Genetic Algorithm Based on Support Vector Machines for Computer Vision Syndrome Classification

Eva María Artime Ríos[1], María del Mar Seguí Crespo[2], Ana Suarez Sánchez[3(✉)], Sergio Luis Suárez Gómez[4], and Fernando Sánchez Lasheras[5]

[1] Hospital Universitario Central de Asturias, Oviedo, Spain
[2] Department of Optics, Pharmacology and Anatomy, University of Alicante, Alicante, Spain
[3] Department of Business Administration, University of Oviedo, Oviedo, Spain
suarezana@uniovi.es
[4] Prospecting and Exploitation of Mines Department, University of Oviedo, Oviedo, Spain
[5] Department of Construction and Manufacturing Engineering, University of Oviedo, Gijón, Spain

Abstract. The inclusion in workplaces of video display terminals has introduced multiple benefits in the organization of the work. Nevertheless, it also implies a series of risks for the health of the workers, since it can cause ocular and visual disorders, among others.

In this work, a group of eye and vision-related problems associated to prolonged computer use (known as computer vision syndrome) are studied. The aim is to select the characteristics of the subject most relevant for the occurrence of this syndrome, and then, to develop a classification model for its prediction.

The estimation of this problem is made by means of support vector machines for classification. This machine learning technique will be trained with the support of a genetic algorithm. This provides different patterns of parameters to the training of the support vector machine, improving its performance.

The model performance is verified in terms of the area under the ROC curve, which leads to a model with high accuracy in the classification of the syndrome.

Keywords: Support vector machines · Genetic Algorithms · Computer vision syndrome

1 Introduction

Medical records means all the documents containing data, assessments and information of any kind on a patient's situation and clinical development throughout the care process. They include the identification of all the professionals who have intervened and aim to obtain the fullest possible integration on the clinical documentation of every patient, at least in the area of each center [1]. The current legislation allows processing the clinical information in electronic support, being the only exception the informed consent [1]. Since the introduction of the Electronic Health Records (EHR) in the National Health System, the health personnel have become users of video display terminals (VDT). Most of the workplaces with VDT are designed for a sedentary work. Nevertheless, in health

personnel, the workplace design frequently involves VDT and working standing up; this happens, for example, with the workstations for the administration of medication or with the portable terminal, which allows working beside the bed.

The work with VDT has introduced multiple benefits, but it also implies a series of risks for the health of the workers, since it demands certain requirements of physical and mental load. It can cause ocular and visual disorders, musculoskeletal disorders or other alterations as the mental fatigue. VDT's use during long periods of time relates to visual intense efforts [2, 3] as well as to changes in the ocular surface and in the condition of the tear film [4, 5]. It needs a good coordination of the eye movements in order to merge the images of both eyes and to obtain a suitable binocular vision [2]. In addition, during computer work the frequency of winking diminishes, increasing the evaporating of the tear film, which compromises the good condition of the ocular surface. The American Optometric Association [6] defines the computer vision syndrome (CVS) as a group of eye and vision-related problems that result from prolonged computer use. The most common symptoms associated with CVS are eyestrain, blurred vision, dry eyes and headaches, among others. In Spain, there are two instruments that recently have been developed and validated to measure this syndrome [7, 8].

CVS's prevalence is high, presenting variations in different studies, according to the characteristics of the sample, the method and the instruments of data collection used. It is related to individual and work-related factors. The higher the exposure to VDT, the greater the prevalence of CVS. Ranashinghe et al. [9], estimate CVS's prevalence in 67.4% in a group of 2,210 computer office workers of Sri Lanka. A lower prevalence (53%) has been observed in a study carried out with a sample of 426 office workers of Spain [10] and in a group of 476 call centers' operators in Brazil (54.6%) [11]. According to the bibliographical review, up to the date no studies have been identified that evaluate the effects on the visual health caused by the exposure to VDT in health personnel. The most common occupations of the samples studied in the revised bibliography include office workers, bank employees, high-tech workers, graphical editors and call centers' operators. Only two studies carried out in Turkey included a sample of workers who used computers from two hospitals. The first one [12] included secretaries, computer operators and hospital data management system users. The second one [5] did not specify whether they were health personnel or not.

The aim of the present work was to select the characteristics of the subject most relevant for the occurrence of the CVS, and then, to develop a classification model for its prediction in health personnel.

2   Methods

2.1   Case of Study

This was an observational cross-sectional epidemiological study, based on the completion of two self-administered questionnaires, on health personnel of the Monte Naranco Hospital of Oviedo (Spain). This hospital is specialized in geriatrics and palliative care and it started using the EHR in 2007.

The study included health personnel that were using VDT at work of the following occupational categories: physicians and surgeons, residents, nurses, advanced practice nurses (APNs) in training and auxiliary nurse.

Of a total of 172 workers, 151 (87.79%) took part in the study. The sample that fulfilled the inclusion and exclusion criteria was finally of 139 workers. The reasons of exclusion were: 3 for dry eye, 1 for amblyopia, 2 for non-surgically controlled cataracts, 3 for vitreous and 1 for retinal disease. Also one worker was excluded because he had been in his job for less than one year and another one for the lack of information on their seniority.

All the participants signed an informed consent form accepting their participation, in which data confidentiality was guaranteed during the entire process. The study was authorized by the Health Authority and approved both by the Research Ethics Committee of the Principality of Asturias and by the Ethics Committee of the University of Alicante, Spain (coordinating institution of the study), in accordance with the tenets of the Declaration of Helsinki.

2.2 Data Collection

The sanitary workers who took part in the study answered the following self-administered questionnaires:

1. Anamnesis and History of Exposure Questionnaire, which was specifically developed for this study, to gather information about of the gender, age, history of eye diseases and treatment, previous eye surgeries, job, work schedule, service or unit of work and daily VDT usage at and outside work.
2. The Computer Vision Syndrome Questionnaire (CVS-Q), designed and validated by Seguí et al. in 2015 [7], was used to measure perceived ocular and visual symptoms during or immediately following computer work. This questionnaire evaluates the frequency (never, occasionally or often/always) and the intensity (moderate or intense) of 16 ocular and visual symptoms: burning, itching, feeling of foreign body, tearing, excessive blinking, eye redness, eye pain, heavy eyelids, dryness, blurred vision, double vision, difficulty focusing for near vision, increased sensitivity to light, colored halos around objects, feeling that sight is worsening, and headache. Subjects with a score of 6 or more on the questionnaire are classified as symptomatic (suffering CVS).

2.3 Support Vector Machines

The support vector machines (SVM) are machine learning techniques. Among other mathematical models for similar problems [13–15], SVM are used to model physical systems through the adaptation on their parameters [16–19]. These methods are broadly known for their usage in classification and regression problems [20, 21]. In the case of the present research, the SVM is used as a classificator. This technique has been selected due to its well-known performance. The performance of SVM relies on the adjustment of the model to data previously taken from the system to be modelled, as training data

set. For training a SVM to model a classification problem, the vectors from the training data are used to map hyperplanes that define the separation of classes. The output estimation provided by a trained SVM can be formulated as:

$$\hat{y}_i = a^T \Phi(x_i) + b \tag{1}$$

Where x_i corresponds to the input vectors from the training set. The function $\Phi(x_i)$ linearizes the influences between inputs and outputs. In this scenario, the parameters are a and b, which are a vector of the same dimension as the image of Φ, and a coefficient, respectively. These parameters are determined by finding an optimized solution to the following problem and with restrictions:

$$\min_{a,\varepsilon,\eta_i,\eta_i'} \frac{1}{2} a^T a + c\left(\frac{1}{N} \sum_{i=1}^{N} (\eta_i + \eta_i') + v\varepsilon \right) \tag{2}$$

$$a^T \Phi(x_i) + b - y_i \leq \varepsilon + \eta_i \tag{3}$$

$$y_i - a^T \Phi(x_i) - b \leq \varepsilon + \eta_i' \tag{4}$$

Where c is a regularization parameter, ε is the tolerance error for each input x_i. Both η and η' are the slack variables, that take positive values. Finally, v is a parameter for the adjustment of the tolerance. Therefore, the output of SMV [22] can be expressed as:

$$\hat{y}_i = F(x) = \sum_{i=1}^{N} (\beta_i' - \beta) \Phi(x_i)^T \Phi(x) + b \tag{5}$$

In this expression, β and β' are the Lagrange multiplayers corresponding to the restrictions above. In this context, the kernel function K, can be defined as $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$. Consequently, the estimation of SVM turns into the following expression

$$\hat{y}_i = \sum_{i=1}^{N} (\beta_i' - \beta) K(x_i, x) + b \tag{6}$$

In general, the SVM for classification can be determined with the parameters c and v, since, as said before, a and b can be obtained as the optimal solution to the quadratic problem [20]. Depending on the sort of function chosen as kernel, other parameters should be determined as well.

2.4   Genetic Algorithms

The Genetic Algorithms are procedures developed to simulate the evolution of a population in terms of optimizing the survival of the next generation. These algorithms were first developed for chromosomic studies [23], but now genetic algorithms work with the premise of improving the fitness over the iterations until a solution for an optimization problem is reached. For each generation, the adjustments to the elements are based on four basic genetic operators used as criteria: crossover, mutation, reproduction and

elitism [24]. The optimization problem and the iterations of the genetic algorithm can be formulated as follows [25]:

For a function $f:D \rightarrow R$, and a set of restrictions, the minimization problem consists on the search of the best value x in the domain D such that $f(x) \leq f(y)$ for all y in the domain D. The value x' in the domain D is a local minimum of f, when a neighborhood $N(x')$ of x' exist where for all z in $N(x')$, $f(x') \leq f(z)$.

In this scenario, the genetic algorithm begins with an initial population $P_0 = \{I_0^1, \ldots, I_0^n\}$. Each step of the algorithm, the objective function is calculated along with its correspondent performance measures, then, generates a new population of which the elements are selected with a determined rule from the four genetic operators defined above. After m steps, the population is denoted as P_m. The algorithm stops when the performance measures are not significantly improved in a chosen number of iterations.

The aim of the proposed algorithm is to find the adequate variables and parameters with which the SVM will model properly the classification problem proposed. The iterations of the genetic algorithm are focused in maximizing the value of the AUC (Area Under Curve) of each of the SVM models calculated. The steps of the algorithm are as Fig. 1 shows.
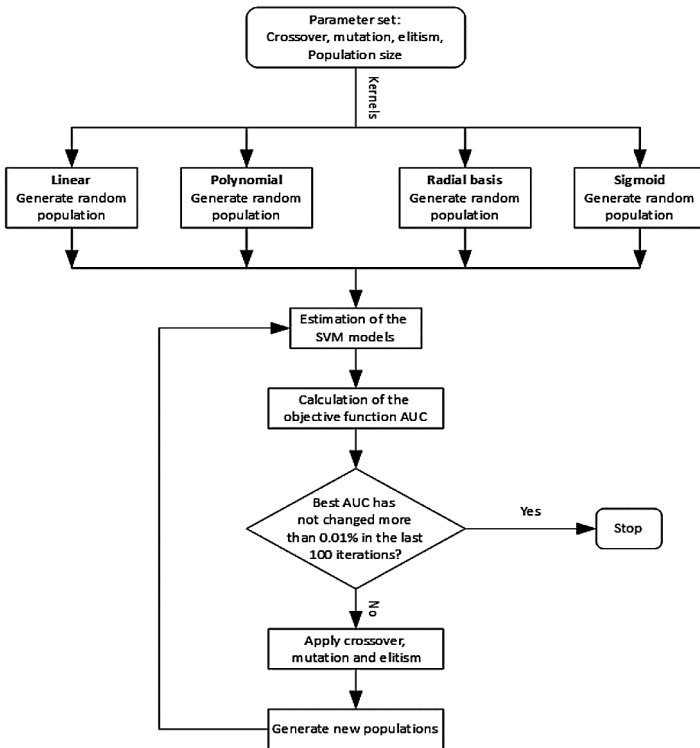

Fig. 1. Algorithm diagram.

The algorithm begins with the setting of the parameters for the genetic algorithm, such as crossover, mutation, elitism and population size. After this, an initial population should be created. As it was stated by Galán et al. [25], although the setting depends strongly in the data considered, a range of optimal parameters can be determined; for the present research, the probability values applied for crossover were those from 0.5 to 1 in steps of 0.1, while the mutation probability employed values of 0.1, 0.2 and 0.3. The elitism probabilities were 0.01, 0.05, 0.1 and 0.2. Different sizes for population of 10, 25 and 50 individuals were considered.

The set of the initial population must be a vector with size of the possible variation of parameters for the SVM. We will consider different types of initial populations, based on the type of kernel. To avoid this problem, we do consider a branching of the algorithm. Concretely, we will consider four ways in parallel; in each branch, the genetic algorithm is performed with each type possible of kernel. Now, the sets of initial populations can be created. Each item of the initial population has the following form, depending on the kernel used: Linear: $(c, v, x_1, \dots, x_k)$. Polynomial: $(\gamma, \alpha_0, \alpha, c, v, x_1, \dots, x_k)$. Radial basis: $(\gamma, c, v, x_1, \dots, x_k)$. Sigmoid: $(\gamma, \alpha_0, x_1, \dots, x_k)$. The parameters common for all the populations are: the cost c of constraints violation which corresponds with the constant of the regularization term in the Lagrange formulation of the SVM, ranging from 10^{-2} to 10^2; the tolerance error v for the determination of the SVM which takes values from $5 \cdot 10^{-4}$ to $5 \cdot 10^{-3}$; variables $x_1, \dots, x_k$, with values 0 or 1, depending if the variable is taken into account for the SVM model or not. For the other parameters, γ is needed in all kernels except linear and it takes values that range from $1/(2 \cdot data\,dimension)$ to $1/(data\,dimension)$; α_0 works as coefficient for the sigmoid and polynomial kernel (from -1 to 1), and finally, α for the degree determination of the polynomial kernel (from 3 up to 5). The SVM model is then estimated with the previous values for each of the branches, and its AUC with a validation subset of the data is calculated. The stop criteria will be satisfied if the AUC does not change more than a 0.01% in the last 100 iterations of the algorithm.

When the stop criteria is not satisfied, a new population has to be created. This is performed through implementing the crossover, mutation and elitism. With the elitism process, the populations were sorted based on their value of the AUC, and only the ones with higher AUC were considered. Crossover will combine sections of the chosen populations, and finally with the mutation process an aleatory modification is introduced to the population. Then the whole process is repeated until the stop criteria are satisfied.

3   Results and Discussion

3.1   Results

The chosen model was selected due to its AUC. The value of this performance metric was 0.9433029, giving a high performance over the validation data. The average AUC of the models for each iteration is shown in Fig. 2. For each iteration the AUC is computed and the average value presented in this Figure. As it can be observed, the performance of the models increases quickly in the firsts iterations, and afterwards

oscillates between the iteration 400 and the 900. Then, it raises again until the perform-ance level is high, and finally the average value remains stable, for, at least, 100 iterations and consequently the algorithm stops, determining the model chosen. This model is set with a sigmoid kernel, and parameters $\gamma = 2$, $1 \cdot 10^{-4}$, $\alpha_0 = 0$, $c = 1$ and $v = 5 \cdot 10^{-4}$.
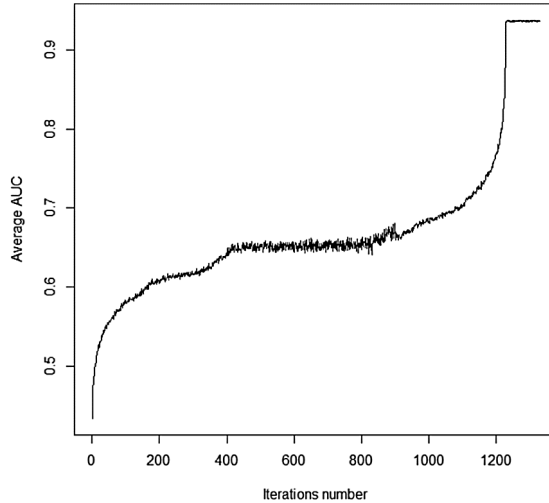

Fig. 2. Average AUC of the estimated models over the genetic algorithm iterations.

The variables chosen are: VDT usage outside work, nurse, hours a day VDT work-place, easy application, shifts including nights, current dryness, gender, current use of eye drops, contact lens and age. Also, the AUC was compared with other performance metric, the Youden index, to corroborate the robustness of the model. The comparison is performed via the correlation between both metrics, having a result of 0.8594404, which implies a high degree of correspondence in the calculated models.

3.2 Discussion and Conclusions

Workers who use the computer outside work, those who work as a nurse and those who are being exposed to VDT more hours per day at the workplace are more likely to suffer CVS. In other studies the total number of hours working on a computer in a day (both at work and outside the workplace) was found to be related to visual discomfort [26]. Also prolonged VDT use at work was connected with an increased risk of CVS [9, 10] and dry eye disease [27].

Female gender was associated with the risk of developing CVS as well. Many studies have reported an association between female gender and the prevalence of CVS [9–11]. In our study, age is also related with CVS; the average age of those with CVS is 45.9 years and the average age of those without CVS is 47.1. In contrast, Ranasinghe et al. [9] reported a mean age higher in those with CVS and Rosenfield [28] indicated that it is unclear whether CVS is associated with age. Portello et al. [29] observed greater

computer-related visual symptoms correlated with the number of hours spent using a computer, in females and patient with dry eye; however, they did not vary significantly with age.

Those who use eye drops for dryness are more likely to suffer CVS, as well as contact lens wearers. These results are consistent with previous studies that found higher prevalence of CVS [10] and also higher prevalence of dry eye disease [30] in contact lens wearers.

Work schedule with shifts including nights was also associated with the prevalence of CVS. Previous studies [31] suggest an association between rotating night shift work and several diseases, including cardiovascular disease, cancer risk, diabetes, hypertension, chronic fatigue, sleeping problems and early spontaneous pregnancy loss.

On the other hand, the prevalence of CVS was lower among those workers who considered the software application as "easy".

Nevertheless, these results should be interpreted with caution given the limitations of our study. The first one is that it is a cross-sectional design, and we cannot be sure that the cause precedes the effect. A second one is that we do not include ophthalmic examinations that inform us of the workers' refractive state. Finally, we do not take into consideration the use of mobile devices at and outside work, which could be a confounding factor.

Despite these limitations, the use of a validated questionnaire to measure CVS is a particular strength in this work, and this is the first study that describes the relationship of CVS and its associated factors in health personnel by means of genetic algorithm based on support vector machines. Both machine-learning techniques has been combined. SVM is used in order to perform classifications due to its well-known performance, while genetics algorithms are employed in order to optimize the SVM parameters. Finally, we would like to remark that the main contribution of the present research is a novel hybrid methodology able to determine the most important variables in a classification problem.

References

1. Gobierno de España: Ley 41/2002, de 14 de noviembre, básica reguladora de la autonomía del paciente y de derechos y obligaciones en materia de información y documentación clínica. 2002 (2008). http://www.boe.es/boe/dias/2002/11/15/pdfs/A40126-40132.pdf. Accessed 15 Mar 2016
2. Scheiman, M.: Accommodative and binocular vision disorders associated with video display terminals: diagnosis and management issues. J. Am. Optom. Assoc. **67**(9), 531–539 (1996)
3. Bergqvist, U.O., Knave, B.G.: Eye discomfort and work with visual display terminals. Scand. J. Work Environ. Health **20**(1), 27–33 (1994)
4. Fenga, C., Aragona, P., Di Nola, C., Spinella, R.: Comparison of ocular surface disease index and tear osmolarity as markers of ocular surface dysfunction in video terminal display workers. Am. J. Ophthalmol. **158**(1), 41–48 (2014)
5. Unlü, C., Güney, E., Akçay, B., Akçali, G., Erdogan, G., Bayramlar, H.: Comparison of ocular-surface disease index questionnaire, tearfilm break-up time, and Schirmer tests for the evaluation of the tearfilm in computer users with and without dry-eye symptomatology. Clin. Ophthalmol. **6**, 1303–1306 (2012)

 6. American Optometric Association. Computer Vision Syndrome. Accessed 16 Apr 2017
 7. del Mar Seguí, M., Cabrero-García, J., Crespo, A., Verdú, J., Ronda, E.: A reliable and valid questionnaire was developed to measure computer vision syndrome at the workplace. J. Clin. Epidemiol. **68**(6), 662–673 (2015)
 8. González-Pérez, M., Susi, R., Antona, B., Barrio, A., González, E.: The computer-vision symptom scale (CVSS17): development and initial validation. Invest. Ophthalmol. Vis. Sci. **55**(7), 4504–4511 (2014)
 9. Ranasinghe, P., et al.: Computer vision syndrome among computer office workers in a developing country: an evaluation of prevalence and risk factors. BMC Res. Notes **9**(1), 150 (2016)
10. Tauste, A., Ronda, E., Molina, M.-J., Seguí, M.: Effect of contact lens use on computer vision syndrome. Ophthalmic Physiol. Opt. **36**, 112–119 (2016)
11. Sa, E.C., Ferreira Junior, M., Rocha, L.E.: Risk factors for computer visual syndrome (CVS) among operators of two call centers in São Paulo, Brazil. Work **41**(Supplement 1), 3568–3574 (2012)
12. Yazici, A., et al.: Change in tear film characteristics in visual display terminal users. Eur. J. Ophthalmol. **25**(2), 85–89 (2014)
13. Antón, J.C.Á., Nieto, P.J.G., de Cos Juez, F.J., Lasheras, F.S., Viejo, C.B., Gutiérrez, N.R.: Battery state-of-charge estimator using the MARS technique. IEEE Trans. Power Electron. **28**(8), 3798–3805 (2013)
14. De Cos Juez, F.J., Sánchez Lasheras, F., García Nieto, P.J.: Suárez Suárez, M.A.: A new data mining methodology applied to the modelling of the influence of diet and lifestyle on the value of bone mineral density in post-menopausal women. Int. J. Comput. Math. **86**(10–11), 1878–1887 (2009)
15. Sánchez-Lasheras, F., de Andrés, J., Lorca, P., de Cos Juez, F.J.: A hybrid device for the solution of sampling bias problems in the forecasting of firms' bankruptcy. Expert Syst. Appl. **39**(8), 7512–7523 (2012)
16. Osborn, J., et al.: Using artificial neural networks for open-loop tomography. Opt. Express **20**(3), 2420–2434 (2012)
17. Guzmán, D., de Cos Juez, F.J., Myers, R., Guesalaga, A., Sánchez Lasheras, F.: Modeling a MEMS deformable mirror using non-parametric estimation techniques. Opt. Express **18**(20), 21356–21369 (2010)
18. Guzmán, D., de Cos Juez, F.J., Sánchez Lasheras, F., Myers, R., Young, L.: Deformable mirror model for open-loop adaptive optics using multivariate adaptive regression splines. Opt. Express **18**(7), 6492–6505 (2010)
19. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
20. de Cos Juez, F.J., García Nieto, P.J., Martínes Torres, J., Taboada Castro, J.: Analysis of lead times of metallic components in the aerospace industry through a supported vector machine model. Math. Comput. Model. **52**(7), 1177–1184 (2010)
21. Sánchez Lasheras, F., Vilán Vilán, J.A., García Nieto, P.J., del Coz Díaz, J.J.: The use of design of experiments to improve a neural network model in order to predict the thickness of the chromium layer in a hard chromium plating process. Math. Comput. Model. **52**(7), 1169–1176 (2010)
22. Vapnik, V.N., Vapnik, V.: Statistical Learning Theory, vol. 1. Wiley, New York (1998)
23. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge (1992)
24. De Falco, I., Della Cioppa, A., Tarantino, E.: Mutation-based genetic algorithm: performance evaluation. Appl. Soft Comput. **1**(4), 285–299 (2002)

25. Ordóñez Galán, C., Sánchez Lasheras, F., de Cos Juez, F.J., Bernardo Sánchez, A.: Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. J. Comput. Appl. Math. **311**, 704–717 (2017)

26. Robertson, M.M., Huang, Y., Larson, N.: The relationship among computer work, environmental design, and musculoskeletal and visual discomfort: examining the moderating role of supervisory relations and co-worker support. Int. Arch. Occup. Environ. Health **89**(1), 7–22 (2016)

27. Uchino, M., et al.: Prevalence of dry eye disease and its risk factors in visual display terminal users: the Osaka study. Am. J. Ophthalmol. **156**(4), 759–766 (2013)

28. Rosenfield, M.: Computer vision syndrome: a review of ocular causes and potential treatments. Ophthalmic Physiol. Opt. **31**(5), 502–515 (2011)

29. Portello, J.K., Rosenfield, M., Bababekova, Y., Estrada, J.M., Leon, A.: Computer-related visual symptoms in office workers. Ophthalmic Physiol. Opt. **32**(5), 375–382 (2012)

30. Kojima, T., et al.: The impact of contact lens wear and visual display terminal work on ocular surface and tear functions in office workers. Am. J. Ophthalmol. **152**(6), 933–940 (2011)

31. Suárez Sánchez, A., Riesgo Fernández, P., Sánchez Lasheras, F., de Cos Juez, F.J., García Nieto, P.J.: Prediction of work-related accidents according to working conditions using support vector machines. Appl. PPL Math. Comput. **218**(7), 3539–3552 (2011)