IHJPAS. 37 (1) 2024
Ibn Al-Haitham Journal for Pure and Applied Sciences
Journal homepage: jih.uobaghdad.edu.iq
PISSN: 1609-4042, EISSN: 2521-3407

Enhanced Support Vector Machine Methods Using Stochastic Gradient Descent and Its Application to Heart Disease Dataset

[1]Seror Faeq Mohammed **[2]Ghadeer Jasim Mohammed Mahdi*** **[3]Md Kamrul Hasan Khan**

[1,2]Department of Mathematics,College of Education for the Pure Science Ibn Al-Haitham , University of Baghdad, Iraq,.

[3]Department of Mathematical Sciences, Fulbright College of Arts & Sciences, 1University of Arkansas, USA

*Corresponding Author. gmahdi@ihcoedu.uobaghdad.edu.iq,

Received 7 May 2023, Received 8 June 2023, Accepted 12 June 2023, Published 20 January 2024

doi.org/10.30526/37.1.3467

Abstract

Support Vector Machines (SVMs) are supervised learning models used to examine data sets in order to classify or predict dependent variables. SVM is typically used for classification by determining the best hyperplane between two classes. However, working with huge datasets can lead to a number of problems, including time-consuming and inefficient solutions. This research updates the SVM by employing a stochastic gradient descent method. The new approach, the extended stochastic gradient descent SVM (ESGD-SVM), was tested on two simulation datasets. The proposed method was compared with other classification approaches such as logistic regression, naive model, K Nearest Neighbors and Random Forest. The results show that the ESGD-SVM has a very high accuracy and is quite robust. ESGD-SVM is used to analyze the heart disease dataset downloaded from Harvard Dataverse. The entire analysis was performed using the program R version 4.3.

Keywords SVM, classification, reduction of dimensions, variables selection, gradient descent, heart disease.

1. Introduction

Suppose we have a data set where for each subject we have information about an □-dimensional covariate vector , $x_{p\times 1}$ and a response y that has two possible categories. Our goal is to develop an algorithm that allows us to predict for each new observation the category of its response based on its □-information. The Support Vector Machine (SVM) is one of the many methods to do this, such as: Logistic Regression [1], Random Forest [2], k nearest neighbors [3], Naïve Bayes [4] and LDA [5]. SVM is a supervised learning technique, i.e., we create a classifier based on a training dataset and use this classifier for future observations.

For group one, let $y = +1$ and for group two, let $y = -1$. The goal of SVM is to design a classifier $f(x)$ such that, the classifier rule: $y = +1$ if $f(x) > 0$ and $y = -1$ if $f(x) < 0$, can be used to determine the response category given the covariate information. Unlike LDA, SVM does not use any distribution for x given its category. Instead, it is a geometric procedure that finds the classifier according to some optimization criterion.

In this work, we will discuss linear and nonlinear SVMsSuppose $f(x)$=,β-0.+,β-T.X for unknown parameters (,β-0.,β) is a function, a hyperplane in the space, that acts as a separator between the two response categories. .

Linearity is a simplifying assumption, and it is not reasonable to always assume that *f* is linear, because in general, non-linear SVMs are a popular tool for many real-world applications. In some cases, linear separation may work sufficiently so that we do not need to consider nonlinear assumptions.

We organize the paper as follows: In Sections 2 and 3, we introduce the hard (linear) boundary and the soft (nonlinear) boundary. The kernel transformation and its properties are discussed in Section 4. The improved SVM with stochastic gradient descent is explained in Section 5. Simulation studies with three data sets and the application to heart disease are described in Sections 6 and 7, respectively. Finally, the results and discussion are discussed in Section 8.

2. Hard Margin

Assume that two categories are linearly separable, so there exists a hyperplane $f(x) = \beta_0 + \beta^T = 0$ that separates the categories. Our task is to find out estimates of β_0 and β. Suppose $(x^{(1)}, y_1), (x^{(2)}, y_2), \ldots (x^{(n)}, y_n)$ are n data points from above setting. Consider any hyper plane $\beta_0 + \beta^T = 0$ in the x-space. The perpendicular distance of the i^{th} covariate point $x^{(i)}$ from this line is $d_i = y_i \frac{\beta_0 + \beta^T X^{(i)}}{\|\beta\|}$, and for all training data points $y_i(\beta_0 + \beta^i x) > 0$ [6]. The minimum of these distance is called the margin, i.e., there is no data point within this distance on either side of this line with respect to this training dataset [7].

The aim is to find a line that has maximum margin among all candidate lines. That line is going to be our estimate of classifier $f(X)$. The intuition behind keeping the margin maximum is a supervised learning procedure. We want to ensure that, for new observations (test data), we accurately determine category of y. We want to maximize the margin because it gives us a room for allowing for variation between training and test datasets.

Assume the optimization problem is

$$min_{\beta_0,\beta} \frac{\|\beta\|^2}{2} \quad \text{subject to } y_i(\beta_0 + \beta^T x^{(i)}) \geq 1 \; for \; i = 1,2,\ldots,n \qquad (1)$$

If $(\hat{\beta}_0, \hat{\beta})$ represent the solution to this optimization problem, some statements according to eq(1) can be hold.

First, $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}^T X = 0$ represents the estimated linear classifier. For any new data point with covariate information x_0, evaluate $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}^T X_0$. If $\hat{f}(x_0) > 0$, decide category1 for that

observation, otherwise if $\hat{f}(x_0) < 0$, decide category 2. Second, the margin of the classifier is: $\frac{1}{\|\hat{\beta}\|}$. Geometrically, there will not be any training data points between the lines

$\hat{\beta}_0 + \hat{\beta}^T X = +1$ and $\hat{\beta}_0 + \hat{\beta}^T X = -1$.

As shown in Fig.1, many lines can separate two groups ($H_1, H_2, H_3,$ and H_4), but there is only one optimal separating hyperplane (H) with two boundary hyperplanes ($H_{m1}\ and\ H_{m2}$) see Fig.2. Third, the training data points lie exactly on one of the above two lines (margins) are referred to as support vectors, so the estimates $\hat{\beta}_0, \beta$ depend only on the data points that are support vectors. That means all data points inside the correct margins have no role in determining the form of the classifier. This property makes SVM very useful for massive data classification problems [7].

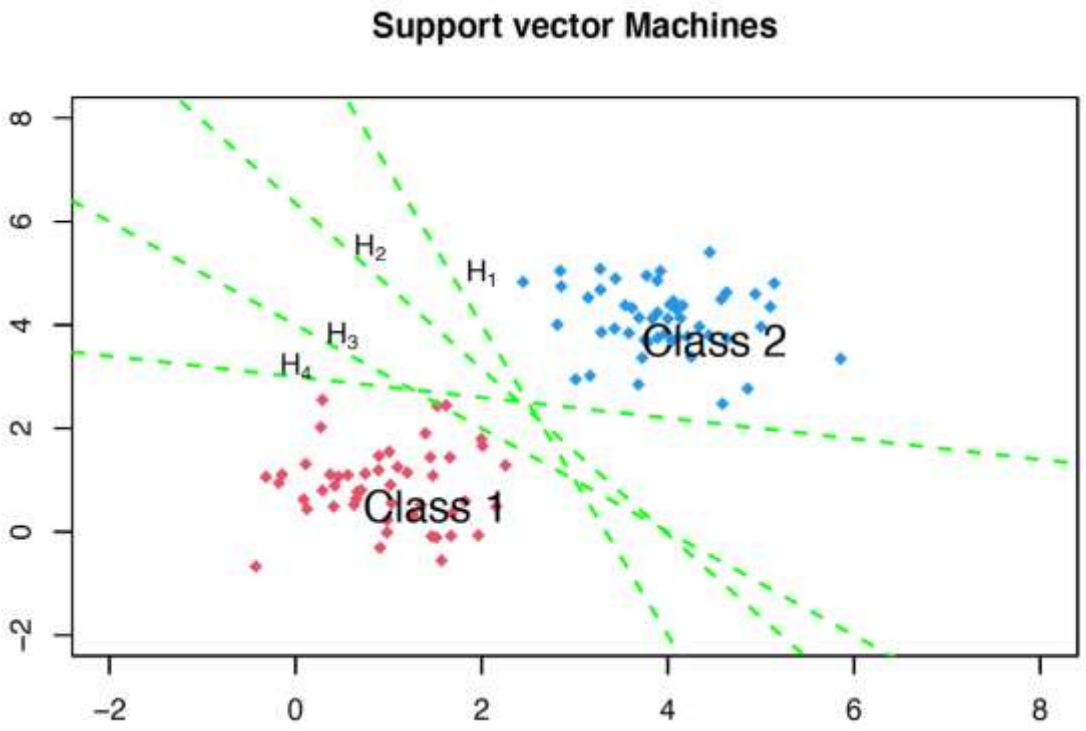


Figure 1. Some lines that separate two groups: $H_1, H_2, H_3, and\ H_4$.

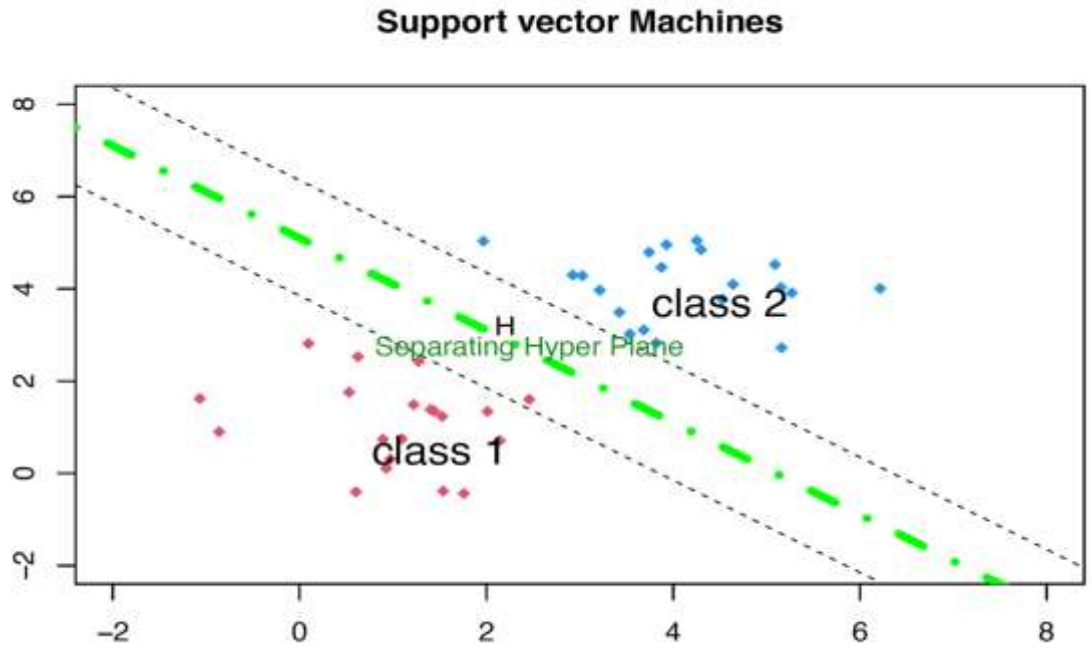


Figure 2. separating hyperplane, H, with two boundary hyperplanes : H_{m1} *and* H_{m2}.

3. Soft Margin

Suppose the assumption of linear separability does not hold between the two categories. That implies, no matter what hyper plane $\beta_0 + \beta^T X = 0$, there will always be points (x, y) such that:

$$\beta_0 + \beta^T x > 0 \; but \; y = -1 \; and \; \beta_0 + \beta^T x < 0 \; but \; y = +1.$$

Consider a nonnegative variable K_i for each observation I =1,2, …, n. K_i denotes the amount of push an observation needs to go the correct side of the margin. We have shown in class that [8]:

a. For points already obeying correct margin, $K_i = 0$.

b. For points violating the margin but staying on the correct side of the line, $0 < K_i < \frac{1}{\|\beta\|}$.

c. For points violating the margin and moving to the other side of the line $K_i > \frac{1}{\|\beta\|}$.

By solving the optimization problem, the estimates of (β_0, β) can be found:

$$\underset{\beta_0, \beta, s_i}{min} \frac{\|\beta\|^2}{2} + C \sum_{i=1}^{n} s_i \;\; subject \; to \; y_i(\beta_0 + \beta^T X^{(i)}) \geq 1 - s_i, s_i \geq 0 \; for \; i = 1,2,\ldots,n \qquad (2)$$

In eq (2), *C* is a pre-fixed turning parameter that balances the relative importance of maximizing the margin and minimizing the total amount of push required for points violating the margin.

If $(\hat{\beta}_0, \beta)$ represent the solution to this optimization problem, then the following statements hold:

1. $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}^T x = 0$ represents the estimated linear classifier. For any new data point with covariate information $x_{0,}$ evaluate $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}^T x_0 . If \; \hat{f}(x_0) > 0$, decide category 1 for that observation, otherwise if $\hat{f}(x_0) < 0$, decide category 2.

2.The Margin of the classifier is: $\frac{1}{\|\widehat{\beta}\|}$.

3. In machine learning literature $\{s_i\}$ are called slack variables.

4. In this case, the support vectors are defined as those observations in the training data lies exactly on any one of the two correct margins, $y_i(\hat{\beta}_0 + \hat{\beta}^T x^{(i)}) = +1$, or violate the correct margins, $y_i(\hat{\beta}_0 + \hat{\beta}^T x^{(i)}) < +1$.

5. It turns out that the estimations $\hat{\beta}_0, \hat{\beta}$ depend only on the data points which are support vectors. That means all data points inside the correct margins have no role in determining the form of the classifier. This property makes SVM very useful for massive data classification problems.

For any two nonlinearly separable classes (e.g., because of the noise), the optimal hyper-plane condition including an extra term can be formalized as follows:

$$\boldsymbol{y_i}(x_i^T \mathrm{w} + \mathrm{b}) \geq 1 - \zeta_i, \qquad , i = 1, \ldots, n.$$

The objective function should be minimized, i.e., $\zeta_i \geq 0$ should be minimized as well as $\|w\|$, as follows:

$$minimize \quad w^T w + c \sum_{i=1}^{n} \zeta_i^k \text{ subject to } y_i(x_i^T w + b) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0; i = 1, \ldots, n. \tag{3}$$

In eq (3), *C* is a regularization parameter that controls the balance between making the margin as big as possible and making the training error as small as possible. Small C tends to focus on the margin and ignore outliers in the training data, while large C may cause the training data to be too well fit (see **Figure 3**).

3.1 Second Norm Soft Margin

The optimization problem is called the second soft (nonlinear) margin problem when $k = 2$, i.e.,

$$minimize \quad w^T w + c \sum_{i=1}^{m} \zeta_i^2.$$

subject to $y_i(x_i^T w + b) \geq 1 - \zeta_i, \quad (i = 1, \ldots, m)$.

Not that if $\zeta_i \geq 0$ is declined, similarly if $\zeta_i < 0$, we can set it to zero and the above function can be further reduced.

The initial Lagrangian for the above 2-norm problem is

$$L_p(w, b, \zeta, \alpha) = \frac{1}{2} w^T w + \frac{c}{2} \sum_{i=1}^{n} \zeta_i^2 - \sum_{i=1}^{n} \alpha_i \, [y_i(w^T x + b) - 1 + \zeta_i].$$

Substituting

$$\frac{\partial L}{\partial w} = \mathrm{w} - \sum_{i=1}^{n} y_i \alpha_i x_i = 0; \ \frac{\partial L}{\partial \zeta} = c\zeta - \alpha = 0; \ \frac{\partial L}{\partial b} = \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Into the initial Lagrangian, we get the dual problem

$$maximize\ L_d(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j x_j^T x_i - \frac{1}{2c}\sum_{i=1}^{m} \alpha_i^2$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \,(x_j^T x_i + \frac{1}{c}\delta_{ij}).$$

Subject to $\alpha_i \geq 0, \sum_{i=1}^{m} \alpha_i\, y_i = 0.$

The above quadratic programming can be solved for α_i. As a result, all support vectors that correspond to $x_i > 0$ satisfy the following condition [8]:

$$y_i(X_i^T w + b) = 1 - \xi_i.$$

Substituting $w = \sum_{j\epsilon s} y_j y_j x_j$ into this equation (where S is the set of support vector), we get:

$y_i(\sum_{j\in S} y_j \alpha_j (X_i^T X_j) + b) = 1 - \xi_i,\ i.e.,\ y_i \sum_{j\in S} y_j \alpha_j\, (x_i^T x_j) = 1 - \xi_i - y_{i^b}.$

For the optimal weight w, we have :

$$\|w\|^2 \ :\ w^T w = \sum_{i\in S} \alpha_i y_i x_i^T \sum_{j\in S} \alpha_j y_j x_j = \sum_{i\in S} \alpha_i y_i \sum_{j\in S} \alpha_i y_i x_i^T x_j.$$

$\sum_{i\in S} \alpha_i (1 - \xi_i - y_i^b) = \sum_{i\in S} \alpha_i - \sum_{i\in S} \alpha_i \xi_i - b \sum_{i\in S} y_i \alpha_i.$

Since $\xi_i = \alpha_i / C$, we have:

$\sum_{i\in S} \alpha_i - \sum_{i\in S} \alpha_i \xi_i = \sum_{i\in S} \alpha_i - \frac{1}{c}\sum_{i\in S} \alpha_i^2.$

Therefor the optimal separation margin becomes:

$\frac{1}{\|w\|} = (\sum_{i\in S} \alpha_i - \frac{1}{c}\sum_{i\in S} \alpha_i^2)^{-1/2}.$

3.2 First Norm Soft Margin

The optimization problem is called the first soft (nonlinear) margin problem when k =1, i.e.,

minimize $w^T w + c \sum_{i=1}^{m} \zeta_i$

Subject to $y_i(x_i^T \mathrm{w} + \mathrm{b}) \geq 1 - \zeta_i, \zeta_i \geq 0$; i=1,…, n.

The first norm algorithm is less complex compared with the second norm algorithm. It is powerful when the training dataset has outliers. The first norm method should be used to ignore the outliers when the data is noisy [9,10].

The primal Lagrangian for first norm problem above is:

$$L_p(w, b, \xi, \alpha, \Upsilon) = \frac{1}{2} w^T w + c \sum_{i=1}^{n} \Upsilon_i \xi_i - \sum_{i=1}^{n} [y_i(w^T x + b) - 1 + \xi_i] - \sum_{i=1}^{n} \Upsilon_i \xi_i.$$

with $\alpha_i \geq 0 \quad and \quad \Upsilon_i \geq 0$. substituting

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} y_i \alpha_i x_i = 0; \quad \frac{\partial L}{\partial \xi} = c\xi - \alpha = 0; \quad \frac{\partial L}{\partial b} = \sum_{i=1}^{n} y_i \alpha_i = 0.$$

Into the initial Lagrangian, we get the dual problem:

maximize $L_d(\alpha, \Upsilon) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j x_j^T x_i - \sum_{i=1}^{n} \alpha_i \xi_i - \sum_{i=1}^{n} \Upsilon_i \xi_i + c\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} y_i y_j \alpha_i \alpha_j x_j^T x_i.$

subject to $0 \le \alpha_i \le c$, $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Remarkably the dual problem objective function is the same as that of the linearly separable case that was discussed previously. This is because the cancellation depends on $c = \alpha_i + y_i$. Now, when $\alpha_i \ge 0$ and $\Upsilon_i \ge 0$, we get $0 \le \alpha_i \le c$. When we solve the Quadratic Programming (QP) problem for α_i , the following optimal decision plane with the margin can be gotten:

$$(\sum_{i \in S} \sum_{j \in S} \alpha_i \alpha_j y_i y_j x_i^T x_j)^{-1/2}$$

4. Kernels Transformation

The kernel transformation techniques can be applied when, the condition set of Karush-Kuhn-Tucker (KKT) are satisfied. Eq (1) is linear in terms of the new space that ϕ (x) maps the data to non-linear in the space, see **Figure 3**. The most common kernels are: linear, polynomial, sigmoid or Multi-Layer Perceptron (MLP) and Gaussian or Radial Basis Function (RPF) [11-13]. Their expressions are as follows:

Linear kernel:$K(x_i, x_j) = x_i^T x_j$.

Polynomial kernel:$K(x_i, x_j) = (1 + x_i^T x_j)^p$.

Sigmoid (MLP) kernel: $K(x_i, x_j) = tanh(k_1 x_i^T x_j + k_2)$.

Gaussian (RBF) kernel: $K(x_i, x_j) = exp[\frac{-(x_i - x_j)^T (x_i - x_j)}{2\sigma^2}]$

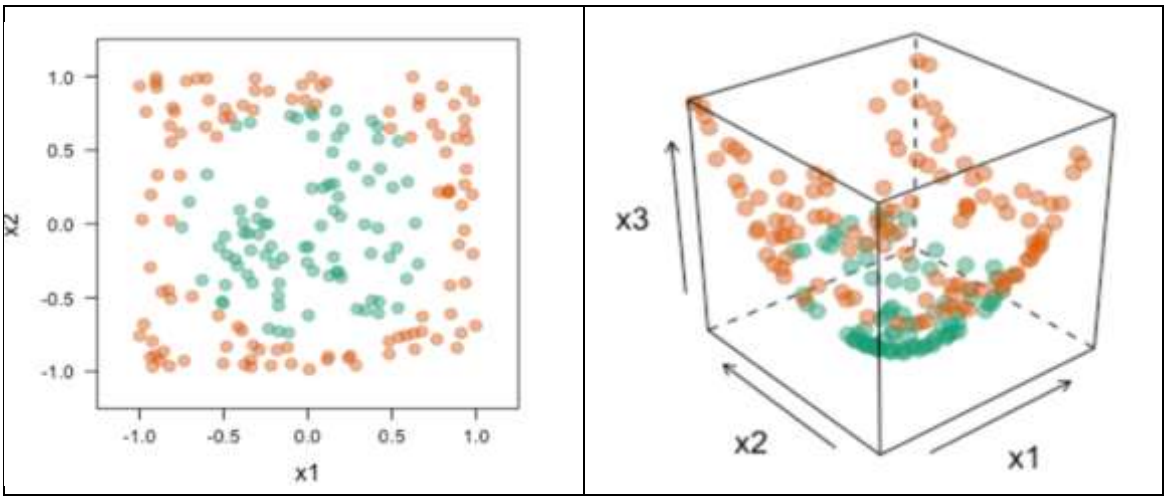
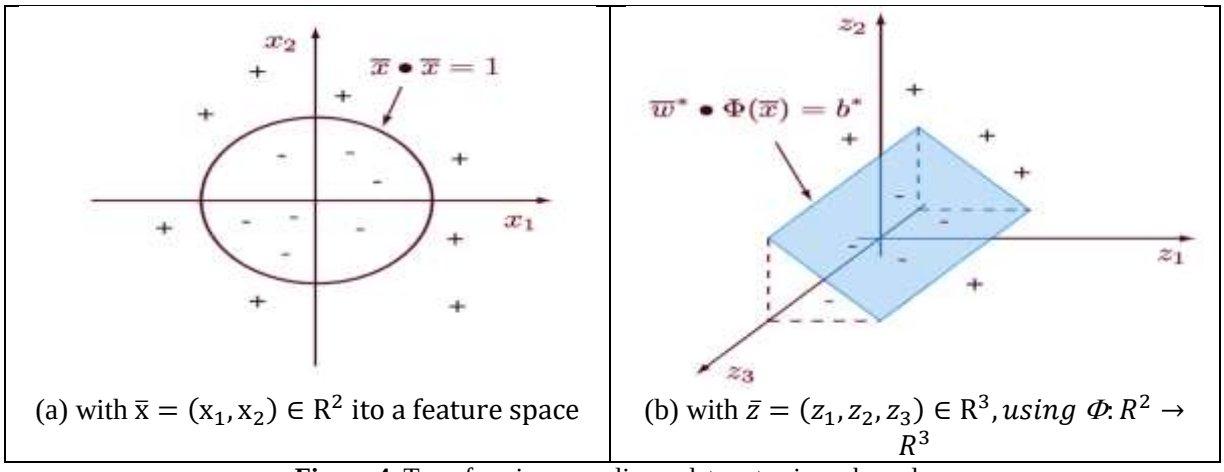


Figure 3. Tow-dimensional space verses three-dimensional space.

we define the kernel function as $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >= \phi(x_i)^T \phi(x_j)$ where ϕ is a mapping from input space to output space, see **Figure 4**.



(a) with $\bar{x} = (x_1, x_2) \in R^2$ ito a feature space

(b) with $\bar{z} = (z_1, z_2, z_3) \in R^3, using\ \Phi: R^2 \rightarrow R^3$

Figure 4. Transforming a nonlinear data set using a kernel.

Now, the corresponding dual form is

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j). \quad \text{Subject to } \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \alpha_i \geq 0, \quad i = 1, \ldots, n. \tag{4}$$

Eq (4) is called a cost function, and it is convex and quadratic in terms of the unknown parameters. It can be solved using quadratic programming. The final decision rule for classification using KKT conditions is [14,15]:

$$L(x, \alpha^*, \beta_0) = \sum_{i=1}^{N_S} y_i \alpha_i^* K(x_i, x) + \beta_0.$$

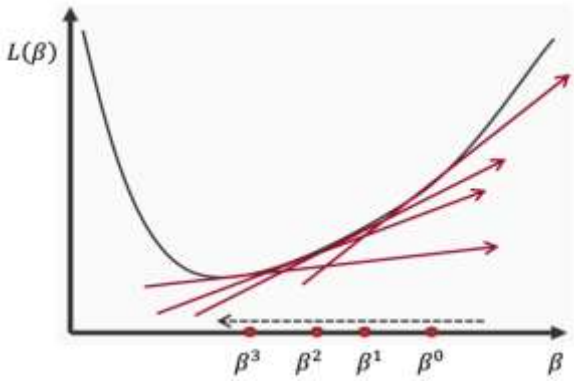
Where N_S is the number of support vectors, and α_i the non-zero Lagrange multipliers that associated with the support vectors.

5. Enhanced Stochastic Gradient Descent SVM

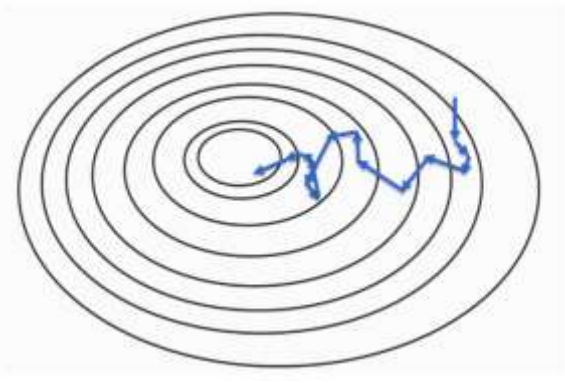
The goal of this section is to minimize the following function,

$$L(\beta) = \frac{1}{2} \beta_0^T \beta_0 + K \sum_i \max(0, 1 - y_i \beta^T x_i) \tag{5}$$

Equation(5) is a quadratic optimization problem and it is convex in □. In the previous section, the QP technique was used, but it is very slow. If there are no constraints, gradient descent can be used [16,17]. In general, the gradient goes in the opposite direction to get to the minimum, as shown in **Figure 5**.a, because the function is in the direction of the steepest slope.



a. Minimizing $L(\beta)$ using a general strategy



b. Minimizing $L(\beta)$ using SGD

Figure 5. Minimizing $L(\beta)$ using stochastic gradient descent.

The general gradient descent SVM strategy (GD-SVM) for minimizing Eq.(5) starts with an initial value for β, say β^0, then iterate until convergence. GD-SVM is faster than QP, but it is still slow because computing $\nabla(\beta^j)$ takes □(□) time in complexity, where □ is the size of the training dataset. If □ is large, GD-SVM is slow [18,19]. In Stochastic Gradient Descent (SGD), the value of the objective function is improved at each step. Evaluating the gradient for each training sample instead of evaluating it for all samples speeds up the process. This pressure is called Enhanced stochastic gradient descent SVM (ESGD-SVM). As can be seen in Fig.5.b, ESGD-SVM requires many more steps than the GD-SVM method, but it is less computationally intensive at each update; and ESGD-SVM is faster than the GD-SVM method. The ESGD-SVM algorithm (Algorithm 1) is guaranteed to converge to the minimum of □ if ,□-□. is small enough [20-22].

Algorithm 1: SVM using Stochastic Gradient Descent (ESGD-SVM)

Given a training set $S = \{(x_i, y_i): x \in \mathbb{R}^n \; and \; y \in \{-1, +1\}\}$

Repeat until convergence

1. Initial value: β^0.

2. For $i = 1, \ldots,$ k:

i. compute: $\nabla(\beta^j) = \frac{\partial f(\beta,\beta^0)}{\partial \beta^j} = \beta^j + R\sum_{i=1}^{n} \frac{\partial L(x_i,y_i)}{\partial \beta^j}$, where R is a regularization factor.

ii. recompute β as follows: $\beta^j \leftarrow \beta^j - \eta\nabla(\beta^j)$, where η is the learning rate value.

iii. $\nabla J^t(\beta^t) = \frac{1}{2}\beta_0^T\beta_0 + K \cdot N \cdot \max(0, 1 - y_i\beta^T x_i)$, where N is the number of training examples.

iv. Update β as follows: $\beta^t \leftarrow \beta^{t-1} - \gamma_t \nabla J^t(\beta^{t-1})$.

3. Repeat (x_i, y_i) to make a full dataset and take the derivative of the SVM objective at the current β^{t-1} to be $\nabla J^t(\beta^{t-1})$.

4. Return final β.

6. Simulation Studies

To test the GD-SVM and ESGD-SVM methods, two simulation datasets with 100 and 200 observations are created. Each dataset contains a set of variables and a response comprising two classes. In real life, it is not known whether the data sets are linearly or non-linearly separable, so complex data sets (non-linearly separable) are generated. GD-SVM and ESGD-SVM are applied to the two data sets by using different types of kernels. In both methods, the MLP kernel provides the best accuracy in the two datasets [23,24]. In **Table 1**, GD-SVM and ESGD-SVM are compared in terms of the best value of □, the number of support vectors, using sensitivity (also called true-positive rate) and specificity (also called true-negative rate) [25,26].

Table 1. Comparison between GD-SVM & ESGD-SVM for a data set with 100 and 200 observations.

Sample size	Method	Kernel Type	Best *K* Value	Number of Support Vectors	Sensitivity%	Specificity%	Accuracy
100 Observations	**GD-SVM**	Linear	0.78	13	82.30	89.43	86.43
		Polynomial	0.45	31	94.40	92.58	93.56
		RBF	0.32	19	92.74	94.01	93.01
		MLP	0.23	22	90.95	94.33	92.33
	ESGD-SVM	Linear	0.96	12	92.30	97.43	93.45
		Polynomial	0.55	19	95.40	96.53	95.53
		RBF	0.37	15	96.74	97.01	96.53
		MLP	0.33	16	98.95	96.33	97.33
200 Observations	**GD-SVM**	Linear	0.83	23	81.31	87.43	85.43
		Polynomial	0.26	41	89.44	91.55	90.54
		RBF	0.39	28	89.74	84.01	87.01
		MLP	0.24	33	89.94	91.73	90.72
	ESGD-SVM	Linear	0.92	14	80.33	83.83	82.81
		Polynomial	0.57	29	87.43	88.58	86.54
		RBF	0.48	25	88.44	87.97	87.99
		MLP	0.37	28	86.41	85.96	85.66

7. Real Dataset

The modification of the SVM using stochastic gradient descent is the main topic of this paper. For real data applications, we have used South African heart disease data. The dataset was downloaded from [18]. In this dataset, a historical sample of men in the Western Cape of South Africa, a region with a high incidence of cardiovascular disease, is described. The following patient characteristics were recorded for each high-risk individual: Factors such as age, type A behavior, family history of heart disease, systolic blood pressure, cumulative cigarette consumption, low-density lipoprotein cholesterol, body fat percentage and obesity. The total number of samples in this data collection is 462. Obesity refers to a high percentage of body fat, while obesity is defined by a high weight-to-height ratio (body mass index, BMI) [27] Excessive antagonism, aggression, and competitiveness are hallmarks of the Type A personality. We will see if we can extrapolate ldl from the available data. Since low-density lipoprotein (ldl) cholesterol is the "bad" cholesterol, elevated levels are thought to be associated with obesity and adiposity. The aim is to create a predictive model for ldl by selecting the most important factors [28].

In **Table (2)**, we examine the Pearson correlation coefficient between the groups. It can be seen that there is a high significant correlation (marked with **), a significant correlation (marked with *) and a weak correlation between the variables [29,30].

Table 2. Pearson correlation matrix between variables of heart disease dataset

Pearson Correlation		sbp	tobacco	ldl	adiposity	typea	obesity	alcohol
tobacco	**R**	0.212						
	P	0.092						
ldl	**R**	0.158	0.159					
	P	0.958	0.203					
adiposity	**R**	0.357*	0.287	0.440**				
	P	**0.043**	0.349	**0.002**				
typea	**R**	-0.057	-0.015	0.440*	-0.043			
	P	0.945	0.503	**0.039**	0.203			
obesity	**R**	0.238*	0.125	0.331*	0.717**	0.074		
	P	**0.043**	0.349	**0.049**	**0.014**	0.249		
alcohol	**R**	0.140	0.201	-0.033	0.123	0.039	0.052	
	P	0.4l3	0.459	0.953	0.034	0.254	0.359	
age	**R**	0.389*	0.450**	0.312	0.626**	-0.103	0.692**	0.101
	P	**0.042**	**0.011**	0.059	**0.023**	0.539	**0.024**	0.239

R: Pearson correlation, P: p value
****High significant correlation between variables (p value < 0.01).**
***Significant correlation between variables (p value < 0.05).**

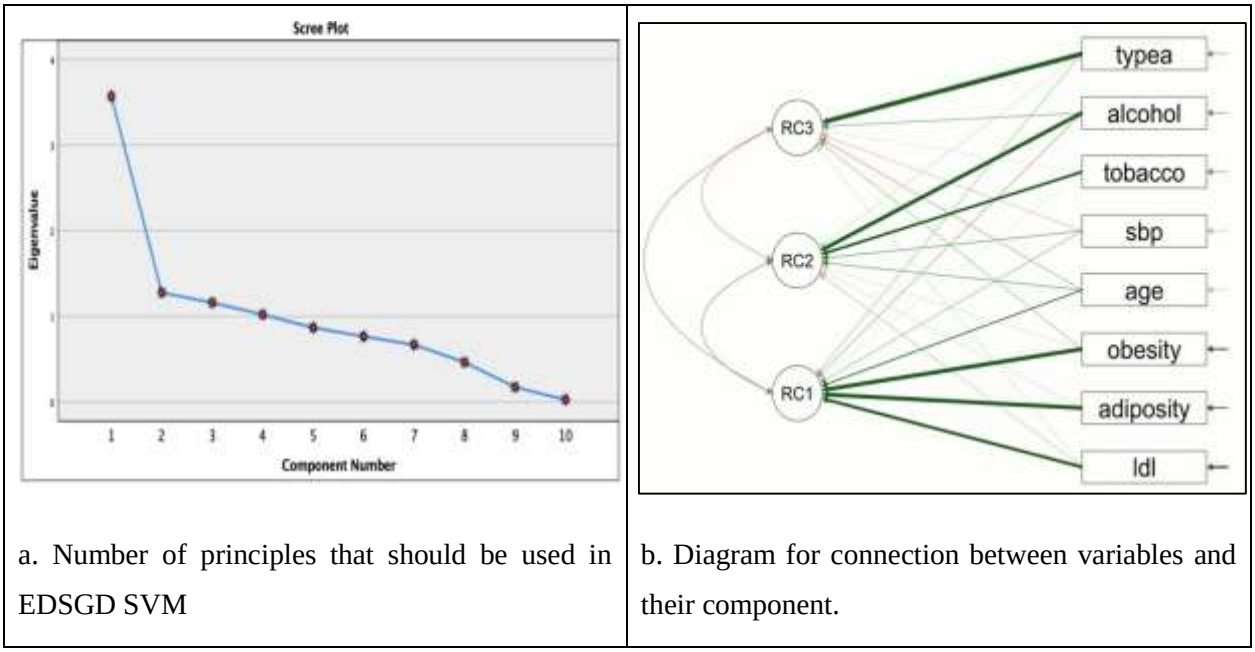
The estimated coefficient for the model was calculated in **Table (3)**, some variables were highly significant since p value was less than 0.02, and some were significant where p value < 0.05. In general, 6 variables are selected as important variables in the dataset.

Table 3. Estimated coefficient for model using heart disease dataset.

Variables in the Equation	$\widehat{\boldsymbol{\beta}}$	S.E.	Wald	P-value	Exp(B)
Constant	-5.225	1.315	15.782	<0.001**	0.005
sbp	0.007	0.006	1.288	0.256	1.007
tobacco	0.379	0.027	8.903	0.003*	1.083
ldl	0.174	0.06	8.498	0.004*	1.19
adiposity	0.019	0.029	0.403	0.526	1.019
famhist	-0.925	0.228	16.488	<0.001**	0.396
typea	0.340	0.012	10.329	0.001*	1.04
obesity	-0.063	0.044	2.021	0.155	0.939
alcohol	0.002	0.004	0.001	0.978	1
age	0.453	0.012	13.901	<0.001**	1.046

***Significant effect for the parameter in the equation.**
****High significant effect for the parameter in the equation.**

Instead of using 6 variables in the model as it was illustrated in **Table (3)**, the principal component variable can be used to reduce the components. In **Figure 6.a**, we draw a curve as a scree plot to show the number of components depends on their eigenvalues. Using an eigenvalue equal to 1 as a cut off, three components (RC1, RC2, RC3) are satisfactory to represent all the variables in the dataset. **Figure 6.b** shows the connection between the variables and their components.



a. Number of principles that should be used in EDSGD SVM

b. Diagram for connection between variables and their component.

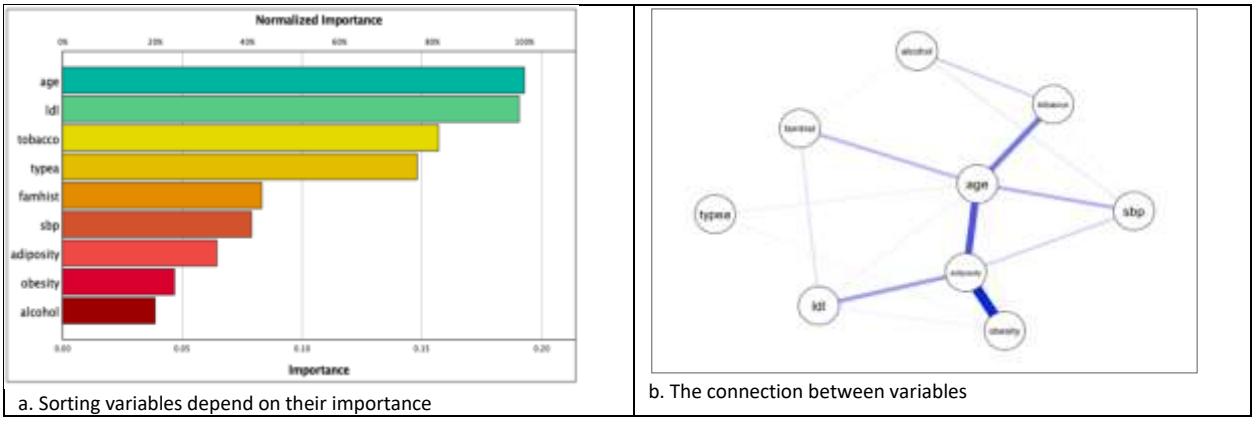
Figure 6. Importance variables for heart disease dataset using PCA technique.

The summary of the main principles and their weight according to the variable that is used in the real dataset is shown in **Table 4**.

Table 4. Principle components for the dataset.

	Component Loadings		
Variables	**RC1**	**RC2**	**RC3**
sbp			
tobacco		0.683	
ldl	0.747		
adiposity	0.868		
typea			0.959
obesity	0.877		
alcohol		0.853	
age	0.477		

Figure 7 shows the important variables and their connections. In **Figure 7.a** the important variables are sorted in order. It shows that age and ldl are the most important variables followed by tobacco and typea. **Figure 7.b** shows the connection between variables. We see the strong connections are marked with thick blue lines, and the weak connection were marked with thin blue lines.



a. Sorting variables depend on their importance

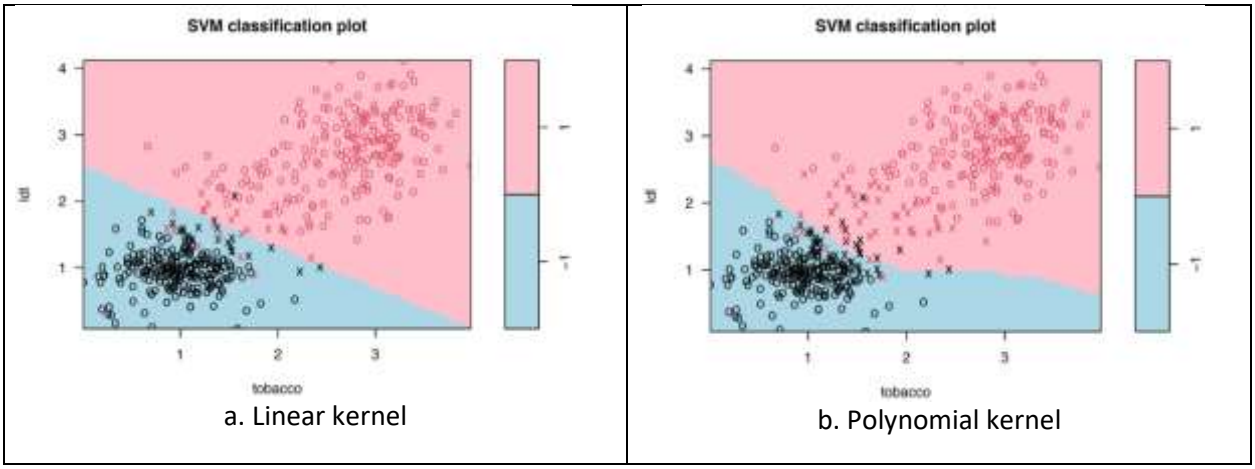
b. The connection between variables

Figure 7. Path diagram for the relation between the variables and the component.

8. Results and Discussion

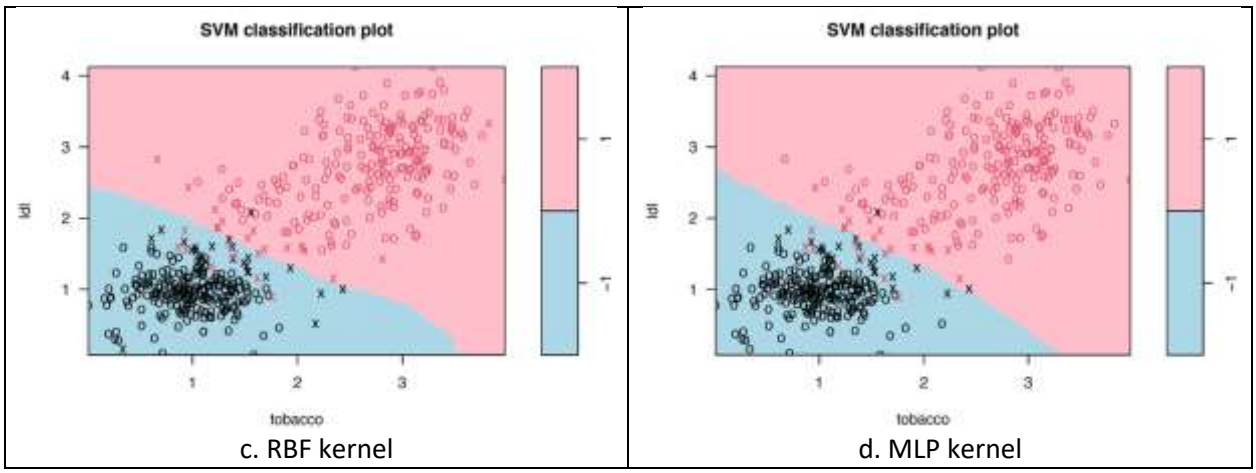
The enhanced version of SVM (ESGD-SVM) is applied to the heart disease dataset. A comparison between some different kernel functions had been shown. In application, we applied the most common kernel which is linear, polynomial, RBF and MLP kernels. In addition, the proposed method was compared with some common methods. These methods used the same dataset for leukemia classification which are k-nearest neighbor random forest and naïve Bayes. In **Figure10**, ESGD-SVM classification method for the heart disease dataset is plotted. The two most important variables, which are tobacco and ldl are used for visualization. As it is shown from the plot (**Figure 8**) and the table (**Table 5**) the best version for ESGD-SVM method is satisfied when the RBF kernel is applied.

The MLP kernel gets 98.10% accuracy which is the highest performance compared with other kernels. SGD-SVM performed 96.53% accuracy for the linear kernel, 98.03% accuracy for the polynomial kernel, and 96.53% accuracy for RBF kernel. ESGD-SVM performed much better than other methods where logistic regression performed 87.42% accuracy, k-nearest neighbor performed 86.70% accuracy, random forest performed 87.33% accuracy, and naïve Bayes performed 84.34% accuracy.



a. Linear kernel

b. Polynomial kernel



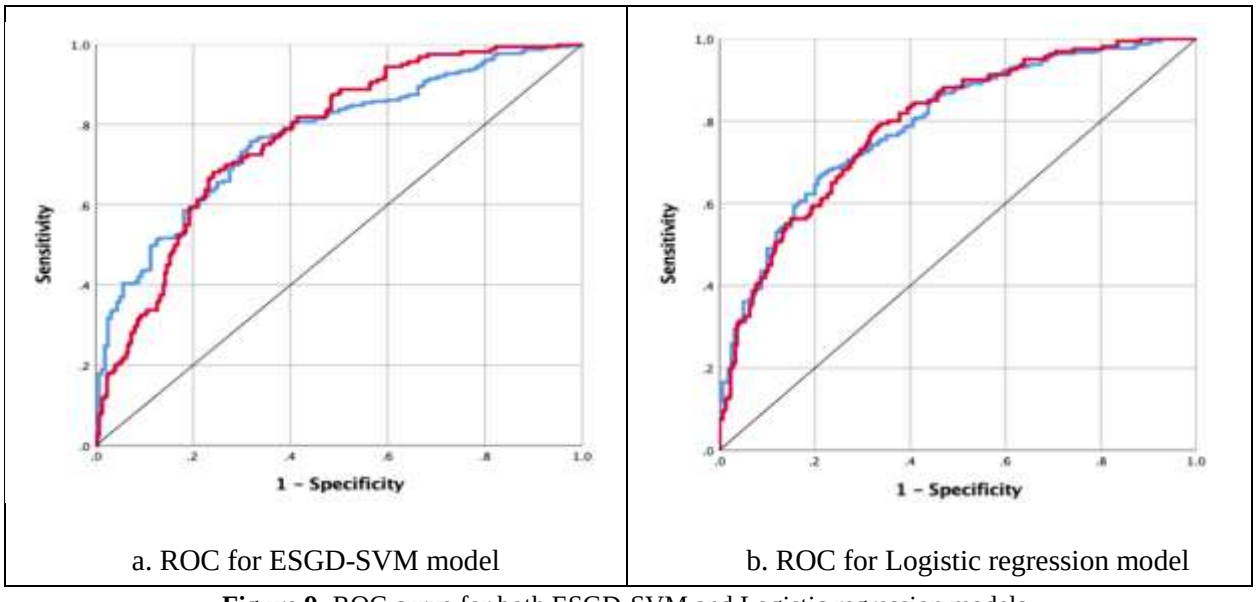
c. RBF kernel d. MLP kernel

Figure 8. Classification for heart disease dataset using ESGD-SVM.

Table 5. Comparison between ESGD-SVM, logistic regression, k-nearest neighbors, random forest, and naive Bayes for classification heart disease dataset.

Methods		Number of Support Vectors	Sensitivity %	Specificity %	Accuracy Rate %
ESGD-SVM	**Linear kernel**	42	96.61	98.67	96.53
	Polynomial kernel	96	98.19	94.19	98.03
	RBF kernel	51	96.61	98.67	96.53
	MLP kernel	48	97.21	99.10	98.10
Logistic Regression			86.39	89.48	87.42
k-nearest neighbors			85.20	88.40	86.70
random forest			89.33	84.87	87.33
naive Bayes			86.49	82.71	84.34

The Receiver Operating Characteristic curve, or ROC curve, is shown. ROC is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied for both the Logistic regression model and ESGD-SVM models.



a. ROC for ESGD-SVM model b. ROC for Logistic regression model

Figure 9. ROC curve for both ESGD-SVM and Logistic regression models.

9. Conclusion

This paper presents the ESGD-SVM method. The stochastic gradient descent process is used to develop the method. The kernel transformation technique and dimensionality reduction for variables are used to achieve the best classification accuracy with ESGD-SVM. Two simulation datasets are used to test the implementation of the method. The results show that ESGD-SVM has higher accuracy compared to some other classification methods: logistic regression, k-nearest neighbors, random forest and naive Bayes. When the method was applied to a real dataset (heart disease), it was found that the highest accuracy (98.10%) was achieved by applying the MLP kernel.

Acknowledgments

The authors are very grateful to the reviewers for their valuable comments and suggestions to improve the paper

Conflict of Interest

The authors declare that they have no conflicts of interest.

Funding

There is no financial support in preparation for the publication.

References

1. Zou, X.; Hu, Y.; Tian, Z.; Shen, K. Logistic regression model optimization and case analysis. IEEE 7th international conference on computer science and network technology (ICCSNT) **2019**, *7*, 135-139.
2. Liaw, A.; Wiener, M. Classification and regression by random Forest. R news. **2002**, *3*,18-22.
3. Khorshid, S.F.; Abdulazeez, A.M. Breast cancer diagnosis based on k-nearest neighbors: a review. *PalArch's Journal of Archaeology of Egypt/Egyptology*. **2021**, *18*, 1927-51.
4. Chen, S.; Webb, G.I.; Liu, L.; Ma, X. A. Novel selective naïve Bayes algorithm. Knowledge-Based Systems. **2020**, 192, 105361
5. Choubey, D.K.; Kumar, M.; Shukla, V.; Tripathi, S.; Dhandhania, V.K.; Comparative analysis of classification methods with PCA and LDA for diabetes. Current diabetes reviews. **2020**, *16*, 833-50.
6. Tawfiq, L.N.; Rashid, T.A. On Comparison Between Radial Basis Function and Wavelet Basis Functions Neural Networks. *Ibn AL-Haitham Journal For Pure and Applied Science*. **2017**, *23*, 184-92.
7. Zhi, J.; Sun, J.; Wang, Z.; Ding, W. Support vector machine classifier for prediction of the metastasis of colorectal cancer. *Int J Mol Med*. **2018**, *41*, 1419-26.
8. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing. **2020**, *408*, 189-215.
9. Hekmatmanesh, A.; Wu, H.; Jamaloo, F.; Li, M.; Handroos, H. A combination of CSP-based method with soft margin SVM classifier and generalized RBF kernel for imagery-based brain computer interface applications. Multimedia Tools and Applications. **2020**, *79*, 17521-49.
10. Wang, Y.; Yu, W.; Fang, Z. Multiple kernel-based SVM classification of hyperspectral images by combining spectral, spatial, and semantic information. Remote Sensing. **2020**, *12*, 120.

11. RAHEEM, S. H; KALAF, B. A.; SALMAN, A. N. Comparison of Some of Estimation methods of Stress-Strength Model: R= P (Y< X< Z). *Baghdad Science Journal*, **2021**, *18.2*, 1103-1103.

12. JEBUR, I. G.; KALAF, B. A.; SALMAN, A. N. An efficient shrinkage estimators for generalized inverse rayleigh distribution based on bounded and series stress-strength models. In: Journal of Physics: Conference Series. IOP Publishing, **2021**, 012054.

13. Mahdi, G.J.; Mohammed, N.J.; Al-Sharea, Z.I. Regression shrinkage and selection variables via an adaptive elastic net model. In Journal of Physics: Conference Series **2021**, 1879, 032014.

14. Qingyang, Z.; Ghadeer, M.; Jian, T.; Hao, C.; A graph-based multi-sample test for identifying pathways associated with cancer progression. Computational Biology and Chemistry, **2020**, 87: 107285.

15. ZHANG, Q.; DAO, T. A distance based multisampling test for high-dimensional compositional data with applications to the human microbiome. BMC bioinformatics, **2020**, *21*, 1-17.

16. Mahdi, G.J, Kalaf, B.A.; Khaleel, M.A. Enhanced supervised principal component analysis for cancer classification. *Iraqi Journal of Science*. **2021**, 1321-33.

17. Mseer, H.A.; Mahdi, G.J. Comparison among variable selection models and its application to health dataset. InAIP Conference Proceedings **2023**, *1*, 2414.

18. Jabbar, A.K. New transform Fundamental properties and its applications. *Ibn Al-Haitham Journal for Pure and Applied Sciences.* **2018**, *31*, 1-10.

19. Mahdi, G.J.; A Modified Support Vector Machine Classifiers Using Stochastic Gradient Descent with Application to Leukemia Cancer Type Dataset. *Baghdad Science Journal*. **2020**,*17*,1255-69.

20. Raheem, S.H.; Kalaf, B.A.; Salman A.N. Comparison of Some of Estimation methods of Stress-Strength Model: R= P (Y< X< Z). *Baghdad Science Journal*. **2021**,*18*,1103-17.

21. Salah, O.M.; Mahdi, G.J.; Al-Latif, I.A. A modified ARIMA model for forecasting chemical sales in the USA. *In Journal of Physics*: Conference Series **2021**, 1879, 032008.

22. AL-NOOR, N. H.; KHALEEL, M. A.; MOHAMMED, G. J. Theory and applications of Marshall Olkin Marshall Olkin Weibull distribution. In: *Journal of Physics: Conference Series*. **2021**,20, 012101.

23. SHEAH, R. H.; ABBAS, I. T. Using multi-objective bat algorithm for solving multi-objective non-linear programming problem. *Iraqi Journal of Science*, **2021**, 997-1015.

24. MOHAMMED, M. J.; MOHAMMED, A. T. Analysis of an Agriculture Data Using Markov Basis for Independent Model. In: *Journal of Physics: Conference Series*. **2020**, 012071.

25. MOHAMMED, M. J.; MOHAMMED, A. T. Parameter estimation of inverse exponential Rayleigh distribution based on classical methods. *International Journal of Nonlinear Analysis and Applications*, **2021**, *12*, 935-944.

26. Bartley, C. Replication Data for: South African Heart Disease" Available online: https://doi.org/10.7910/DVN/76SIQD Harvard Dataverse, V1, **2016**.

27. Bayda, A. Abdul Jabbar, K. B.; Iraq, T. A. Mohd, R. A.; Lee, L. S. Application of simulated annealing to solve multi-objectives for aggregate production planning. In: *AIP Conference Proceedings*. **2016**, 1739, 020086.

28. Bogatinovski, J.; Ljupčo, T.; Sašo, D.; Dragi, Kocev. Comprehensive comparative study of multi-label classification methods. Expert Systems with Applications. **2022**, 203, 117215.

29. FJELLSTRÖM, C.; NYSTRÖM, Kaj. Deep learning, stochastic gradient descent and diffusion maps. *Journal of Computational Mathematics and Data Science*. **2022**, *4*, 100054.

30. HASSAN, A. S.; KHALEEL, M. A.; MOHAMD, R. E. An extension of exponentiated Lomax distribution with application to lifetime data. *Thailand Statistician*. **2021**, *19*, 484-500.