Chapter 8

Modeling a second-generation glucose oxidase biosensor with statistical machine learning methods

Livier Rentería-Gutiérrez[1], Lluís A. Belanche-Muñoz[2], Félix F. González-Navarro[1*], Margarita Stilianova-Stoytcheva

[1]Instituto de Ingeniería. Blvd. Benito Juárez y Calle de la Normal S/N 21280 Mexicali, Mexico.

[2]Universitat Politècnica de Catalunya. Dept. de Llenguatges i Sistemes Informàtics C/Jordi Girona, 1-3, 08034 Barcelona, Spain.

livier.renteria@uabc.edu.mx, belanche@lsi.upc.edu, fernando.gonzalez@uabc.edu.mx, margarita.stoytcheva@uabc.edu.mx

*Corresponding author

Doi: http://dx.doi.org/10.3926/oms.166

Referencing this chapter

Rentería-Gutiérrez, L., Belanche-Muñoz, L.A., González-Navarro, F.F., & Stilianova-Stoytcheva, M. (2014). Modeling a Second-Generation Glucose Oxidase Biosensor with Statistical Machine Learning Methods. In M. Stoytcheva & J.F. Osma (Eds.). *Biosensors: Recent Advances and Mathematical Challenges*. Barcelona: España, OmniaScience. pp. 163-183.

1. Introduction

Biosensors are analytic compact devices that embody a biological piece of detection called a bio-receptor, usually formed by enzymes, microorganisms, immunoreceptors, cell receptors or chemoreceptors in current technology. They are coupled to a physical-chemical transducer that translates the biological signal to a measurable electrical signal, that is proportional to the concentration of the target compound or group of compounds to be assessed. Enzymes are mainly favored in biosensor construction because they have the capability to recognize a specific molecule (Thévenot, Toth, Durst & Wilson, 2001). One of the most attractive advantages of this sensing technology is its capacity to provide electrochemical readings in a fast, continuous and highly sensitive way. Moreover, they are susceptible to be miniaturized and its electrical response potential (or electric current) can be easily processed by cheap and compact instrumentation devices (Morrison, Dokmeci, Demirci & Khademhosseini, 2008).

The potential use of biosensors has been extended to several fields of science and engineering. Contaminant detection of the water resources (Saharudin & Asim, 2006); pathogen agent detection (Pohanka, M., Skládal & Kroca, M. (2007); drug detection in the food industry (Elliott, 2006) are only a few examples. In particular, the impact and benefits in the medical field is beyond doubt. The monitoring of lactate, urea, cholesterol or glucose are some of the body-essential features related to this technology (Department of Trade and industry, 2012; Malhotra & Chaubey, 2003). Disease monitoring and diagnosis require well-trained and qualified personnel for data acquisition and testing. It is worth mentioning that these medical tasks are highly delicate, in both sensitivity –the possibility to measure a false positive– and specificity –a false negative detection (Malhotra & Chaubey, 2003). In an ordinary scenario, a patient requires a physical examination and laboratory tests needing a few days to obtain medical results. Such a delay can sometimes derive in complications due to the lack of the proper medical treatment. The fast response, low cost and design simplicity of the biosensor approach makes it a very promising technological device in public health applications.

In the following sections, the modeling of Second-generation Electrochemical Glucose Oxidase Amperometric Biosensors (GOABs from now on) will be discussed. The importance of such devices will be contextualized in one of the most critical and prevalent diseases nowadays, the Diabetes Mellitus (DM). Machine Learning (ML) is a field within computer science and a very active area, playing an important role in science, finance and industry. It consists of a wide spectrum of methods, techniques and algorithms that aim at learning from data to find useful information or predictive models of a phenomenon. Classical and statistical ML techniques for regression are used for the modeling of the response of a GOAB. The rest of this chapter is organized as follows: In section 2, a few general ideas about DM are given, to introduce the reader into the importance and convenience to deal with the DM in its collateral consequences with this arising technology. Some general concepts about GOABs and GOAB modeling are given in section 3; section 4 describes the specific GOAB dataset and the statistical machine learning techniques used in this study. The experimental results are presented and discussed in section 6. The chapter ends with the conclusions and final thoughts.

2. Diabetes Mellitus

The DM is a serious condition where patients present high levels of blood glucose. This is due to two possible causes: a deficient insulin production or a malfunction in a particular type of cells called islets. These cells do not respond properly to insulin in the blood-glucose regulation process. In the food consumption process by humans, the body converts the inputs into glucose. A critical organ in the human body, the pancreas, produces insulin to convert this glucose into energy. When a patient is diagnosed with the DM disease, the whole process presents an erratic dynamics (Diabetes Research Institute, 2012).

There exist two types of DM: Type I includes patients that are insulin-dependent, where the islet cells are not recognized as part of the body by the Immune System and are consequently destroyed; as a result, insulin is not produced anymore; Type II embraces those diagnosed patients that produce part of their insulin needs, but not enough to maintain acceptable blood-glucose levels.

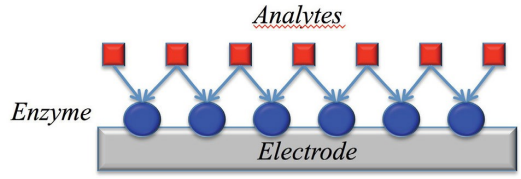

Figure 1. General scheme of an enzimatic biosensor

World Health Organization (WHO) statistics report that around 347 million of people worldwide have diabetes. Mortality estimates hits around 3.4 million deaths from consequences of high fasting blood sugar. Low- and middle-income countries have the worst mortality scenario, where more than 80% of diabetes patients die today. The WHO foresees the DM to be the 7th leading cause of death by 2030 (World Health Organization, 2013). Thus the continuous glucose monitoring by biosensors can be very helpful to diagnosed patients to prevent acute or chronic complications; however, accuracy and stability issues are still under development, preventing a more widespread use in the market (Keneth, 2007).

3. Modeling Glucose Oxidase Biosensors

3.1. Glucose Oxidase Amperometric Biosensors

Nowadays there exist diverse techniques to measure glucose levels in the blood. The most common are spectrophotometric using small devices called *glucometers*. Some of these take advantage of the oxidation of glucose to gluconolactone catalyzed by glucose oxidase. Others use a similar reaction but with another enzyme, the glucose dehydrogenase. This latter brings up the advantage of more sensitivity but is less stable in presence of other substances. Some alternatives in glucose monitoring include control of the ketones in the urine, in case that glucose detection in blood becomes difficult (American Diabetes Association, 2013).

Currently, there exists an alternative monitoring technology based on electrochemical enzymatic sensors. They consist of the elements, biochemical and physical, assembled in direct contact, or close enough to establish a relationship with the analyte, to produce a measurable signal, as indicated in Figure 1. An enzymatic amperometric sensor works due to oxygen consumption, hydrogen peroxide production, or b-nicotinamide adenine obtaining during the process of the catalytic conversion of the substrate (Equations 1 and 2) (Prodromidis & Karayannis, 2002). The occurring electrochemical reactions are commented below.

$$\text{Substrate} + O_2 \xrightarrow{\text{Enzyme}} \text{Product} + H_2O_2 \tag{1}$$

$$\text{Substrate} + NAD^+ \xrightarrow{\text{Enzyme}} \text{Product} + NADH + H^+ \tag{2}$$

The enzyme useful for measuring and assessing blood-glucose levels in biological fluids is known as *glucose oxidase*. The enzyme Glucose Oxidase belongs to the oxidoreductase class that catalyzes the β-D-glucose oxidation to D-glucone-1, 5-lactone and hydrogen peroxide (Equation 3). It is produced by the *Penicillium Notatum* and other fungi in the presence of glucose and oxygen. It is used to measure the glucose concentration in blood and urine samples. The general reaction equations can be described as:

$$\begin{gathered} \beta - D - Glucose + E_{ox} \rightleftharpoons E_{red} + Gluconic\ Acid \\ E_{red} \rightleftharpoons E_{ox} \end{gathered} \tag{3}$$

where E_{ox} and E_{red} are the oxidized and the reduced form of the glucose oxidase enzyme. Given this, there exist three generations of GOAB:

1st Generation. With O_2 added to the equation:

$$\begin{gathered} E_{red} + O_2 \rightleftharpoons E_{ox} + H_2O_2 \\ H_2O_2 \rightleftharpoons 2\mathrm{H}^+ + O_2 + 2\,e^- \left(on\ the\ electrode\right) \end{gathered} \tag{4}$$

2nd Generation. A mediator is present:

$$\begin{aligned} E_{red} + M_{ox} &\rightleftharpoons E_{ox} + M_{red} \\ M_{red} &\rightleftharpoons M_{ox} + ne - (on\,the\,electrode) \end{aligned} \tag{5}$$

3rd Generation. Without oxygen or other mediator:

$$E_{red} \rightleftharpoons E_{ox} \tag{6}$$

It is known that 1st Generation GOABs have a number of disadvantages, mainly oxygen concentration fluctuations; H_2O_2 inhibits the glucoseoxidase, among others. A solution that circumvents these problems was found in the 2nd and 3rd Generation GOABs (Stoytcheva, Nankov & Sharkova, 1995). This study focuses precisely in the 2nd generation GOABs.

3.2. GOAB Mathematical modeling

Mathematical modeling represents a powerful tool in the electrochemical biosensors design cycle, allowing for considerable reductions in costs and time (Wang, 2011). Despite the high benefits that GOAB technology brings to real life applications, it still bears some delicate design issues that make a better analysis and understanding still necessary (Petrauskas & Baronas, 2009). The GOAB response and its performance is affected by a number of factors, including electrode construction features (material, area and geometry), electrical factors (potential, current, charge, impedance), electrolytic factors (pH, solvents), and reaction variables (thermodynamic and kinetic parameters) (Borgmann, Schulte, Neugebauer & Schuhmann, 2011).

Given that Amperometric Electrochemical Biosensors (AEB) are used in combinatory synthesis procedures, the chemical reactions end-product might be scarce and limited, being measured from micrograms to miligrams scale; moreover, substrates inside enzymes can not be measured with analytical devices. Since the late 70s, several mathematical models have been used and proposed to this end, as an important tool looking for higher precision and simplicity (Malhotra & Chaubey, 2003). First AEB modeling efforts dealt with quantitative descriptions of biosensors kinetic behavior of simple idealized enzymes (Blaedel, Kissel & Boguslaski, 1972). Steady-state flux and distribution equations were used to show how enzymes fixed in gels may be used for immobilized enzyme kinetics analysis. The urease electrode potentiometric response were used in the experiment. Steady-state modeling of the current response by means of digital devices was one of the first digital simulation in glucose oxidase biosensor analysis (Mell & Maloy, 1975). Current modeling approaches range from analytical solutions of partial differential equations applied to simple biocatalytic processes to complex computer simulations of catalytic conversions and multiple transducer geometry (Baronas, 2010). Recent advances contemplate the use of *machine learning* (ML) algorithms, such as Artificial Neural Networks or Support Vector Machines. The use of these tools in AEB modeling is an emerging topic in the specialized scientific literature (Rangelova, Tsankova & Dimcheva, 2010; Alonso, Istamboulie, Ramírez-García Noguer, Marty & Muñoz, 2010). In this sense, model construction by ML techniques becomes a reasonable strategy, given that its black-box point of view liberates the modeler to fully and clearly express the mathematical laws underlying the physical phenomenon – in our case, the amperometric response of a biosensor.

4. Materials and Methods

4.1. Available data

Second Generation biosensors incorporate a mediator, in this case a p-benzoquinone mediated amperometric graphite sensor with covalently linked glucoseoxidase. This mediator is responsible for the electronic transfer between the enzyme and the electrode surface. Additionally, the following reagents were used: glucose oxidase (E.C. 1.1.3.4. from *Aspergillus*, 1000 U/mg), N-cyclohexyl-N'-[2-(methylmorpholino)ethyl]carbodiimide-4-toluenesulphonate (Merk) and glucose.

Amperometric data acquisition was achieved using a Radelkis OH-105 polarograph. The enzymatic working electrode used was a rotating disk electrode with a diameter of 6 mm, prepared from a spectrally pure graphite with glucoseoxidase immobilized on its surface. Saturated calomel electrode was used as reference electrode. The auxiliary electrode was a glassycarbon electrode.

The amperometric response was analyzed under different conditions of the Glucose *(Glucose),* pH *(PH),* temperature *(T)* and concentration of the mediator, the p-benzoquinone *(Benzoquinone)*.

Values for these input parameters (used as predictors) were Glucose (in mM) $\in$ {4, 8, 12, 16, 20}; p-benzoquinone (in mM) $\in$ {1, 0.8, 0.4, 0.2}; pH (dimensionless) $\in$ {4, 5, 6, 7} and Temperature (in Celsius scale) $\in$ {20, 37, 47, 57}. The response or faradaic current (I) was measured in mA.

The resulting data file consists of 320 rows (observations) and 5 columns (4 predictive variables and a continuous target variable, which corresponds to the biosensor response). As stated above, the predictive variables are: *Glucose, Benzoquinone, T* and *PH*. These predictive variables are standardized to zero mean, unit standard deviation. Finally the data file is shuffled to avoid predefined ordering biases.

4.2. Regression Methods

Suppose we are given training data of the form $\{(x_n, t_n)\}_{n=1}^{N} \subset X \times R$, where $X = R^d$ denotes the space of input vectors.

4.2.1. Classical regression analysis

Generalized linear models (GLMs) are a commonly used form of model for regression modelling. A GLM takes the form:

$$f_{GLM}(x) = \sum_{m=0}^{M} \beta\, m \Phi\, m(x) \tag{7}$$

where φ is a set of M basis functions (which can be arbitrary real-valued functions) with $\varphi_0(\cdot) = 1$ and β is a vector of coefficients.

The number and form of the basis functions has to be decided beforehand. In this work we consider two useful GLM settings: linear and restricted polynomial regression:

Linear regression We consider the choices $\varphi_m(x) = x_m$ and $M = d$

Polynomial regression We consider the choices $\varphi_m(x)$ to be polynomials in x of limited degree (three, at most); in this case, every monomial will have its own β coefficient and M will be the total number of monomials.

The basis functions define a projection of the input data into a higher-dimensional space where the data is more likely to be linear. Since the obtained expressions are linear in the coefficients, these coefficients are optimized using standard least squares methods.

In order to assess how well these models fit the data, the leave-one-out cross-validation (LOOCV) method can be used. Every observation is excluded from the training set, the model is fit using the remaining points, and then is made to predict the left out observation. The process is repeated over the entire training set, and the LOOCV error is computed by taking the average over these predictions. This method provides an almost unbiased estimate of the generalization error and has the added advantage of being fast to compute for linear models. Typically one has to maximize:

$$R^2 cv = 1 - \frac{GE_{cv}(\hat{y})}{\frac{1}{N}\sum_{i=1}^{N}(\hat{y} - \overline{y})^2}$$

where $GE_{cv}(\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e_i}{1-h_{ii}}\right)^2$ is the LOOCV error, and e_i, h_{ii} are the residual and the leverage of observation x_i. A related measure of performance (to be minimized) is given by the normalized root mean-square error (NRMSE) (Bishop, 1996):

$$\sqrt{\frac{\sum_{i=1}^{N}\left(\frac{e_i}{1-h_{ii}}\right)^2}{\sum_{i=1}^{N}(\hat{y} - \overline{y})^2}}$$

which can be interpreted as the fraction of output standard deviation that is explained by the model. Note that we can obtain the corresponding cross-validation NRMSE as $\sqrt{1 - R^2_{cv}}$

4.2.2. Support Vector Machines

Support Vector Machines for regression (SVMR) have become a popular tool for the modelling of non-linear regression tasks (Smola, & Schölkopf, 2004). The SVMR is one of the several kernel-based techniques available in machine learning. These are methods based on implicitly mapping the data from the original input space to a feature space of higher dimensionality and then solving a linear problem in that space. The used E-insensitive loss function $|z|_E = \max\{0, |z| - E\}$

penalizes errors that are greater than a threshold *E*, usually leading to sparser representations, entailing algorithmic and representational advantages, (Vapnik, 1998).

Let *H* be a real RKHS with kernel κ. The input data is transformed with a feature map $\Phi: X \rightarrow H$, to obtain the new data set $\{(\Phi(X_n), t_n)\}^N_{n=1}$. In a SVMR, the aim is to find a function f_{SVMR}: $(\Phi(x), w)_H + a_0$, for some $w \in H$ and $a_0 \in R$, which is as flat as possible and deviates a maximum of *E* from the given target values t_n, for all $n = 1,..., N$.

The usual formulation of the optimization problem is as the dual of the convex quadratic program:

$$\begin{aligned} &\min_{w \in H,\, a_o \in R} \quad \frac{1}{2}\|w\|\frac{2}{h} + \frac{c}{N}\sum_{n=1}^{N}\left(\xi_n + \hat{\xi}\right) \\ &Subject\ to \begin{cases} \langle \Phi(x_n), w\rangle H + a_o - t_n \le \varepsilon + \xi_n \\ t_n - \langle \Phi(x_n), w\rangle H - a_0 \le \varepsilon + \xi_n \\ \xi_n, \hat{\xi}_n \ge 0 \end{cases} \end{aligned} \tag{8}$$

for $n = 1, ..., N$. To solve (8), one considers the dual problem derived by the Lagrangian:

$$\begin{aligned} \max_{a,\hat{a}} \Bigg\{ & -\frac{1}{2}\sum_{n,m=1}^{N}(\hat{a}_n - a_n)(\hat{a}_m - a_m)k(x_n, x_m) \\ & -\epsilon\sum_{n=1}^{N}\left(\hat{a}_n - a_n\right) + \sum_{n=1}^{N} t_n\left(\hat{a}_n - a_n\right) \end{aligned} \tag{9}$$

Subject to $\sum_{n=1}^{N}\left(\hat{a}_n - a_n\right) = 0$ and a_n , $\hat{a}_n \in [0, C/N]$

Exploiting the saddle point conditions, it can be proved that $W = \sum_{n=1}^{N}\left(\hat{a}_n - a_n\right)\Phi(x_n)$; given that $k(x, x') = \langle \Phi(x), \Phi(x')\rangle H$, the solution becomes'

$$fSVMR(x) = \sum_{n=1}^{N}\left(\hat{a}_n - a_n\right)k\left(x_n, x\right) + a_0, x \in X \tag{10}$$

4.2.3. Relevance Vector Machines

The Relevance Vector Machine (RVM) is a sparse Bayesian method for training GLMs which has the same functional form as the SVMR (Tipping, 2001). It is a kernel-based technique that typically leads to sparser models than the SVMR, and may also perform better in many cases. The RVM introduces a prior over the weights governed by a set of hyperparameters, one

associated with each weight, whose most probable values are iteratively estimated from the data. The RVM has a reduced sensitivity to hyperparameter settings than the SVM.

In the RVM, a zero mean Gaussian prior with independent variances (acting as hyperparameters) $\alpha_j \equiv 1/\sigma^2_{w_j j}$ is defined over each weight:

$$p(w|\alpha) = \prod_{j=1}^{M} \sqrt{\frac{\alpha_j}{2\pi}} \exp\left(\frac{-1}{2}\alpha_j w_j^2\right)$$

As in standard regression, assuming an independent zero-mean Gaussian noise model of variance σ^2 for the targets, the likelihood of a target vector *t* is:

$$p(t|w, \sigma^2) = \left(2\pi\sigma^2\right)^{\frac{-N}{2}} \exp\left(\frac{-1}{2\sigma^2}\|t - \Phi w\|^2\right)$$

where Φ is the Gram or kernel matrix of the inputs. In these conditions, the posterior over the weights $p(w|t, \alpha, \sigma^2)$ is also a Gaussian $N(\mu, \Sigma)$ that can be obtained using the Bayes rule:

$$p(w|t, \alpha, \sigma^2) = \frac{p(t|w, \sigma^2)\, p(w|\alpha)}{\int p(t|w, \sigma^2)\, p(w|\alpha)\, d\, w}$$

where $\mu = \sigma^{-2}\Sigma\Phi^T t$ and $\Sigma = (\sigma^{-2}\Phi^T \Phi + \Lambda)^{-1}$, being $\Lambda = \mathrm{diag}(\alpha_1,\ldots,\alpha_M)$. Now the likelihood distribution over the training targets can be calculated by integrating out the weights to obtain the marginal likelihood for the hyperparameters:

$$p(t|\alpha, \sigma^2) = \int p(t|\mathrm{w}, \sigma^2) p(\mathrm{w}|\alpha) d\,\mathrm{w}$$

This marginal distribution is again Gaussian $N(\mathbf{0}, A)$, where $A = \sigma^2 I + \Phi\Lambda^{-1}\Phi^T$. For computational efficiency, the logarithm of the evidence is maximized:

$$L(\alpha) = \log p\,(t|\alpha, \sigma^2) = -\frac{1}{2}\,(M \log(2\pi) + \log|A| + t^T A^{-1}\, t)$$

The estimated value of the model weights is given by their maximum a posteriori (MAP) estimate, which is the mean of the posterior distribution $p(w|t, \alpha, \sigma^2)$. This MAP estimate depends on the hyperparameters $\boldsymbol{\alpha}$ and σ^2. These variables are obtained by maximizing the marginal likelihood $L(\boldsymbol{\alpha})$. Sparsity is achieved because in practice many of the hyperparameters $\boldsymbol{\alpha}_j$ tend to infinity, yielding a posterior distribution of the corresponding weight w_j that is sharply peaked around zero. These weights can then be deleted from the model, as well as their associated basis functions.

5. Experimental setup

The experimental part explores the modeling of the biosensor output from two different points of view. In the first set of experiments, we treat the inputs as they are, namely continuous regressors. In the second set of experiments, we consider the possibility of treating the regressors as categorical instead of continuous. This is supported by the fact that all four predictors take on a limited number of values (to be precise, four for the *Benzoquinone*, *T* and *PH*, and five for the *Glucose*). Categorical predictor variables cannot be entered directly into a regression model (and be meaningfully interpreted). Typically, a categorical variable with *c* modalities will be transformed into *c* − 1 binary variables (each with two modalities).

For example, if a categorical variable had five modalities, then four binary variables are created that would contain the same information as the single categorical variable. In particular, all the distances between the modalities are equal, regardless of the specific coding chosen. These variables have the advantage of simplicity of interpretation and are sometimes preferred to correlated predictor variables. They are also useful to assess non-linearities between the regressors and the output.* In the present situation, the new values for the *Glucose* are "very low", "low", "medium", "high" and "very high" (64 observations each), whereas new values for the other three predictors are "low", "low-medium", "medium-high" and "high" (80 observations each).

5.1. Optimization of the SVMR and the RVM

First, the available data were randomly split into two sets: 220 observations (68.75%) for training and the remaining 100 observations (31.25%) for testing.

The standard regression methods need no additional parameter specification. In order to obtain the solution for a kernel method, one has to choose the kernel function, and determine appropriate values for the associated hyperparameters. In the *SVMR*, the *E* parameter controls the width of the *E*-insensitive zone. The cost parameter *C* determines the trade-off between model complexity (flatness of the solution) and the degree to which deviations larger than E are tolerated. The value of *E* can affect the number of support vectors used to construct the regression function. The bigger the *E*, the fewer support vectors are selected and smoother regression functions. An effective approach is to estimate the generalization error –usually through cross-validation– and then optimize for these parameters so that this estimation is minimized. For the *SVMR*, *C* is varied logarithmically between $10^{-1.5}$ and $10^{1.5}$ (20 equally-spaced values for the exponent), and E is varied logarithmically between 10^{-3} and 10^{0} (10 equally-spaced values for the exponent).

Among the kernels that are available in the literature, we select the polynomial kernel: the function

$$k_{Poly}(x, y) = (\langle x, y \rangle + R)^m$$

is called the polynomial kernel, with $R \geq 0$ and integral $m \geq 1$.

* In rigour these are categorical *ordinal* predictors, although we will treat them as categorical *nominal* ones, given the absence of specific methods for ordinal predictors (Agresti, 2002).

Another kernel that can be used is the Gaussian Radial Basis Function (RBF) kernel, known to be a safe default choice for kernel methods working on real vectors (Schölkopf & Smola, 2001):

$$k_{RBF}(x, y) = \exp\left(-\gamma \|x - y\|^2\right) \tag{11}$$

where $\gamma > 0$ is the smoothing parameter. The γ parameter in the RBF kernel is estimated using the *sigest* method, based upon the 10% and 90% quantiles of the sample distriubion of $\|x - x'\|^2$ (Caputo, Sim, Furesjo & Smola, 2002). We try different polynomial kernels, given by degrees *m* from 1 to 5 and *R* = 1.

The parameters were optimized through 30 times 10-fold cross-validation (30 x 10 cv) usingthe training set; a model is then refit in the training set using the best parameter configuration; this model is now made to predict the held out test set. The error reported in all cases is the normalized root mean-square error (NRMSE) in the test set. A model showing a NRMSE of 1 corresponds to the best *constant* regression; good models should then have a NRMSE considerably smaller than 1 (and reasonably close to 0).

For the RVM, the same considerations apply for the kernels and their parameters (the RVM needs no specification of *C* or the *E* parameter). Theoretically, the whole training set could be used to fit the RVM (without cross-validation). However, resampling is still needed to choose the best kernel configuration; therefore, the same 30 x 10 cv procedure is used to evaluate performance in the training set.

6. Results and Discussion

6.1. Basic statistical analysis

After the pre-process described in section 4.1, we get a data set with the summary described in Table 1. The 'Target' variable refers to the biosensor output.

Glucose	***Benzoquinone***	***T***	***PH***	**Target**
Min.: -1.412	Min.: -1.2629	Min.: -1.4797	Min.: -1.3395	Min.: 0.2848
1st Qu.: -0.706	1st Qu.: -0.7893	1st Qu.: -0.5481	1st Qu.: -0.6698	1st Qu.: 1.9575
Median: 0.000	Median: 0.0000	Median: 0.1279	Median: 0.0000	Median: 4.2178
Mean: 0.000	Mean: 0.0000	Mean: 0.0000	Mean: 0.0000	Mean: 12.3994
3rd Qu.: 0.706	3rd Qu.: 0.7893	3rd Qu.: 0.6759	3rd Qu.: 0.6698	3rd Qu.: 14.9122
Max.: 1.412	Max.: 1.2629	Max.: 1.2240	Max.: 1.3395	Max.: 75.5506

Table 1. Descriptive statistics after pre-processing

We can see that all the predictive variables are perfectly symmetrical (the mean and median are equal), with the exception of *T*, whose distribution is skewed to the left (negative skew), since its mean is smaller than its median (see the set of boxplots in Figure 2).

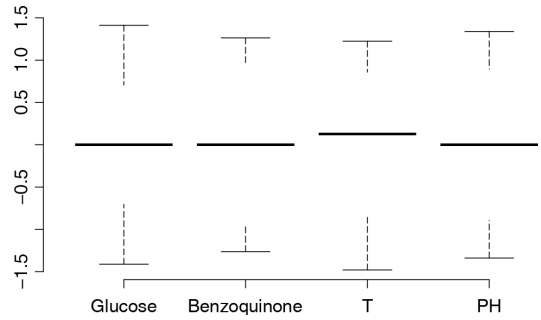

Figure 2. Box plots of predictive variables

A convenient first step is to take the natural log of the target variable. The effect of this change is clearly visible (Figure 3).

Now we compute the Pearson's product-moment correlations (Table 2).

First variable	Second variable	Correlation
PH	Target	-0.332
Glucose	Target	0.267
T	Target	0.167
Benzoquinone	Target	0.063

Table 2. Pearson's variable correlations; only the four largest are shown

These very small correlations suggest that only *PH* bears some linear relation with the target variable,* in addition, there is no linear relation concerning the predictive variables with one another. In order to refine this result, we compute Spearman's ρ correlation coefficient. Although this coefficient does not detect general quadratic non-linearities, it is good for detecting possible outliers and monotonic non-linearities (Table 3).

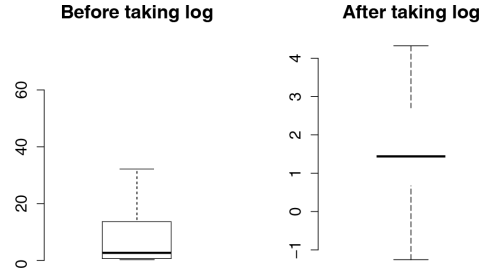

Figure 3.Box plots of the target (output) variable

* Although true correlations are not equal to 0 (*PH*-Target: t = -6.2685, df = 318, p-value = 1.187e-09, *Glucose*-Target: t = 4.9337, df = 318, p-value = 1.303e-06, *T*-Target: t = 3.0135, df = 318, p-value = 0.00279).

First variable	Second variable	Correlation
PH	Target	-0.569
Glucose	Target	0.405
T	Target	0.195
Benzoquinone	Target	0.082

Table 3. Spearman's variable correlations; only the four largest are shown

The correlations are larger for some variables –notably *PH* and the *Glucose*, suggesting a relation of non-linear nature with the biosensor output.

6.2. Regression with continuous predictors

6.2.1. Standard Linear Regression

	Estimate	Std. Error	*t* value	Pr(>|*t*|)
(Intercept)	1.6182	0.0653	24.7764	6.3e-65
Glucose	0.5072	0.0666	7.6171	8.1e-13
Benzoquinone	0.1151	0.0658	1.7480	8.2e-02
T	0.3436	0.0650	5.2883	3.0e-07
PH	-0.6896	0.0647	-10.6570	1.5e-21

Table 4.Coefficients and significance of the linear regression analysis with continuous predictors

The results of linear regression are shown in Table 4. These results show that all the coefficients are significantly different from zero, as given by the negligible p-values, with the exception of that for *Benzoquinone*, which is barely significant (p-value = 0.082). Since the predictive variables are standardized, the coefficients can be related to the relevance of the corresponding variables. The most important predictor is *PH*, followed by *Glucose* and *T*; the *Benzoquinone* is by far the less important predictor. The importance of variables can be further assessed by an Analysis of Variance (ANOVA) (Table 5).

	Df	Sum Sq	Mean Sq	*F* Value	*Pr*(>*F*)
Glucose	1	54.6828	54.6828	58.5062	6.7e-13
Benzoquinone	1	1.0242	1.0242	1.0958	3.0e-01
T	1	24.7466	24.7466	26.4768	6.0e-07
PH	1	106.1500	106.1500	113.5718	1.5e-21
Residuals	215	200.9500	0.9347		

Table 5. ANOVA in the training set with continuous predictors

According to the ANOVA, PH is by large the most important predictor, followed by *Glucose* and *T;* the *Benzoquinone* is by far the less important predictor. However, an AIC analysis does not suggest to remove any of the regressors, and therefore we keep all of them. We may now wish to see how well the model fits and predicts the training data. We get a relatively low R^2_{CV} = 0.4724 (corresponding to a NRMSE of 0.7264), which indicates a rather poor model.

The previous results using linear regression again suggest that the possible relation between the output of the biosensor and the predictive variables is a non-linear one. Therefore polynomial regression as described in section 4.2 is considered. After preliminary modeling trials in the training set, one ends up with a set of regressors formed by a third-degree polynomial on *PH* and a second-degree polynomial on both *T* and the *Glucose*. The addition of other terms does not increase the model quality and adds further complexity. The results of this polynomial regression are shown in Table 6.

	Estimate	**Std. Error**	***t* value**	**Pr(>|*t*|)**
(Intercept)	1.6468	0.0103	159.1228	1.0e-221
poly(*PH*, 3)-1	-12.9545	0.1831	-70.7585	5.8e-149
poly(*PH*, 3)-2	-7.6297	0.1857	-41.0931	1.2e-102
poly(*PH*, 3)-3	13.3168	0.1871	71.1572	1.9e-149
poly(*T*, 2)-1	5.7589	0.1839	31.3108	3.0e-81
poly(*T*, 2)-2	-5.1793	0.1836	-28.2150	1.5e-73
Benzoquinone	0.1012	0.0105	9.6714	1.5e-18
poly(*Glucose*, 2)-1	9.3897	0.1882	49.8934	9.0e-119
poly(*Glucose*, 2)-2	-3.5097	0.1842	-19.0517	9.7e-48

Table 6. Coefficients of the non-linear (polynomial) regression analysis with continuous predictors. The notation poly(V, r)-s *stands for the* s*-degree monomial of an* r*-degree polynomial on regressor* V

The results show that all the coefficients are significantly different from zero, as given by the negligible *p*-values, with no exception. Indeed, the AIC analysis does not suggest to remove any of the regressors. Variable importance can again be further assessed by the ANOVA (Table 7).

We may now wish to see how well the model fits and predicts the training data. We get an excellent R^2_{CV} = 0.9872 (corresponding to a NRMSE of only 0.1132), which indicates a very promising model, remarkably better than the linear one.

	Df	**Sum Sq**	**Mean Sq**	***F* Value**	***Pr*(>*F*)**
poly(*PH*, 3)	3	273.3378	91.1126	3899.0107	1.9e-184
poly(*T*, 2)	2	40.3336	20.1668	863.0042	2.6e-102
Benzoquinone	1	2.3375	2.3375	100.0308	1.6e-19
poly(*Glucose*, 2)	2	66.6140	33.3070	1425.3161	2.8e-123
Residuals	211	4.9307	0.0234		

Table 7. ANOVA in the training set with continuous predictors. The notation poly (V, r) *stands for an* r*-degree polynomial on regressor* V

6.2.2. Regression with the SVMR

We now turn to results obtained with the kernel-based methods (Table 8). Using a linear kernel, the best set of parameters for the SVMR (optimized through 30 x 10 cv) is E = 0.464 and C = 0.695. The cross-validation NRMSE of this choice is 0.7387. Actually all the results are in the range 0.73-0.80, no matter the value of C and E. Something similar happens for the quadratic

kernel, although in this case the range is 0.60-0.80. These readings suggest a general poor model, in consonance with the linear regression model previously reported. Using higher degree kernels, the results are much better, in accordance to those found for the polynomial regression model. Indeed, the result with a cubic polynomial is similar to the polynomial regression (which used cubic polynomials in the *PH*). The best result corresponds to the quartic polynomial, markedly better than the polynomial regression.

Kernel	degree	*C*	*E*	cv NRMSE
Linear	1	0.695	0.464	0.7387
Quadratic	2	0.036	0.464	0.6017
Cubic	3	1.623	0.046	0.0867
Quartic	4	0.215	0.046	0.0491
Quintic	5	0.183	0.046	0.0942
RBF	-	0.069	0.215	0.4441

Table 8. Results for the SVMR with polynomial kernels of different degrees and the RBF kernel with continuous predictors; cv NRMSE is the cross-validation NRMSE

The results seem to confirm the need for at least third-order information extracted from the original regressors; however, higher-order terms start to overfit the data. The relatively poor results obtained by the RBF kernel can also be explained in this light. A nice aspect of the results is that, for all kernels, for same values of *E*, predictive performance is tied for many values of the *C* parameter; in these cases, we selected the smallest value of *C*, in accordance with Statistical Learning Theory (Vapnik, 1998).

6.2.3. Regression with the RVM

The results for the RVM with the different kernels are displayed in Table 9. They are in consonance with those for the SVMR previously reported, although they are consistently better. Remarkably, there is a coincidence with the SVMR in that the best result corresponds to the quartic polynomial.

Kernel	degree	cv NRMSE
Linear	1	0.7356
Quadratic	2	0.5973
Cubic	3	0.0845
Quartic	4	0.0301
Quintic	5	0.0392
RBF	-	0.2148

Table 9. Results for the RVM with polynomial kernels of different degrees and the RBF kernel with continuous predictors; cv NRMSE is the cross-validation NRMSE

6.3. Regression with categorical predictors

6.3.1. Standard Linear Regression

The results of linear regression are shown in Table 10.

	Estimate	**Std. Error**	***t* value**	**Pr(>|*t*|)**
(Intercept)	1.1474	0.0364	31.5440	8.3e-81
Glucose (low)	-0.6432	0.0295	-21.8069	2.0e-55
Glucose (medium)	-0.1810	0.0286	-6.3219	1.6e-09
Glucose (very high)	0.1067	0.0298	3.5778	4.3e-04
Glucose (very low)	-1.4215	0.0298	-47.7724	2.1e-113
Benzoquinone (low)	-0.2697	0.0268	-10.0788	1.1e-19
Benzoquinone (low-medium)	-0.0804	0.0260	-3.0866	2.3e-03
Benzoquinone (medium-high)	0.0220	0.0267	0.8220	4.1e-01
T (low)	-0.7716	0.0260	-29.6715	2.6e-76
T (low-medium)	0.1902	0.0257	7.4008	3.4e-12
T (medium-high)	0.3650	0.0271	13.4747	4.3e-30
PH (low)	1.2709	0.0258	49.2501	6.4e-116
PH (low-medium)	2.8093	0.0268	104.6867	1.4e-180
*PH (*medium-high)	0.1746	0.0265	6.5778	3.9e-10

Table 10. Coefficients of the linear regression analysis with categorical predictors in the training set

These results show that all the coefficients are significantly different from zero, as given by the negligible p-values, with the exception of that for *Benzoquinone* (medium-high), which is not significant. The importance of variables can be further assessed by the ANOVA on the previous regression (Table 11).

According to the ANOVA, PH is by large the most important predictor, followed by Glucose and T; the *Benzoquinone* is by far the less important predictor. Again, AIC analysis does not suggest to remove any of the regressors, and therefore we keep all of them. We may now wish to see how well the model fits and predicts the training data. We get an excellent R^2_{CV} = 0.9887 (corresponding to a NRMSE of 0.1062). This result vastly improves that of linear regression with continuous predictors (NRMSE of 0.7264). Sadly, there is no possibility of developing a standard polynomial regression model using categorical predictors.

	Df	**Sum Sq**	**Mean Sq**	***F* value**	***Pr*(>*F*)**
Glucose	4	66.2744	16.5686	891.6448	8.5e-129
Benzoquinone	3	4.1249	1.3750	73.9946	1.6e-32
T	3	52.4748	17.4916	941.3168	6.1e-120
PH	3	260.8515	86.9505	4679.2708	3.6e-189
Residuals	206	3.8279	0.0186		

Table 11. ANOVA in the training set with categorical predictors

6.3.2. Regression with the SVMR

We now turn to results obtained with the kernel-based methods (Table 12). The best set of parameters is $E = 10^{-3}$ and C around 1, using the quadratic kernel. Actually all the results are markedly better than those obtained with the SVMR in the same conditions but using continuous predictors. This result is in consonance with that previously reported for the standard regression. Again, using non-linear (higher degree) kernels, the results are better than with linear ones, although in this case the technique starts to overfit at cubic polynomials. The RBF kernel performs much better too, and is comparable to the cubic polynomial.

Kernel	degree	*C*	*E*	cv NRMSE
Linear	1	3.793	0.1	0.1077
Quadratic	2	1.062	0.001	0.0026
Cubic	3	0.455	0.001	0.0307
Quartic	4	0.195	0.0022	0.1065
Quintic	5	0.127	0.001	0.2000
RBF	-	5.796	0.001	0.0371

Table 12. Results for the SVMR with polynomial kernels of different degrees and the RBF kernel with categorical predictors; cv NRMSE is the cross-validation NRMSE

The results make perfect sense in the light of model complexity, the linear kernel being too simple, and polynomials beyond the cubic one being too complex. In addition, it was observed that E was the critical parameter; for all kernels, for similar values of E, predictive performance varies smoothly with the C parameter, and many times it is rather independent; again, in these cases, we selected the smallest value of C.

6.3.3. Regression with the RVM

The results for the RVM with the different kernels are displayed in Table 13. They are in consonance with those for the SVMR previously reported, although this time those for the SVMR seem slightly better. Again, there is a coincidence with the SVMR in that the best result corresponds to the quartic polynomial.

6.4. Discussion

In view of the results reported so far, two methods stand out from the rest: the two nonlinear kernel methods of moderate complexity. Specifically, both the SVMR and the RVM with quadratic kernel and using categorical predictors deliver very good 30 x 10 cv errors, around or below 10^{-3} of NRMSE. The decision among these two methods is not an easy one, given that the errors are similar (with a slight advantage of the SVMR). On the other hand, the RVM is expected to deliver a sparser model. In order to decide, we proceeded to refit both methods in the training set using the best parameter configuration. The SVMR ($C = 1.062$ and $E = 10^{-3}$) delivers a modeling NRMSE of 0.0023 with 142 support vectors (a 64.5% of the training data); the RVM delivers a modeling NRMSE of 0.0020 with 63 relevance vectors (a 28.6% of the training data). Now the decision is clear: given that the models now have the right complexity (because they have been optimized towards minimizing predictive error), the modeling error is a relevant quantity. The RVM then achieves a smaller NRMSE with less than half of the regressors than the SVMR does.

Kernel	degree	cv NRMSE
Linear	1	0.1066
Quadratic	2	0.0030
Cubic	3	0.0371
Quartic	4	0.1133
Quintic	5	0.2239
RBF	-	0.0454

Table 13. Results for the RVM with polynomial kernels of different degrees and the RBF kernel with categorical predictors; cv NRMSE is the cross-validation NRMSE

Therefore we make the choice of the RVM with a quadratic kernel and categorical predictors. This model is then made to predict the held out test set, yielding a predictive NRMSE of 0.0022. This is a very nice result for two reasons:

- The predictive NRMSE of 0.0022 is equivalent to a residual (non-explained) variance of only 0.22% of the total variance of the (test) predicted target values; therefore the model is indeed a very accurate one.
- The training (or modeling) NRMSE of 0.0020 and the predictive NRMSE of 0.0022 are in very good agreement, and an indication of a model of the right complexity.

The predictive results can also be displayed. In Figure 4 the predictions are plotted against the true values. It can be seen that the predictions are very good even when expressed in the original units (exponentiating the prediction). To be precise, the prediction error expressed in the original units amounts to an NRMSE of 0.0027. In order to obtain a final model that could be used in the future, we refit the RVM with a quadratic kernel (using categorical predictors) in the entire dataset. The obtained model has a modeling NRMSE of 0.00184 with 62 relevance vectors (a 19.4% of the training data). Expressed in the original units, this error corresponds to a NRMSE of 0.00176.

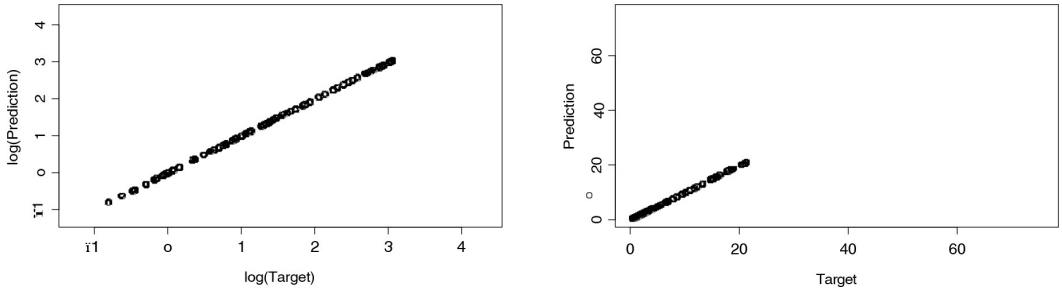

Figure 4. Final test set predictions. Left: in the log units; right: in the original units

At this point it is of great interest to make a comparison of all the results for the different regression methods and types of modeling.

Linear vs. non-linear methods. Given the poor (and consistent) results obtained by the three linear methods (linear regression, linear SVMR and linear RVM), it can be concluded that the true relation between the predictive variables and the biosensor output is a non-linear one.

Classical regression vs. kernel methods. Using continuous predictors, polynomial regression is able to give a fairly good result (NRMSE of 0.1132); kernel regression with the SVMR is able to improve this to half the error (NRMSE of 0.0491); kernel regression with the RVM delivers an NRMSE of 0.0301, both using a fourth-degree polynomial.

Continuous vs. categorical predictors. Although the non-linear methods make good use of the continuous variables, the limited amount of information they give (only 4 to 5 different values) makes the modeling a non-trivial undertaking; in contrast, standard regression and both the SVMR and the RVM deliver consistently better results using the categorized version of the dataset.

SVMR vs. the RVM Both methods can be seen as sparse GLM trainers, and indeed the RVM was initially presented as an alternative (and direct competitor) to the SVMR. The results follow similar paths for same kernels and it is not clear which one is performing better; the SVMR shows slightly lower errors but the RVM needs no parameter specification and delivers a sparser solution.

7. Conclusions

The continuous amperometric response of a GOAB has been successfully modelled by means of several classic and statistical learning methods. Specifically, kernel-based regression techniques have been used. The reported experimental results show a very low prediction error of the biosensor output, obtained using a relevance vector machine (RVM) with a quadratic kernel and categorical predictors. This constitutes a rather simple model of the biosensor output, because it is sparse and uses only four regressors with a limited number of values. We have also found that the pH is the most important predictor, followed by the Glucose and the temperature. An alternative technique could be grounded in the support vector machine for regression (SVMR). While SVMR can only be applied to the subset of generalized linear models (GLM) that can be defined by a valid kernel function, the RVM can train a GLM with any collection of basis functions.

Continuous glucose monitoring by means of a GOAB can constitute a remarkable ally to diabetic patients involved in serious collateral chronic complications. However, their design is still under development, in order to improve both accuracy and stability. In electrochemical biosensors design, mathematical modeling is a highly recurrent tool, given that it facilitates the computational simulation saving design and testing time and resources. The experimental proposal and conditions developed in this chapter could be applied for other scenarios in the wide spectrum of biosensing technology.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics, 2nd edition. Wiley-Interscience.

Alonso, G., Istamboulie, G., Ramírez-García A., Noguer, T., Marty, J., & Muñoz, J. (2010). Artificial neural network implementation in single low-cost chip for the detection of insecticides by modeling of screen-printed enzymatic sensors response. C*omputers and Electronics in Agriculture, 74(2)*, 223-229. http://dx.doi.org/10.1016/j.compag.2010.08.003

American Diabetes Association. (2013). Available at http://www.diabetes.org

Baronas, R. (2010). *Mathematical modeling of biosensors an introduction for chemists and mathematicians*. Springer. http://dx.doi.org/10.1007/978-90-481-3243-0

Bishop, C. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press, USA.

Blaedel, W.J., Kissel, T.R., & Boguslaski, R.C. (1972). Kinetic behavior of enzymes immobilized in artificial membranes. *Analytical Chemistry, 44(12)*, 2030-2037. PMID: 4657296. http://dx.doi.org/10.1021/ac60320a021

Borgmann, S., Schulte, A., Neugebauer, S., & Schuhmann, W. (2011). Amperometric biosensors. In R. Alkire, D. Kolb & J. Lipkowski (Eds.). *Bioelectrochemistry: Fundamentals, Applications and Recent Developments*. Wiley-VCH.

Caputo, B., Sim, K., Furesjo, F., & Smola, A. (2002). Appearance-based Object Recognition using SVMs: Which Kernel Should I Use. *Proc. of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision.*

Department of Trade and industry. (2012). *Biosensors for Industrial Applications, A review of Biosensor Technology*. United Kingdom Goverment.

Diabetes Research Institute. (2012). Available at http://www.diabetesresearch.org

Elliott, C. (2006). Biosensor detects toxic drugs in food. *Trac-trends in Analytical Chemistry, 25*.

Keneth, W. (2007). How to design a biosensor. *Journal of Diabetes Science and Technology, 1(2)*, 201-204. http://dx.doi.org/10.1177/193229680700100210

Malhotra, B., & Chaubey, A. (2003). Biosensors for clinical diagnostics industry. *Sensors and Actuators B: Chemical, 91(1-3)*, 17-127. http://dx.doi.org/10.1016/S0925-4005(03)00075-3

Mell, L.D., & Maloy, J.T. (1975). Model for the amperometric enzyme electrode obtained through digital simulation and applied to the immobilized glucose oxidase system. *Analytical Chemistry, 47(2)*, 299-307. http://dx.doi.org/10.1021/ac60352a006

Morrison, D., Dokmeci, M., Demirci, U., & Khademhosseini, A. (2008). Clinical applications of micro- and nanoscale biosensors. In K. Gonsalves, C. Halberstadt, C. Laurencin & L. Nair (Eds.). *Biomedical Nanostructures*. Wiley.

Petrauskas, K., & Baronas, R. (2009). Computational modelling of biosensors with an outer perforated membrane. *Nonlinear Analysis: Modelling and Control, 14(1)*, 85-102.

Pohanka, M., Skládal, P., & Kroca, M. (2007). Biosensors for biological warfare agent detection. *Defence Science Journal, 57.*

Prodromidis, M., & Karayannis, M. (2002). Enzyme based amperometric biosensors for food analysis. *Electroanalysis, 14(4)*, 241-261. http://dx.doi.org/10.1002/1521-4109(200202)14:4<241::AID-ELAN241>3.0.CO;2-P

Rangelova, V., Tsankova, D., & Dimcheva, N. (2010). Soft computing techniques in modelling the influence of ph and temperature on dopamine biosensor. In V. Somerset (Ed.). *Intelligent and Biosensors*. INTECH. http://dx.doi.org/10.5772/7029

Saharudin, H., & Asim, R. (2006). Optical biodetection of cadmium and lead ions in water. *Medical engineering & physics, 28(10)*, 978-981. http://dx.doi.org/10.1016/j.medengphy.2006.04.004

Schölkopf, B., & Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press.

Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14(3)*, 199-222. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88

Stoytcheva, M., Nankov, N., & Sharkova, V. (1995). Analytical characterisation and application of a p-benzoquinone mediated amperometric graphite sensor with covalently linked glucoseoxidase. *Analytica Chimica Acta, 315(12)*, 101-107. http://dx.doi.org/10.1016/0003-2670(95)00314-P

Thévenot, D., Toth, K., Durst, R., & Wilson, G. (2001). Electrochemical biosensors: recommended definitions and classification. *Biosensors and Bioelectronics, 16(1-2)*, 121-131. http://dx.doi.org/10.1016/S0956-5663(01)00115-4

Tipping, M. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research, 1*, 211-244.

Vapnik, V. (1998). *Statistical learning theory*. Wiley.

Wang, Q (2011). *Mathematical Methods for Biosensor models*. PhD thesis, Dublin Institute of Technology.

World Health Organization. (2013). Available at http://www.who.int