

Semi-Supervised Learning For Quality Control Of High-Value Wood Products

Mark Schubert (✉ mark.schubert@empa.ch)

EMPA - Materials Science and Technology <https://orcid.org/0000-0002-6900-0370>

Walter Sonderegger

ETH-Bereich Hochschulen

Mirko Luković

Empa Materials Science and Technology

Oliver Kläusler

ETH-Bereich Hochschulen

Research Article

Keywords: machine learning, semi-supervised learning, quality control, random forest, support vector machine, wood products

Posted Date: March 10th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1380640/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The quality control of wood products is often only checked at the end of the production process so that countermeasures can only be taken with a time delay in the event of fluctuations in product quality. This often leads to unnecessary and cost-intensive rejects. Furthermore, since quality control often requires additional procedural steps to be performed by a skilled worker, testing is time-consuming and costly. While traditional machine learning (ML) methods based on supervised learning have been used in the field with some success, the limited availability of labeled data is the major hurdle for further improving model performance. In the present study, the potential of enhancing the performance of the ML methods random forest (RF) and support vector machines (SVM) for quality classification by using semi-supervised learning (SSL) was investigated. Labelled and unlabeled data were provided by Swiss Wood Solutions AG, which produces densified wood for high-value wood products such as musical instruments. The developed approach includes labeling of the unlabeled data using SSL, training and 10k cross-validation of the ML algorithms RF and SVM, and determining the generalization ability using the hold-out test set. Based on the evaluation indices such as accuracy, F1-score, recall, false-positive-rate (FPR) and confusion matrices, it was shown that SSL could enhance the prediction performance of the quality classification of ML models compared to the conventional supervised learning method. Despite having a small data set, the work paves the way for future applications of semi-supervised ML for wood quality assessment.

1. Introduction

The processing of raw material wood into materials, semi-finished products and products is characterized by various sequential production steps of discrete manufacturing including cutting, sorting, joining, forming, etc., each of which is geared to a specific product requirement or product design. Due to the many starting materials (different types of wood and grading classes, various wood-based materials, etc.) and design variants, the production process of many wood products is extremely complex. In order to increase the value added and ensure competitiveness, the quality control (QC) of the wood products created in the course of industrial production is of paramount importance. Particularly in high-price segments, where the quality requirements for the products are especially high, such as in musical instrument manufacturing. QC is the task of assuring that the products produced reach a certain standard that is set either by the company or by the customers. The field developed rapidly during the second half of the 20th century and is today an integral part of most manufacturing companies. The three major methods of QC are 'Acceptance Sampling', 'Statistical Process Control', and 'Experimental Design' (Fountoulaki et al. 2011). Acceptance Sampling, where only a sample of products is tested to draw conclusions about the entire batch, is used when testing is expensive, time-consuming and/or destructive, which is mainly the case in the wood industry. The quality control of wood products is often only checked at the end of the production process, so that countermeasures can only be taken with a time delay in the event of fluctuations in product quality. This often leads to unnecessary and cost-intensive rejects, sometimes even of an entire day's production quantity. Furthermore, since quality control often

requires additional procedural steps to be performed by a skilled worker, testing of wood products is time-consuming and costly. This results in there being only small-labeled datasets available for comprehensive analysis.

Machine learning (ML) is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention (Bishop 2006). ML algorithms can be classified into different groups based on the way they “learn” about data to make predictions: supervised, unsupervised and semi-supervised learning. The more commonly used ML models are based on supervised learning, where the algorithm learns from a labeled dataset, which provides an answer key that the algorithm can use to evaluate its accuracy and provide feedback. An unsupervised model, in contrast, uses unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own.

Semi-supervised learning (SSL) falls between unsupervised learning and supervised learning and combines a small amount of labeled data with a larger amount of unlabeled data during training (Ouali et al. 2020). More formally, the goal of SSL is to leverage the unlabeled data (D_u) to produce a prediction function (f_θ) with trainable parameters (θ), that is more accurate than what would have been obtained by only using the labeled data (D_l). For instance, D_u might provide additional information about the structure of the data distribution ($p(x)$) to better estimate the decision boundary between the different classes (Ouali et al. 2020).

While traditional ML methods based on supervised learning have been used with some success to predict the quality of wood products (Barnes 2001; Gupta et al. 2007; André et al. 2008; Esteban et al. 2011; Bardak et al. 2016b, a; Schubert and Kläusler 2020; Schubert et al. 2020; Ehrhart et al. 2021), the limited availability of labeled data is the major hurdle for further improving ML model performance.

To the best of our knowledge, this work is the first time that semi-supervised learning has been used to improve the QC of real data from the wood products industry. The aim of the present work was to: (1) use a SSL method for labeling the unlabeled dataset and combine it with traditional ML algorithms, namely, random forest (RF) and support vector machines (SVM). (2) Determine the prediction accuracy of quality classification using evaluation indices and confusion matrices. (3) To compare the generalization capability of the SSL-ML algorithms with the RF and SVM algorithms trained in a supervised manner only.

2. Materials And Methods

2.1 Production and quality control of high-value wood products

Swiss Wood Solutions AG is a business and technology incubator for sustainable, wood-based products, which produces densified wood for high-value wood products such as musical instruments. The

production includes several steps. After the first incoming control of the wooden raw material (QC I) the accordant square wood enters the climatization process (Wood Climatization I) as illustrated in Fig. 1. In this phase, wood moisture and wood temperature are adjusted. The subsequent step is the thermo-mechanical densification, which takes place in a hydraulic pressing machine. As a next step, the wood again undergoes climatization (Wood Climatization II) to slowly adjust the wood moisture content to the ambient air conditions (rel. humidity and temperature) and slowly relieve stresses in the wood. After this modification procedure, the final quality control (QC II) is carried out (Fig. 1).

2.1.1 Quality control

A first incoming control (QC I) takes place before the wood modification process. The squared timbers are sorted visually according to the following parameters:

- alignment parallel to the fibre (length) and straight in tangential (width) and radial (height) directions
- without knots
- without compression or tension wood

The mass, length, width, height, and density of the squared timber are also measured which is used for the arrangement of the modification charges.

After the densification process, the second quality control (QC II) checkpoint takes place which involves a visual inspection of each timber concerning cracks, deformations and discoloration. Of great importance is the re-swelling of the wood after densification, which occurs mainly under humid conditions and can reach almost 90–100% in extreme cases (Sandberg et al. 2013). To ensure high quality products are produced, a maximal *irreversible swelling of 2%* is set as an internally tolerable threshold after a water uptake of 5 days. Two specimens of 1 cm in length are cut at each test location, one for the water uptake testing and one reference specimen. The samples for water uptake are stored in water for 5 days, afterwards they are re-climatized in the storage room and oven-dried together with the reference specimens. The differences in the relative heights of the two adjacent specimens are compared with each other and the irreversible swelling is determined. Timber batches with values above the threshold ($> 2\%$) are then classified as failed (= 1) and batches below the threshold of 2% are classified as passed (= 0).

2.2 Data

The data was collected from the production process of the Swiss Wood Solutions AG described above. Since special attention was paid to the influence of the raw material on the quality of the final product, process parameters of the densification process were deliberately not included in the analysis. Thus, the labeled (D_i) and unlabeled data set (D_{ij}) included the following nine input features (x): Wood species (1_Maple; 2_Spruce; 3_Walnut; 4_Fir), and raw material properties (5_Mass [g]; 6_Length [mm]; 7_Width [mm]; 8_Height [mm]; 9_Density [kg/m³]). The response (y) was the binary quality classification (0 = pass, 1 = fail) of the densified wood according to the quality control protocol described in the previous chapter.

All the data sets used in this study, such as the training and hold-out test sets, had a binary class distribution of around 10% (1 = fail) and 90% (0 = pass).

The input features (x) were normalized between 0 and 1 in order to fit the requirements of the learning algorithms and to avoid the problem of prevailing high values which would have a greater influence on the loss function because of their scale and not for being more important than the other values.

2.3 Machine learning

2.3.1 Supervised learning

Since the pass-or-fail compliance determined during final inspection in manufacturing is a discrete variable, classification techniques should be used to predict such an outcome (Kotsiantis 2007).

As illustrated in Fig. 2, the limited labeled data set ($N = 166$) was split into a training data set ($n = 120$) and a hold-out test set ($n = 46$). The training dataset was used to train and adjust the hyperparameters of the random forest (RF) and support vector machine (SVM) ML algorithms using 10k cross-validation, whereas the hold-out test set was used to evaluate the generalization ability of the ML models. No detailed mathematical description of the common ML techniques applied in this study are given and reference is made to the corresponding literature (Bishop 2006).

Introduced by Vapnik (2000), SVMs are supervised learning models for classification and regression. The objective of the SVM algorithm is to find a hyperplane (decision boundary) in an N -dimensional space (N - number of features) that can distinctly classify the data points (Vapnik 2000; Schölkopf and Smola 2018).

RF is an ensemble learning method for classification that operates by constructing a multitude of decision trees trained with the bagging (bootstrapping aggregation) method. Generally, RF consists of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, \dots\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x (Breiman 2001; Meinshausen 2006).

2.3.2 Semi-supervised learning

SSL methods can improve learning performance by using additional unlabeled instances compared to supervised learning algorithms, which can use only labeled data. The definition of SSL was given by Chapelle et al. (2006): "Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. In addition to unlabeled data, the algorithm is provided with some supervision information – but not necessarily for all examples. Often, this information will be the targets associated with some of the examples. In this case, the data set $X = (x_i); i \in [n]$ can be divided into two parts: the points $X_L = (x_{L1}, \dots, x_L)$, for which labels $Y_L = (y_{L1}, \dots, y_L)$ are provided, and the points $X_U = (x_{U1}, \dots, x_{Uu})$, the labels of which are not known."

In the present work, the graph-based algorithm 'label spreading' was used which was first introduced by Zhou et al. 2003. The algorithm is inspired by a technique from experimental psychology called spreading activation network (Anderson 1983; Shrager Jeff et al. 1987) and diffusion kernels (Kondor and Lafferty 2002), and also from the published work on semi-supervised learning and clustering (Ng et al. 2002; Chapelle et al. 2003). Points in the dataset are connected in a graph based on their relative distances in the input space. The algorithm minimizes a loss function that has regularization properties and is often more robust to noise than other semi-supervised algorithms, such as label propagation (Zhu and Ghahramani 2002). The keynote of the method is to let every point iteratively spread its label information to its neighbors until a global stable state is achieved.

To spread labels across the nodes in the similarity graph, the iterative spreading propagation algorithm follows these steps (Zhou et al. 2003):

1. Form the affinity matrix W defined by $W_{ij} = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
2. Construct the matrix $S = D^{-1/2} W D^{1/2}$ in which D is a diagonal matrix with its (i, i) -element equal to the sum of the i -th row of W .
3. Iterate $F(t+1) = \alpha S F(t) + (1 - \alpha) Y$ until convergence, where α is a parameter in $(0,1)$.
4. Let F^* denote the limit of the sequence $\{F(t)\}$. Label each point x_i as label $y_i = \arg \max_{j \leq c} F^*_{ij}$.

As shown in Fig. 2, after labeling the unlabeled dataset, the new training dataset included $n = 400$. The same hold-out test set was used to ensure better comparability in terms of generalization ability of the ML algorithms.

2.4 Performance assessment and feature selection

In this work, neighborhood component feature selection analysis was carried out to identify relevant input parameters (Yang et al. 2012). The evaluation of the ML algorithms' performances was based on the following metrics:

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN)$$

$$\text{F1-score} = 2 \times TP / (2 \times TP + FP + FN)$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall (true positive rate)} = TP / (TP + FN)$$

$$\text{False positive rate (FPR)} = FP / (TN + FP)$$

Where TP = true positive, TN = true negative, FP = false positive, FN = false negative.

In addition, confusion matrices were used to visualize the performance of the ML algorithms and the corresponding Cohen-Kappa value (κ) was calculated (Cohen 1960):

$$\kappa = p_A - p_E / 1 - p_E$$

Where p_A is the observed relative agreement between two annotators, and p_E is the hypothetical probability of agreement by chance (with data labels randomly assigned). In particular, $\kappa = 1$ corresponds to the case of perfect agreement, whereas $\kappa = 0$ indicates no agreement other than what would be expected by chance.

The SSL algorithm 'label spreading' and the ML algorithms RF and SVM, as well as performance assessment and feature selection, were implemented with MATLAB R2020b using the Statistics and Machine Learning Toolbox.

3. Results And Discussion

Ensuring high product quality is essential for the long-term success of a producing company, particularly where the quality control is time-consuming and expensive. Therefore, the economic implementation of a comprehensive and reliable quality inspection is of utmost importance. This is especially true for small and medium-sized companies that do not have a highly automated production processes and produce small batches. In the tension between ever increasing requirements and new technological opportunities, the contribution of this work is the development of a semi-supervised learning (SSL) method using unlabeled data for improving predictive model-based quality inspection in industrial wood-based material production.

In recent years, SSL has emerged as an exciting new research direction in ML (Ouali et al. 2020; van Engelen and Hoos 2020). Such methods deal with the situation where few labeled training examples are available together with a significant number of unlabeled samples. In such a setting, SSL methods are more applicable to real-world applications where unlabeled data is readily available and easy to obtain, while labeled instances are often difficult, expensive, and time-consuming to collect, which is often the case in wood product manufacturing.

In this work, a particularly small dataset was used to determine the effect of SSL on the performance of the algorithms random forest (RF) and support vector machines (SVM), two common machine learning techniques for classification.

Choosing correct metrics when evaluating the performance of a classification model is crucial since each metric places varying emphasis on the overall accuracy, precision, recall, or agreement between model and ground truth for different class values. Therefore, for a comprehensive analysis, several metrics should be included depending on the user's requirements. The standard performance measure for classification models is accuracy. However, in cases where there is a class imbalance, which is common for quality-related industrial applications, accuracy can be misleading as a performance measure since it does not penalize misclassification of the minority (Schmitt et al. 2020). Therefore, complementary evaluation indices, such as recall and false positive rates as well as confusion matrices and Cohen-Kappa, were used in this study.

The comparison of the results based on the statistical performance measures showed that the augmentation of the training data with the SSL method had a positive impact on the result. As shown in Table 1, SSL could improve the performance of RF and SVM algorithms based on the evaluation indices as indicated by the results obtained for validation. In addition, the slightly lower standard deviations show that the results of the 10-fold cross-validation of SSL are more consistent, indicating that the learning process was more effective.

Table 1
Descriptive statistics for the training and validation (10-fold cross-validation).

	Training					Validation				
	Acc.	F1	Prec.	Recall	FPR	Acc.	F1	Prec.	Recall	FPR
RF	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.0 ± 0.0	0.90 ± 0.1	0.94 ± 0.0	0.92 ± 0.1	0.97 ± 0.0	0.60 ± 0.5
SSL-RF	0.99 ± 0.0	0.99 ± 0.0	0.99 ± 0.0	1.0 ± 0.0	0.0 ± 0.0	0.93 ± 0.0	0.96 ± 0.0	0.95 ± 0.0	0.97 ± 0.0	0.29 ± 0.2
SVM	0.93 ± 0.0	0.96 ± 0.0	0.93 ± 0.0	0.98 ± 0.0	0.54 ± 0.1	0.92 ± 0.1	0.95 ± 0.0	0.93 ± 0.1	0.98 ± 0.0	0.60 ± 0.5
SSL-SVM	0.96 ± 0.0	0.97 ± 0.0	0.95 ± 0.0	1.00 ± 0.0	0.24 ± 0.1	0.93 ± 0.0	0.96 ± 0.0	0.94 ± 0.0	0.99 ± 0.0	0.35 ± 0.2
Matlab parameters:										
RF: 'Trees': 100; 'Sample size': size(X_{train} , 1); 'Splitting rule': Randomly selected m_{try} ; 'MinLeafSize': 1; 'NumPredictorsToSample': 9										
SVM: 'KernelFunction': RBF; 'IterationLimit': $1e^6$; 'Solver': SMO;										
SSL: 'Alpha': 0.01; 'IterationLimit': $2e^3$; 'Tolerance': $1e^{-3}$; 'SimilarityGraph': 'knn'; 'NumNeighbors': 3'										
The default values were used for those parameters not indicated therein										

Generalization is the main goal of a ML algorithm, as the trained and validated algorithm should perform similarly when using new data during operation. In the present work, the hold-out test set was used to evaluate the generalization ability of ML algorithms. It was kept separate from the training and validation data sets and used only in the final evaluation of the model's performance. This independence is important to avoid bias and to properly represent the behavior of the model when new input data is used in the future.

Table 2
Descriptive statistics for the testing (hold-out test set).

Testing					
	Acc.	F1	Prec.	Recall	FPR
RF	0.87	0.93	0.90	0.95	0.67
SSL-RF	<i>0.96</i>	<i>0.98</i>	<i>0.98</i>	<i>0.98</i>	<i>0.17</i>
SVM	0.89	0.94	0.89	<i>1.00</i>	0.83
SSL-SVM	<i>0.93</i>	<i>0.96</i>	<i>0.95</i>	0.98	<i>0.33</i>
Matlab parameters:					
RF: 'Trees': 100; 'Sample size': size(X_{train} ,1); 'Splitting rule': Randomly selected m_{try} ; 'MinLeafSize': 1; 'NumPredictorsToSample': 9					
SVM: 'KernelFunction': RBF; 'IterationLimit': $1e^6$; 'Solver': SMO;					
SSL: 'Alpha': 0.01; 'IterationLimit': $2e^3$; 'Tolerance': $1e^{-3}$; 'SimilarityGraph': 'knn'; 'NumNeighbors': 3					
The default values were used for those parameters not indicated therein. Best values are highlighted.					

The results of the final testing are shown in Table 2 and Fig. 3. As judged by the indices, the SSL method outperforms the existing supervised learning techniques on the binary classification of wood product quality. As shown in Table 2, most of the evaluation indices are better in SSL compared to the traditional supervised learning method. In particular, the results in the underrepresented class (1 = fail) could be improved, as can be seen from the significantly lower value of the false positive rate. Furthermore, these results are confirmed by the confusion matrices and the Cohen-Kappa values (κ) (Fig. 3).

For example, κ was increased from 0.25 for SVM to 0.69 (SSL-SVM) and from 0.33 for RF to 0.81 (SSL-RF), respectively. A comparison of the two ML algorithms shows that SSL-RF provides better performance than SSL-SVM (Table 2 and Fig. 3). Based on the results of the testing with the hold-out test set, it can be clearly shown that the generalization performance of the ML algorithms RF and SVM could be enhanced by the SSL method 'label spreading'.

Besides 'label spreading', other SSL algorithms such as 'label propagation' (Zhu and Ghahramani 2002) and the rule-based 'Yarowsky algorithm' (Abney 2004) were also tested in this work. However, none of these algorithms revealed better results than the 'label spreading' method (data not shown). Such graph-based SSL algorithms have always been a popular subject for research with a vast number of successful models because of its wide variety of applicability (Chen et al. 2020; Li et al. 2020). In the present work, the graph-based 'label spreading' method significantly increased the information content of the training data set using the unlabeled dataset, especially in the underrepresented class where the supervised models performed poorly. The augmentation of the training data with information was also confirmed by the feature selection method neighborhood component analysis (NCA) (Fig. 4).

Whereas NCA with the non-augmented dataset could only identify two relevant features, such as the wood species maple and density, NCA with the larger dataset identified more relevant features (Fig. 4). NCA with the extended dataset revealed that within the different wood species, maple, spruce and walnut are the most relevant. This agrees well with the real values which show a failure rate of 25% for maple, followed by 10.4% for spruce, 10% for walnut, and 0% for fir. In addition, mass, height and density were identified as relevant features as was confirmed by the analysis of the real values. These values show that the wood products classified as low quality had, on average, about 26% less mass, 21.6% less height and 5.4% less density than the wood products classified as good.

Overall, it could be clearly shown that SSL can be used to augment small data sets to improve the generalization ability of ML algorithms such as RF. Although further research is needed to verify the results and scale the application of predictive model-based inspection for deployment, this work paves the way for future applications of semi-supervised ML for quality assessment. This is especially the case for small and medium-sized companies in the wood industry who can take advantage of this method because it achieves very good results even with very small batch sizes.

4. Conclusion

This paper discussed the need for an efficient quality inspection method for the wood industry, especially for small and medium sized companies where quality control is often time consuming, costly, and limited by small available datasets. We explored the potential of using semi-supervised learning (SSL) to improve the predictive performance of the random forest (RF) and support vector machines (SVM) for quality classification of real wood product data. Despite the use of a small dataset, two important findings were obtained:

- The SSL algorithm 'label spreading' was able to significantly increase the information content of the data, with the result that the generalization performance of the ML algorithms RF and SVM was considerably improved. This was especially true in the underrepresented class, as was shown by several evaluation indices such as F1-score, recall, precision and confusion matrices.
- It has been shown that a few wood-specific parameters such as wood species, density, and dimension are sufficient to train an ML algorithm based on SSL in order to make it capable of predicting the final product quality. This enables efficient quality control and would not only save time and costs, but also allow the raw material wood to be used for other applications with lower quality requirements.

In summary, although more work is needed in this direction, predictive model-based quality inspection based on SSL is a promising approach to make quality control processes more efficient and economical.

Declarations

Acknowledgments

The present study emerged from the research project: 37659.1 IP-ENG 'Machine learning for high-value wood products'. The authors express their gratitude to the Innosuisse – Swiss Innovation Agency for its financial support.

Conflicts of interest/Competing interests

On behalf of all authors, the corresponding author states that there is no conflict of interest and that there are no competing interests.

References

1. Abney S (2004) Understanding the Yarowsky Algorithm. *Comput Linguist* 30:365–395. <https://doi.org/10.1162/0891201041850876>
2. Anderson JR (1983) *The Architecture of Cognition*. Harvard University Press, USA
3. André N, Cho H-W, Baek SH et al (2008) Prediction of internal bond strength in a medium density fiberboard process using multivariate statistical methods and variable selection. *Wood Sci Technol* 42:521–534. <https://doi.org/10.1007/s00226-008-0204-7>
4. Bardak S, Tiryaki S, Bardak T, Aydın A (2016a) Predictive Performance of Artificial Neural Network and Multiple Linear Regression Models in Predicting Adhesive Bonding Strength of Wood. *Strength Mater* 48:811–824. <https://doi.org/10.1007/s11223-017-9828-x>
5. Bardak S, Tiryaki S, Nemli G, Aydın A (2016b) Investigation and neural network prediction of wood bonding quality based on pressing conditions. *Int J Adhes Adhes* 68:115–123. <https://doi.org/10.1016/j.ijadhadh.2016.02.010>
6. Barnes D (2001) A model of the effect of strand length and strand thickness on the strength properties of oriented wood composites. *For Prod J* 51:36
7. Bishop CM (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg
8. Breiman L (2001) Random Forests. *Mach Learn* 45:5–32. <https://doi.org/10.1023/A:1010933404324>
9. Chapelle O, Scholkopf B, Zien A (eds) (2006) *Semi-Supervised Learning*. The MIT Press
10. Chapelle O, Weston J, Schölkopf B (2003) Cluster Kernels for Semi-Supervised Learning. In: Becker S, Thrun S, Obermayer K (eds) *Advances in Neural Information Processing Systems*. MIT Press
11. Chen P, Ma T, Qin X et al (2020) Data-Efficient Semi-Supervised Learning by Reliable Edge Mining. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 9189–9198
12. Cohen J (1960) A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 20:37–46. <https://doi.org/10.1177/001316446002000104>
13. Ehrhart T, Palma P, Schubert M et al (2021) Predicting the strength of European beech (*Fagus sylvatica* L.) boards using image-based local fibre direction data. *Wood Sci Technol*. <https://doi.org/10.1007/s00226-021-01347-w>

14. Esteban LG, Fernández FG, de Palacios P (2011) Prediction of plywood bonding quality using an artificial neural network. 65:209–214. <https://doi.org/doi:10.1515/hf.2011.003>
15. Fountoulaki A, Karacapilidis N, Manatakis M (2011) Augmenting statistical quality control with machine learning techniques: an overview. *Int J Bus Syst Res* 5:610–626. <https://doi.org/10.1504/IJBSR.2011.043162>
16. Gupta A, Jordan P, Pang S (2007) Modelling of the Development of the Vertical Density Profile of MDF during Hot Pressing. *Chem Prod Process Model*. <https://doi.org/doi:10.2202/1934-2659.1075.2>:
17. Kondor RI, Lafferty JD (2002) Diffusion Kernels on Graphs and Other Discrete Input Spaces. In: *Proceedings of the Nineteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp 315–322
18. Kotsiantis SB (2007) Supervised Machine Learning: A Review of Classification Techniques. In: *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real World AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies*. IOS Press, NLD, pp 3–24
19. Li S, Liu B, Chen D et al (2020) Density-Aware Graph for Deep Semi-Supervised Visual Recognition. *CoRR abs/2003.13194*
20. Meinshausen N (2006) Quantile Regression Forests. *J Mach Learn Res* 7:983–999
21. Ng A, Jordan M, Weiss Y (2002) On Spectral Clustering: Analysis and an algorithm. In: Dietterich T, Becker S, Ghahramani Z (eds) *Advances in Neural Information Processing Systems*. MIT Press
22. Ouali Y, Hudelot C, Tami M (2020) An Overview of Deep Semi-Supervised Learning. *ArXiv abs/2006.05278*
23. Sandberg D, Haller P, Navi P (2013) Thermo-hydro and thermo-hydro-mechanical wood processing: An opportunity for future environmentally friendly wood products. *Wood Mater Sci Eng* 8:64–88. <https://doi.org/10.1080/17480272.2012.751935>
24. Schmitt J, Bönig J, Borggräfe T et al (2020) Predictive model-based quality inspection using Machine Learning and Edge Cloud Computing. *Adv Eng Inform* 45:101101. <https://doi.org/10.1016/j.aei.2020.101101>
25. Schölkopf B, Smola AJ (2018) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press
26. Schubert M, Kläusler O (2020) Applying machine learning to predict the tensile shear strength of bonded beech wood as a function of the composition of polyurethane prepolymers and various pretreatments. *Wood Sci Technol* 54:19–29. <https://doi.org/10.1007/s00226-019-01144-6>
27. Schubert M, Luković M, Christen H (2020) Prediction of mechanical properties of wood fiber insulation boards as a function of machine and process parameters by random forest. *Wood Sci Technol* 54:703–713. <https://doi.org/10.1007/s00226-020-01184-3>
28. Shrager Jeff H, Tad, Huberman Bernardo A (1987) Observation of Phase Transitions in Spreading Activation Networks. *Science* 236:1092–1094. <https://doi.org/10.1126/science.236.4805.1092>

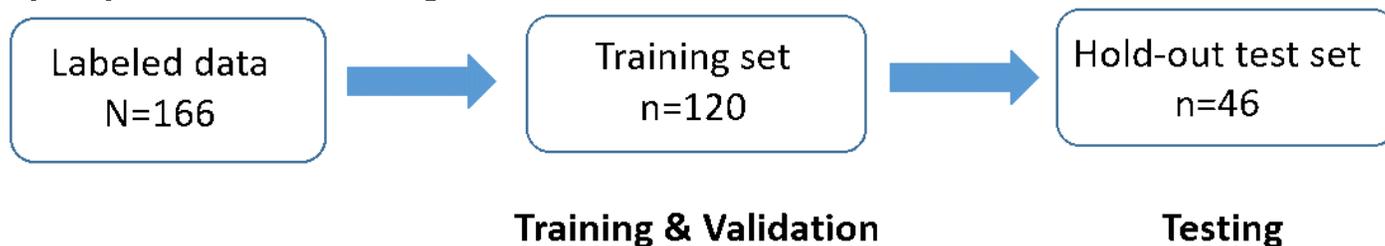
29. van Engelen JE, Hoos HH (2020) A survey on semi-supervised learning. Mach Learn 109:373–440. <https://doi.org/10.1007/s10994-019-05855-6>
30. Vapnik V (2000) The Nature of Statistical Learning Theory. Springer, New York
31. Yang W, Wang K, Zuo W (2012) Neighborhood Component Feature Selection for High-Dimensional Data. J Comput. <https://doi.org/10.4304/jcp.7.1.161-168>. 7:
32. Zhou D, Bousquet O, Lal T et al (2003) Learning with Local and Global Consistency. In: Thrun S, Saul L, Schölkopf B (eds) Advances in Neural Information Processing Systems. MIT Press
33. Zhu X, Ghahramani Z (2002) Learning from labeled and unlabeled data with label propagation. In: CMU CALD tech report CMU-CALD-02-107

Figures

Figure 1

Schematic procedure detailing the manufacturing process of the densified wood.

a) Supervised learning



b) Semi-Supervised learning



Figure 2

Schematic procedure of the machine learning workflow using **a)** supervised learning and **b)** semi-supervised learning.

Figure 3

Confusion matrices with Cohen-Kappa values (κ) using **a)** RF, **b)** SSL-RF, **c)** SVM, and **d)** SSL-SVM. Binary classification: 0=pass; 1=fail.

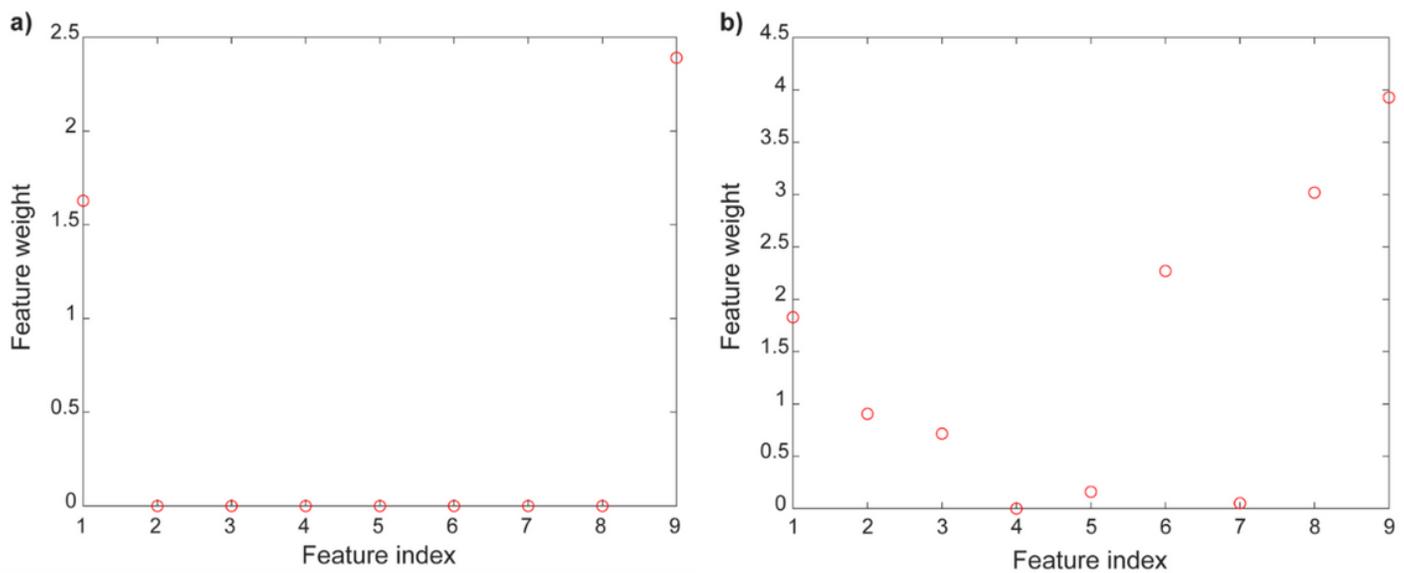


Figure 4

Neighborhood component analysis (NCA) using **a)** RF and **b)** SSL-RF. Feature index: 1_Maple; 2_Spruce; 3_Walnut; 4_Fir; 5_Mass [g]; 6_Length [mm]; 7_Width [mm]; 8_Height [mm]; 9_Density [kg/m³].