Development of an Efficient Cement Production Monitoring System Based on the Improved Random Forest Algorithm

Hanane Zermane (h.zermane@univ-batna2.dz)
University of Batna

Abbes Drardja
University of Batna

Research Article

Keywords: Intelligent automation, Smart manufacturing, Machine learning, Random Forest, Prediction, Classification.

Posted Date: September 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-914830/v1

Version of Record: A version of this preprint was published at The International Journal of Advanced Manufacturing Technology on February 16th, 2022. See the published version at https://doi.org/10.1007/s00170-022-08884-z.

Abstract

Strengthening production plants and process control functions contribute to a global improvement of manufacturing systems because of their cross-functional characteristics in the industry. Companies established various innovative and operational strategies and there is increasing competitiveness among them and increase companies' value. Machine Learning (ML) techniques have become an intelligent enticing option to address industrial issues in the current manufacturing sector since the emergence of Industry 4.0, and the extensive integration of paradigms such as big data, cloud computing, high computational power, and enormous storage capacity. Implementing a system that can identify faults early to avoid critical situations in the line production and environment is crucial. Therefore, one of the powerful machine learning algorithms is Random Forest (RF). The ensemble learning algorithm is performed to fault diagnosis and SCADA real-time data classification and predicting the state of the line production. Random Forests proved to be a better classifier with a 95% accuracy. Comparing to the SVM model, the accuracy is 94.18%, however, the K-NN model accuracy is about 93.83%, an accuracy of 80.25% is achieved using the logistic regression model, finally, about 83.73% is obtained by the decision tree model. The excellent experimental results achieved on the Random Forest model showed the merits of this implementation in the production performance, ensuring predictive maintenance, and avoid wasting energy.

1. Introduction

For all countries, manufacturing is a major sector and is a vital gauge of the financial status. Despite sophisticated production, many developed countries are trying to discover new opportunities and redesign the manufacturing industries to acquire unconquerable positions. It is possible in the presence of technical progress and advancement of automation and computing to modern factories.

Technology and intelligent methods, in turn, help us achieve enterprise goals through the use of artificial intelligence. A variety of statistical and artificial intelligence (AI) approaches are developed for modeling production line processes in different fields of industry. This vast domain contains several branches, including machine learning. Machine learning Algorithms are increasingly more common in many applications and are extremely beneficial for a typical operator to utilize.

Machine learning is the knowledge of making computers comprehend and act like people, and provide facts and information without having to program precisely. It may be classed in supervised models of learning, semi-supervised models of learning, and non-supervised models. To execute a given task, machine learning employs several algorithms or models. These algorithms are mostly used in several fields, including medical prediction (1–3), psychology (4, 5), object recognition (6–8), quality monitoring (9) industry (10–12), and many other domains.

However, our technical work focus is on the application of several algorithms to anticipate and classify diverse sorts of industrial process failures utilizing machine learning approaches. Random forest, SVM,

and K-NN algorithms are performed to innovate the supervised monitoring system based on an efficient predictive model. The methodology combines SCADA real-time data as inputs to the machine learning model to predict the state of the line production if it is in good functioning or bad.

The remainder of the article is structured as follows. Section 2 contains the contribution and the motivation of the study. The previous relevant studies concerning the technique of prediction are discussed in Sect. 3 of methods. Section 4 provides the materials and the data utilized in the proposed approach, followed by Sect. 5 of results and measuring performance. The last comments and the future works of this study are provided in the concluding part.

2. Contribution And Motivation

Today's industrial production system is highly complicated due to the enormous demand for industrial products, which have become an integral aspect of consumers' lives. Manufacturers were compelled to utilize technology and intelligent methods for these systems to improve production, minimize production disruptions, simplify the supervisory process, decrease maintenance costs as much as possible, satisfy customers, ensure equipment prevention, and save human lives.

We are in favor of using Random Forest models in this article. Random forest achieves higher forecasting performance than conventional regression. It protects against overfitting and detects interactions between variables. Due to its benefits in comparison to other statistical methods, Random forest is a popular instrument in a wide range of sectors including image recognition, banking, disease prevention, and patient health planning. However, Random forest is utilized somewhat less commonly within real-time data in complex and critical industrial processes.

3. Methods

Because it is relevant to such a wide range of use cases, machine learning is generating a lot of interest. Classification is a supervised learning method in machine learning in which the computer program learns from the data input given to it and then utilizes this learning to categorize new observations. Choosing an algorithm is a key stage in the machine learning process, so ensure it genuinely matches the problem's use case (13, 14).

3.1 Ensemble learning techniques

Ensemble learning techniques are originally proposed for classification tasks in a manner of supervised learning in 1965 (15, 16). The decision tree algorithm is a supervised learning method that is used for categorizing problems and is one of the most common machine learning algorithms in use today. Given a set of previously classified data, a decision tree is used to categorize subsequent observations. Decision trees are a sequential model that quickly and cohesively connects a series of fundamental tests in which a numeric characteristic is compared to a threshold value in each test.

Learning trees are a frequent starting point for ensemble techniques. Strong learners made up of several trees are referred to be "forests." The trees that make up a forest might be shallow with few depths or deep with many depths, while if it is not fully grown it is a lot of depths. Deep trees, on the other hand, have low bias but a large variance and so are appropriate options for bagging methods that are primarily concerned with lowering variance.

Random Forests (RF) are ensemble learning algorithms that rely on the combination of several decision tree-based components grown from a certain amount of randomness. Each tree in the forest utilizes a different and randomly selected set of predictor variables, which is where the term “random” comes from. The idea has been exploited during the ’90s as a random subspace method for constructing ensembles of decision trees (Bagging, Boosting, and randomization) (17–20). However, the formal definition and use of Random Forests (RF) have been announced by Breiman in 2001 for classification and regression problems.

Their robustness and flexibility made them useful in modeling the input-output relationship (21). The division criteria and optimization of tree sizes are important to a great much of the prior attention given to decision trees. Rarely is a problem resolved between overfitting and maximum accuracy. A strategy for building the decision-tree board classification is provided, maintaining the highest accuracy on training data and increasingly increasing complexity in terms of generalization accuracy (19).

The random forest algorithm is robust against overfitting compared to many other classifiers, including discriminant analysis, support vector machines, and neural networks (22). This idea is proved generally by improved results achieved by the random forest technique. Its improving and satisfying results make it applied in different fields including industry and manufacturing, pattern recognition, risk identification, and several other fields (22–26).

3.2 Support Vector Machine

The Support Vector Machine (SVM) is a supervised learning algorithm proposed by Vapnik (27, 28). SVM is built on decision planes by constructing hyperplanes in two or multidimensional space, which determine decision boundaries that divide and distinguish between a collection of instances belonging to various classes. It may be utilized for regression as well as classification (29, 30). It has a strong theoretical foundation and achieved excellent empirical success.

The accurate categorization of new objects or test instances based on the available train instances is referred to as optimal separation. The mapping process is the mathematical function known as kernels that are utilized to map objects (31). SVM employs an iterative method to minimize the error function to identify the best separable hyperplane and maximizes the margin between classes. Several studies applied SVM in different other fields (3, 32, 33), especially, in industry and automation (13, 34–37).

3.3 K-Nearest Neighbor

The k-nearest neighbor technique or, sometimes, memory-based methods (15). They are useful for classification and regression problems. In reality, it is more commonly employed to address classification problems in the Data Science field (1). It's a straightforward algorithm that saves all existing instances and classifies any new cases based on a majority vote of its *k* neighbors. The K-NN obtains *k* neighbors of the test pattern data from the training data. If a majority of these *k* neighbors are from a class then the object is assigned to it. Otherwise, it is assigned to the other class and so on (38). The nearest-neighbor technique has been used for several purposes (3, 30, 39).

3.4 Logistic Regression

The sigmoid function is used in logistic regression to evaluate data and predict discrete classes that exist in a dataset. Although logistic regression appears to be a kind of linear regression, it is a classification approach. Logistic regression predicts discrete classes, whereas linear regression handles numerical equations and generates numerical predictions to detect connections between variables. Logistic regression is used to draw a borderline in the input vector's feature distribution domain, and the region formed by the borderline(s) has a certain class and forms a group.

The parameters forming the boundary line(s) are produced by learning, where the input data fits into the region indicates its class at the inference (40). It is generally used for binary classification to predict two discrete groups after applying a transformation function. The sigmoid function is used to determine the output and transform numerical values into a probability expression between 0 and 1 (41). Several studies discussed logistic regression applications (10, 42, 43).

4. Materials And Data

In this study, in an East Algerian cement plant of Ain Touta (SCIMAT), the workshop of a raw mill is selected. Throughout the production line, the product passes via a collection of electrical, mechanical, automated equipment and a large number of other devices to process and maintain this operation and keep it on functionality mode if system needs. The overall procedure of the workshop on raw mills is shown in Fig. 1.

The dataset is collected from the cement factory. It contains 20 features and one target class that indicates if the process line is good (1) or in a non-functioning state. The number of samples is about 38,187 collected during the running of the production line for 6 trimesters in 2018/2019. Dataset classes in our case are the existence or absence of an alarm default. We recorded all sensor settings that can be utilized as training data for the machine learning system. The line production has 76.93% of good functioning, however about 23.06%, is in degradation mode. Accordingly, the economic policies of society might have a catastrophic impact.

The implementation of this work is based on Python language (version 3.8) under the Anaconda environment. Python incorporates several libraries and packages including Scikit-learn that makes use of this rich environment to deliver cutting-edge implementations of several well-known machine learning

techniques, all while retaining an easy-to-use interface that is strongly linked with the Python language. This addresses the rising need for statistical data analysis by non-specialists in the software and online sectors, as well as areas outside than computer science, such as biology or physics (44).

5. Results And Discussion

5.1 Data analysis

There are different features that the system utilizes as associated factors. The most important ones which have a high effect on the supervision system are Transporter Tape flow M01I01 (Sum of the workshop feeders quantity (A02, D02, and E02)), M01P1, M01T1, M01P3, M01T3, M01X1, J01J1, and S01M1I01. The distribution of all studied factors according to the state of the system is displayed in Fig. 2.

The heatmap presented in Fig. 3 illustrated the correlations between the different attributes of the selected dataset. All characteristics/features given in the dataset are very less correlated with each other. This implies that we have to include all the characteristics because we can only eliminate the characteristics where the correlation of two or more characteristics is very high.

Regarding the plot displayed in Fig. 2, several observations are noted. Factors indicators that influence negatively the state of the line production are the crusher acoustic indicator (M01X1) and the operator sp03 (QCXH20). Figure 4 displayed the different influential factors on the functioning of the line production.

5.2 Model Construction and Results

In our experiments, the data set is split into two parts, respectively as the training set (67%) and testing set (33%). The training set is used to train the prediction model while the testing set is used to validate the performance of the trained model. More specifically, the accuracy of predictions on the testing set, the core and key of further applications, plays an essential part in the validation and directly affects whether it could be used. During the first stage, the algorithms were applied to a training dataset and the performance was evaluated. Later, the algorithms were applied to a testing dataset to make predictions.

In the first model, the process of building decision trees involves asking a question of each instance and then continuing to split. When multiple features decide the target value of a particular instance, which feature should be chosen as the root node to start the split process, in what order should we continue to choose features with each new division of a node. Hither is a need to measure the informative character of the features and to use the feature with the most information as a feature to divide the data. This information is given by a measure called "information gain". Therefore, understanding the entropy of the dataset is indispensable.

The labeled dataset is trained using Random Forest Classifier for 200 decision trees per estimate. Figure 5 displayed a part of the decision forest that displays a decision about the elevator load that indicates that with its value (57.5%), the system state is in good functioning based on the majority of “class 1” which is 11 against 7 belonging to “class 0”.

5.3 Model evaluation

Results demonstrate the overall system performance enhancement in predicting bearing failure when modeled data are included with SCADA data. Based on data from the cement plant, the performances of different machine-learning models on unseen data are then evaluated using industry-standard metrics including training accuracy, testing accuracy, sensitivity, and specificity. Evaluation results are collected in Table 1.

Table 1
The evaluation metrics of the predictive model.

Metrics	Value
Training Accuracy	0.9498058899650994
Testing Accuracy	0.9484913621526948
Sensitivity	0.9796095444685466
Specificity	0.9414976599063962
AUC Score	0.9844503878559198

Other metrics such as accuracy, precision, recall, and F1 score, and AUC (area under the receiver operating characteristic curve) are utilized to evaluate the model. The improvement is in terms of accuracy, precision, recall, F1 score, and AUC Score based on the best modeling case in this study. The accuracy is the proportion of right predictions made by the entire model divided by the total number of samples used to test the model. It is the total number of correct predictions divided by the total number of assessment samples.

The recall is the percentage of items accurately predicted in a class. It is a relationship between the number of instances properly predicted and the total of correct predictions and missed right predictions for that class. The precision is the proportion of valid predictions for each class divided by the number of evaluation samples for each class. The weighted average of accuracy and recall is the F1-score. The support is the number of occurrences of each class in the true output. The model's evaluation report is collected in Table 2.

Table 2
The random forest model evaluation report.

	precision	recall	f1-score	support
Class 0	0.98	0.79	0.87	2858
Class 1	0.94	1.00	0.97	9703
accuracy			0.95	12561
macro avg	0.96	0.89	0.92	12561
weighted avg	0.95	0.95	0.95	12561

After validating our model we check the confusion matrix to understand how the model performs for each label. The matrix revealed to present values and percentage of prediction illustrated in the matrix in Fig. 6. (a). The ROC curve of the implemented model and every point is above the no-skill line is showed in Fig. 6. (b).

In terms of evaluation indicators, we test the classification accuracy in different classifiers to evaluate the performance of the proposed scheme. Comparing to the SVM model, the accuracy is 94.18%, however, the K-NN model accuracy is about 93.83%, an accuracy of 80.25% is achieved using the logistic regression model, finally, about 83.73% is obtained by the decision tree model.

6 Conclusion

The concept is that machine learning has been integrated into the industry and that this theory applies to a practical industrial project. For the learning model, we adopted the RF algorithm and for the industrial actual project, SCIMAT society. This paper provides a system with the ability to classify data using the Random Forest classifier. A comparison is carried out with Support Vector Machine (SVM) and K-NN classifiers. After analyzing the results of several experiments of compared machine learning algorithms that were applied to the dataset, it was observed that overall Random Forest was the best algorithm to be used.

When the results of different classifiers were examined, the accuracies of these classifiers ranged between 80% and 95%. The accuracies were found to be for the Random Forests proved to be a better classifier with a 95% accuracy of correct classification rate, SVM model about 94%, however, the K-NN model accuracy is about 93.83%. an accuracy of 80.25% is achieved using the logistic regression model and about 83.73% is obtained by the decision tree model.

The learning model and architecture presented improve control flexibility. The capacity to handle data and a great deal of information to boost productivity, minimize maintenance costs, and several other advantages. In the future, we can use test the presented dataset with other improved machine learning algorithms to provide better efficiency.

Declarations

Acknowledgments

The authors would like to take this opportunity to express their profound gratitude and deep regard for everyone that helped in this work. This paper and the research behind it would not have been possible without the exceptional support of colleagues with provided insight and expertise that greatly assisted the research and the technical assistance. The authors wish to extend their special thanks and greet everyone, who has supported the creation of this work. Great thanks to The DGRSDT (General Directorate of Scientific Research and Technological Development of Algeria).

Funding No funds, grants, or other support were received.

Conflicts of interest/Competing interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

Availability of data and material Not available.

Code availability Not available.

Ethics approval Not applicable.

Consent to participate/Consent for publication Not applicable.

Authors' contributions HZ and AD led the study by being involved in data collection, performing the statistical analysis, coding, interpreting the results, conceiving, and designing the article, and providing intellectual content revisions and suggestions for clarity and precision on the subject matter.

References

1. Reddy DN, Priyanka R, Sanjana S, Bagali SM, Sadiya S. Machine Learning Algorithms for Detection : A Survey and Classification. 2021;12(10):3468–74.
2. Elaziz MA, Hosny KM, Salah A, Darwish MM, Lu S, Sahlol AT. New machine learning method for image-based diagnosis of COVID-19. PLoS One [Internet]. 2020;15(6). Available from: http://dx.doi.org/10.1371/journal.pone.0235187
3. Kale R, Shitole S. Analysis of Crop disease detection with SVM, KNN and Random forest classification. Inf Technol Ind. 2021;9(1):364–72.
4. Rozek DC, Andres WC, Smith NB, Leifker FR, Arne K, Jennings G, et al. Using Machine Learning to Predict Suicide Attempts in Military Personnel. Psychiatry Res [Internet]. 2020;294(October):113515. Available from: https://doi.org/10.1016/j.psychres.2020.113515
5. Fife DA, Onofrio JD. Common , Uncommon , and Novel Applications of Random Forest in Psychological Research.

6. Li H, Lin Z, Shen X, Brandt J, Hua G. A convolutional neural network cascade for face detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2015;07-12-June:5325–34.
7. Li K, Jin Y, Akram MW, Han R, Chen J. Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis Comput [Internet]. 2020;36(2):391–404. Available from: https://doi.org/10.1007/s00371-019-01627-4
8. Lin K, Zhao H, Lv J, Li C, Liu X, Chen R, et al. Face Detection and Segmentation Based on Improved Mask R-CNN. Discret Dyn Nat Soc. 2020;2020.
9. Ismail M, El-assal A. Utilization of Machine Learning Techniques for Quality Monitoring and Prediction. 2021;(Ml):4830–9.
10. Jung H, Jeon J, Choi D, Park AJY. Application of machine learning techniques in injection molding quality prediction: Implications on sustainable manufacturing industry. Sustain. 2021;13(8).
11. Mokhtari S, Abbaspour A, Yen KK, Sargolzaei A. A machine learning approach for anomaly detection in industrial control systems based on measurement data. Electron. 2021;10(4):1–13.
12. Hemalatha N-R, Karishma M, Shanmuga Sundari N, Anandhavalli D. Performance Analysis of Spinning Machines Using Machine Learning. In: International Conference on Smart Data Intelligence (ICSMDI 2021) [Internet]. 2021. p. 1–10. Available from: https://ssrn.com/abstract=3851252
13. Usuga Cadavid JP, Lamouri S, Grabot B, Pellerin R, Fortin A. Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. J Intell Manuf [Internet]. 2020;31(6):1531–58. Available from: https://doi.org/10.1007/s10845-019-01531-7
14. Mohana-Priya T, Punithavall M, Rajesh-Kanna R. Conceptual Review on Machine Learning Algorithms for Classification Techniques. Int J Sci Res Comput Sci Eng Inf Technol. 2021;7(1):215–22.
15. Nilsson NJ. Introduction To Machine learning. Machine Learning. Stanford: Stanford University; 1998. 188 p.
16. Yang Y. Ensemble Learning. Temporal Data Min Via Unsupervised Ensemble Learn. 2017;35–56.
17. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
18. Amit Y, Geman D. Shape Quantization and Recognition with Randomized Trees. Neural Comput. 1997;9(7):1545–88.
19. Ho TK. The Random Subspace Method for Constructing Decision Forests. IEEE Trans Pattern Anal Mach Intell. 1998;20(8):832–44.
20. Dietterich TG. Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn. 2000;40(2):139–57.
21. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
22. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.
23. Kiangala SK, Wang Z. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in

an Industry 4.0 environment. Mach Learn with Appl [Internet]. 2021;4(December 2020):100024. Available from: https://doi.org/10.1016/j.mlwa.2021.100024

24. Dantone M, Gall J, Fanelli G, Van Gool L. Real-time facial feature detection using conditional regression forests. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2012;2578–85.
25. Jia J, Xu Y, Zhang S, Xue X. The facial expression recognition method of random forest based on improved PCA extracting feature. ICSPCC 2016 - IEEE Int Conf Signal Process Commun Comput Conf Proc. 2016;(1):0–4.
26. Liu LJ, Shen WK, Zhu JM. Research on Risk Identification System Based on Random Forest Algorithm-High-Order Moment Model. Complexity. 2021;2021.
27. Corinna C, Vapnik V. Support-Vector Networks. Mach Leaming. 1995;20:273–97.
28. Vapnik V. Pattern Recognition-Statistical Learning Theory. Canada: Wiley; 1998. 1–760 p.
29. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15:3133–81.
30. Rozenberg G, Thomas B, Kok JN. Handbook of Natural Computing. Handbook of Natural Computing. Berlin, Heidelberg: Springer; 2012. 2105 p.
31. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Support Vector Machines (SVM). Cambridge University Press. London; 2000. 93–124 p.
32. Lazri M, Labadi K, Brucker JM, Ameur S. Improving satellite rainfall estimation from MSG data in Northern Algeria by using a multi-classifier model based on machine learning. J Hydrol [Internet]. 2020;584(July 2019):124705. Available from: https://doi.org/10.1016/j.jhydrol.2020.124705
33. Rahab H, Zitouni A, Djoudi M. SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments. Adv Intell Syst Comput. 2018;662:139–49.
34. Zermane H, Kasmi R. Intelligent industrial process control based on fuzzy logic and machine learning. Int J Fuzzy Syst Appl. 2020;9(1).
35. Ahmadi SH, Khosrowjerdi MJ. Fault detection Automation in Distributed Control Systems using Data-driven methods : SVM and KNN. TechRxiv Prepr. 2021;0–7.
36. Yin G, Zhang YT, Li ZN, Ren GQ, Fan HB. Online fault diagnosis method based on Incremental Support Vector Data Description and Extreme Learning Machine with incremental output structure. Neurocomputing. 2014;128:224–31.
37. Suykens JAK. Support vector machines and kernel-based learning for dynamical systems modeling [Internet]. Vol. 42, IFAC Proceedings Volumes. IFAC; 2009. 1029–1037 p. Available from: http://dx.doi.org/10.3182/20090706-3-FR-2004.00171
38. Murty MN. SPRINGER BRIEFS IN COMPUTER SCIENCE Support Vector Machines and Perceptrons Learning, Optimization, Classification, and Application to Social Networks. 2016. 103 p.
39. Dino HI, Abdulrazzaq MB. Facial Expression Classification Based on SVM, KNN and MLP Classifiers. 2019 Int Conf Adv Sci Eng ICOASE 2019. 2019;70–5.

40. Takano S. Thinking Machines Machine Learning and Its Hardware Implementation. Japan: Academic Press is an imprint of Elsevier; 2017. 324 p.
41. Theobald O. Machine Learning For Absolute Beginners. 2nd ed. Scatter Plot Press. Brooklyn: Scatter Plot Press; 2017. 1–128 p.
42. Desai A, Guo Y, Sheng S, Phillips C, Williams L. Prognosis of wind turbine gearbox bearing failures using SCADA and modeled data. Proc Annu Conf Progn Heal Manag Soc PHM. 2020;12(1):1–10.
43. Merghadi A, Abderrahmane B, Tien Bui D. Landslide susceptibility assessment at Mila basin (Algeria): A comparative assessment of prediction capability of advanced machine learning methods. ISPRS Int J Geo-Information. 2018;7(7).
44. Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn. GetMobile Mob Comput Commun. 2015;19(1):29–33.

Figures

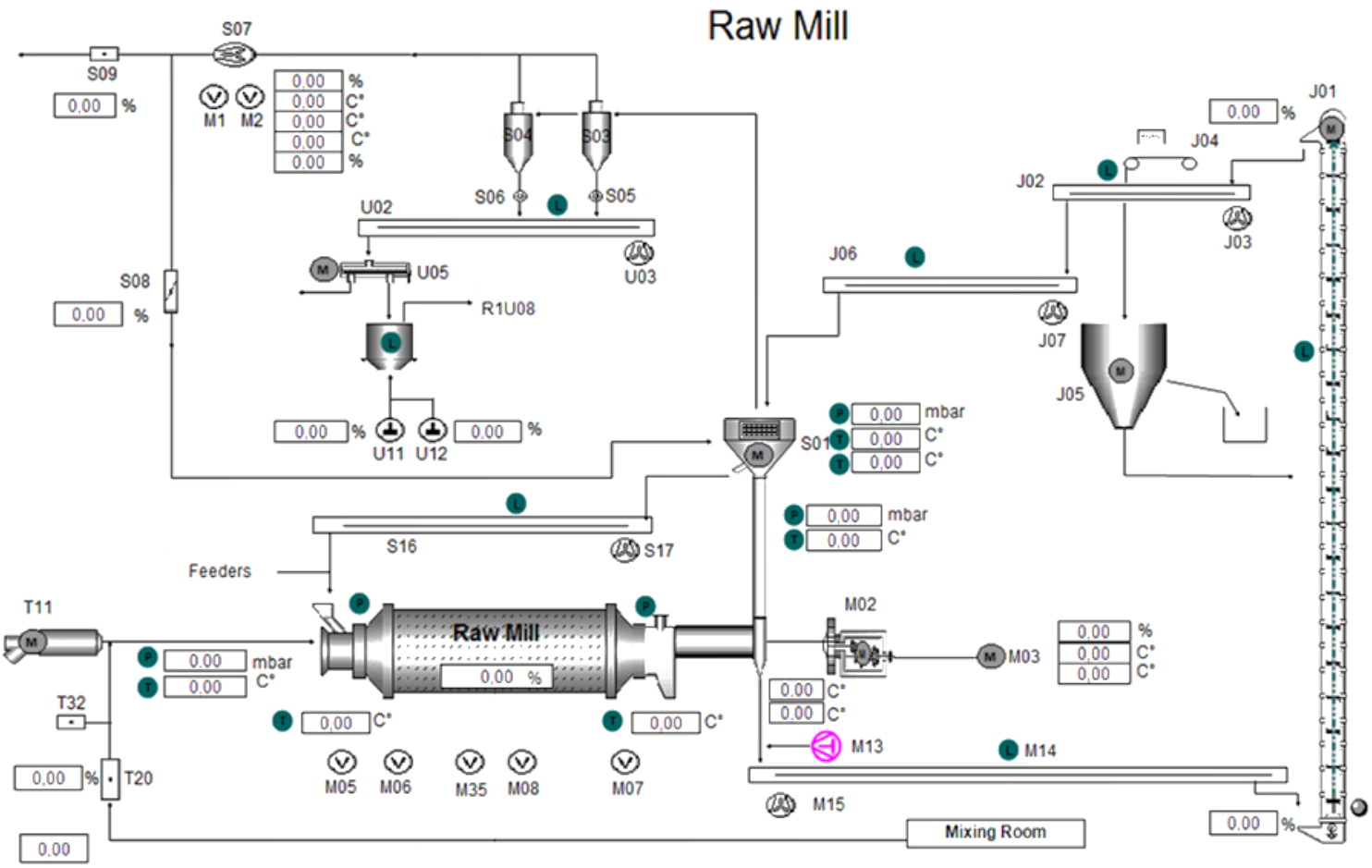

Figure 1

Process of the raw mill workshop.

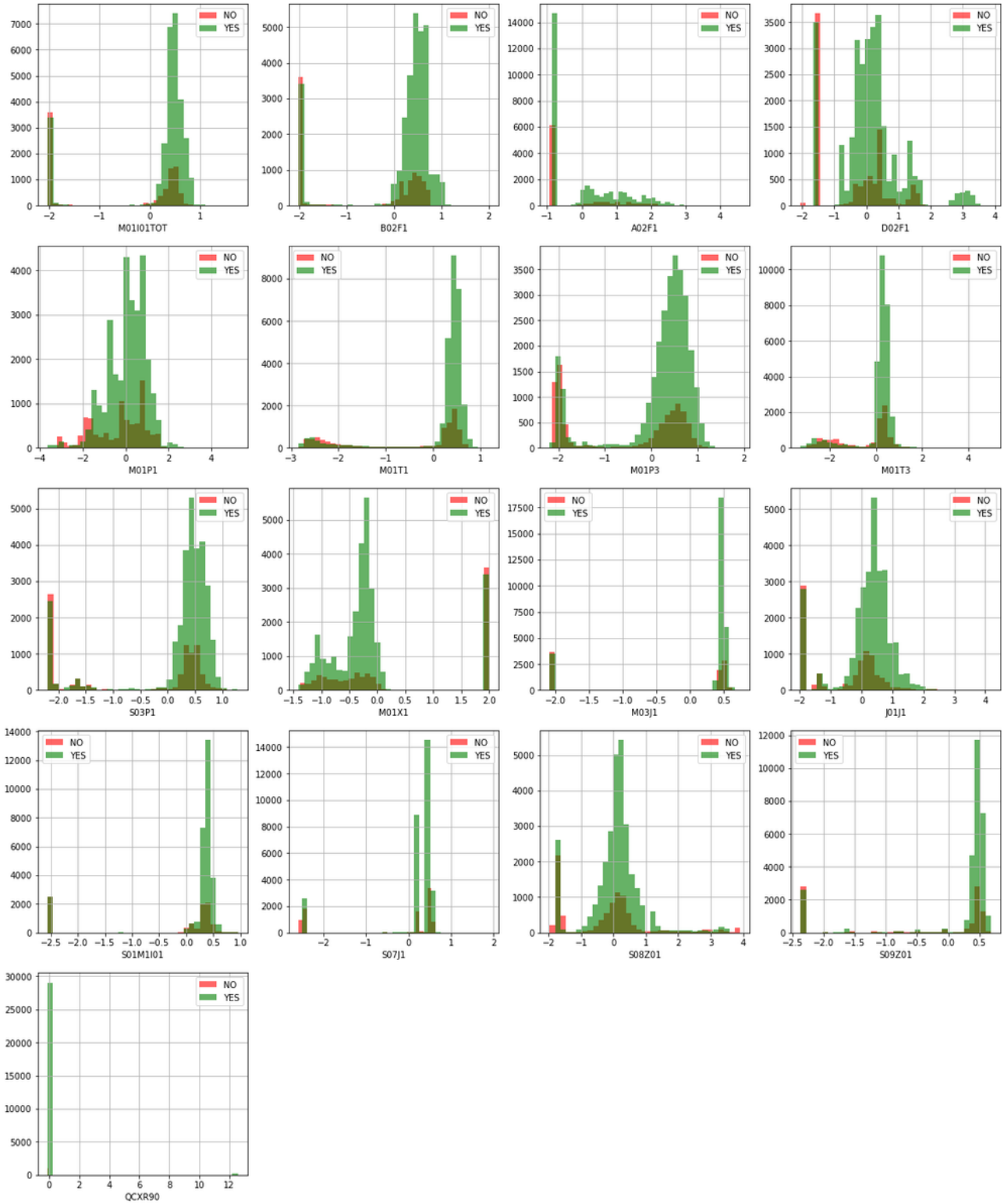

Figure 2

Distribution of factors according to the state of the system.

Figure 3

The heatmap of features.

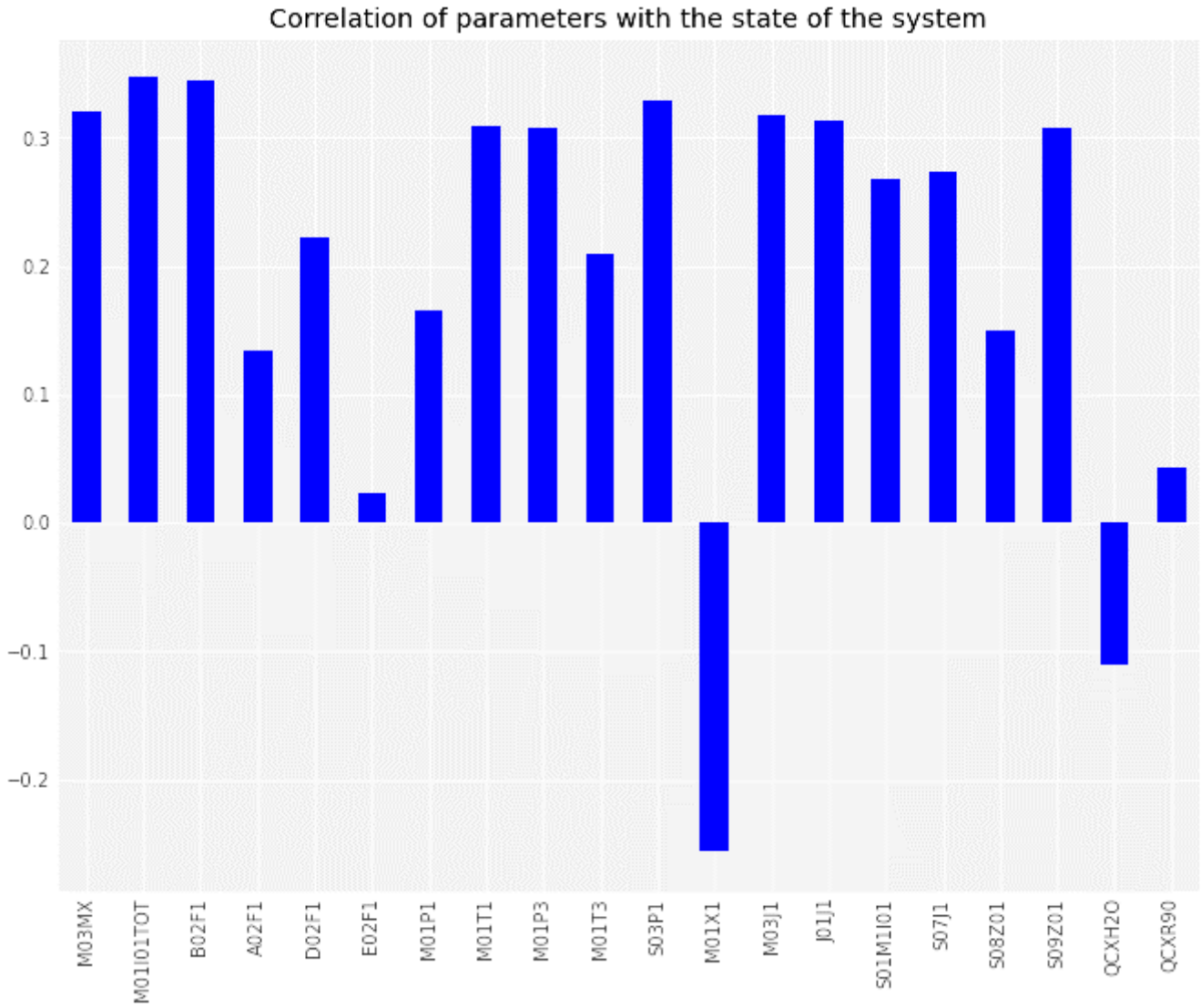

Figure 4

Influential factors.

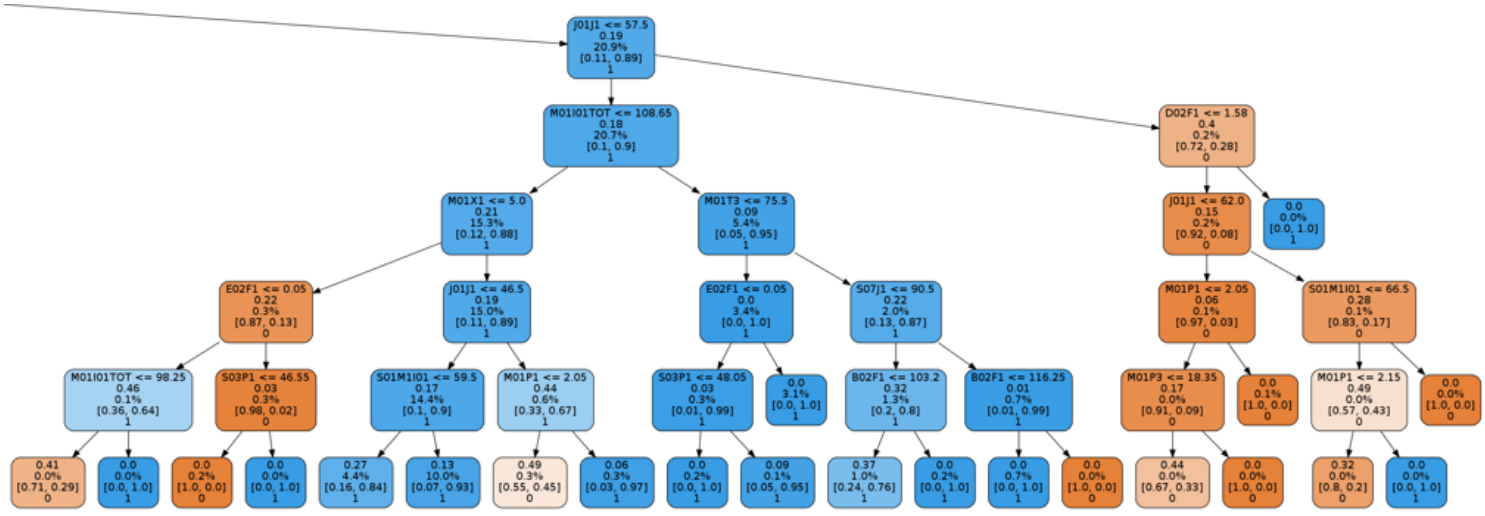

Figure 5

Part of the random forest.

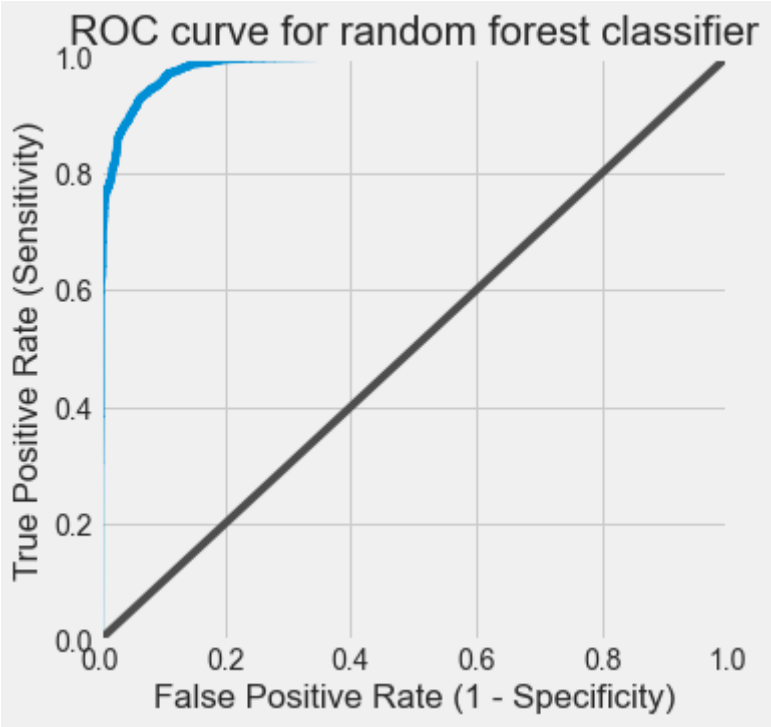

Figure 6

Evaluation of the random forest model.