Development of an efficient cement production monitoring system based on the improved random forest algorithm

Hanane Zermane1 · Abbes Drardja1

Received: 17 September 2021 / Accepted: 4 February 2022 / Published online: 16 February 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

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; there is increasing competitiveness among them and increasing companies’ value. Machine learning (ML) techniques 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 and high computational power. Implementing a system able to identify faults early to avoid critical situations in the production line and its environment is crucial. Therefore, powerful machine learning algorithms are performed for fault diagnosis, real-time data classification, and predicting the state of functioning of the production line. Random forests proved to be a better classifier with an accuracy of 97%, compared to the SVM model’s accuracy which is 94.18%. However, the K-NN model’s accuracy is about 93.83%. An accuracy of 80.25% is achieved by the logistic regression model. About 83.73% is obtained by the decision tree’s model. The excellent experimental results reached on the random forest model demonstrated the merits of this implementation in the production performance, ensuring predictive maintenance and avoiding wasting energy.

Keywords Intelligent automation · Smart manufacturing · Machine learning · Random forest · Prediction · Classification

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 (AI). A variety of statistical and 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 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, semi-supervised models, and non-supervised models. To execute a given task, machine learning employs several algorithms or models. Machine learning algorithms are increasingly more common in many applications and are extremely beneficial for a typical operator to utilize. 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.

In this context, our challenge is considering the classification of real-time data in the SCADA system as inputs to the machine learning model to predict the state of the production line process, if it is in good functioning or bad one. 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, logistic regression, and K-NN algorithms are performed to innovate the supervised monitoring system based on an efficient predictive model. A comparative discussion revealed to select the random forest as the best classifier for the predictive model.
The advantages of the random forest algorithm are several, it is more interpretable, and feature importance can be estimated during training for little additional computation. It has the possibility of plotting the sample proximities and the visualization of output decision trees. Random forest readily handles larger numbers of predictors, and it grows trees in parallel independently of one another; nevertheless, it is faster to train with fewer parameters. The cross-validation is unnecessary because it generates an internal unbiased estimate of the generalization error (test error) as the forest building progresses.

The remainder of the article is structured as follows. Section 2 contains the contribution and the motivation of the study followed by a literature review section that contains relevant papers in this field and the knowledge gap of machine learning application in industrial processes’ faults prediction. The previous relevant studies concerning the technique of prediction are discussed in Sect. 4 of methods. Section 5 provides the materials and the data utilized in the proposed approach, followed by Sect. 6 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 predicting performance compared to other techniques including SVM, K-NN, Decisions trees, boosting, neural networks, and logistic regression. It protects against overfitting and detects interactions between variables [13–16]. Moreover, for the reason that random forest only considers a subset of predictors for each split, it can tackle much bigger problems before slowing down. If a high-performance computer (HPC) (e.g., many cores) is available, RF can run in embarrassingly parallel without the need for shared memory because all trees are independent [17, 18].

Due to all 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 industrial real-time data in complex and critical industrial processes including cement manufacturing.

Our study makes a valuable contribution to the industrial field such as cement production, because of the several advantages that the work highlights. Details about the data, the industrial process, and the machine learning techniques applied in the study are highlighted, in addition to the focus on findings and their economic impact on the real system. The developed model proved its efficiency in terms of constructing intelligent models that allow optimized participation more than human operators.

Accordingly, we believe that the following research directions are required for the next generation of prognostic and health management systems, especially in complicated industrial processes with enormous real-time alarms and faults. The final objective is to obtain an autonomous system able to supervise the factory in real time. The proposed approach is described in the flowchart of Fig. 1.

3 Literature review

Machine learning is briefly introduced in more detail concerning the manufacturing domain, especially in complex manufacturing environments where detection of the causes
of problems is difficult. It is based on knowledge extraction from data. It is a research field at the intersection of statistics, artificial intelligence, and computer science and is also known as predictive analytics or statistical learning that could analyze industrial data to acquire insights into how to improve the efficiency of specific assets as well as the overall production operation, reaching by that a smart manufacturing [19, 20].

The area of machine learning is quite broad, with several techniques, theories, and accessible methodologies. For many industrial practitioners, this constitutes a hurdle to the use of these sophisticated technologies, which may impede the usage of the large volumes of data that are becoming more available. However, the lack of a guiding theoretical framework of ML technology in manufacturing, the amount of redundant data, and the complexity of processes present a knowledge gap. Consequently, many problems are facing the applications of data analytics and machine learning in the industry at first [21–23].

Though, given their sophisticated and improved results, machine learning techniques are viable options for overcoming some of today’s key issues in complicated production systems. These data-driven techniques can discover extremely complex and non-linear patterns in a variety of data kinds and sources, and then turn raw data into feature spaces, or models, which may subsequently be used for prediction, regression, detection, forecasting, or classification [24, 25]. Along with the industry 4.0 revolution, in several industries, machine learning has been effectively used in process optimization [26], monitoring and control applications in production, and predictive maintenance [27].

There are several machine learning approaches, tools, and techniques accessible, each with its own set of benefits and drawbacks. Many of these techniques have successful applications of machine learning in manufacturing, which are available and are already in daily use in industrial applications worldwide. In the next paragraphs, a panoply of brilliant applications of techniques in industry and manufacturing fields.

Several studies applied SVM in different other fields [3, 28, 29], especially in industry and automation [25, 30–33]. In manufacturing, this can be utilized to identify and classify damaged products like surface roughness [34]. A major application area of SVM in manufacturing is monitoring [35]. Especially tool/machine condition monitoring, fault diagnosis, and tool wear are domains where SVM is continuously and successfully applied [36]. Similarly, statistical process monitoring in manufacturing is a field where SVMs were successfully applied [37].

The K-nearest neighbor technique has been used for several purposes [3, 38, 39], for structural health monitoring applications [40], for failure mode classification and bearing capacity prediction [41], and for many other industrial applications [25, 42]. However, K-NN has high computational complexity for high-dimensional data because the performance of K-NN relies on the number of dimensions [43].

Logistic regression applications are discussed in several studies dedicated to industry [10, 44, 45]. It is applied for the efficient energy consumption prediction [46], for the prediction of loss of position during the dynamic positioning drilling operations [47], and the analysis on the application of advanced manufacturing technologies [21, 48]. The logistic regression technique is applied also for the interpretation of dissolved gas analysis for power transformers [49].

Machine learning approaches mentioned above are built to examine vast volumes of data and can handle high dimensionality effectively. However, factors like probable overfitting must be considered throughout the application process; there are methods available to solve this issue, including random forest.

For many machine learning problems, it is demonstrated that the ensemble leads to a better model generalization compared to a single base classifier. On the other hand, parallel adjustment of base classifiers leads to independent models, which is also named bagging. One famous example of bagging methods is random forest [50], which is a combination of randomly sampled tree predictors. In a first step, random forest randomly selects a subset of the features space and then performs a conventional split selection procedure within the selected feature subset [24].

The improving and satisfying results of the random forest technique make it applicable in different fields [51], including industry and manufacturing [52], pattern recognition, risk identification, and several other fields [53–57]. It is applied in the welding process [58], in the hard metal industry [59], and for detecting failures and optimizing the performance of lift systems [60]. Based on the random forest technique, Chen et al. [61] developed an acoustic signal-based tool condition monitoring in belt grinding of nickel-based superalloys.

4 Methods

The area of machine learning is quite broad, with several techniques, theories, and methodologies accessible. For many industrial practitioners, this constitutes a hurdle to the use of these sophisticated technologies, which may impede the usage of the large volumes of data that are becoming more available.

Because it is relevant to such a wide range of use cases, machine learning is generating a lot of interest. Prediction by 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 [25, 62].

## 4.1 Support vector machine

The support vector machine (SVM) is a supervised learning algorithm proposed by Vapnik [63, 64]. 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 [15, 38]. 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 [65]. SVM employs an iterative method to minimize the error function to identify the best separable hyperplane and maximize the margin between classes.

## 4.2 K-nearest neighbor

The k-nearest neighbor technique or, sometimes called, memory-based methods [66] 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 is 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 [67].

## 4.3 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 [68]. 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 [69].

## 4.4 Ensemble learning techniques

Ensemble learning techniques are originally proposed for classification tasks in a manner of supervised learning in 1965 [66, 70]. The decision tree 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 1990s as a random subspace method for constructing ensembles of decision trees (Bagging, Boosting, and randomization) [71–74]. However, the formal definition and use of random forests 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 [50]. 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 [73].

The random forest algorithm is robust against overfitting compared to many other classifiers, including discriminant analysis, support vector machines, and neural networks [54]. This idea is proved generally by significant results achieved by the random forest technique. The flowchart of the random forest approach is illustrated in Fig. 2.
5 Materials and data

In this study, in the 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, and automated equipment and a large number of other devices to process and maintain this operation and keep it on functionality mode if the system needs. The overall procedure of the workshop of the raw mill is displayed in Fig. 3.
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. All sensor settings recorded can be utilized as training data for the machine learning system. The production line 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 which 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 [75].

### 6 Results and discussion

#### 6.1 Data analysis

The dataset is collected from the cement factory of Ain Touta in East of Algeria (SCIMAT). It contains 21 features and one class that indicates if the process line is good (class “state = 1”) or in a bad-functioning state (class “state = 0”). The number of instances is about 38,187 collected during the running of the production line for 6 trimesters in 2018–2019. All the sensor settings are configured to be used in the training of the machine learning model.

There are different features that the system utilizes as associated factors. The most important ones which have a high effect on the supervision system in addition to other parameters, are Transporter Tape flow M01I01 (Sum of the workshop feeder’s quantity (A02, D02, and E02)), M01P1, M01T1, M01P3, M01T3, M01X1, J01J1, and S01M1I01. Parameter settings’ (features) are reported in Table 1.

The production line as mentioned above has 23.06% of bad functioning, and, thus, could have a terrible effect on the economical development of the enterprise. Each equipment has a set of parameters that the industrial system needs to run, halt, and control all actions about this one for driving it into a good functioning. The distribution of all studied features according to the state of the system is displayed in Fig. 4.

The correlations between the different attributes of the selected dataset are illustrated as a heatmap presented in Fig. 5. 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 it is possible to only eliminate the characteristics where the correlation of two or more characteristics is very high.

Features can affect the functioning of the system positively or negatively. Those that influence negatively the state of the production line are the crusher acoustic indicator

| Features   | Setpoint range and unit | Description                                      |
|------------|-------------------------|--------------------------------------------------|
| A02F1      | [0–24] (tones/hour)     | Flow rate of the Feed Transporter A02            |
| B02F1      | [100–150] (tones/hour)  | Flow rate of the Feed Transporter B02            |
| D02F1      | [0–3] (tones/hour)      | Flow rate of the Feed Transporter D02            |
| E02F1      | [0–5] (tones/hour)      | Flow rate of the Feed Transporter E02            |
| J01J1      | [0–60] (%)              | Elevator’s power                                 |
| M01I01     | [100–150] (tones/hour)  | Crusher Total feed’s rate                        |
| M01P1      | [1–2.1] (mbar)          | Pressure in Crusher’s input                      |
| M01P3      | [17–24] (mbar)          | Pressure Crusher’s output                        |
| M01T1      | [220–400] (°C)          | Temperature Crusher’s input                      |
| M01T3      | [70–120] (°C)           | Temperature Crusher’s output                     |
| M01X1      | [0–50] (%)              | Acoustic equipment of the Crusher                |
| M03J1      | [0–92] (%)              | Crusher Engine’s load                            |
| QCXH2O     | [0.8–15.8] (%)          | Laboratory Operator sp03                        |
| QCXR9O     | [0.9–18] (%)            | Laboratory Operator sp02                        |
| S01M1I01   | [0–120] (%)             | Separator’s Current                              |
| S03P1      | [0–60] (mbar)           | Separator’s Pressure                             |
| S07J1      | [0–94] (%)              | Separator Fan’s load                             |
| S08Z01     | [0–95] (%)              | Separator Butterfly register’s position         |
| S09Z01     | [0–50] (%)              | Separator Butterfly register’s position         |
Fig. 4 Distribution of factors according to the state of the system
After several experiments, the data set is split into two parts, respectively, as the training set (67%) and testing set (33%). The given division reached the best prediction accuracy. The training set is used to train the prediction model, while the testing set is used to validate the performance of the trained model.

The OOB_score (OOB sampling) is the random forest cross-validation method. In this sampling, about one third of the data is not used to train the model and can be used to evaluate its performance. These samples are called the out-of-bag samples. It is very similar to the leave-one-out-cross-validation method, but almost no additional computational burden goes along with it.

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 utilized. 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 the random forest classifier for 200 decision trees per an estimate, which is the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions. A higher number of trees increases the performance and makes the predictions more stable.

Random forests are utilized to train a large number of decision trees at the same time. These trees are trained on subsets of the observations, and then these predictions are combined creating better-generalized predictions. Because each tree is trained on a random set of data, a random state is defined when initializing the model to achieve repeatable results.

To control the depth and complexity of the individual trees and avoid overfitting, two hyperparameters are set by limiting the number of levels to reduce the number of grown trees used during classification (“max_depth” = 5), and the minimum number of leaf nodes required to split an internal node before reaching the prediction to optimize the number of predictors involved in the learning process (“min_samples_leaf” = 5). The complexity of the model is defined by $O(M \times (m \times n \log n))$. The parameters include the number of instances in the training data ($n = 25,501$), the number of features ($m = 20$), and the number of build trees ($M = 200$). Consequently, the computational cost of building a tree is $O = 449,486,461.777$.

Concerning the model parameter uncertainty, Breiman (2001) demonstrated in his paper that two elements impact the variance of a random forest’s predictions, including the variation of each tree in the forest and the correlation between trees. Implicitly, trees that are confident in their predictions and are less associated with one another indicate that the forest is generally certain.
To get the uncertainty of the random forest model, we need to collect the prediction of each tree separately and obtain the probability of each tree. The uncertainty of the model is the standard deviation which is 0.37 and the mean which is 0.76 obtained from the list of probabilities for each prediction.

A part of the decision forest is displayed in Fig. 7, it illustrates 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.”

### 6.3 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 2.

Other metrics such as accuracy, precision, recall, 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 3.

After validating our model, we check the confusion matrix to understand how the model performs for each label. The matrix revealed how to present values and percentage of prediction illustrated in the matrix in Fig. 8a. The ROC curve of the implemented model and every point is above the no-skill line displayed in Fig. 8b, which proves that the model did not overfit.

In terms of evaluation indicators, we test the classification accuracy in different classifiers to evaluate the performance of the proposed scheme. Compared to

### Table 2 The evaluation metrics of the predictive model

| Metrics            | Value                           |
|--------------------|---------------------------------|
| Training accuracy  | 0.9984706482098741              |
| Testing accuracy   | 0.9701456890374970              |
| Sensitivity        | 0.9542459736456809              |
| Specificity        | 0.9745650625699461              |
| AUC score          | 0.9938163695076756              |

### Table 3 The random forest model evaluation report

|                  | Precision | Recall | f1-score | Support |
|------------------|-----------|--------|----------|---------|
| Class 0          | 0.95      | 0.91   | 0.93     | 2857    |
| Class 1          | 0.97      | 0.99   | 0.98     | 9703    |
| Accuracy         | 0.97      |        |          |         |
| Macro avg        | 0.96      | 0.95   | 0.96     | 12,561  |
| Weighted avg     | 0.97      | 0.97   | 0.97     | 12,561  |
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, and finally, about 83.73% is obtained by the decision tree model.

**7 Conclusion**

The concept is that machine learning has been integrated into the industry and that this theory applies to a practical industrial project. Machine learning is already proved as a strong instrument and successful approach for a wide range of applications in intelligent industrial systems and smart manufacturing, and its significance will widely grow in the future and will increase further at a rapid pace.

Modern manufacturing systems are challenging with increasing complexity, dynamic, high dimensionality, and chaotic structures. Because the majority of manufacturing applications can supply labeled data, it was argued that supervised learning techniques, including random forest, decision trees, SVM, K-NN, and logistic regression, are a suitable fit for most manufacturing applications.

One of the most significant advantages of random forest is its adaptability. It may be used for both regression and classification problems, and it is simple to observe how much weight it gives to the input characteristics. Random forest is also a useful technique since the default hyperparameters it employs frequently yield good prediction results. Understanding the hyperparameters is rather simple, and there are not many of them. Overfitting is one of the most serious issues in machine learning, although the random forest classifier prevents it most of the time.

Accordingly, this paper provided a system with the ability to classify data using the identified techniques for the industrial manufacturing of cement in SCIMAT society. For the developed learning model, we adopted the Random Forest algorithm justified by a comparison which is carried out with SVM, decision tree, logistic regression, 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 97%. The accuracies were found to be for the random forests proved to be a better classifier with a 97% 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 and 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.

**Acknowledgements** 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).

Author contributions HZ and AD led the study by being involved in data collection, performing the statistical analysis, coding, interpreting the results, conceiving, designing the article, and providing intellectual content revisions and suggestions for clarity and precision on the subject matter. HZ and AD contributed in major correcting of the recommended revisions of the paper.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

1. Reddy DN, Priyanaka R, Sanjana S et al (2021) Machine learning algorithms for detection: a survey and classification 12:3468–3474
2. Elaziz MA, Hosny KM, Salah A et al (2020) New machine learning method for image-based diagnosis of COVID-19. PLoS One 15:1–18. https://doi.org/10.1371/journal.pone.0235187
3. Kale R, Shitole S (2021) Analysis of Crop disease detection with SVM, KNN and Random forest classification. Inf Technol Ind 9:364–372. https://doi.org/10.17762/iti.v9i1.140
4. Rozeck DC, Andres WC, Smith NB et al (2020) Using Machine Learning to Predict Suicide Attempts in Military Personnel. Psychiatry Res 294:113–155. https://doi.org/10.1016/j.psychres.2020.113515
5. Fife DA, Onofrio J (2021) Common, uncommon, and novel applications of random forest in psychological research
6. Li H, Lin Z, Shen X et al (2015) A convolutional neural network cascade for face detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit (07–12 June): 5325–5334. https://doi.org/10.1109/CVPR.2015.7299170
7. Li K, Jin Y, Akram MW et al (2020) Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. Vis Comput 36:391–404. https://doi.org/10.1007/s00371-019-01627-4
8. Lin K, Zhao H, Lv J et al (2020) Face detection and segmentation based on improved mask R-CNN. Discret Dyn Nat Soc. https://doi.org/10.1155/2020/924917
9. Ismail M, El-asser A (2021) Utilization of machine learning techniques for quality monitoring and prediction pp. 4830–4839
10. Jung H, Jeon J, Choi D, Park AJY (2021) Application of machine learning techniques in injection molding quality prediction: Implications on sustainable manufacturing industry. Sustainability 13. https://doi.org/10.3390/su13084120
11. Mokhtari S, Abbaspour A, Yen KK, Sargolzaei A (2021) A machine learning approach for anomaly detection in industrial control systems based on measurement data. Electron 10:1–13. https://doi.org/10.3390/electronics10040407
12. Hemalatha N-R, Karishma M, ShanmugaSundari N, Anandhavalli D (2021) Performance analysis of spinning machines using machine learning. In: International Conference on Smart Data Intelligence (ICSMDI 2021) pp 1–10
13. Saberion M, Cisař P, Labbé L et al (2018) Comparative performance analysis of support vector machine, random forest, logistic regression and k-nearest neighbors in rainbow trout (oncorhynchus mykiss) classification using image-based features. Sensors (Switzerland) 18:1–15. https://doi.org/10.3390/s18041027
14. Kremic E, Subasi A (2016) Performance of random forest and SVM in face recognition. Int Arab J Inf Technol 13:287–293
15. Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 15:3133–3181. https://doi.org/10.1117/1.JRS.11.015020
16. Hulten G (2019) Building intelligent systems: a guide to machine learning engineering. Apress, Lynnwood, Washington, USA
17. Speiser JL, Miller ME, Tooze J, Ip E (2019) A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl 134:93–101. https://doi.org/10.1016/j.eswa.2019.05.028
18. Montillo AA (2009) Random Forests. Temple University, USA
19. Wuest T (2015) ThorstenWuest-Identifying product and process state drivers in manufacturing systems using supervised machine learning. Springer
20. Rai R, Triware MK, Ivanov D, Dolgui A (2021) Machine learning in manufacturing and industry 4.0 applications. Int J Prod Res 59:4773–4778. https://doi.org/10.1080/00207543.2021.1956675
21. Sharma A, Zhang Z, Rai R (2021) The interpretive model of manufacturing: a theoretical framework and research agenda for machine learning in manufacturing. Int J Prod Res 59:4960–4994. https://doi.org/10.1080/00207543.2021.1930234
22. Kusiak A (2018) Smart manufacturing, Int J Prod Res 56:508–517. https://doi.org/10.1080/00207543.2017.1351644
23. Van Heerden I, Bas A (2021) Viewpoint: AI as author - bridging the gap between machine learning and literary theory. J Artif Intell Res 71:175–189. https://doi.org/10.1613/jair.1.12593
24. Wuest T, Weimer D, Irgens C, Thoben KD (2016) Machine learning in manufacturing: Advantages, challenges, and applications. Prod Manuf Rev 4:23–45. https://doi.org/10.1016/j.ijmr.2016.1192517
25. Usuga Cadavid JP, Lamouri S, Grabot B et al (2020) Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. J Intell Manuf 31:1531–1558. https://doi.org/10.1007/s10845-019-01531-7
26. Li Y, Carabelli S, Fadda E et al (2020) Machine learning and optimization for production rescheduling in Industry 4.0. Int J Adv Manuf Technol 110:2445–2463. https://doi.org/10.1007/s00170-020-05850-5
27. Wang L, Liu Z, Liu A, Tao F (2021) Artificial intelligence in product lifecycle management. Int J Adv Manuf Technol 114:771–796. https://doi.org/10.1007/s00170-021-06882-1
28. Lazi M, Labadi K, Brucker JM, Ameur S (2020) Improving satellite rainfall estimation from MSG data in Northern Algeria by using a multi-classifier model based on machine learning. J Hydrod 584:124705. https://doi.org/10.1016/j.jhysod.2020.124705
29. Rahab H, Zitouni A, Djoudi M (2018) SIAAC: Sentiment Polarity Identification on Arabic Algerian Newspaper Comments. Adv Intell Syst Comput 662:139–149. https://doi.org/10.1007/978-3-319-67621-0_12
30. Zermane H, Kasmi R (2020) Intelligent industrial process control based on fuzzy logic and machine learning. Int J Fuzzy Syst Appl 9:92–111. https://doi.org/10.4018/IFSSA.2020010104
31. Ahmadi SH, Khosrowjerdi MJ (2021) Fault detection Automation in Distributed Control Systems using Data-driven methods: SVM
and KNN. TechRxiv Prepr 0–7. https://doi.org/10.36227/techrxiv.15029739.v2

32. Yin G, Zhang YT, Li ZN et al (2014) Online fault diagnosis method based on incremental support vector data description and extreme learning machine with incremental output structure. Neurocomputing 128:224–231. https://doi.org/10.1016/j.neucom.2013.01.061

33. Suykens JAK (2009) Support vector machines and kernel-based learning for dynamical systems modelling. IFAC

34. Abu-Maftouz I, El Ariss O, Esfakuran AHM, Banerjee A (2017) Surface roughness prediction as a classification problem using support vector machine. Int J Adv Manuf Technol. https://doi.org/10.1007/s00170-017-0165-9

35. Guo J, Li A, Zhang R (2020) Tool condition monitoring in milling process using multifractal detrended fluctuation analysis and support vector machine. Int J Adv Manuf Technol 110:1445–1456. https://doi.org/10.1007/s00170-020-05931-5

36. Kong D, Chen Y, Li N, Tan S (2017) Tool wear monitoring based on kernel principal component analysis and v-support vector regression. Int J Adv Manuf Technol 89:175–190. https://doi.org/10.1007/s00170-016-9070-x

37. Cuestas P, Peñabaena-Niebles R, Garcia E (2017) Support vector machine in statistical process monitoring: a methodological and analytical review. Int J Adv Manuf Technol 91:485–500. https://doi.org/10.1007/s00170-016-9693-y

38. Rozenberg G, Thomas B, Kok JN (2012) Handbook of Natural Computing. Springer, Berlin, Heidelberg

39. Dino HI (2019) Abdulrazzaq MB (2019) Facial expression classification based on SVM, KNN and MLP classifiers. Int Conf Adv Sci Eng ICOASE 2019:70–75. https://doi.org/10.1109/ICOASE.2019.8723728

40. Vitola J, Pozo F, Tibaduiza DA, Anaya M (2017) A sensor data fusion system based on k-nearest neighbor pattern classification for structural health monitoring applications. Sensors (Switzerland) 17:1–26. https://doi.org/10.3390/s17020417

41. Feng DC, Liu ZT, Wang XD et al (2020) Failure mode classification and bearing capacity prediction for reinforced concrete columns based on ensemble machine learning algorithm. Adv Eng Informatics 45:101126. https://doi.org/10.1016/j.aei.2020.101126

42. Wong QY, Chu YB (2021) A mobile production monitoring system based on internet of thing (IoT) and random forest classification. Int J Electr Electron Eng Telecommun 10:243–250. https://doi.org/10.10187/fjelect.10.4.243-250

43. Liu C, Cichon A, Królczyk G, Li Z (2021) Technology development and commercial applications of industrial fault diagnosis system: a review. Springer, London

44. Desai A, Guo Y, Sheng S et al (2020) Prognosis of wind turbine gearbox bearing failures using SCADA and modeled data. Proc Annu Conf Progn Heal Manag Soc PHM 12:1–10. https://doi.org/10.36001/phmconf.2020.v1211.1292

45. Merghadi A, Abderrahmane B, Tien Bui D (2018) Landslide susceptibility assessment at Mila basin (Algeria): A comparative assessment of prediction capability of advanced machine learning methods. ISPRS Int J Geo-Information 7. https://doi.org/10.3390/ijgi7070268

46. V E S, Shin C, Cho Y, (2021) Efficient energy consumption prediction model for a data analytic-enabled industry building in a smart city. Build Res Inf 49:127–143. https://doi.org/10.1080/09613218.2020.1809983

47. Sanchez-Varela Z, Boullosa-Facles D, Barrena JLL, Gomez-Solaech MA (2021) Prediction of loss of position during dynamic positioning drilling operations using binary logistic regression modeling. J Mar Sci Eng 9:1–18. https://doi.org/10.3390/jmse9020139

48. Wang G, Zhang L, Guo J (2021) Analysis on application level based on ordinal logistic regression and best of advanced manufacturing technologies (AMT) selection based on fuzzy- TOPSIS integration approach. J Intell Fuzzy Syst 40:8427–8437. https://doi.org/10.3233/JIFS-189663

49. Almoalliem YD, Taha IBM, Mosaad MW et al (2021) Application of logistic regression algorithm in the interpretation of dissolved gas analysis for power transformers. Electron 10. https://doi.org/10.3390/electronics10102106

50. Breiman L (2001) Random forests. Mach Learn 45:5–32

51. D’Almeida AL, Bergiante NCR, de Souza FG et al (2022) Digital transformation: a review on artificial intelligence techniques in drilling and production applications. Int J Adv Manuf Technol. https://doi.org/10.1007/s00170-021-08631-w

52. He Z, Xu J, Tran KP et al (2021) Modeling of textile manufacturing processes using intelligent techniques: a review. Int J Adv Manuf Technol 116:39–67. https://doi.org/10.1007/s00170-021-07444-1

53. Kiangala SK, Wang Z (2021) 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 Appl 4:100024. https://doi.org/10.1016/j.mlwa.2021.100024

54. Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2:18–22

55. Dantone M, Gall J, Fanelli G, Van Gool L (2012) Real-time facial feature detection using conditional regression forests. Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit 2578–2585. https://doi.org/10.1109/CVPR.2012.6249796

56. Jia J, Xu Y, Zhang S, Xue X (2016) The facial expression recognition method of random forest based on improved PCA extracting feature. ICSPCC 2016 - IEEE Int Conf Signal Process Commun Conf Proc 0–4. https://doi.org/10.1109/ICSPCC.2016.7753643

57. Liu LJ, Shen WK, Zhu JM (2021) Research on risk identification system based on random forest algorithm-high-order moment model complexity. https://doi.org/10.1155/2021/5588018

58. Xing B, Xiao Y, Qin QH, Cui H (2018) Quality assessment of resistance spot welding process based on dynamic resistance signal and random forest based. Int J Adv Manuf Technol 94:327–339. https://doi.org/10.1007/s00170-017-0899-6

59. Kotsiopoulos T, Leonaris L, Dimitriou N et al (2021) Deep multi-sensorial data analysis for production monitoring in hard metal industry. Int J Adv Manuf Technol 115:823–836. https://doi.org/10.1007/s00170-020-06173-1

60. Pennel M, Hsiung J, Putcha VB (2018) Detecting failures and optimizing performance in artificial lift using machine learning models. SPE West Reg Meet Proc 2018-April. https://doi.org/10.2118/190090-ms

61. Chen J, Chen H, Xu J et al (2018) Acoustic signal-based tool condition monitoring in belt grinding of nickel-based superalloys using RF classifier and MLR algorithm. Int J Adv Manuf Technol 98:859–872. https://doi.org/10.1007/s00170-018-2270-9

62. Mohana-Priya T, Punithavall M, Rajesh-Kanna R (2021) Conceptual review on machine learning algorithms for classification techniques. Int J Sci Res Comput Sci Eng Inf Technol 7:215–222. https://doi.org/10.32628/IJSRCEIT

63. Corinna C, Vapnik V (1995) Support-vector networks Mach Learning 20:273–297

64. Vapnik VN (1998) Pattern recognition-statistical learning theory. Wiley, Canada

65. Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Support Vector Machines ( SVM ), London

66. Nilsson NJ (1998) Introduction To Machine learning. Stanford University, Stanford

67. Murti MN (2016) Support vector machines and perceptrons learning, optimization, classification, and application to social networks. Springer Briefs in Computer Science
68. Takano S (2017) Thinking machines machine learning and its hardware implementation. Academic Press is an imprint of Elsevier, Japan
69. Theobald O (2017) Machine learning for absolute beginners, 2nd edn. Scatter Plot Press, Brooklyn
70. Yang Y (2017) Ensemble Learning Temporal Data Min Via Unsupervised Ensemble Learn pp. 35–56. https://doi.org/10.1016/b978-0-12-811654-8.00004-x
71. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1007/bf00058655
72. Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Comput 9:1545–1588. https://doi.org/10.1162/neco.1997.9.7.1545
73. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20:832–844
74. Dietterich TG (2000) Experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 40:139–157. https://doi.org/10.1023/A:1007607513941
75. Varoquaux G, Buitinck L, Louppe G et al (2015) Scikit-learn. GetMobile Mob. Comput Commun 19:29–33. https://doi.org/10.1145/2786984.2786995

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.