*Advances in Manufacturing Technology XXXIV*
*M. Shafik and K. Case (Eds.)*

*doi:10.3233/ATDE210065*

# Machine Learning Algorithms Comparison for Manufacturing Applications

Mohammed ALMANEI[a], Omogbai OLEGHE[b], Sandeep JAGTAP[a]
and Konstantinos SALONITIS[a,1]
[a] *Manufacturing Department, Cranfield University, Cranfield, MK43 0AL, UK*
[b] *Systems Engineering, Department of Engineering, University of Lagos, Nigeria*

**Abstract.** With the vast amount of data available, and its increasing complexity in manufacturing processes, traditional statistical approaches have started to fall short. This is where machine learning plays a key role, addressing the challenges by bringing the ability to analyse large and complex datasets from multiple sources, finding non-linear and intricate patterns on data, relationships between several factors and their influence on the manufacturing process outputs. This paper demonstrates the advantages and applications of using supervised machine learning techniques in the manufacturing industry. It focuses on binary classification and compares the performance of three different machine learning algorithms: logistic regression, support vector machine, and neural networks. A case study has been conducted on a manufacturing company, using the techniques and algorithms mentioned. The case study focuses on analysing the relationship between different manufacturing process variables and their impact on one key output variable of a product, which in this case is the result of a quality test that measures product performance. The modelling problem has been oriented towards a Boolean goal to predict whether the parts will pass this test.

**Keywords.** Machine learning, logistic regression, support vector machine, neural networks.

## 1. Introduction

Industry 4.0 capability has led to an extreme expansion on the sensors and data collected for diverse parts of the processes, which has resulted in both the increase in the amount of data stored and transmitted substantially and the complexity of handling such data sources. Therefore, traditional statistical approaches have started to fall short in capabilities to analyse and trend variables with non-linear patterns and more complex relationships between them [1]. Machine learning (ML) can play a key role in manufacturing process analysis and improvements, with proven results in many industrial sectors. Nevertheless, the problems behind this powerful tool and techniques rely on the quality of the data currently collected by the industry [2].

Even though most of the manufacturing industries can handle, collect and analyse many data sources, most manufacturing processes are not ready to implement ML techniques and models right away. There are many different pre-processing steps to clean, format, segregate and structure the data sets before using them to create predictive models [3]. This paper focuses on  analysing the benefits of using ML techniques to

[1] Corresponding Author. k.salonitis@cranfield.ac.uk

improve manufacturing processes and describe current challenges, constraints and limitations faced while trying to do so by applying ML algorithms in a practical case study.

## 2. Machine Learning

ML is a subfield of computer science, considered a branch of artificial intelligence, that can be divided into two phases, training and testing, with the former related to finding the causes of the goal variable, and the latter to the capability of making predictions on further data [4]. Several definitions of ML are available, relating data patterns and how these can be used for problem-solving. Being a broad area, the preferred definition by the authors of the present paper is the capability of predicting future data, based on historical data [5], finding patterns and relationships inherent in the datasets analysed that would not be found by other techniques to make appropriate predictions on features of future data. Alternative solutions would require an excessive amount of work.

Two main types of ML are identified [6], namely supervised and unsupervised methods. A number of authors [7] classify reinforcement learning as a third type of ML. Supervised ML is characterised by knowing the desired state of the goal variable and using this knowledge to analyse the features and specific conditions that generate this state in order to make future predictions [8]. Two data sets (or two sub sets of the data) are required, one for training and one for testing the model. But both datasets must represent the same set of features or variables and the same system. Unsupervised ML, on the other hand, do not make the distinction between training and test data. In unsupervised, the output variable is not labelled e.g. good or bad. In supervised, it is labelled e.g. pass or fail. Simple K-means clustering is an example of unsupervised learning, where the clusters can be used to reveal the labels. Its main goal is to find patterns or clusters in subgroups of the data analysed to learn or identify a structure of the data where no feedback or labels are provided; in other words, the outcome is unknown [9]. Reinforcement learning refers to a set of algorithms that ca be used in both supervised and unsupervised learning; less feedback is provided to the algorithm and consists mainly of evaluating the action recommended by the algorithm [10]. It focuses on dynamic and complex environments, where unforeseeable stresses and external constraints to the system are present [11].

Supervised ML is widely used in process analysis and improvement in the manufacturing industry since it is expected that the goal state of the process is known, therefore, labelled data is available or can be inferred to train the ML models. For the present research because of the objective of the study and the type of system variables and dataset, supervised ML was used.

### 2.1. *Supervised Machine Learning*

The learning process of ML models is classified by the amount of labelled data available to train the model [10]. Supervised ML algorithms can only be used if there is at least part of the dataset labelled according to a specific target. In order to use supervised ML, both desired, and undesired states of the goal variable have to be known [12]. This allows using a classification algorithm to calculate process states and associate them with the goal variable. Once the features causing the final state of the goal variable are known, they can be used to predict future process behaviour on unknown or unlabelled data.

There are two main objectives for algorithms on supervised ML: classification and regression, and while some algorithms can do both, others can only perform one of these tasks [7]. In classification, the goal is to determine the category to which a certain output belongs; here, the labelled data belongs to a certain limited number of values. In order to separate the different states of these values, a decision boundary is created. A decision boundary is an imaginary line created as far as possible of each of the states. The form of this boundary will determine the accuracy of the model; algorithms differ in the way they create the decision boundary. The labelled data represents a real value for regression objectives, also called target, and future values are predicted by the ML model based on unlabelled examples used as input.

The classification problem can be divided into binary and multiclass classification. The first one focuses on classifying the group of inputs into only two different states, either yes or no, false or true, good or no good. Multiclass classification refers to placing an example into a category of many provided [5]. For the present paper, binary classification has been selected to solve the classification problem since all the characteristics required to use it can be fulfilled. The aim of the project is aligned with the capabilities of this strategy. Labelled data is available and only corresponds to two states, good products and bad products. The end goal of the ML model is to identify the different features influencing this output and classify them into these two groups to make future predictions on unlabelled data.

### *2.2. Machine Learning Algorithms*

ML technology can be applicable to a variety of problems and provides different algorithms for their solution [13]. These algorithms have different objectives but a similar end goal: learning and creating predictions on future states based on known states of a process. As mentioned earlier in the paper, there are different algorithms able to target both classification and regression problems, as well as many only able to target one of them. For this research and case study, the focus has been on binary classification algorithms for supervised ML. A thorough literature review was undertaken with more than 30 scientific papers reviewed on the topic of ML algorithms appropriate for such a binary classification problem. Three different ML techniques were selected, namely Support Vector Machine (SVM), Logistic Regression (LR) and Neural Networks (NN).

As part of the classification algorithms, SVM requires the labelled data to belong to either positive or negative groups; usually, it is required that the data is labelled as +1 and -1 [7]. The algorithm then solves an optimisation problem, finding an imaginary hyperplane that splits positive from negative examples.

Regardless of its name, LR is a classification algorithm rather than a regression. The origin of the name is from statistics and is due to its similarity with the mathematical formulation for linear regression [7]. The standard logistic function is aligned with the project's classification objective. If the values of the function are properly optimised, the output can be interpreted as the probability of the value to be positive.

NNs have been around for quite a long time. The focus for this project has been on the efficient use of multilayer perceptron NNs for statistical pattern recognition [14]. A perceptron takes several binary inputs and produces a single binary output using a weighting system to assign real numbers to each of the inputs to represent their importance for the model [15]. The output value is determined by whether the weighted sum of the inputs is lower or greater than a threshold value, which is another real number part of the parameters of the neuron.

## 3. Adopted method for developing ML models

The methodology used for this research has been based on the "CRISP-DM (CRoss-Industry Standard Process for Data Mining)" model [16]. It has been chosen for its validity and wide uses across industries for data analysis projects. This methodology consists of six steps conducted from the start to the end of the project: Business understanding, data understanding, data preparation, modelling, evaluation, and deployment [17]. This model provides a structured and methodological approach to conducting data mining projects, which in this case has been emphasised on the use of ML since both types of projects are closely related and follow similar steps.

## 4. Case study and models development

The product analysed is composed of four parts that are assembled. Three of them are manufactured in-house, whereas the fourth one is purchased from a supplier. Three different objectives were set that were to be modelled using the different ML algorithms to predict the impact of specific issues on the key process output variable (KPOV):

- Identify if part X has a higher influence compared to part Y on the KPOV.
- The interaction of the key component dimensions to the KPOV.
- The impact of the processing parameters of the three parts on the KPOV.

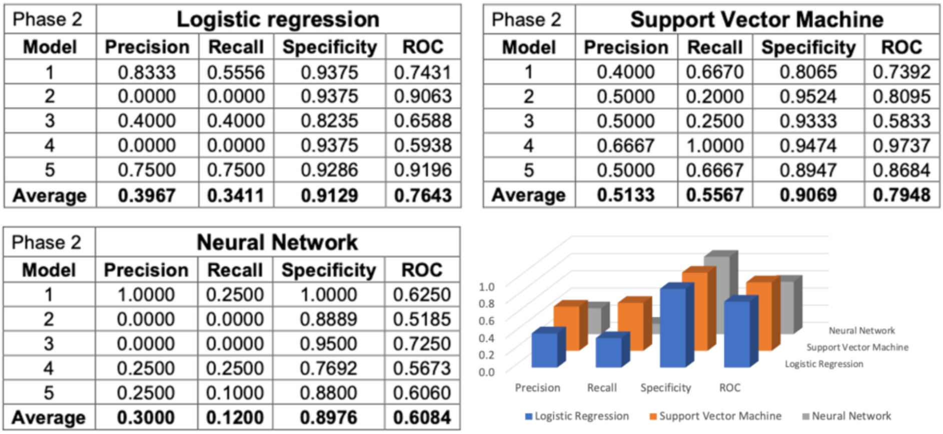
Each objective set was associated with specific data sources, such as product performance results, product dimensions, machine performance, processing conditions, and even external factors as environmental temperature. Data preparation, including deciding the time frame segregation, the data conversion, cleaning, and normalisation, are amongst the steps requiring most of the resources and time.

As described in the literature review, supervised ML techniques were used along with a Boolean feature goal for this project. A feature is the variable of interest that is intended to be predicted by other relevant factors. In Table 1, the terms used for the modelling and validation phases are explained. A description of the modelling process performed is shown in this section, as well as the insights from signals and profiles.

**Table 1.** Modelling characteristics

| Characteristic | Description |
|---|---|
| Signal | Identify the most statistically correlated features from the dataset. |
| Profile | A combination of two or more features from the dataset has a higher probability, when combined, of maximising or minimising the goal. |
| Models | Composed by the previous characteristics and predictive capability statistics. These depend on the nature of the goal feature and are different if the feature is a continuous or Boolean variable. In this case, since a Boolean value is used to determine if the part will fail the final quality test, a variety of metrics are calculated, such as precision, recall, and specificity. |
| Precision | Fraction of data-points classified as positive that are positive. |
| Recall | Fraction of all the positive instances that are classified as positive. |
| Specificity | Fraction of all the negative data-points that were correctly classified as negative. |
| ROC curves | Plots of the true-positive rates against false-positive rates at various discrimination thresholds. |
| Confusion matrix | The number of correct and incorrect predictions produced by the model, compared to the actual outcomes. |

For modelling the case study, three supervised ML algorithms were used for this project: LR, SVM, and NN. Models were trained five times with random sampling for the train and test data to validate their accuracy. In the following tables, the results of the training and validation of the models are presented. For the economy of space, only the results for the second objective are presented in Figure 1. SVM proved better performance for this project with a higher average ROC than the rest. Nonetheless, one of the training runs from this algorithm showed a ROC of 0.5833, which is barely better than random classification. In addition to this, this algorithm showed poor precision and recall overall. These problems are due to the data leakage problem experienced for aggregating the datasets before the analysis. Conclusions from this phase are that it is necessary to find a better approach to join datasets from the dimensional test to the KPOV test, to improve the resolution of the data, and eliminate or reduce the data leakage. This will considerably enhance the accuracy of the models to predict product quality based on its dimensions. Similar findings were observed for the other objectives set.



| Phase 2 | Logistic regression | | | |
|---|---|---|---|---|
| **Model** | **Precision** | **Recall** | **Specificity** | **ROC** |
| 1 | 0.8333 | 0.5556 | 0.9375 | 0.7431 |
| 2 | 0.0000 | 0.0000 | 0.9375 | 0.9063 |
| 3 | 0.4000 | 0.4000 | 0.8235 | 0.6588 |
| 4 | 0.0000 | 0.0000 | 0.9375 | 0.5938 |
| 5 | 0.7500 | 0.7500 | 0.9286 | 0.9196 |
| **Average** | **0.3967** | **0.3411** | **0.9129** | **0.7643** |

| Phase 2 | Support Vector Machine | | | |
|---|---|---|---|---|
| **Model** | **Precision** | **Recall** | **Specificity** | **ROC** |
| 1 | 0.4000 | 0.6670 | 0.8065 | 0.7392 |
| 2 | 0.5000 | 0.2000 | 0.9524 | 0.8095 |
| 3 | 0.5000 | 0.2500 | 0.9333 | 0.5833 |
| 4 | 0.6667 | 1.0000 | 0.9474 | 0.9737 |
| 5 | 0.5000 | 0.6667 | 0.8947 | 0.8684 |
| **Average** | **0.5133** | **0.5567** | **0.9069** | **0.7948** |

| Phase 2 | Neural Network | | | |
|---|---|---|---|---|
| **Model** | **Precision** | **Recall** | **Specificity** | **ROC** |
| 1 | 1.0000 | 0.2500 | 1.0000 | 0.6250 |
| 2 | 0.0000 | 0.0000 | 0.8889 | 0.5185 |
| 3 | 0.0000 | 0.0000 | 0.9500 | 0.7250 |
| 4 | 0.2500 | 0.2500 | 0.7692 | 0.5673 |
| 5 | 0.2500 | 0.1000 | 0.8800 | 0.6060 |
| **Average** | **0.3000** | **0.1200** | **0.8976** | **0.6084** |

**Figure 1.** Results comparing the three different ML algorithms.

## 5. Conclusions

The aim of this paper has been fulfilled by the effective identification of benefits, constraints and limitations of using ML techniques in a manufacturing environment. By the correct application of ML techniques to a specific case study, where valuable insights have been provided to the company as a result of the project, regarding their KPOV and the impact of the different factors analysed on it.

ML projects are an iterative process. It is crucial to optimise and increase the datasets used for their analysis while exploring different approaches to obtain substantial benefits for the industry. The technology is readily available, and can perform powerful analysis, but the industry must be prepared to use it. As demonstrated by the case study, the main constraints and limitations were not due to modelling problems or a slow learning curve of the different software required and ML techniques themselves. But due to the quality of the data used for the project and the lack of both traceability and standardisation among all the manufacturing processes.

## References

[1] Mohammadi, P. and Wang, Z.J. (2016) 'Machine learning for quality prediction in abrasion-resistant material manufacturing process', *Canadian Conference on Electrical and Computer Engineering*, 2016-Octob, pp. 4–7.

[2] Danubianu, M. and Stefan, " (2014) 'Step By Step Data Preprocessing for Data Mining. a Case Study', International Conference on Information Technologies (InfoTech-2015)

[3] Singh Malik, J., Goyal, P. and Sharma, M.K. (2007) 'A Comprehensive Approach Towards Data Preprocessing Techniques & amp; Association Rules', Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM), , p. 12.

[4] Liu, H. and Cocea, M. (2018) Granular Computing Based Machine Learning.

[5] Daume, H. (2012) A course in machine learning Todo.

[6] Shalev-Shwartz, S. and Ben-David, S. (2013) Understanding machine learning: From theory to algorithms.

[7] Burkov, A. (2019) 'The Hundred - Page Machine Learning' p. 159

[8] Shin, C.K. and Park, S.C. (2000) 'A machine learning approach to yield management in semiconductor manufacturing', International Journal of Production Research, 38(17), pp. 4261–4271.

[9] Wuest, T., Weimer, D., Irgens, C. and Thoben, K.D. (2016) 'Machine learning in

[10] Monostori, L. (2002) 'AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing', IFAC Proceedings Volumes (IFAC-PapersOnline), 15(1), pp. 119–130.

[11] Aissani, N., Beldjilali, B. and Trentesaux, D. (2008) 'Use of machine learning for continuous improvement of the real time heterarchical manufacturing control system performances', International Journal of Industrial and Systems Engineering, 3(4), p. 474.

[12] Wuest, T., Weimer, D., Irgens, C. and Thoben, K.D. (2016) 'Machine learning in manufacturing: Advantages, challenges, and applications', Production and Manufacturing Research, 4(1), pp. 23–45.

[13] Witkowski, T., Antczak, P. and Antczak, A. (2011) 'Machine learning Based classification in manufacturing system', Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2011, pp. 580–585.

[14] Jordan, M., Kleinberg, J. and Scho, B. (2006) Pattern Recognition and Machine Learning.

[15] Neapolitan, R.E. and Neapolitan, R.E. (2018) 'Neural Networks and Deep Learning', Artificial Intelligence, pp. 389–411.

[16] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000) Crisp-dm 1.0.

[17] Nadali, A., Kakhky, E.N. and Nosratabadi, H.E. (2011) 'Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system', ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology, 6(April), pp. 161–165.