# Principled and Data Efficient Support Vector Machine Training Using the Minimum Description Length Principle, with Application in Breast Cancer Prediction

**Harsh Singh**[1], **Ognjen Arandjelović**[2]
[1]International Institute of Information Technology
Naya Raipur, India
harshs17101@iiitnr.edu.in

[2] University of St Andrews
St Andrews, United Kingdom
ognjen.arandjelovic@gmail.com

## Abstract

Support vector machines (SVMs) are established as highly successful classifiers in a broad range of applications, including numerous medical ones. Nevertheless, their current employment is restricted by a limitation in the manner in which they are trained, most often the training-validation-test or $k$-fold cross-validation approaches, which are wasteful both in terms of the use of the available data as well as computational resources. This is a particularly important consideration in many medical problems, in which data availability is low (be it because of the inherent difficulty in obtaining sufficient data, or because of practical reasons, e.g. pertaining to privacy and data sharing). In this paper we propose a novel approach to training SVMs which does not suffer from the aforementioned limitation, which is at the same time much more rigorous in nature, being built upon solid information theoretic grounds. Specifically, we show how the training process, that is the process of hyperparameter inference, can be formulated as a search for the optimal model under the minimum description length (MDL) criterion, allowing for theory rather than empiricism driven selection and removing the need for validation data. The effectiveness and superiority of our approach are demonstrated on the Wisconsin Diagnostic Breast Cancer Data Set.

## Introduction

Support vector machines (SVMs) are supervised machine learning models which have been used extensively in medical and biomedical applications (Yue, Dimitriou, and Arandjelovic 2019; Caie, Dimitriou, and Arandjelovic 2020; Gavriel et al. 2021). This popularity of SVMs stems, first and foremost, from their often highly competitive performance, but also their mathematically well-understood behaviour and explainability, contrasting many types of neural networks. The major methodological limitation associated with SVMs concerns the setting of their hyperparameters. All research to date employs one of the following approaches as regards the setting of the hyperparameter values: (i) they are set to default values which are sensible in the absence of any domain specific knowledge, or (ii) they are adopted from previous successful work on similar problems, or (iii) they are learnt using the standard training-validation-test protocol (Dimitriou et al. 2018) so as to ensure model specificity while preventing overfitting (Dimitriou et al. 2018). All of these approaches are rather unsatisfactory. Specifically, the first two are failing to make any use of training data and adapt to the particular problem at hand. The last, most principled one, is unattractive in that it is purely empirical (though this may seem to be practically unavoidable in some instances, herein we show that it is not) and wasteful both in terms of data, as a validation data set, separate from the training and test data sets, has to be set aside. This is particularly problematic in applications where data is scarce, which is often the case in medical applications (Barracliffe, Arandjelovic, and Humphris 2017).

In this paper we propose a novel approach of training support vector machines and setting their hyperparameter values which avoids the aforementioned problems. In particular, our idea is to make use of the minimum description length (MDL) principle which can be seen as a formalization of Occam's razor. MDL allows for the quality of fit to be assessed and balanced against the complexity of a model (in our case the number of support vectors of a trained model, given the values of hyperparameters) on principled, theoretical grounds, i.e. without the need for follow-up empirical assessment. Thus, using a publicly available Breast Cancer Wisconsin (Diagnostic) Data Set (Wolberg, Street, and Mangasarian 1992), we show how MDL can be employed in the training of SVMs and demonstrate that the outcome is not only more methodologically appealing and principled, but due to its more efficient use of data that it also achieves superior results when compared with the traditional training-validation-test approach.

## Existing SVMs model selection approaches

Selecting the appropriate model from a large pool of models trained with different hyperparameters values is a difficult task. Yet, it is important to get a sense of the model's reliability and generalization ability before it is applied in practice. A naïve selection approach can result in the adoption of models which exhibit overfitting or indeed underfitting (Shalev-Shwartz and Ben-David 2014). Unsurprisingly, there has been plenty of previous work (Cawley and Talbot 2010; Heckerman and Meek 1997; Raschka 2018; Shalev-Shwartz and Ben-David 2014) in this realm and the techniques proposed vary somewhat based on the type of problem considered, that is on whether one is dealing with classification or regression.

In the context of regression, the Hold Out technique (Blum, Kalai, and Langford 1999) estimates the empirical error of a model using unseen data, and uses the estimate to make the best model choice. In contrast, the Model Selection Curve (Murata, Yoshizawa, and Amari 1993) method involves drawing predicted points of trained models on the training and validation set, and the model which exhibits the best consistency between the two sets is selected as the best one.

In the context of classification, the training-validation-test approach is probably the most widely used one (Galvao et al. 2005). The available data is split into three subsets – namely training, validation, and test, with the size of the first of these usually being much larger than that of the other two. Models with different hyperparameter values (or indeed models using different learning algorithms altogether) are trained on the training data set. The trained model is then queried to predict the output of all data in the validation set. The best model is selected based on the trained models' performances on the validation set, whereas the performance of the said selected model is finally assessed on the test data set (thus ensuring a lack of bias).

The $k$-fold validation approach can be seen as a modified version of training-validation-test. It involves firstly the splitting of the available data into $k$ subsets, referred to as *folds* in this context (Kohavi et al. 1995). In each iteration, one particular fold is selected and kept aside (withheld) while the model is trained on the union of all others. The performance of the trained model is assessed on the withheld fold. The process is iterated with a different fold being withheld in each iteration. The final performance metric is the average of the performance metrics obtained at each fold. The process is applied to different hyperparameter values or learning algorithms. The model with the best mean performance metric is selected as the best fit model. As before, the performance of the model is ultimately assessed on the test data set.

## Proposed method

As we briefly noted earlier, there are two major drawbacks to the existing model selection methods. The first of these lies in the inefficient use of data. In particular, since validation and test sets need to be entirely disjoint from the training data set (and of course with one another), less data is available to actually train the model, and the lesser amount of training data translates to less well trained models. This is particularly important in many medical applications wherein data scarcity poses a significant challenge. This scarcity may be inherent, e.g. because a specific condition of interest is rare, or it may be of a practical nature, e.g. because relevant data cannot be easily accessed or shared due to privacy concerns or regulations.

Our approach surmounts the drawbacks of these limitations by preserving the validation/test sets and using them for training. In particular, we are interested in the MDL (Grünwald and Grunwald 2007) criterion, which can be thought of as a formalization of Occam's razor (Blumer et al. 1987), allowing one to make a principled compromise between the complexity of a model and the explanatory power of the model in the context of training data. This means that no validation data is needed, which can instead be used for training, thus improving the model.

### Description length and the MDL, and SVMs

In the context of the present work the description length is the length in bits needed to encode the parameters of a model and the data given the model. Formally:

$$DL(M, D) = L(M) + L(D|M) \tag{1}$$

where $DL(M, D)$ is the description length corresponding to the model $M$ and data $D$. For a statistical model defined as a parametrized family of probability distributions, as in the case of SVMs, the description length can be further written as:

$$DL(M, D) = \frac{1}{2} N_M \log_2 N - \sum_{i=0}^{N-1} \log_2 P(d_i) \tag{2}$$

where $d_i$ are individual data points from $D$, $N$ their count, and $N_M$ the number of free model parameters.

**Support vector based learning** Support vector machines perform classification through linear separation in a high (possibly infinitely) dimensional space into which the original input data is mapped (Schölkopf, Smola, and Müller 1998). Importantly, the seemingly intractable task of mapping data into the high dimensional space is achieved efficiently by performing the aforesaid mapping implicitly rather than explicitly. This is done by employing the so-called 'kernel trick' which ensures that dot products in the high dimensional space can be readily computed using the variables in the original space. Given labelled training data (input vectors and the associated labels) in the form $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, a support vector machine aims to find a mapping which minimizes the number of misclassified training instances, in a regularized fashion. The mapping $x \rightarrow \Phi(x)$ is performed implicitly by employing a Mercer-admissible kernel (Schölkopf, Smola, and Müller 1998) $k(x_i, x_j)$ which allows for the dot products between mapped data to be computed in the input space: $\Phi(x_i) \cdot \Phi(x_j) = k(x_i, x_j)$. The classification vector in the transformed, high dimensional space of the form

$$w = \sum_{i=1}^{n} q_i y_i \Phi(x_i) \tag{3}$$

is sought by minimizing

$$\sum_{i=1}^{n} q_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i q_i k(x_i, x_j) y_j q_j \tag{4}$$

subject to the constraints $\sum_{i=1}^{n} q_i y_i = 0$ and $0 \leq q_i \leq 1/(2nc)$, with the parameter $c$ penalizing prediction errors.

The key insight we introduce lies in the modelling of the distribution of data in the target high dimensional space of a SVM. In particular, we assume that class data points are normally distributed, which is an assumption consistent with the linear separability goal of support vector based learning:

$$P(\hat{d}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \tag{5}$$

where $\mu$ and $\sigma$ are respectively the mean and the standard deviation of the normal distribution, and $\hat{d}_i$ the data, all in the target high dimensional space of the SVM. Then, the data description length term in (2) becomes:

$$\sum_{i=0}^{N-1} \log_2 P(\hat{d}_i) = \sum_{i=0}^{N-1} \log_2 \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\hat{d}_i-\mu)^2}{2\sigma^2}} \right] \tag{6}$$

$$= \log_2 \left[ \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{\sum_{i=0}^{N-1} \frac{-(\hat{d}_i-\mu)^2}{2\sigma^2}} \right] \tag{7}$$

$$= -\frac{N}{2} \log_2(2\pi\sigma^2) - \frac{1}{2\sigma^2 \ln 2} \sum_{i=0}^{N-1} (x_i - \mu)^2 \tag{8}$$

and since $\sum_{i=0}^{N-1} (\hat{d}_i - \mu)^2 = \sigma^2 N$:

$$\sum_{i=0}^{N-1} \log_2 P(\hat{d}_i) = -\frac{N}{2} \log_2(2\pi\sigma^2) - \frac{N}{2\ln 2} \tag{9}$$

$$= -\frac{N}{2} \left( \log_2 2\pi\sigma^2 + \log_2 e \right) \tag{10}$$

$$= -\frac{N}{2} \log_2(2\pi e\sigma^2) \tag{11}$$

Thus in our case the description length in (2) becomes:

$$DL(M, D) = \frac{1}{2} N_M \log_2 N + \frac{1}{2} N \log_2(2\pi\sigma^2 e) \tag{12}$$

where $N_M$, the number of free model parameters, is equal to the number of support vectors of the trained model.

# Experimental analysis

## Experimental data

We carried out our experiments on the Breast Cancer Wisconsin (Diagnostic) Data Set (Wolberg, Street, and Mangasarian 1992) – a popular and publicly freely available corpus. By way of summary, the data set comprises exemplars (569 in total) of two classes corresponding to the two diagnostic decisions as regards cancer malignancy (malignant or benign), characterized by 30 features extracted from images of fine-needle aspirates of breast masses. These features describe the characteristics of the cell nuclei morphology (captured through the mean and extreme values, and standard deviations of relevant characteristics across a slide) present in the images and are obtained after performing cytological analysis and simple image processing (curve-fitting for cell delineation) (Bennett 1992; Bennett and Mangasarian 1992).

Table 1: Summary of experimental parameters.

| Parameter | Value |
|---|---|
| Number of features | 30 |
| Number of samples | 569 |
| Min & max values of $c$ | 1 & 1000 |
| Min & max values of $\gamma$ | 0.001 & 1 |
| Number of equidistant $c$ samples | 100 |
| Number of equidistant $\gamma$ samples | 100 |

## Methodology

The following summarizes the process used to apply the proposed method and compute the corresponding performance metrics:

1. The data was divided into two subsets, training and test, with the split ratio (in terms of data sample counts) of 4:1.
2. Radial basis function (RBF) SVM models were trained using 10,000 combinations of $c$ and $\gamma$ (100 $c$ and 100 $\gamma$) as per Table 1, where $\gamma$ is the reciprocal of the RBF standard deviation.
3. Using the training set only, description lengths were computed for models trained with different values of hyperparameters. The performance of the model selected according to the MDL criterion was assessed on the test data.

Comparison was made with the standard $k$-fold based approach:

1. The data was was divided into training, validation, and test set with split ratio 3:1:1. Then, $k$-fold cross-validation was applied with 4 folds on the union of training and validation sets. As before, the performance of the final selected model was assessed on the test data.

## Results and discussion

We start our discussion with the baseline approach first, that is the $k$-fold cross-validation based model selection. The plot in Figure 1 shows the dependency of the final, trained classifier accuracy on the values of the parameters $\gamma$ and $c$, which is useful in demonstrating that the model actually selected is indeed sound i.e. one which does not exhibit over- or under-training. It is worth observing that expectedly the accuracy is greatly affected by varying $\gamma$ but not by the regularizing penalty parameter $c$. The latter suggests good separability of classes in the implicit high dimensional space of the trained SVMs.

Turning our attention to the proposed method now, the variation in the description length attained with the models trained with different values of the hyperparameters $\gamma$ and $c$ is shown in Figure 2. As anticipated based on the behaviour observed using $k$-fold cross-validation training, here too we observe that in this case the description length is highly affected by $\gamma$ but not by $c$. Importantly, it is also the case that

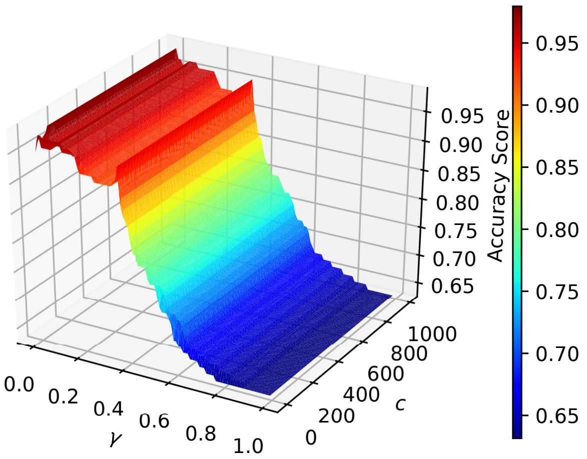

Figure 1: Trained classifier accuracy as a function of the parameters $\gamma$ & $c$, using $k$-fold cross-validation.

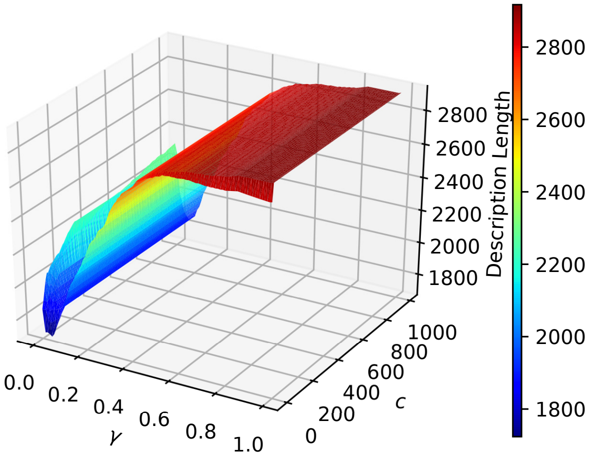

Figure 2: Description length as a function of SVM parameters $\gamma$ & $c$.

a clear local minimum is present in the plot, corresponding to the ultimately selected model i.e. the ‘best’ model under the minimum description length criterion.

Having ensured that both approaches are making sensible decisions in the context of the given data, we proceeded by evaluating the two selected models – one using $k$-fold cross-validation and another with the proposed approach– and comparing them with one another. A summary is shown in Table 2, demonstrating that the proposed method achieves superior results according to all metrics considered. This is despite the absence of empirically, verification guided model selection – or rather, precisely because of, as we have argued, with our approach being based on fundamental theoretical grounds which allows for a more efficient use of available data.

For completeness, we also include the plot showing the variation of classification accuracy of different trained models compared in terms of their description length in Figure 3, which shows that the parameter combination selected by our MDL guided selection process indeed does achieve best performance too.

Table 2: Comparison of the models selected using traditional $k$-fold cross-validation (xV) and the method proposed in the present paper.

| **Method** | **Accuracy** | **Recall** | **F1-Score** | **Precision** |
|---|---|---|---|---|
| $k$-fold xV | 0.9649 | 0.9286 | 0.9512 | 0.9750 |
| Proposed | 0.9737 | 0.9524 | 0.9639 | 0.9756 |

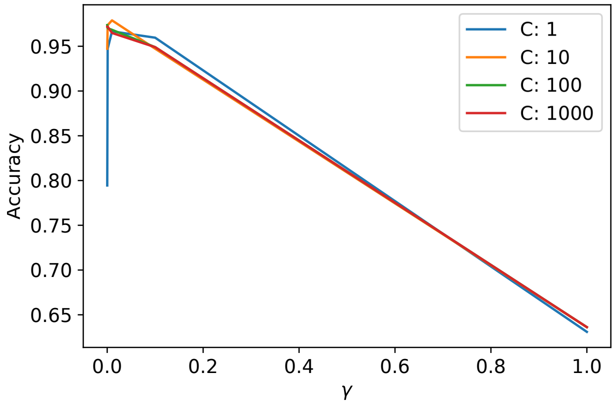

Figure 3: Accuracy as a function of $\gamma$ for fixed values of the penalty $c$, namely $c = 1, 10, 100, 1000$.

## Summary and Future Work

In this paper we introduced a novel, theoretically rigorous framework for SVM model selection which overcomes the inefficiencies of the current, empirically driven approaches used in practice. This contribution is especially important in numerous medical applications, where the effect of the aforementioned inefficiencies can be a limiting factors in applicability of machine learning (e.g. due to limited data). The effectiveness of the proposed approach was demonstrated on the Wisconsin Diagnostic Breast Cancer Data Set, on which it is shown to outperform the existing alternatives.

## References

Barracliffe, L.; Arandjelovic, O.; and Humphris, G. 2017. A pilot study of breast cancer patients: Can machine learning predict healthcare professionals’ responses to patient emotions. In *Proceedings of the International Conference on Bioinformatics and Computational Biology*, 20–22.

Bennett, K. 1992. *Decision tree construction via linear programming*. Number 1067 in Computer Sciences Technical Report, University of Wisconsin-Madison. University of Wisconsin-Madison, Computer Sciences Department.

Bennett, K. P.; and Mangasarian, O. L. 1992. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1): 23–34.

Blum, A.; Kalai, A.; and Langford, J. 1999. Beating the hold-out: bounds for $k$-fold and progressive cross-validation. In *Proceedings of the Conference on Computational Learning Theory*, 203–208.

Blumer, A.; Ehrenfeucht, A.; Haussler, D.; and Warmuth, M. K. 1987. Occam's razor. *Information Processing Letters*, 24(6): 377–380.

Caie, P. D.; Dimitriou, N.; and Arandjelovic, O. 2020. Precision medicine in digital pathology via image analysis and machine. *Artificial Intelligence and Deep Learning in Pathology*, 149.

Cawley, G. C.; and Talbot, N. L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11: 2079–2107.

Dimitriou, N.; Arandjelović, O.; Harrison, D. J.; and Caie, P. D. 2018. A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *NPJ Digital Medicine*, 1(1): 1–9.

Galvao, R. K. H.; Araujo, M. C. U.; José, G. E.; Pontes, M. J. C.; Silva, E. C.; and Saldanha, T. C. B. 2005. A method for calibration and validation subset partitioning. *Talanta*, 67(4): 736–740.

Gavriel, C. G.; Dimitriou, N.; Brieu, N.; Nearchou, I. P.; Arandjelović, O.; Schmidt, G.; Harrison, D. J.; and Caie, P. D. 2021. Assessment of Immunological Features in Muscle-Invasive Bladder Cancer Prognosis Using Ensemble Learning. *Cancers*, 13(7): 1624.

Grünwald, P. D.; and Grunwald, A. 2007. *The minimum description length principle*. MIT press.

Heckerman, D.; and Meek, C. 1997. Models and selection criteria for regression and classification. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 223–228.

Kohavi, R.; et al. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2, 1137–1145. Montreal, Canada.

Murata, N.; Yoshizawa, S.; and Amari, S.-i. 1993. Learning curves, model selection and complexity of neural networks. In *Advances in Neural Information Processing Systems*, 607–614.

Raschka, S. 2018. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5): 1299–1319.

Shalev-Shwartz, S.; and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Wolberg, W. H.; Street, W. N.; and Mangasarian, O. L. 1992. Breast cancer Wisconsin (diagnostic) data set. *UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/]*.

Yue, X.; Dimitriou, N.; and Arandjelovic, O. 2019. Colorectal cancer outcome prediction from H&E whole slide images using machine learning and automatically inferred phenotype profiles. In *Proceedings of the International Conference on Bioinformatics and Computational Biology*, 139–149.