

# Support Vector Machine Classification on a Biased Training Set: Multi-Jet Background Rejection at Hadron Colliders

F. Sforza, V. Lippi, S. Leone

---

## Abstract

This paper describes an innovative way to optimize a multivariate classifier, in particular a Support Vector Machine algorithm, on a problem characterized by a biased training sample. This is possible thanks to the feedback of a signal-background template fit performed on a validation sample and included both in the optimization process and in the input variable selection. The procedure is applied to a real case of interest at hadron collider experiments: the reduction and the estimate of the multi-jet background in the  $W \rightarrow e\nu$  plus jets data sample collected by the CDF experiment. The training samples, partially derived from data and partially from simulation, are described in detail together with the input variables exploited for the classification. At present, the reached performance is superior to any other prescription applied to the same final state at hadron collider experiments.

*Keywords:* QCD rejection, SVM, Monte Carlo Models, CDF, Tevatron

---

## 1. Introduction

A multivariate classifier is an adaptive algorithm trained to identify a signal of interest against other background events on the basis of a set of input variables. Therefore the understanding of the training samples and the input variables selection are two key elements to obtain optimal results.

In this paper we apply the previous paradigm in an innovative way to both the training and the input variable selection of a Support Vector Machine [1] (SVM) algorithm. In particular we obtain an excellent signal-background multivariate classifier when one of the training samples is biased (i.e. it does not correctly reproduce all characteristics of the signal or of the background samples) and statistically limited to few thousands of events.

We decided to explore the use of the SVM algorithm (described in Section 2) because of several advantages with respect to other multivariate techniques. For example, Artificial Neural Networks, commonly used in High Energy Physics [2], require to arbitrarily set the complexity of the classifier (i.e. the number of neurons and layers of the net), the training may converge to local minima and, usually, large training sets are needed to finely map the input space. On the

other hand the SVM algorithm, whose basic idea is the identification of the best hyper-plane separating two classes of vectors, has unique solution of the training algorithm, a small number of free parameters and good performance on low statistics training sets, as only a small number of training vectors are exploited in the final solution [3]. Other promising results using SVMs in High Energy Physics analyses are reported in Ref. [4–6].

In this work, developed in the framework of the analysis searching for the Higgs boson with the Collider Detector at Fermilab (CDF) experiment [7], we deal with three machine-learning challenges: the reduction of the effect of a bias in the training set, a robust evaluation of the performance of the trained SVM, and the optimal input variable selection. The key point to achieve all of the above is the scan of the free parameters of the training procedure together with a cross check of the efficiency of each resulting SVM (described in Section 4.1). The efficiency cross check is performed both on the training set, with a  $n$ -fold cross validation, and on unclassified events, with a template fitting procedure over the SVM output value distribution (described in Section 3).

We tested the developed methodology on a toy model (Section 3.1) and finally we applied it to a real physics case (Section 4). In this latest part we exploited the described template fitting procedure to identify a reliable and optimal input variable selection. A wide literature discusses the topic of input variable selection but we focused on the identification of the best, minimal set of inputs. It is clear that a highly discriminative variable will improve the classification power but it has been shown in Ref. [8] that also the performance of the SVM algorithm itself improves when the variables are well chosen, especially if they have very different discrimination power [9]. Intuitively, the introduction of too many not-significant inputs introduces a noise term to the algorithm, decreasing its performance. Furthermore, a reduced set of variables identifies the most important characteristics of the analyzed process and may decrease the number of needed validation checks. The complete procedure (described in Section 4.3) involves the automatic training and performance evaluation with several variables configurations.

The physics case under exam is the reduction and estimate of the multi-jet background contamination in a dataset enriched in leptonic  $W$  bosons decays, selected in association with hadronic jets.

The dataset, collected at a proton-antiproton collider which was operating at  $\sqrt{s} = 1.96$  TeV (the Tevatron), is one of the main investigation channels at hadron collider experiments. Several interesting, but rare, processes (i.e. associate  $WH$  production, diboson or top quark production, etc.) produce a  $W$  boson in the final state. The  $W$  leptonic decay is used to identify a clear event signature out of the overwhelming multi-jet background produced by generic QCD interactions between the proton and antiproton constituents. Hadronic jets faking the lepton and the neutrino identification introduce a significant multi-jet background contamination (especially for electrons identification algorithms). Because of its nature, the multi-jet background is a mixture of detector effects and physics processes. Commonly, data-driven models, obtained with a specific selection enriched in multi-jet events, are used to estimate this contam-

ination. These models may be statistically limited and the use of a different selection criteria often introduces unexpected biases in the simulated variables. All this makes the application of multivariate techniques particularly challenging.

## 2. Support Vector Machines

The SVM is a supervised learning binary classifier whose basic function is the identification of the *best separating hyper-plane* between two classes of  $n$ -dimension vectors.

Given a training set made by two classes of vectors (i.e. signal and background) linearly separable, the SVM algorithm produces, as a solution, a unique plane defined by the vectors at the boundary of the two classes; those are the so called *support vectors*. In the case of non-linear separation, the plane is found in an abstract space, defined by a transformation of the input vectors. Although the transformation can be very complex, it is not necessary to know it exactly, but we just need to know its effect on the scalar product between the vectors, named *Kernel*. Finally the cases of not perfect separability of the two samples are solved by introducing a penalty parameter accounting for the contamination.

It is possible to find more details in Ref. [3] and [1], but, for sake of clarity, a short overview of the algorithm is also given in the following. For the actual, numerical, implementation of the SVM algorithm we relied on the LIBSVM open source library [10].

### 2.1. The Linear Case

Figure 1 shows how the linear classification problem can be formalized in the minimization of  $|\vec{w}|^2$  (with  $\vec{w}$  = vector normal to a plane) with the constraint:

$$y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \begin{cases} y_i = +1, & i \in \text{signal}; \\ y_i = -1, & i \in \text{background}; \end{cases} \quad (1)$$

the problem has a unique solution obtained by the maximization of:

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j, \quad (2)$$

which are derived by the application of the Lagrange multipliers to Eq. 1.

The solution identifies  $\alpha_i > 0$  for some  $i$ . The associated vectors, i.e. the support vectors, are a subset of the training sample that defines the best hyper-plane separating the two input classes (see Figure 1).

When not completely separable classes of vectors are present, a *penalty parameter*  $C$  is added to account for the contamination. The new minimization condition is:

$$|\vec{w}|^2 + C \sum_i \xi_i; \quad (3)$$

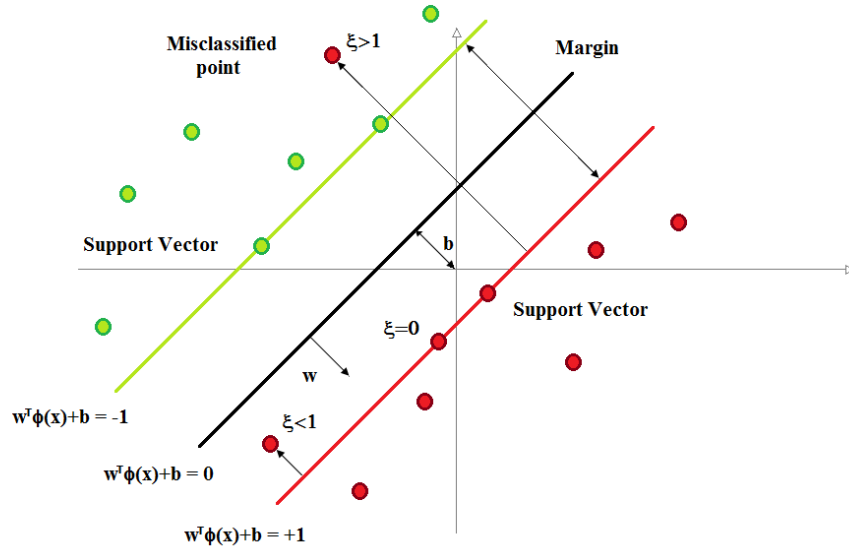


Figure 1: An example of SVM: two linearly separable classes of vectors are represented with red and blue dots. The plane leading to a maximum separation is defined by the weight vector  $\vec{w}$  and the constant term  $b$ .

with the new constraint (derived from Eq. 1):

$$y_i(\vec{x}_i \cdot \vec{w} + b) \geq 1 - \xi_i \quad \text{with} \quad \xi_i \geq 0. \quad (4)$$

The parameter  $C$  defines the SVM implementation before the training, therefore it represents a hyper parameter of the SVM.

For any new vector  $\vec{X}$  considered for classification, we evaluate its position  $D(\vec{X})$  with respect to the plane defined by the support vectors  $\vec{x}_i$  and the parameters  $\alpha_i$ :

$$D(\vec{X}) = \sum_i \alpha_i y_i \vec{x}_i \cdot \vec{X} - b, \quad (5)$$

where  $b$  is a constant term of the solution.

The variable  $D(\vec{X})$  is the final output of the SVM: its sign gives the signal-background classification and, if the geometry is simple, its value is the distance of a test vector  $\vec{X}$  from the classification plane. As we are going to see in the next paragraph, a non-linear classification is possible only thanks to a not-explicit transformation to a different vector space, where  $D$  may lose the immediate geometrical meaning.

Commonly SVMs are used as binary classifiers but, here, we add a large degree of flexibility by exploiting the full information of the continuous variable  $D$ . The SVM is used as a dimensionality reducer of the classification problem: the separation power and the correlations among several significant input variables are summarized into one continuous distribution.

## 2.2. Kernel Methods

Non-linearly-separable classes of vectors can be classified by transforming them with an appropriate function,  $\Phi(\vec{x})$ , that maps the elements into another space, usually with higher dimension, where the linear separation is possible.

However the identification of  $\Phi(\vec{x})$  is not trivial, therefore the so called *Kernel trick* is often used. A Kernel function,  $\mathbf{K}(x_i, x_j)$ , generalizes the scalar product appearing in Eq. 2 (or Eq. 4) without the need of explicitly knowing  $\Phi(\vec{x})$ . The Kernel is the composition of the mapping  $\Phi(\vec{x})$  with the inner product:

$$\mathbf{K}(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \quad \text{with} \quad \Phi : \mathfrak{R}^n \mapsto \mathcal{H}. \quad (6)$$

The function  $\mathbf{K}$  should satisfy to a general set of rules to be a Kernel, but we describe only the *Gaussian* Kernel we used in this work. It is expressed as:

$$K(x_i, x_j) = e^{-\gamma|\vec{x}_i - \vec{x}_j|^2}; \quad (7)$$

The corresponding  $\Phi(x)$  is unknown and it maps the input vectors to an infinite dimension space. The Kernel is defined only by one hyper-parameter,  $\gamma$ , that should be set before the training of the SVM.

## 3. SVM Training on a Biased Sample

Several multivariate techniques are based on the assumption that the labelled samples used for the classifier training are drawn from the same probability distribution of the unclassified events. In our case of study, where only an approximate and statistically limited model of the background processes is available (see Section 4 and in particular Section 4.1 for the multi-jet background description), we do not expect the previous assumption to hold for every portion of the phase space. To cope with this problem, we developed an original methodology to evaluate the SVM training performance.

Section 2 shows that, for each choice of hyper-parameters and training vectors, only one optimal SVM solution exists and we need to evaluate the performance of it.

As a performance estimator we use the *confusion matrix* of the classifier: the element  $(i, j)$  of the matrix is the fraction of the class  $i$  classified as member of class  $j$ . Figure 2 shows a representation of it in the two-classes case, where one class is labelled as *background* and the other as *signal*. We obtain a reliable estimate of the classifier quality by filling the confusion matrix in two independent ways and combining all the available information.

The first performance evaluation method is the *k-fold cross-validation*: the training set is divided into  $k$  sub-samples of which one is used as a validation set and the remaining  $k - 1$  are used in the training; the confusion matrix is then evaluated applying the trained discriminant to the validation set. The cross-validation process is repeated  $k$  times (the *folds*) and the final performance is the average of them. This method is solid against over fitting but it has no protection against biases on the training sample.

<i>Sgn</i> classified as <i>Sgn</i>	<i>Bkg</i> classified as <i>Sgn</i>
<i>Sgn</i> classified as <i>Bkg</i>	<i>Bkg</i> classified as <i>Bkg</i>

Figure 2: Definition of confusion matrix for a two classes (*Sgn* and *Bkg*) classification problem. This reproduces the case of an algorithm used to discriminate signal *vs* background: the elements of the matrix are the signal and background classification performance and the cross contamination.

The second method, a key feature of this work, exploits a signal plus background template fit to extract the off-diagonal terms of the confusion matrix. The fit is performed on a validation sample of unclassified events (i.e. the *data*) using a significant variable that allows some signal to background discrimination. While the signal and background templates are derived directly from the training samples, the unclassified data events are composed by an unknown mixture of true signal and background events.

The fitting routine, implemented in the ROOT [11] analysis package and derived from Ref. [12], maximizes a binned likelihood function,  $\lambda$ , over the significant variable distribution. The fractions of the signal and of the background templates are the free parameters from which we derive the elements of the confusion matrix.

If the variable chosen for the fit is not well reproduced by the simulation then we expect that the fitted fractions are going to largely differ from the results obtained with the  $k$ -fold cross-validation. At the same time we can quantitatively evaluate the agreement between the data shape and the fitted templates as the quantity:

$$\chi^2 = -2 \ln(\lambda) \tag{8}$$

follows a  $\chi^2$  probability distribution (under general assumptions).

The last critical point is the identification of a sensitive variable to be used in the fit. In a previous work [13] we used the distribution of the imbalance of the total transverse momentum of the particles produced in a hadron collision (also referred as missing transverse energy or  $\cancel{E}_T$ ), as it is sensitive to the background we were interested in (i.e. multi-jet contamination). Here we moved to a more general approach, by the machine learning point of view, with the direct usage of the SVM output value  $D$ , defined by Eq. 5. We suppose that, if the SVM training performance evaluation is reliable, the variable  $D$  is highly sensitive to the background and signal composition of the data sample, therefore it can be used in the template fit procedure. To verify the reliability of the training, we evaluate the  $\chi^2$  of the template fit: this ensures a good shape agreement between the data and signal and background templates. We also cross checked the validity of the fit procedure over a toy example discussed in the following.

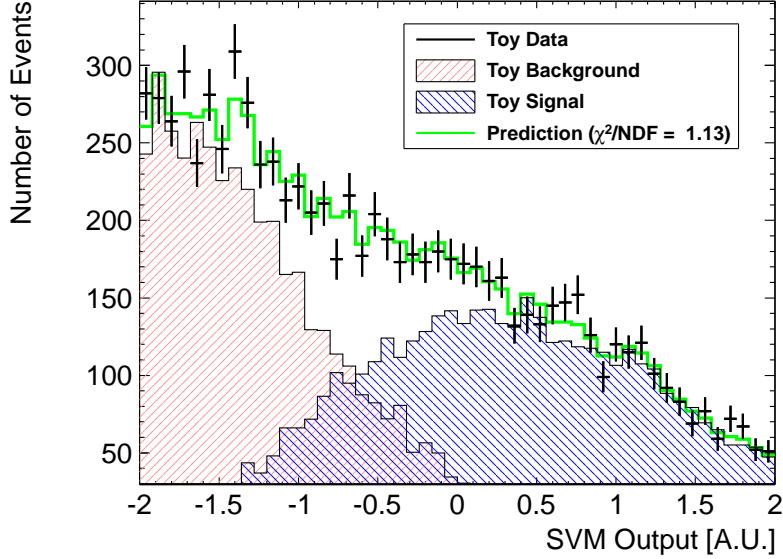


Figure 3: A two-component fit of signal (blue) and background (red) templates is performed for the distribution of the SVM variable  $D$  (Eq. 5), over toy generated data.

### 3.1. A Toy Example

We built a toy example in order to verify the robustness of the proposed method for an SVM performance evaluation when partially biased samples are present. The toy is composed by three data-sets generated with known probability distributions:

**signal model:**  $10^5$  vectors generated from a 2–dim Gaussian distribution with the following mean,  $\vec{\mu}_{Sgn}$ , and standard deviation,  $\tilde{\sigma}_{Sgn}$ :

$$\vec{\mu}_{Sgn} = \begin{bmatrix} -3 \\ 0 \end{bmatrix}, \quad \tilde{\sigma}_{Sgn} = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}. \quad (9)$$

**Background model:**  $10^5$  vectors generated from a 2–dim Gaussian distribution with the following mean,  $\vec{\mu}_{Bkg}$ , and standard deviation,  $\tilde{\sigma}_{Bkg}$ :

$$\vec{\mu}_{Bkg} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}, \quad \tilde{\sigma}_{Bkg} = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}. \quad (10)$$

**Data:** a mixture of  $5 \cdot 10^4$  vectors generated from the same distribution of the signal model (Eq. 9) and  $5 \cdot 10^4$  vectors generated from a background

distribution similar to the background model (Eq. 10) but with  $\hat{\sigma}_{Bkg}$  increased by 20% in one direction to simulate a mismatch between the real background and the model.

We tested several combinations of the hyper-parameters  $C$  and  $\gamma$  using the signal and background model in the training. For each obtained SVM we evaluated the  $k$ -fold cross validation and we performed the template fit on the SVM variable  $D$  evaluated on the data sample. Figure 3 shows an example of the fit.

The result is reported in Figure 4 and, as we know the true label of the toy data vectors, the real performances of the SVM are reported on the  $x$  axis of the diagram. The evaluation of the performance obtained from the fit lies on the diagonal of the plane. It gives a much more realistic estimate of the SVM classifier performance with respect to the direct  $k$ -fold evaluation which may bias the result by a sizable amount also in this simple case.

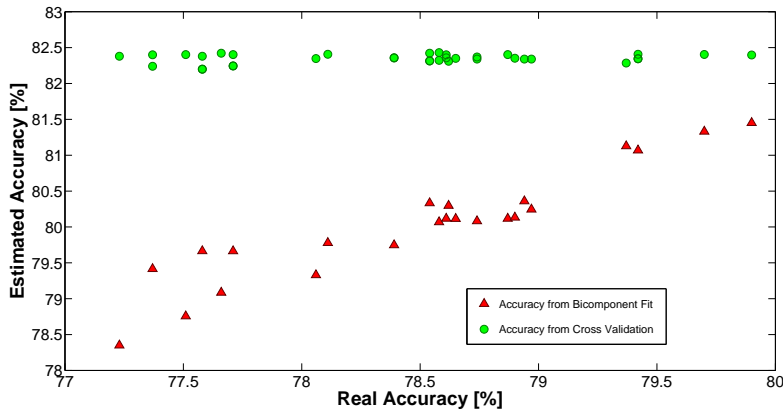


Figure 4: SVM performance estimate with a  $k$ -fold cross-validation (green circles) and with a two-component signal and background template fit on the SVM output variable (red triangle) of toy data of known composition. The true performances of the SVM classifier are reported on the  $x$  axis. The fit evaluation appears on the diagonal of the plane, signaling a realistic estimate of the true performance of the classifier with respect to the direct  $k$ -fold evaluation which biases the result by a sizable percentage.

#### 4. Multi-jet Background Rejection in the $W$ plus Jets Data Sample

The algorithm described in previous sections can be applied to the reduction of the multi-jet background in the  $W$  plus jets channel at hadron collider experiments.

This channel is the basis of many relevant analyses. Events are selected by the identification of the leptonic decay of a  $W$  boson ( $W \rightarrow e\nu$  or  $W \rightarrow \mu\nu$  in

our case) together with one or more jets (i.e. final states of a quark hadronization). Several interesting processes are detectable in such final state but they are characterized by a tiny production cross section ( $\mathcal{O}(1)$  pb) if compared to the total  $p\bar{p}$  inelastic collision cross section ( $\mathcal{O}(1)$  mb) at  $\sqrt{s} = 1.96$  TeV. A few examples: Higgs boson production in association with a  $W$  boson ( $WH$ ), single-top production, diboson production ( $WW, WZ$ ). The selection of a leptonic  $W$  boson decay is the key of the signal identification as it reduces the rate of uninteresting processes by a factor of  $10^5$  with respect to the total inelastic cross section of  $\approx 10^8$  pb, typical of the  $p\bar{p}$  interactions at the,  $\sqrt{s} = 1.96$  TeV, Tevatron center of mass energy.

A mixture of physics processes and detector effects may produce the identification of a false  $W$  boson thus allowing contamination of the selected sample by multi-jet events. In general [14], such background is reduced by a more accurate lepton identification or by rejecting events with kinematic not compatible with a  $W$  boson decay. The remaining contamination is then estimated with data-driven models where some of the lepton selection requirements are inverted to obtain a multi-jet enriched sample. The modeling and the understanding of the sample remain a challenge because of two main reasons: first the multi-jet models are statistically limited to the actual selected data, second, the inversion of some selection criteria may bias the sample. Both these effects pose strong constraints on the applicability of multivariate techniques on the multi-jet rejection.

Our SVM classifier can overcome these difficulties because the optimization and the training take into account a cross check on an unclassified data control sample. The preliminary idea of this algorithm [13, 15] proved to be successful in several analyses performed by the CDF collaboration [7, 17, 18, 37]. In preceding works the SVM was used, following its original concept, as a binary classifier. In this paper we discuss a more powerful and innovative use of the algorithm. We use the continuous distribution  $D$  (Eq. 5) explicitly in the optimization. The improvement is dual: although the training algorithm sets the optimal separation at  $D = 0$ , now the level can be varied to increase the signal efficiency or decrease the background contamination, depending on the physics analysis needs. Furthermore the agreement of the SVM output distribution with data can be used to extract information about the multi-jet modeling and normalization. We heavily relied on this second feature in the variable selection process. As described in section 4.3 and schematized in Figure 5, we trained a new, optimal SVM for each variable configuration, then we tested the performance and the agreement of the SVM using the  $D$  variable and the two-component fit described in Section 3. The process was automatically iterated until stable performance was achieved.

To prove the robustness and the quality of the complete algorithm, we applied it to two datasets, collected by the CDF II experiment and characterized by different kinematic and background contamination. The first contains a high energy electron identified in the central region of the detector and the second contains a high energy electron identified in the forward region of the detector. We chose to train the algorithm only on the electron sample because the multi-

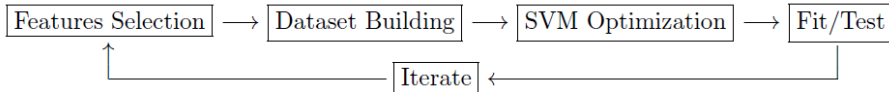


Figure 5: Flowchart of feature selection - training - test procedure.

jet contamination is expected to be larger than in the muon sample. However, as we base the discrimination on kinematic variables, the same algorithm proved to be optimal also for several other lepton categories [16].

#### 4.1. Training Sample Description and Selection Criteria

We built the training sets using electron plus jets events selected in data and Monte Carlo (MC) samples with the standard particle identification algorithms used by the CDF Collaboration [19].

A brief description of the experimental setup is helpful to understand the selection criteria. The CDF II is a general-purpose particle detector placed at one of the two collision points of the Tevatron  $p\bar{p}$ ,  $\sqrt{s} = 1.96$  collider (in operation from 2001 to September 2011). Different information on the data is collected by several subdetectors: a tracking system (silicon detector [20] plus drift chamber [21] in a 1.4 Tesla solenoid), a calorimeter system [22] (composed by an electromagnetic and a hadronic section) and an outer muon identification system (composed by drift chambers and scintillators [23]). The subdetectors have azimuthal and forward-backward symmetry with respect to the geometrical center of the detector corresponding to the nominal collision point. Positions and angles are expressed in a cylindrical coordinate system, with the  $z$  axis along the proton beam, azimuthal angle  $\phi$  and polar angle  $\theta$ . The following variables are defined according to these principles<sup>1</sup>: the pseudorapidity  $\eta = -\ln[\tan(\theta/2)]$ , the transverse energy  $E_T = E \sin \theta$  (as measured by the calorimetry), the transverse momentum  $p_T = p \sin \theta$  (as measured by the tracking systems) and the angular distance between two particles in the  $\eta-\phi$  space,  $R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ .

The identification of central electron candidates ( $|\eta| < 1.1$ ) is based on the following criteria [24]: a good quality track pointing to a significant energy deposit in the electromagnetic section of the central calorimeter ( $E_T > 20$  GeV) and the compatibility of the electromagnetic shower shape and composition (according to five variables) with test-beam data and  $Z \rightarrow e\bar{e}$  events. Forward electron candidates (identified in the calorimeter region with  $1.2 < |\eta| < 2.0$ ) are identified in a similar way but, due to poor tracking chamber coverage, no track matching is required and a different strategy, named Phoenix matching scheme [25], is used to reject fake-electrons.

The  $W \rightarrow e\nu$  identification is completed by requiring the electrons to be isolated from nearby activity, as measured in the electromagnetic calorimeter,

<sup>1</sup>They are relativistic invariant in the case of massless particles.

within a cone of  $R = 0.4$  ( $Iso = E_{R=0.4}/E_{R=0.1} < 0.1$ ) and the presence of an imbalance in the total transverse energy measured by the calorimeter system greater than 15 GeV ( $\cancel{E}_T > 15$  GeV). The isolation requirement derives from the expected kinematic behaviour of the  $W$  decay while the  $\cancel{E}_T$  signals the presence of a neutrino in the event<sup>2</sup>. Events with another lepton candidate are rejected as they introduce  $Z \rightarrow e\bar{e}$  contamination.

The final step of the lepton plus jets selection is the reconstruction of two or more central ( $|\eta| < 2.0$ ) jets using a fixed cone ( $R = 0.4$ ) identification algorithm [26]. The transverse energy of the jets should be greater than 18 GeV, after correcting for detector effects. The per-jet correction, named  $Cor_j$  in the following, is also taken into account in the evaluation of  $\cancel{E}_T$ . Large jet activity and fluctuations in their energy measurement can originate false  $\cancel{E}_T$  thus increasing the probability of  $W$  mis-identification. In the following we indicate with *raw* superscript all the quantities calculated before jet energy correction: i.e.  $\cancel{E}_T^{raw}$  and  $E_{T,Jet}^{raw}$ . An additional requirement of  $\cancel{E}_T^{raw} > 20$  GeV is applied in the forward electron sample to remove a bias caused by the online event selection.

The described selection criteria are used to build the training and test samples according to the following specifications:

**Signal:** both for the central and the forward electron selection, we used a MC-based training set of  $W \rightarrow e\nu$ +jets events. A reliable and simple estimate of the expected signal kinematic and hadronization properties is obtained by combining the  $W \rightarrow e\nu$  plus one and plus two partons MCs where the events are generated by ALPGEN [27] and the hadronic showering is performed by PYTHIA [28]. We used for the training only 7000 events, out of the approximately  $10^5$  generated, while we kept the rest for validation purposes. The small size of the signal training samples is forced by the statistical limits of the background samples and by the need to balance them in the training phase.

**Background (Central):** the multi-jet model specific to the central electrons selection is obtained by enriching the samples with fake electrons. In particular, the electromagnetic shower shape and composition comparison should fail at least two out of the five requirements.

**Background (Forward):** an appropriate multi-jet background training set is obtained, for the forward electron sample, by selecting non-isolated electrons ( $Iso > 0.1$ ).

**Unclassified Data:** as explained in Section 3, we also need a sample of unclassified events for the cross check of the SVM training performance.

---

<sup>2</sup>The total transverse momentum of the particles produced in a hadron collider can be considered exactly zero in the center-of-mass system, therefore an imbalance signals the presence of at least one escaping undetected particle as a neutrino.

To this purpose we used about one fifth of the CDF data ( $\mathcal{L} \approx 2 \text{ fb}^{-1}$ ) corresponding to average run conditions and luminosity profile.

The different prescriptions used for the multi-jet background models were developed in several CDF analyses performed in  $W$  plus jets data sample (two relevant examples are in Ref. [19] and [14]) and they are essentially obtained by inverting one or more selection criteria with respect to the standard  $W \rightarrow e\nu$  selection.

Before the training, the behaviour of the background models has been studied and improved in few aspects. The key point of the study is the similarity of the electron candidates, used in the multi-jet models, with the standard jet candidates. The central background model was calibrated on a data region dominated by multi-jet events, obtained with full event selection except for a reversed  $\cancel{E}_T$  selection ( $\cancel{E}_T < 15 \text{ GeV}$ ) to suppress the presence of real  $W$  bosons. The kinematic properties of the multi-jet sample have been improved by correcting the  $\cancel{E}_T$  with a function of the candidate electron  $E_T$  (for more details see also Ref. [16]). In the case of the forward electron sample, we found that the simple  $E_T$  spectrum of the non-isolated lepton was too soft. An improved multi-jet model was obtained by adding to the  $E_T$  of the candidate electron the energy of the outer isolation cone ( $E_T^{0.1 < \Delta R < 0.4}$ ).

#### 4.2. Input Variable Description

A multivariate algorithm relies on a given set of input variables. The *feature selection* problem is fundamental in machine-learning and, if possible, even more in the present case where the background sample does not guarantee a good model of all the variables.

We started with a large set of variables (twenty-four) identified according to two basic criteria: no evident correlation with the lepton identification variables and the use of the kinematic difference between the simulated  $W + \text{jets}$  events and the multi-jet background model. At the end of the optimization, these requirements allowed the use of the multi-jet rejection procedure on several different lepton identification algorithms.

All the variables are listed in Table 1. They involve the kinematic properties of the lepton, the leading and second leading jet and the  $\cancel{E}_T$  module and direction. In the following we describe the few more complex variables entering in the set:

- $\cancel{p}_T$  is the missing momentum defined as the momentum imbalance on the transverse plane. It is computed adding all the reconstructed charged tracks transverse momenta,  $\vec{p}_i$ :

$$\vec{\cancel{p}}_T \equiv - \sum_i \vec{p}_i^T \quad \text{with } |\vec{p}_i^T| > 0.5 \text{ GeV}/c; \quad (11)$$

- $M_T^W$  is the *transverse mass* of the reconstructed  $W$  boson:

$$M_T^W = \sqrt{2(E_T^{lep} \cancel{E}_T - E_x^{lep} \cancel{E}_x - E_y^{lep} \cancel{E}_y)}. \quad (12)$$

Possible Input Variables							
1	$p_T^{lep}$	7	$E_T^{raw,jet1}$	13	$\Delta\phi(\cancel{p}_T, lep)$	19	$\Delta R(lep, jet2)$
2	$\cancel{E}_T$	8	$E_T^{raw,jet2}$	14	$\Delta\phi(\cancel{p}_T, \cancel{E}_T)$	20	$\Delta R(\nu^{min}, jet1)$
3	$\cancel{E}_T^{raw}$	9	$E_T^{cor,jet1}$	15	$\Delta\phi(\cancel{p}_T, \cancel{E}_T^{raw})$	21	$\Delta R(\nu^{min}, jet2)$
4	$\cancel{p}_T$	10	$E_T^{cor,jet2}$	16	$\Delta\phi(lep, \cancel{E}_T)$	22	$\Delta R(\nu^{min}, lep)$
5	$M_T^W$	11	$\Delta\phi(jet1, \cancel{E}_T)$	17	$\Delta\phi(lep, \cancel{E}_T^{raw})$	23	$\Delta R(\nu^{max}, jet1)$
6	$MetSig$	12	$\Delta\phi(jet2, \cancel{E}_T)$	18	$\Delta R(lep, jet1)$	24	$\Delta R(\nu^{max}, jet1)$

Table 1: All the possible input variables used for the SVM training and optimization. See also Sections 4.1 and 4.2 for a detailed description.

- $MetSig$  is the  $\cancel{E}_T$  significance, a variable that relates the reconstructed  $\cancel{E}_T$  with the detector activity (jets and unclustered energy):

$$MetSig = \frac{\cancel{E}_T}{\sqrt{\Delta E^{jets} + \Delta E^{uncl}}}, \quad (13)$$

where:

$$\Delta E^{jets} = \sum_j^{jets} \left( Cor_j^2 \cos^2 \left( \Delta\phi(\vec{p}_j, \cancel{E}_T) \right) E_T^{raw,j} \right), \quad (14)$$

$$\Delta E^{uncl} = \cos^2 \left( \Delta\phi(\vec{E}_T^{uncl}, \cancel{E}_T) \right) E_T^{uncl}, \quad (15)$$

$uncl$  refers to the calorimeter energy not clustered into electrons or jets and  $Cor_j$  is the total correction applied to each jet.

- $\nu^{Min}$ ,  $\nu^{Max}$  are the two possible reconstructions of the neutrino momenta. As the  $p_z^\nu$  component is not directly measurable we infer it from the  $W$  boson mass and the lepton momentum. The constraints lead to a quadratic equation which may have two real solutions, one real solution, or two complex solutions<sup>3</sup>. The reconstructed  $\nu^{Min}$ ,  $\nu^{Max}$  derive from the distinction of  $p_z^{\nu,Max}$  and  $p_z^{\nu,Min}$ .

#### 4.3. Optimization and Variable Selection

There is a wide literature about variable selection (see for example [31] for a review of some methods). In Ref. [29] the variable set is optimized during the training while, in Ref. [38], multiplicative weights, ranging from 0 to 1 and tuned between successive training cycles, are associated to each input variable. In our case, as we use a figure of merit based on a statistical test and performed on the

<sup>3</sup>The real part is chosen in this case.

trained classifier, we were not able to use a feature selection system integrated into the training. We solved the problem by exploring the space of possible variables: we started from a reduced subset of all the possible inputs and then we increase it adding more variables to the most performing sets, as also done in Ref. [? ].

Unluckily the brute-force search over all the possible combinations of variables across all the  $C, \gamma$  phase space of a given SVM training is computationally unfeasible. Just the search over all the possible combinations of twenty-four variables requires a total of 16777215 different SVM configurations.

To scan the most relevant sectors of the phase space we applied a factorized and incremental optimization:

- for all the configurations of three variables, we evaluate a grid of  $C, \gamma$  values in the intervals:

$$\log_2 C \in [-2, 9] \quad \text{and} \quad \log_2 \gamma \in [-4, 7], \quad (16)$$

where the use of a logarithmic scale allows to scan the parameters across several orders of magnitude. For each variable configuration we select only the best SVM training configuration, according to the result of the confusion matrix.

- Then, for each *best* SVM and variable configuration, we perform a two-component template fit on the SVM variable  $D$ . We evaluate the  $\chi^2$  of the fit, reduced by the Number of Degrees of Freedom ( $NDoF$ ), and we compare the fitted background contamination,  $f_{Bkg}^{Fit}$ , against the one obtained from the  $n$ -fold cross-validation,  $f_{Bkg}^{n-fold}$ . The SVM under exam is rejected if:

$$\frac{\chi^2}{NDoF} > 3 \quad \text{or} \quad \frac{f_{Bkg}^{Fit}}{f_{Bkg}^{n-fold}} > 2. \quad (17)$$

Notice that the quality of the fit is not directly optimized by the SVM training, so we are performing a consistency check of our classifier in an unbiased sample (data) with an independent technique.

- After the first iteration of the template fit cross check, we display all the remaining SVMs on a *signal-efficiency vs background-contamination* scatter plot like the one in Figure 6. The five best variable combinations are selected for further processing and then again, we add all the possible combinations of 1, 2 and 3 variables to them for a second iteration.

After few iterations the *best result* does not improve significantly. In figure 7 the performances are shown as a function of three significant quantities: the distance from the ideal performance (i.e. 100% signal efficiency, 0% background contamination) as returned by the template fit, the fraction of mis-classified training vectors as returned by the  $k$ -fold cross validation, and the total background contamination obtained by the two-component fit (this is a helpful quantity in an actual physics analysis). Notice how all the curves of performance flatten within  $1 \div 2\%$  when the number of variables approaches eight.

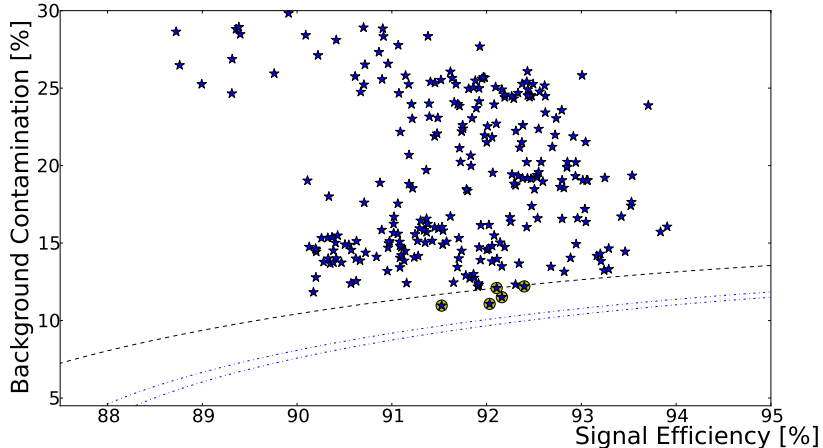


Figure 6: Performances of different SVM configurations, obtained by the combination of tree (out of twenty-four) input variables in the central region training, are displayed as blue stars on a *signal-efficiency vs background-contamination* scatter plot. The signal efficiency is estimated from simulation while the background contamination is obtained from the two-component template fit of the SVM distance,  $D$ , described in Section 3. The five configurations closest to  $\epsilon_{Sig} = 1$ ,  $f_{Bkg} = 0$  (the best possible result) are circled in yellow and selected for further iterations of the training with an increased number of input variables. We performed three iterations of the training, with three, six and eight input variables and the three dotted lines represent the distance, from  $\epsilon_{Sig} = 1$ ,  $f_{Bkg} = 0$ , enclosing the five best configurations in each case. After the third iteration no more appreciable improvement occurs.

The total number of SVM optimizations to be performed is the product of the number of explored  $C$ , the number of explored  $\gamma$ , and the number of combinations of variables. This equals approximately to 300000, in the case of the three training steps explained before. The SVM optimizations have been divided in equal numbers between 121 CPUs and performed on a distributed analysis grid system [32, 33]. For the first training step (3 variables) the computation took 6 hours, 12 minutes and 22 seconds.

#### 4.4. Final SVM Configuration

The two optimal SVMs obtained from the central (superscript  $c$ ) and forward (superscript  $f$ ) electron training sets are defined by the  $C$  and  $\gamma$  hyper-parameters and by a combination of input variables. In particular the values of the hyper-parameters are:

$$C^c = 7, \gamma^c = -1 \quad \text{and} \quad C^f = 8, \gamma^f = -1. \quad (18)$$

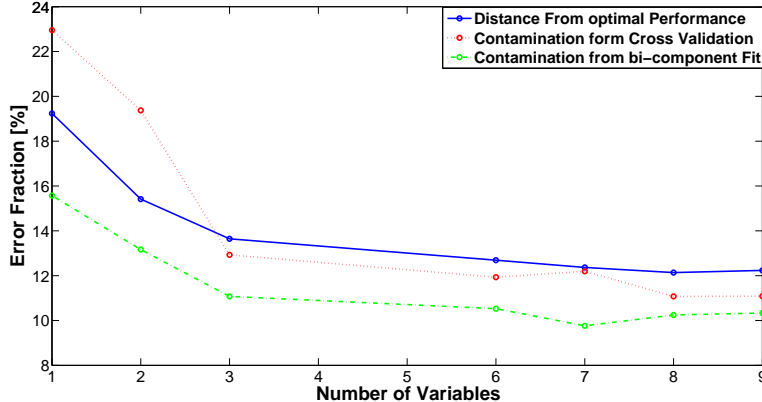


Figure 7: Improvement of performance increasing the number of variables as a function of three significant quantities: the distance from the ideal performance (i.e. 100% signal efficiency, 0% background contamination) as returned by the template fit, the fraction of mis-classified background vectors as returned by the  $k$ -fold cross validation and the total background contamination obtained again by the two-component fit (this is a helpful quantity in an actual physics analysis). Notice how all the curves of performance flatten within  $1 \div 2\%$  when the number of variables approaches eight.

The specific input variables giving the best performances are reported in Table 2, eight for central-electron selection and six for the forward electron selection.

Although the two training sets are selected independently and they present different kinematics, a certain degree of similarity arises in the final configurations as five of the final input variables are in common between the two SVMs. One of them, the  $M_T^W$  is closely related to the kinematic of the  $W$  decay, others, like the  $\cancel{E}_T^{raw}$  and the  $\cancel{p}_T$ , give independent information about the reconstructed neutrino. With the  $MetSig$  and the  $\Delta\phi(\cancel{p}_T, \cancel{E}_T)$ , the angular information between the reconstructed quantities is exploited.

As none of the input variables relies on the specific electron identification algorithm definition, the same SVM classifier can be applied in channels where the  $W$  decay is selected by different lepton reconstruction criteria. An example of this usage is reported in Ref. [16].

Another advantage of the approach presented in this paper is the possibility to exploit the full shape information of the SVM output variable  $D$ . Figures 8 and 9 show the shape of the central and forward SVM discriminants for the multi-jet background models and the  $W$ +jets signal. A lower background contamination can be achieved increasing the SVM selection threshold from the zero value. Table 3 reports the evaluation of the performance as given by the two elements of the confusion matrix useful in a multi-jet background rejection problem: the signal efficiency (obtained from the training set) and the background

Final SVM Input Variables			
Central SVM:	$M_T^W$	$\cancel{E}_T^{raw}$	$\cancel{p}_T$
	$MetSig$	$\Delta\phi(\cancel{p}_T, \cancel{E}_T)$	$\Delta\phi(lep, \cancel{E}_T)$
	$\Delta R(\nu^{Min}, lep)$	$\Delta\phi(jet1, \cancel{E}_T)$	
Forward SVM:	$M_T^W$	$\cancel{E}_T^{raw}$	$\cancel{p}_T$
	$MetSig$	$\Delta\phi(\cancel{p}_T, \cancel{E}_T)$	$\Delta\phi(\cancel{p}_T, \cancel{E}_T^{raw})$

Table 2: Final input variables used for the configuration of the *central* and *forward* SVM multi-jet discriminants.

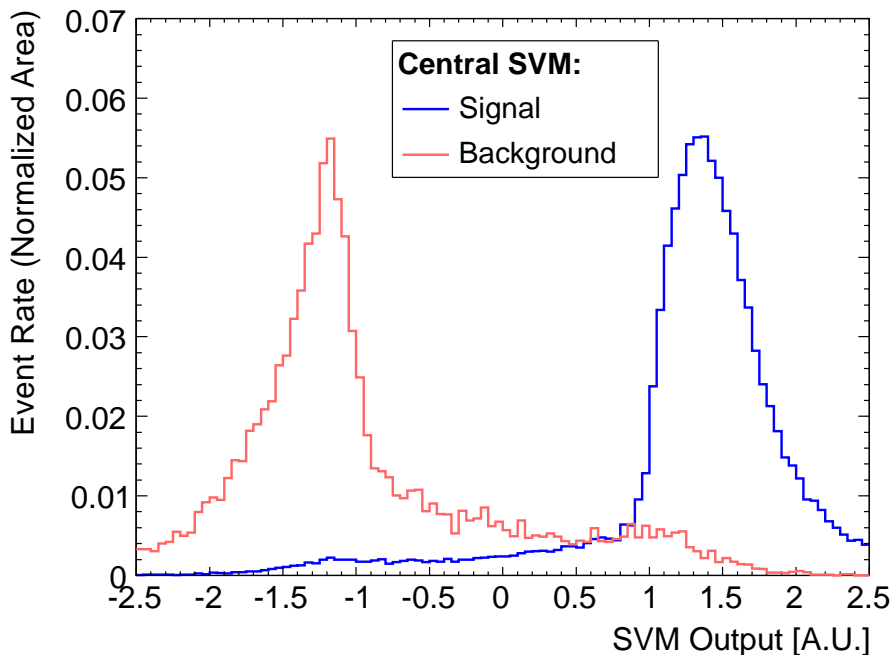


Figure 8: Distribution of the SVM variable  $D$ , described in Eq. 5, for the central SVM discriminant obtained from the optimization process. Multi-jet background model is shown in red,  $W$ +jets MC signal is shown in blue.

contamination (obtained from the template fit procedure) both evaluated for a SVM selection threshold of  $D = 0$ .

The performance of our multi-jet rejection algorithm can also be compared to other methods used in similar contexts. The  $WH \rightarrow e\nu$  plus two jets associate production and decay (with a Higgs boson mass of  $115 \text{ GeV}/c^2$ ) can be used as a common signal reference process to evaluate the effective per-

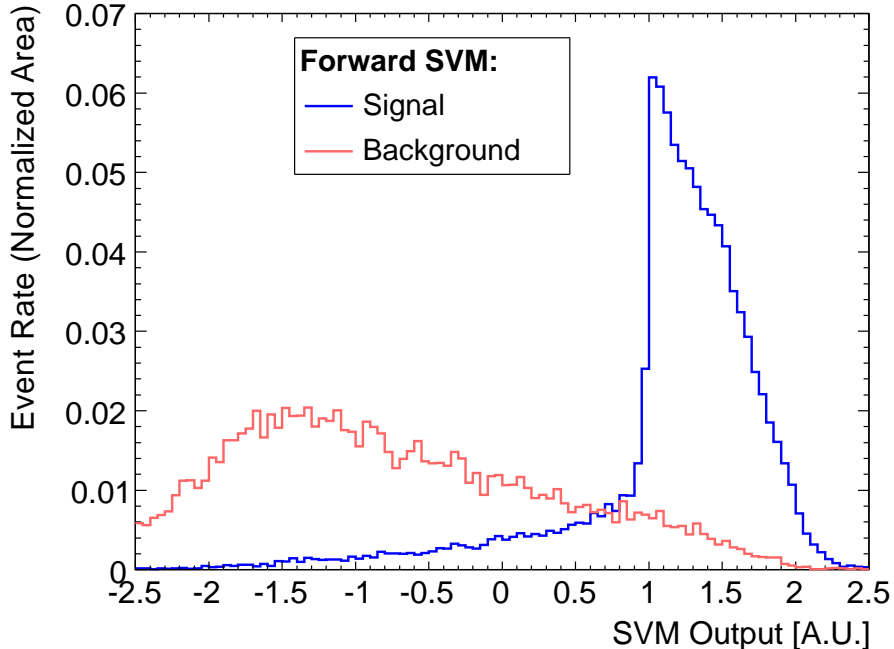


Figure 9: Distribution of the SVM variable  $D$ , described in Eq. 5 for the forward SVM discriminants obtained from the optimization process. Multi-jet background model is shown in red,  $W$ +jets MC signal is shown in blue.

formance of the algorithms. A first direct comparison is possible within the CDF collaboration against a previous SVM based method [13] used in several analyses [7, 17, 18, 37]. The new multi-jet rejection strategy improves the  $WH$  signal acceptance by 5% allowing the same multi-jet background contamination. Then it is possible to compare results presented in this paper to the multivariate strategy applied by the D0 collaboration in the same decay channel. Ref. [34] describes a multi-jet rejection power of 75% for a very loose operating point of the algorithm which allows a  $WH$  signal efficiency of 97%. For the same signal efficiency, the approach described in this paper rejects approximately 80% of the multi-jet background. The last comparison is done against the recent results of the Atlas [35] and CMS [36] collaborations. Although the reported multi-jet background contamination is comparable, a lower signal efficiency is expected from the use of tight selection criteria: for example the presence of large  $\cancel{E}_T$  or large  $M_T^W$  in the event.

It is possible to conclude that, at today, the presented result is superior to any other multi-jet rejection strategy applied for the same final state (i.e.  $W \rightarrow e\nu$  plus two jets) at hadron collider experiments.

	Signal Efficiency	Background Contamination
Central SVM:	$0.935 \pm 0.006$	$0.141 \pm 0.005$
Forward SVM:	$0.908 \pm 0.009$	$0.25 \pm 0.01$

Table 3: Performance of the *central* and *forward* SVMs final configurations evaluated for a selection threshold of  $D = 0$ . The signal efficiency is derived the MC selection while the background contamination is derived from the template fit procedure. Errors are statistical only.

#### 4.5. Example of Use for Background Estimate

An example of the use of the developed SVM discriminant is reported in Ref. [16]. In particular both the possibility to move the SVM selection threshold and the fit of the multi-jet contamination over SVM output variable  $D$  are used.

Figure 10 shows how the different physics processes contribute to the shape of the SVM discriminant used in the selection of the forward electron sample. The multi-jet, or QCD, fraction is extracted from the fit together with the total  $W$  plus jets component. The SVM threshold used for the final signal region identification is  $D = 1$  instead of the standard value of  $D = 0$ . This further decreases the multi-jet contamination and reduces the impact of the systematics related to the estimate of such background.

## 5. Conclusion

In this paper we present an innovative method for training optimization and variable selection for a SVM multivariate classifier in the problematic case of biased and statistically limited training samples.

The optimization of the classifier is possible thanks to the feedback of a signal-background template fit performed on a validation sample of unclassified events over the SVM output distribution  $D$  (defined in Eq. 5). By construction, the SVM output  $D$  is a variable sensitive to the signal and background contamination.

The optimization algorithm was then applied to an actual hadron collider physics case. The problem of multi-jet rejection in the  $W$  plus jets channel has been analyzed in two different cases: a  $W$  boson decaying to an electron identified in the central and in the forward region of the CDF detector and selected together with two or more hadronic jets.

The optimization algorithm allowed to scan a pool of twenty-four input variables to obtain the minimal combination of them giving the lower background contamination and the higher signal efficiency. The resulting classifiers exploit eight and six input variables combinations, respectively for the *central* and *forward* SVM classifiers, not directly related to the lepton identification algorithm. The performance is superior to any present multi-jet rejection algorithm applied to the same final state.

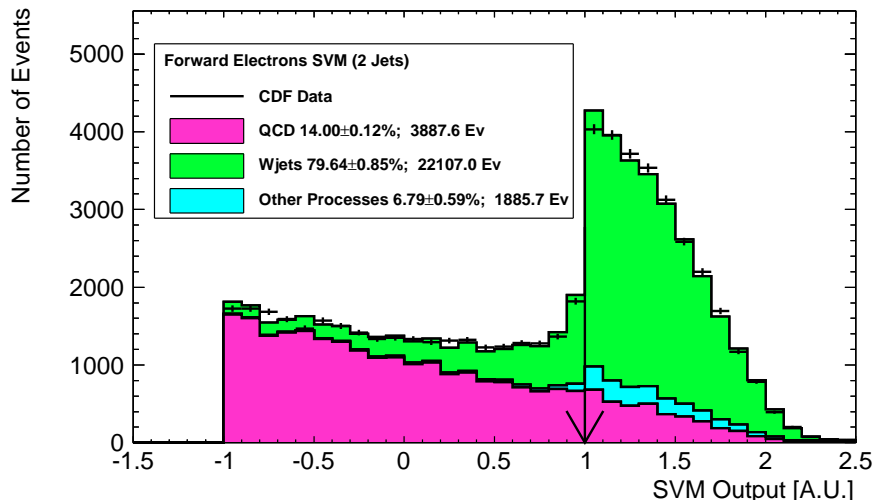


Figure 10: Contribution of the different physics processes to the shape of the SVM output distribution  $D$  used during the *forward* electron sample selection. The multi-jet, or QCD, background fraction (in magenta) is extracted from the fit together with the total  $W$  plus jets component (in green). The remaining physics processes (in light blue) are normalized to the expected production cross sections and acceptances. The SVM selection threshold for the final signal region identification is  $D = 1$  instead that the standard value of  $D = 0$ . This further decreases the QCD contamination in the final analysis sample.

The CDF II dataset and the specific multi-jet rejection analysis have been a perfect test-bench for the SVM optimization algorithm. However the procedure can be exported to any other problem where the use of multivariate techniques is required but no accurate simulation is available or possible.

## Acknowledgments

We would like to thank the University Research Association for the support given to F. Sforza with the Visiting Scholars Program at FNAL of Spring 2010, Prof. G. Chiarelli for the support, the guidance and the discussions that contributed to this paper, and Dr. H. Wolfe for the review of the manuscript.

## References

- [1] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006.

- [2] L. Ametller, L. Garrido, G. Stimpfl-Abele, P. Talavera, and P. Yepes, Discriminating signal from background using neural networks: Application to top-quark search at the Fermilab Tevatron, *Phys. Rev. D* 54, 1233 (1996).
- [3] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*, The MIT Press, 2001.
- [4] A. Vaiciulis, Support vector machines in analysis of top quark production, *Nucl. Instrum. Methods Phys. Res., Sect. A* 502 (2003) 492.
- [5] P. C. Bhat, Run II physics at the Fermilab Tevatron and advanced analysis methods, *Nucl. Instrum. Methods Phys. Res., Sect. A* 502 (2003) 327.
- [6] J. De Sanctis, M. Masotti and A. Bonasera, Backtracing of the Impact Parameter through Pattern Recognition Analysis on Heavy Ion Reaction Data, *AIP Conference Proceedings* 899 (2007) 105.
- [7] T. Aaltonen *et al.* (CDF Collaboration), Search for the standard model Higgs boson produced in association with a  $W^\pm$  boson with  $7.5 \text{ fb}^{-1}$  integrated luminosity at CDF, *Phys. Rev. D* 86, 032011 (2012).
- [8] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, Feature selection for SVMs, in *Advances in Neural Information Processing Systems 13*, The MIT press, 2000.
- [9] Chen, Y.-W. and Lin, C.-J. Guyon, I.; Nikravesh, M.; Gunn, S. and Zadeh, L. (Eds.), *Combining SVMs with Various Feature Selection Strategies, Feature Extraction*, Springer Berlin Heidelberg, 2006, 207, 315.
- [10] C. C. Chang and C. J. Lin, LIBSVM: a library for support vector machines, <http://www.csie.ntu.edu.tw/~%20cjlin/libsvm>, 2001.
- [11] R. Brun, ROOT - An object oriented data analysis framework, *Nucl. Instrum. Methods Phys. Res., Sect. A* 389 (1997) 81.
- [12] R. Barlow, Fitting using finite Monte Carlo samples, *Computer Physics Communications* 77 (1993) 219.
- [13] F. Sforza *et al.*, Rejection of Multi-jete Background in  $p\bar{p} \rightarrow e\nu + j\bar{j}$  Channel through a SVM Classifier, *J. Phys. Conf. Ser.* 331 (2011) 032045.
- [14] T. Aaltonen *et al.* (CDF Collaboration), Observation of single top quark production and measurement of  $|V_{tb}|$  with CDF, *Phys. Rev. D* 82 (2010) 112005.
- [15] F. Sforza, V. Lippi, G. Chiarelli, and S. Leone, Rejection of multi-jet background in a hadron collider environment through a SVM classifier, in: *Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, 2011 IEEE, p1404.

- [16] F. Sforza, Evidence for Diboson Production in the Lepton plus Heavy Flavor Jets Final State at CDF, Ph. D. Thesis, University of Pisa (2012), FERMILAB-THESIS-2013-01
- [17] T. Aaltonen *et al.* (CDF Collaboration), Search for the standard model Higgs boson decaying to a  $b\bar{b}$  pair in events with one charged lepton and large missing transverse energy using the full CDF data set, Phys. Rev. Lett. 109 (2012) 111804.
- [18] F. Sforza (for the CDF Collaboration), Evidence for  $WZ/WW \rightarrow \ell\nu +$  Heavy Flavors Vector Boson Production in  $7.5 \text{ fb}^{-1}$  of CDF Data, Proceedings of the DPF-2011 Conference, Providence, August 2011.
- [19] D. Acosta *et al.* (CDF Collaboration), Measurement of the  $t\bar{t}$  Production Cross Section in  $p\bar{p}$  Collisions at  $\sqrt{s} = 1.96 \text{ TeV}$  Using Lepton + Jets Events with Secondary Vertex b-Tagging, Phys. Rev. D 71, (2005) 052003.
- [20] K. A. Sill *et al.* CDF Run II silicon tracking projects, Nucl. Instrum. Methods A 446 (2000) 1.
- [21] A. Affolder *et al.* CDF Central Outer Tracker, Nucl. Instrum. Methods A 526 (2004) 249.
- [22] L. Balka *et al.*, The CDF central electromagnetic calorimeter, Nucl. Instrum. Methods A 267 (1988) 272.
- [23] C. M. Ginsburg *et al.* CDF Run 2 Muon System, Eur. Phys. J. 33 (2004) 1002.
- [24] A. Abulencia *et al.* (CDF Collaboration), Measurements of Inclusive W and Z Cross Sections in  $p\bar{p}$  Collisions at  $\sqrt{s} = 1.96 \text{ TeV}$ , J. Phys. G 34 (2007) 2457.
- [25] C. Issever, W Charge Asymmetry Measurement in CDF Run 2, AIP Conf. Proc. 670 (2003) 371.
- [26] A. Bhatti *et al.*, Determination of the jet energy scale at the Collider Detector at Fermilab, Nucl. Instr. Meth. A 566 (2006) 375.
- [27] M. L. Mangano, M. Moretti, F. Piccinini, Roberto Pittau, and Antonio D. Polosa, ALPGEN, a generator for hard multiparton processes in hadronic collisions, J. High Energy Phys. 07 (2003) 1.
- [28] T. Sjostrand *et al.*, High-Energy-Physics Event Generation with PYTHIA 6.1, Comp. Phys. Commun. 135 (2001) 238.
- [29] M. H. Nguyen and F. De la Torre, Optimal Feature Selection for Support Vector Machines, Pattern Recognition 43 (2010) 584.
- [30] S. Avidan, Joint Feature-Basis Subset Selection, Proc. IEEE Conf. Computer Vision and Pattern Recognition, (2004) 283.

- [31] Y.-W. Chen and C.-J. Lin, Combining SVMs with Various Feature Selection Strategies, *Feature Extraction*, 207 (2006) 315.
- [32] I. Sfiligoi, CDF computing, *Computer Physics Communications* 177 (2007) 235.
- [33] V. Bartsch *et al.*, Testing the CDF Distributed Computing Framework, *Proceedings of the CHEP/22204 Conference*, Interlaken, Switzerland (2004).
- [34] D0 Collaboration, Search for  $WH$  associated production with  $8.5 \text{ fb}^{-1}$  of Tevatron data, *D0 Conference Note 6220* (2011).
- [35] ATLAS Collaboration, Search for the Standard Model Higgs boson in produced in association with a vector boson and decaying to bottom quarks with the ATLAS detector, *ATLAS CONF Note 161* (2012).
- [36] CMS Collaboration, Search for the standard model Higgs boson produced in association with W or Z bosons, and decaying to bottom quarks for HCP 2012, *CMS PAS HIG 12-044* (2012).
- [37] T. Aaltonen et al. (CDF Collaboration), Search for Resonant Top-antitop Production in the Semi-leptonic Decay Mode Using the Full CDF Data Set, submitted to *Phys. Rev. Lett.* (2012).
- [38] D. Mladenić et al. Feature selection using linear classifier weights: interaction with classification models, *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (2004) 234.