. Baseline visual field characteristics in the ocular hypertension treatment study. Ophthalmology 2002;109(3):432–7.
[43] Keltner JL, Johnson CA, Cello KE, Bandermann SE, Fan J, Levine RA, et al. Visual field quality control in the Ocular Hypertension Treatment Study (OHTS). J Glaucoma 2007;16(8):665–9.
[44] Katz J, Sommer A, Witt K. Reliability of visual field results over repeated testing. Ophthalmology 1991;98(1):70–5.
[45] Drance SM. The glaucomatous visual field. Br J Ophthalmol 1972;56(3):186–200.
[46] Rhee D, Uhler TA, Katz LJ. Psychophysical testing. In: Rhee D, editor. Glaucoma: color atlas and synopsis of clinical ophthalmology. New York: McGraw-Hill Medical Publishing Division; 2003. p. 120–35.