On the selection of decision trees in Random Forests

Simon Bernard, Laurent Heutte, Sébastien Adam

On the Selection of Decision Trees in Random Forests

Simon Bernard, Laurent Heutte, Sébastien Adam
Université de Rouen, LITIS EA 4108
BP 12 - 76801 Saint-Etienne du Rouvray, France.
{simon.bernard,laurent.heutte,sebastien.adam}@univ-rouen.fr

***Abstract*— In this paper we present a study on the Random Forest (RF) family of ensemble methods. In a "classical" RF induction process a fixed number of randomized decision trees are inducted to form an ensemble. This kind of algorithm presents two main drawbacks : (i) the number of trees has to be *a priori* fixed (ii) the interpretability and analysis capacities offered by decision tree classifiers are lost due to the randomization principle. This kind of process where trees are independently added to the ensemble, offers no guarantee that all those trees will well cooperate into the same committee. This statement rises two questions : are there any decision trees in a RF that make the performance of the ensemble decrease? If so, is it possible to form a more accurate committee by removing from the initial ensemble those decision trees? The answer to these questions is tackled as a classifier selection problem, and we thus show that better subsets of decision trees can be obtained even using a sub-optimal classifier selection method. This proves that "classical" RF induction process, for which randomized trees are arbitrary added to the ensemble, is not the best approach to produce accurate RF classifier. We also show the interest in designing RF by adding trees in a more dependent way than it is traditionally done in these "classical" algorithms.**

I. Introduction

One of the Machine Learning issues consists in designing high performance classification systems based on a set of representative samples of a population of data. Among the different approaches to deal with this kind of problematic, combining an ensemble of individual weak classifiers to form a unique classification system — called Classifier Ensemble — has aroused a growing interest in the scientific community. This interest has been fed by recent researches that have shown some combination principles to be particularly efficient, such as Boosting [1] (or Arcing [2]), Bagging [3], Random Subspaces [4], or more recently, Random Forests [5]. The efficiency in combining classifiers leans on the ability to take into account the complementarity between individual classifiers, in order to improve as much as possible the generalization performance of the ensemble. An explanation of this link between complementarity and performance is the diversity property. Although there is no agreed definition for diversity [6], this concept is usually recognized to be one of the most important characteristics for the improvement of the generalization performance in an ensemble of classifiers [7]. One can define it as the ability of the individual classifiers of an ensemble to agree mainly on good predictions and to disagree on prediction errors.

Among the different approaches that aim at building ensembles of diverse classifiers, those using randomization to produce diversity have proven to be particularly efficient, as for Bagging [3] or Random Subspaces methods [4]. These two methods both use randomization in the induction process, in order to build base classifiers different from each others, and thus introducing diversity among them. Recently Leo Breiman has proposed a new family of ensemble methods called Random Forest (RF) [5], based on this randomization concept. RF can be defined as a generic principle of classifier combination that uses L tree-structured base classifiers $\{h(x, \Theta_k),\ k = 1, ...L\}$ where $\{\Theta_k\}$ is a family of independent identically distributed random vectors, and x is an input data. The particularity of this kind of combination is that each decision tree is built from a random vector of parameters. A Random Forest can be built for example by randomly sampling a feature subset for each decision tree (as in Random Subspaces), and/or by randomly sampling a training data subset for each decision tree (as in Bagging).

Since they have been introduced in 2001, RF have been studied in many ways, theoretically as well as experimentally [8], [9], [5], [10], [11], [12], [13], [14], [15]. In most of those works, it has been shown that RF are particularly competitive with one of the most efficient learning principles, *i.e.* boosting [5], [11], [14]. However, the mechanisms that explain the good performance of RF are not clearly identified. For example, it has been theoretically proved in [5] and experimentally confirmed in [13], that above a certain number of trees, adding more trees in the forest does not improve the accuracy. This statement concerns the induction processes that randomly produce trees without any *a priori* knowledge on their intrinsic characteristics. Yet, no research work has studied the effect of the number of trees on the performance of a RF.

In this paper we propose to go one step further in the understanding of RF mechanisms. The goal is to determine whether or not it is possible to select a subset of trees from a forest that is able to outperform this forest. Our aim is not to find the optimal subset of individual classifiers among a large ensemble of trees, but rather to study the extent to

which it is possible to enhance accuracy of a RF by focusing on some particular subsets of trees. The "final" goal of this work is thus to identify some particular properties that are shared by these sub-forests, and the tree selection approach we propose in this paper is a first step toward this direction. Therefore, as we will discuss in section III, there is no need here to apply optimal classifier selection techniques to RF. We have thus decided to use two simple classifier selection techniques, *i.e.* SFS (Sequential Forward Selection) and SBS (Sequential Backward Selection) [16], and to monitor the error rates of each subset obtained during the experiments. Moreover, using a selection process such as SFS allows to highlight the interest in designing RF by adding trees in a more dependent way than it is traditionally done in "classical" RF induction methods, where trees are built strictly independently from each other. We also show that a sequential tree induction approach, in which trees are dependently added to the forest, would considerably "minimize" the number of trees to combine in a RF.

The paper is thus organized as follows: we recall in section 2 the Forest-RI principles; in section 3, we first explain our approach of classifier selection in RF, and then describe our experimental protocol, the datasets used, and the results obtained. We finally draw some conclusions and future works in the last section.

II. The Forest-RI algorithm

One can see Random Forests as a family of methods, made of different decision trees ensemble induction algorithms, such as the Breiman Forest-RI method often cited as the reference algorithm in the literature. In this algorithm the Bagging principle is used with another randomization technique called Random Feature Selection. The training step consists in building an ensemble of decision trees, each one trained from a bootstrap sample of the original training set — *i.e.* applying the Bagging principle — and with a decision tree induction method called Random Tree. This induction algorithm, usually based on the CART algorithm [17], modifies the splitting procedure for each node, in such a way that the selection of the feature used for the splitting criterion is partially randomized. That is to say, for each node, a feature subset is randomly drawn, from which the best splitting criterion is then selected.

To sum up, in the Forest-RI method, a decision tree is grown by using the following process :

- Let N be the size of the original training set. N instances are randomly drawn with replacement, to form the bootstrap sample, which is then used to build the tree.
- Let M be the dimensionality of the original feature space, and K a preliminary fixed parameter so that $K \in [1, M]$. For each node of the tree, a subset of K features is randomly drawn without replacement, among which the best split is then selected.
- The tree is thus built to reach its maximum size. No pruning is performed.

In this process the tree induction is directed by a single hyperparameter, *i.e.* the number K of randomly selected features. This number allows to introduce more or less randomization in the induction. Consequently, except when $K = M$, in which case the tree induction is not randomized at all, each tree of a RF presents structure and properties that can not be foreseen *a priori*. With the introduction of randomization in the RF induction, we hope to take benefits of complementarities of individual trees, but there is no guarantee that adding a tree in a RF will allow to improve the performance of the ensemble. One can even imagine that some trees of a RF make the accuracy of the ensemble be lower. This idea has led us to study how to improve the performance of a RF by selecting a particular subset of its trees.

In the literature, only few research works have focused on the number of trees that have to be grown in a RF. When introducing RF formalism in [5], Breiman demonstrated that above a certain number of trees, adding more trees does not allow to improve the performance. Precisely he stated that for an increasing number of trees in the forest, the generalization error converges to a maximum. This result indicates that the number of trees in a forest does not have to be as large as possible to produce an accurate RF. The work of Latinne et al. in [13], and our work in [8] experimentally confirm this statement. However, noting that above a certain number of trees no improvement can be obtained by adding more "arbitrary" trees in the forest does not mean obviously that the optimal performance has been reached. Thus the idea of our experimental work is to establish whether or not a subset of individual trees is able to outperform the whole ensemble.

Notice that in the rest of this paper, the term Random Forest (RF) will always stand for a forest built with the Forest-RI algorithm.

III. Selecting better subsets of trees from a RF

The principle of our experiments is to apply classifier selection techniques on a RF made up of a large number of trees. For that purpose two main choices have to be made: a selection criterion and a selection method.

Selection criteria for classifier selection can be divided into two main approaches: the filter approach and the wrapper approach [18]. On the one hand the filter approach consists in selecting a subset of classifiers according to an *a priori* evaluation that does not take into account the combination performance. On the other hand, the wrapper approach attempts to select the subset of classifiers that *a posteriori* optimizes the combination performance. As our goal is to establish whether or not a better subset of trees

that outperforms the initial forest can be found, the wrapper principle has been adopted for our experiments. Thus classifiers have been selected by optimizing the accuracy — i.e. minimizing the error rate — of the resulting subsets of trees.

Concerning the selection methods, as mentioned in section I, our aim is not to find the optimal subset of individual classifiers among a large ensemble of trees, but rather to analyze the extent to which RF performance can be improved by removing from the ensemble some particular trees. Thus the optimality of the selection methods is not a priority here. That is the reason why the two well-known classifier selection algorithms, SFS (Sequential Forward Selection) and SBS (Sequential Backward Selection) have been chosen. These two methods are known to be sub-optimal because the sequential process makes each iteration depend on the previous one, and finally not all the possible solutions are explored. However they present the advantage to be fast and simple. Those two selection techniques iteratively build a sub-optimal subset from an ensemble of classifiers according to a given criterion [16]. At each iteration of the SFS process, each remaining classifier is added to the current subset and the one that optimizes the performance of the ensemble is retained. In the same manner, in the SBS process, each classifier of the current subset is removed, and the one for which the remaining ensemble exhibits the best accuracy is definitely discarded. The stopping criterion in such iterative processes is commonly based on the convergence of the accuracy, but it can also be defined for example by a maximum number of iterations that determines the number of classifiers in the final subset [19]. For our experiments we have decided to let the selection algorithms explore all the possible iterations, *i.e.* for a number L', from 1 to L, of trees in the final subset, where L is the size of the original RF. In that way we can also study the evolution of the RF accuracy according to the number of trees retained in the subset, in order to have an idea of how many trees can be removed from the RF to obtain a more accurate classifier, and how the performance can be improved.

We first describe in the following subsection the datasets used. We then detail our experimental protocol and results in the next two subsections.

A. Datasets

The 10 datasets that have been used in these experiments are described in Table I: the first 7 datasets in the table have been selected from the UCI repository [20]; Twonorm and Ringnorm are two synthetic datasets designed by Breiman and are described in [2]; and the MNIST database [21] is a handwritten digit recognition database on which greyscale mean values have been extracted as explained in [8]. Those datasets have been selected because they do not contain any missing value and the features are all numerical features.

Note that for the experiments described in this section we have decided to randomly split each original dataset, with two thirds of the samples used for training, and the last third for testing.

B. Experimental protocol

Our experiments consist in applying the two previously presented classifier selection methods on a large ensemble of trees, and in monitoring the evolution of the error rate of each subset obtained during the selection processes. The full experimental protocol is described below.

First, each dataset has been divided into a training and a testing subset, with respectively two thirds of the samples used for training, and the other third for testing. As explained previously, our goal is to study the evolution of the accuracy of a RF according to the number of trees it contains. Thus only one split of each dataset has been produced. We denote this split by $T = (T_r, T_s)$ where T_r and T_s stand respectively for the training set and the testing set.

Then, a RF is grown from T_r, with a number L of trees fixed to 300. The value of the hyperparameter K has been fixed to $\sqrt{M}$, which is a default value commonly used in the literature. A previous work on the parametrization of RF, presented in [22], has shown that this value of K is a good compromise to induct accurate RF. SFS and SBS methods are applied on the RF, so that at each iteration the tree to add (SFS) or to remove (SBS) is the one that allows to obtained the most accurate sub-forest. For comparison another random selection method has been applied to the RF, that iteratively adds a randomly selected tree from the original RF to the final subset. This selection process, noted SRS (for Sequential Random Selection), allows to simulate the induction of a RF, for an increasing number of trees from 1 to L. Finally three tables of L error rates are obtained for each dataset.

Algorithm 1 summarizes the whole experimental protocol applied to each dataset. This procedure outputs a table of $L \times 3$ error rates (one table for each selection method) for each dataset. Those results are presented and discussed in the next subsection.

C. Results

Figure 1 presents 10 diagrams of our results for the 10 datasets used. For each of them, three curves have been plotted, representing the error rates obtained with the three previously detailed selection processes, according to the number of trees in the subsets. Table II presents the best error rates obtained for each of the three selection processes on each dataset, and the number of trees of the corresponding subsets.

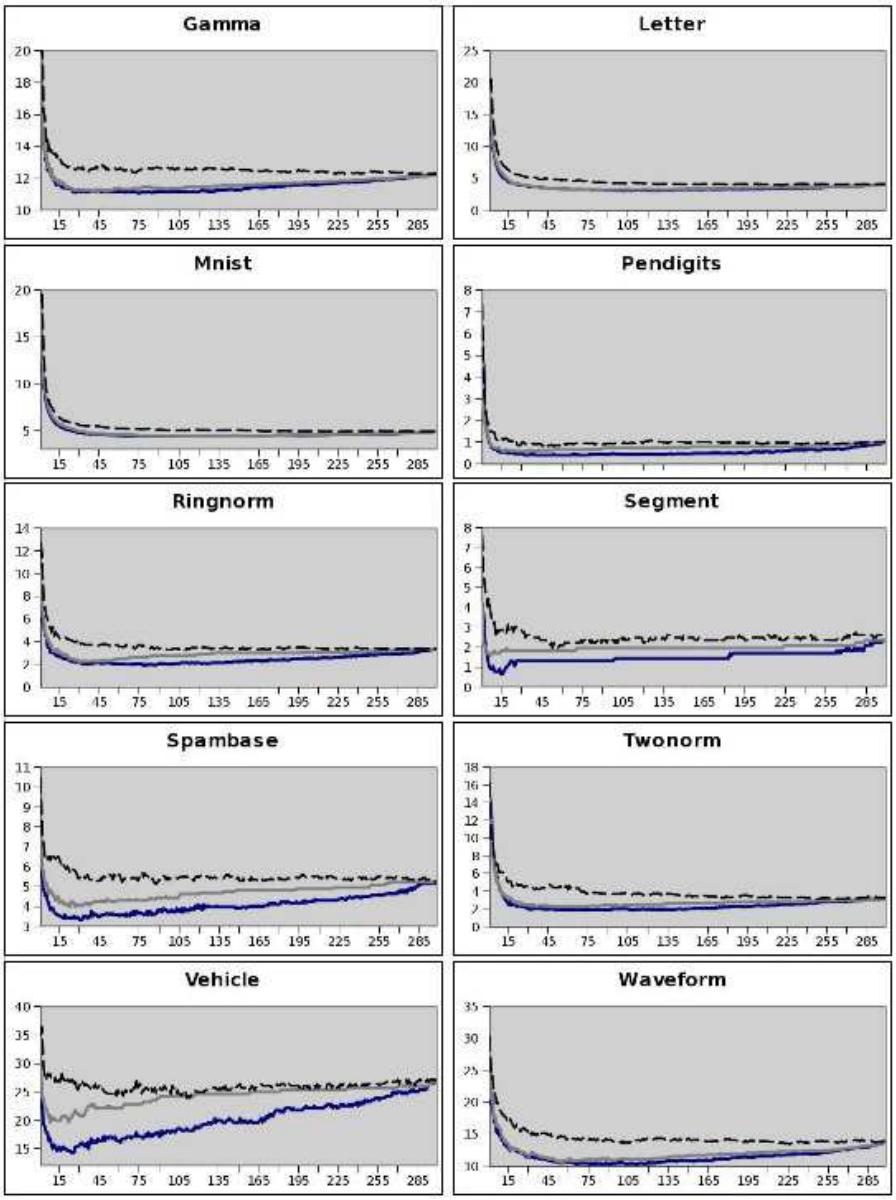
One can first observe from Table II that in spite of the sub-optimality of SFS and SBS, these algorithms always

TABLE I

DATASETS DESCRIPTION

Dataset	Size	Features	Classes	Dataset	Size	Features	Classes
Gamma	19020	10	2	Vehicle	946	18	4
Letter	20000	16	26	Waveform	5000	40	3
Pendigits	10992	16	10	Ringnorm	7400	20	2
Segment	2310	19	7	Twonorm	7400	20	2
Spambase	4610	57	2	Mnist	60000	84	10

Fig. 1. Error Rates obtained during the three selection processes on 10 datasets, according to the number of trees in the subsets. The black curves represent the error rates obtained with SFS, the gray curves the error rates with SBS, and the dashed-line curves the error rates with SRS.



allow to find a subset of trees that outperforms the initial RF, inducted with Forest-RI. This observation highlights the interest of studying the selection of subsets of trees in RF to improve the performance. Therefore one can first conjecture that the performance should be much more improved by searching for the optimal subset of trees, using for example

TABLE II

BEST ERROR RATES AND NUMBER OF TREES OF THE CORRESPONDING SELECTED SUBSETS

Dataset	SFS		SBS		Forest-RI
	error rates	# trees	error rates	# trees	300 trees
Gamma	11.07	79	11.17	50	12.19
Letter	3.07	98	3.20	70	4.09
Pendigits	0.41	32	0.57	28	1,01
Segment	0.66	15	1.57	8	2.49
Spambase	3.33	31	3.98	24	5.22
Vehicle	14.29	25	19.64	9	26.79
Waveform	10.16	86	10.46	56	14
Ringnorm	1.9	34	2.15	31	3.33
Twonorm	1.82	75	2.19	51	3.2
MNIST	4.41	97	4.4	119	4.93

Algorithm 1 Experimental Protocol

Require: N the number of samples in the original dataset. M the number of features in the original dataset.

Randomly draw without replacement $\frac{2}{3} \times N$ samples from the original dataset to form the training subset T_r. The remaining samples form the testing subset T_s.

$h \leftarrow$ Forest-RI($L = 300$,$K = \sqrt{M}$,T_r).

$h^{(0)}_{SFS} \leftarrow \emptyset$.

$h^{(0)}_{SBS} \leftarrow h$.

$h^{(0)}_{SRS} \leftarrow \emptyset$.

for $i = 1$ *to* L **do**

$h^{(i)}_{SFS} \leftarrow h^{(i-1)}_{SFS} \cup h(k)$ where $k = argmin_{h(j) \notin h^{(i-1)}_{SFS}}$ $\{$error$(h^{(i-1)}_{SFS} \cup h(j)$,$T_s)$ $\}$.

$h^{(i)}_{SBS} \leftarrow h^{(i-1)}_{SBS} \setminus h(k)$ where $k = argmin_{h(j) \in h^{(i-1)}_{SBS}}$ $\{$error$(h^{(i-1)}_{SBS} \setminus h(j)$,$T_s)$ $\}$.

$h^{(i)}_{SRS} \leftarrow h^{(i-1)}_{SRS} \cup h(k)$ where $k =$ random$(j), h(j) \notin h^{(i-1)}_{SRS}$.

Store the error rates of $h^{(i)}_{SFS}$, $h^{(i)}_{SBS}$ and $h^{(i)}_{SRS}$.

end for

optimal (Branch and Bound method [23]) or near optimal (Genetic Algorithms [16]) classifier selection methods.

A second observation that can be made from those diagrams is that the lowest error rate, for each dataset, is reached by a subset of decision trees obtained with a small number of trees, *i.e.* almost every time less than 100 trees. This corresponds to less than $\frac{1}{3}$ of the total number of trees in the initial forest. In other words for each RF grown during our experiments, at least $\frac{2}{3}$ of the trees have been removed to reach the best error rates. This number is even sometimes much more important since the best accuracy has been reached for some datasets with less than 30 trees (Segment and Vehicle), which corresponds to only 10% of the total number of trees grown in the initial RF. This shows that among all the trees of a RF, only few of them should be combined to obtain an accurate classifier. Furthermore those results highlight that when a RF is grown with a "classical" RF induction algorithm such as Forest-RI, all the trees do not allow to improve the performance, and some of them even make the ensemble do more prediction mistakes. In addition the fact that the forward search is always the most efficient approach to find a sub-optimal subset of trees, makes us conjecture that it could be useful to design a dynamical RF induction process that would add to the ensemble only decision trees that improve the accuracy of the RF. This would be beneficial in terms of computational and performance gain.

IV. CONCLUSIONS

In this paper, a study on tree selection in Random Forests has been presented. The goal was to highlight that some particular subsets of trees of a RF are able to perform better than this forest. Two well-known selection methods have been used for that purpose : SFS (Sequential Forward Selection) and SBS (Sequential Backward Selection). In spite of the sub-optimality of the SFS and SBS methods, this work has shown that it always exists a subset of well selected trees able to outperform an ensemble grown with a "classical" RF induction algorithm such as Forest-RI. Thus an interesting perpective of those experiments would be to apply some other classifier selection methods known to be more efficient than SFS and SBS, like the Branch and Bound method for example which is an optimal selection method [23], or like genetic algorithms [16]. It would allow to better foresee the extent to which a subset is able to

outperform the whole ensemble of trees.

Moreover using SFS selection process in these experiments, has allowed to highlight the interest in designing RF by adding trees in a more dependent way than it is traditionally done in "classical" RF induction methods, where trees are built strictly independently from each other. However the nature of this dependance still remain an open issue and we believe that the next task to achieve in this research work is to identify some particular properties shared by the best sub-RF found during the selection process, so that those properties could be used as criteria for leading the sequential tree induction process. This could be tackled for example through the study of out-of-bag estimates, strength, correlation [5], diversity, specificities of the tree structure like for example the splitting features, or the bag samples, etc.

Finally we believe that such a sequential approach for inducting RF would be interesting in the way it could allow to decrease the number of trees to be inducted in a RF, since our experiments have shown that the best sub-forests obtained during the selection process contained significantly less trees than the initial RF. However this observation highlights another point of interest that should be studied: which stopping criterion should be used for such a sequential RF induction?

REFERENCES

[1] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," *International Conference on Machine Learning*, pp. 148–156, 1996.

[2] L. Breiman, "Arcing classifiers," *The Annals of Statistics*, vol. 26, no. 3, pp. 801–849, 1998.

[3] ——, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[4] T. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[6] L. Kuncheva, "That elusive diversity in classifier ensembles," *IbPRIA*, pp. 1126–1138, 2003.

[7] ——, *Combining Pattern Recognition. Methods and Algorithms*. John Wiley and Sons, 2004.

[8] S. Bernard, L. Heutte, and S. Adam, "Using random forests for handwritten digit recognition," *International Conference on Document Analysis and Recognition*, pp. 1043–1047, 2007.

[9] P. Boinee, A. D. Angelis, and G. Foresti, "Meta random forests," *International Journal of Computational Intelligence*, vol. 2, no. 3, pp. 138–147, 2005.

[10] L. Breiman, "Consistency of random forests and other averaging classifiers," *Technical Report*, 2004.

[11] A. Cutler and G. Zhao, "Pert - perfect random tree ensembles," *Computing Science and Statistics*, vol. 33, 2001.

[12] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 36, no. 1, pp. 3–42, 2006.

[13] P. Latinne, O. Debeir, and C. Decaestecker, "Limiting the number of trees in random forests," *2nd International Workshop on Multiple Classifier Systems*, pp. 178–187, 2001.

[14] J. Rodriguez, L. Kuncheva, and C. Alonso, "Rotation forest : A new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[15] M. Robnik-Sikonja, "Improving random forests," *European Conference on Machine Learning, LNAI 3210, Springer, Berlin*, pp. 359–370, 2004.

[16] H. Hao, C. Liu, and H. Sako, "Comparison of genetic algorithm and sequential search methods for classifier subset selection." *Seventh International Conference on Document Analysis and Recognition*, vol. 2, pp. 765–769, 2003.

[17] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Chapman and Hall (Wadsworth, Inc.): New York, 1984.

[18] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[19] F. Roli, G. Giacinto, and G. Vernazza, "Methods for designing multiple classifier systems," *Multiple Classifiers Systems*, pp. 78–87, 2001.

[20] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://archive.ics.uci.edu/ml/

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] S. Bernard, L. Heutte, and S. Adam, "Influence of hyperparameters on random forest accuracy," *Technical Report, University of Rouen*, 2008.

[23] P. Somol, P. Pudil, and J. Kittler, "Fast branch and bound algorithms for optimal feature selection," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 900–912, 2004.