Predicting late delivery in Supply chain 4.0 using feature selection: a machine learning model

Houria ABOULOIFA
Mathematics and Computer Science Department, Lab L.M.I.E.T
Hassan Ist University
Faculty of Sciences and Techniques, Settat, Morocco
h.abouloifa@uhp.ac.ma

Mohamed BAHAJ
Mathematics and Computer Science Department, Lab L.M.I.E.T
Hassan Ist University
Faculty of Sciences and Techniques, Settat, Morocco
mohamedbahaj@uhp.ac.ma

2022 5th International Conference on Advanced Communication Technologies and Networking (CommNet) | 978-1-6654-5054-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/COMMNET56067.2022.9993969

***Abstract*— Late delivery is a common issue that can affect organizational performance and growth of any Supply Chain. The ability to predict delivery status would be an invaluable tool for any organization seeking to retain customers and predict their future behavior. This transition of Supply Chain performance in the context of industry 4.0 is called Supply chain 4.0. This study employed machine learning (ML) algorithms to predict whether a customer's order will be delivered late. It is presented as a comparative performance combination of Random Forest Classifier algorithms and three Feature Selection scenarios. In this experiment, the best predictors were identified using the SelectKBest and the predictive models' performance was evaluated using several metrics of Anova method's f_classif() function. The empirical results have demonstrated that feature selection has increased the model accuracy from 97.9% to 99.38%.**

Keywords*— *Machine Learning (ML). Supply Chain 4.0. Late delivery prediction. Feature selection. Industry 4.0.

I. INTRODUCTION

Companies are increasingly facing a variable demand in term of product quantities and technical specifications. Several consumer products are becoming customizable. This variability has a considerable impact on production costs which did forced companies to manage their supply chain in an integrated and dynamic way to achieve the expected performance levels. In such an environment, decision-making becomes complicated. The coordination between the different links remains a difficult task, since each link seeks its local optimum which does not necessarily correspond to the global optimum of the chain. Thus, companies are called upon to equip themselves with a smart decision-making system capable of anticipating and prediction eventual obstacles and problems in order to guarantee communication both internally and externally.

In this context, industry 4.0 envisions a digital transformation in the enterprise, entwining the cyber-physical world and real world to deliver networked production with enhanced process transparency [1]. It has transformed the manufacturing process, paving the way for intelligent manufacturing that promises self-sufficient production processes through the use of machines and gadgets that communicate with one another via digital connectivity [2]. Integrated systems have become smarter as they learn from past experience and decision making is being more reliable: it is the era of machine learning. As a sub domain of artificial intelligence, machine learning has reached its higher level of implementation in Supply chain especially in the last decade. The industrial environment is being more hostile than ever and the client is being more demanding since new technologies have increased performance levels and have smoothened interactions between supply chain links.

In this state of affairs and as one of most feared issues for a Supply chain, late delivery does not only affects the company's image and integrity, but it also compromises the loyalty of a wide range of customers. Predicting late delivery risk beforehand can be a game changing: studies have shown that client are more tolerant about late delay when they are notified in advance. They're less disappointed and more likely to forgive the problem.

The research problem in this paper is to identify an intelligent, yet high performing machine learning model capable of predicting and forecasting late delivery orders. The rest of the sections are organized as follows: Section 2 provides the review of existing literature and analysis of review, followed by Section 3 on the supply chain 4.0 in the actual conjuncture of industry 4.0. Section 4 is the experimental deployment of the sought model and the final section presents the conclusion and future research directions.

II. RELATED WORKS

In recent years, the amount of high-dimensional data that exists has considerably expanded [3]. As a result, machine learning approaches struggle to cope with the enormous amount of input variables, offering a fascinating challenge for researchers. Pre-processing of data is required in order to employ machine learning technologies efficiently. Feature selection is one of the most common and crucial data pre-processing techniques, and it has become an essential part of the machine learning process [4]. It is defined as the process of detecting relevant features and removing irrelevant, redundant, or noisy data that has a negative impact on machine learning systems, caused by inaccuracies in attribute values (wrongly measured variables, missing values) [5]. The purpose of feature selection is to minimize needless complexity in the inferred models and increase the algorithm's performance.

Feature subset selection works by removing features that are not relevant or are redundant. The subset of features selected should give the best performance according to some objective function [6]. If not, the process of subset generated is repeated according to another criterion until getting validated results (Fig.1.). A feature is good in general if it is relevant to the class notion but not redundant to any other relevant features. The outcomes of feature selection are further verified by comparing data with and without feature selection using two different classification algorithms [7].

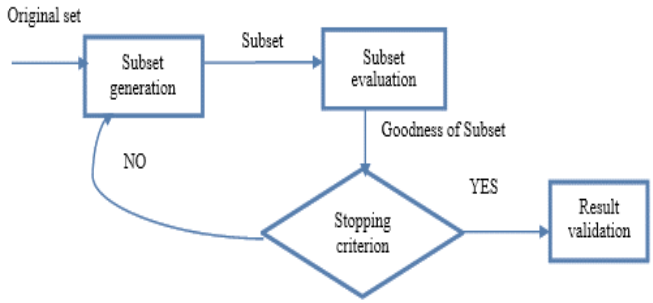

Fig.1. Feature selection process according to [8]

Supply chain is ranked among the top priorities of every industry. It is considered one of the largest areas in term of data volume, dimension and flow due to its significant importance as well as the enormous complexity of processes, functions, and operations [9]. As a consequence, feature selection has been deployed in many disciplines of SCM.

In the financial field for instance, [10] study aims to identify companies that engage in financial statement fraud employing feature selection and data mining. A comparison is done with and without feature selection on a dataset including 202 Chinese firms. Results showed that models combined with feature selection has outperformed all other approaches.

For E-Supply chain, Big Data analytics and machine learning were combined with feature selection [11] to discover the elements that influence online service supply chains and help in the creation of effective pricing strategies. ML's feature selection algorithm is used to identify the important influencing factors of 12,330 customers' online shopping. As a result, the feature selection technique demonstrates an efficiency increasing optimization and shows that client buying behavior is influenced by the level of visualization and the quality of content. Similarly, an as feature selection establishment in Supply chain is often combined with more than one machine learning algorithm, [12] suggest an approach that uses simulated annealing to pick features by combining three random features at each round and applying different machine learning algorithms to each combination of features. The goal of this combination is to determine the most appropriate characteristics for intrusion detection, which can drastically minimize the amount of training and testing required, as well as the detection time. The purpose is to find the best performing model in detecting normal traffic and attack traffic.

In another context, the performance of closed-loop supply chain Network designs is evaluated using Social Network Analysis metrics by [13]. Custom-designed network- level metrics and random forest (RF) feature selection employed in this approach to increase Supply chain design resilience when subjected to disturbances and the balance of flows.

As the related work to this paper are listed, demonstrating the importance of our research that combine new technologies with Supply chain in the actual context of industry 4.0, the question to answer by now is how this ‘new supply chain’ is different than the traditional one and what advantage can it present in the industry 4.0 conjuncture. Therefore, the next section answer those questions and demonstrate the role of Supply chain 4.0 as well as its smart architecture..

III. SUPPLY CHAIN IN THE ERA OF MACHINE LEARNING: SUPPLY CHAIN 4.0

Industry 4.0 is a rapidly growing research topic that has attracted a wide range of interest from both academia and practitioners [14]. Many studies have been deployed to get evidence regarding the influence of Industry 4.0 and its disruptive technologies on the supply chains: Supply Chain 4.0 is described as the relation between Industry 4.0 and supply chains. It is ultimately a direct consequence of the advent of Industry 4.0. This concept, born in the German industrial sector [15], responds to the introduction of new technologies that automate production and eliminate errors throughout the manufacturing process by collecting and analyzing massive amounts of data collected from across the entire supply chain. This data can be used to make informed decisions and predict future scenarios. The challenge in doing so is to provide actionable information to the users in time.

There is a certain consensus that Industry 4.0 in supply chains means much more than only technology adoption. It involves understanding the capabilities required (e.g., infrastructure, people skills, coordination) to effectively implement Industry 4.0’s technologies as well as generate the impact of these technologies on supply chains’ performance criteria (e.g., transparency, responsiveness, efficiency, flexibility) and strategic goals [16].

As the challenges of the actual economic context are becoming intriguing, and unlike a traditional supply chain model, digital supply networks are dynamic, integrated, and characterized by a high-velocity, continuous flow of information and analytics. Supply Chain 4.0 is essentially the combined use of the latest innovations in internet, robotics, and data technology. The initial step toward this new supply chain has mostly involved the creation of digitized and/or automated versions of simple, monotonous, and yet laborious tasks that occur throughout the supply-chain [17]. The gradual transition to supply chain 4.0 is realizing the promise of a more flexible and autonomous organization for companies seeking productivity. They are investing in new tools to empower and automate their means of production: Big data, connected objects and artificial intelligence are present at all levels of the production chain [18].

As a consequence, the way information flows through the supply chain is significantly altered by Supply Chain 4.0. Traditional supply chains go in a straight line from suppliers to customers, with each firm sourcing inputs from suppliers and then delivering its goods to clients. Supply Chain 4.0 is then a more homogeneous version of traditional supply chain ecosystem in which information flows in all directions (Fig.2.), analytics enable supply chain adjustments, and real-time response take place easily as supply links are becoming more dynamic. As for the customer, he is more integrated then ever which improves transparency of the interaction and the information and makes it easily reachable.

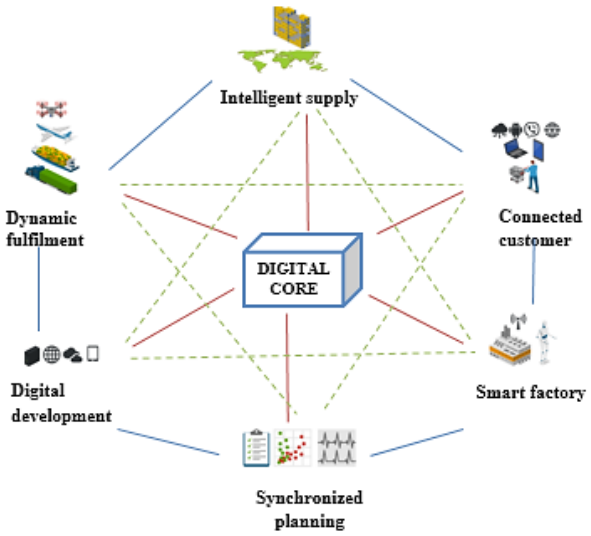

Fig.2. Supply chain 4.0 ecosystem

Companies can benefit from the digital transformation in Supply Chain terms of cost reduction, time savings and reactive decision making among other things. Organizations can now put the client at the center of the supply chain by adapting production to their needs thanks to an optimal alliance of Industry 4.0 solutions. To put it another way, adopting a new digitalization strategy for the supply chain 4.0 is now a crucial step in understanding the developing technology landscape, reaping the benefits, and therefore bolstering its competitive position.

Better communication technologies help companies to engage with one other more smoothly, allowing for faster decision-making. This enhances and accelerates decision-making as well as more others aspects:

- Productivity and efficiency: Using self-driving production equipment, storage systems, order picking, and transportation (such as drones, mobile robots, or futuristic driverless trucks), a corporation can reach its full potential.
- Errors are eliminated: specialist software, such as warehouse management systems (WMS), allows for the elimination of errors in the administration of commodities. As a result, service is much faster and more gratifying.
- Process integration: A digital supply chain allows all organizations to see data about items at any point in the process, ensuring complete product traceability.

In the next section, we treat a very challenging aspect of Supply chain: delivery management. We elaborate a machine learning model of predicting orders that might be delivered with a delay to optimize the operational aspect of supply chain 4.0.

IV. Late Delivery Predicting Model using Machine Learning

A. Proposed methodology

In this paper, we propose a model of late delivery prediction using machine learning. The implementation is executed in Python using scikit-learn. We start from dataset of past orders that has been cleaned in a pre-processing step. After that, we elaborate a training model based on the previous late delivery cases detected all over dataset past period. We use the training results to develop a test model and we deploy to that purpose Random Forest Classifier. We employ afterwards feature selection to reduce dataset and to accentuate information liability and we compare the results of three selection scenarios as we adjust the metrics. The model is tested after that and results shows that feature selection achieve better performance with an accuracy of over 99% comparing to the primary accuracy which was around 97%.

Fig.3. Feature selection process

B. Data collecting

An aerospace manufacturing business in Morocco provided the data for this case study. Data are extracted from the ERP system and are based on information history of the previous four years (from 2018 to 2021). Collected data include information about past orders to 107 order and the main inputs are constructed as follow:

- Order number: The order number assigned by ERP
- Client ID: The ID used to create the supplier in ERP
- Order cost is the price of the order line.
- Order quantity is the number of items in the order.
- Due date: The date by which the items must be delivered;
- Delivery date: The day on which the goods are actually delivered;
- Delay in days: number of days between the due date and the delivery date
- Non-compliance quantity: The quantity of goods judged as non-compliant
- Non-compliance level: The ratio between delivered goods and goods judged as non-compliant
- Delivered quantity: The quantity actually delivered
- Remaining quantity: The difference between the order quantity and the delivered quantity
- Administrator ID: The ID of the administrator of the order
- Delivery ID: The number of the delivery document
- Delivery Status: in advance, on time or late delivery

Provided that the data are imported into the Jupyter notebook database after being exported as an Excel sheet. Python is frequently used in conjunction with Jupyter, an interactive web-based development environment for notebooks, code, and data, to run machine learning programs.

C. Model implementation

The random forest classifier is used for the classification of orders status. An order can be delivered on advance, on time or late. The implementation of this study's model aims to predict if an order will be delivered late on the light on dataset lessons learned, by creating a forecasting algorithm capable of anticipating that status for a better delivery management. The choice of random forest classifier is due to its capability to successfully handle high data dimensionality and multicolinearity, being both fast and insensitive to overfitting [19]

There exists different metrics in machine learning models to measure the performance of the classification methods like sensitivity, precision, F-measure, accuracy and specificity, recall. These performance metrics are generally used to analyze the performance of different models [20]. In this

study, performance of the model will be evaluated using accuracy. It's calculated by dividing the total number of right guesses by the total number of predictions. This is defined as the proportion of data that is appropriately classified to the total amount of data that is categorized.

The mathematic formula for accuracy is given as follows:

Accuracy= (TP +TN) / (TP + FP + TN + FN) *(1)*

In above equation:

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

The primary results using Random forest Classifier demonstrate an accuracy of 97.9% (Fig.4.). This ratio can yet be optimized. For this purpose, we deploy in the following section a feature selection method using SelectKbest from ANOVA.

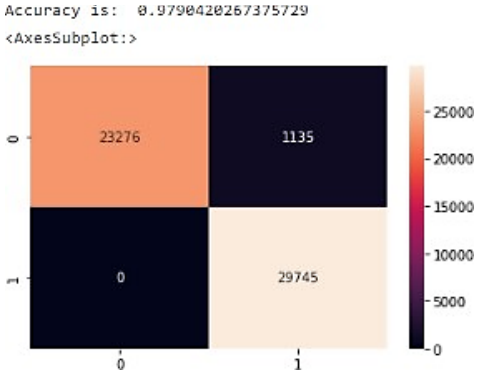

Fig.4. Confusion matrix of primary Random Forest Classifier model

D. *Feature selection*

Feature selection, also known as Variable Selection [21], is an extensively used data preprocessing technique in data mining which is basically used for reduction of data by eliminating insignificant and superfluous attributes from any dataset [22]. This technique improves data comprehension, promotes improved data visualization, decreases learning algorithm training time, and improves prediction performance.

To acquire more accurate results in less time, it's always advisable to remove noisy and inconsistent data before applying any model to the data. In real-world applications, reducing the dimensionality of a dataset is critical. Furthermore, selecting the most critical traits reduces the complexity significantly [22].

Feature selection in this study in established using ANOVA f-test in the f_classif() function, provided by the scikit-learn machine library. ANOVA stands for "analysis of variance," and it is a parametric statistical hypothesis test that determines whether the means of two or more samples of data are from the same distribution or not. An F- statistic, also known as an F-test, is a class of statistical tests that use a statistical test like ANOVA to calculate the ratio between variance values, such as the variance from two separate samples or the explained and unexplained variance. An ANOVA f-test is a sort of F-statistic that uses the ANOVA approach.

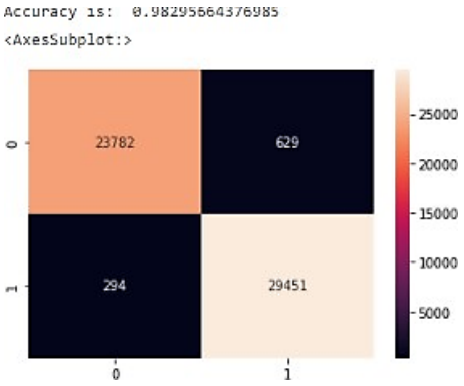

Fig.5. K=5 Confusion Matrix

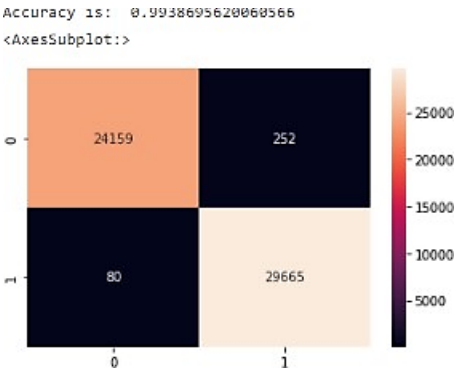

Fig.6. K=7 Confusion Matrix

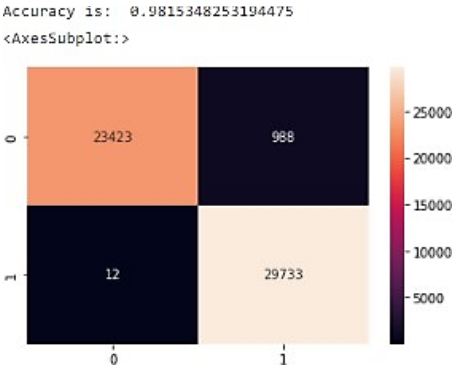

Fig.7. K=10 Confusion Matrix

To establish our feature selection model, we adjust features number to get the best possible prediction. As so, we employ three scenario of feature selection starting by K=5: five top most relevant features (Fig.5.), k=7: seven top most relevant features (Fig.6.), and k=10: ten top most relevant features (Fig.7.)

Results reveal that the best accuracy is reached using K=7, the model is clearly optimized comparing to the primary model where only random forest classifier was used (Tab.1.)

	Primary Random Forest Classifier Model	**Feature selection with f_classif() function**		
		K= 5	K= 7	K=10
Model accuracy	0.979	0.982	0.993	0.981

Tab.1. comparison of models accuracy

The elaborated model allows us to lower noise and boost model accuracy by choosing only the most important elements and consequently reducing the length of training. The model is in this case trained more quickly since features that are not relevant for the prediction are excluded. Otherwise, over excluding features can have an opposite effect (the case of K= 10). Choosing the biggest K value does not mean that the model will have a better performance.

The next and last section is a summarizing of the work done in this paper as well as some further perspectives to our study.

V. CONCLUSION

The main focus of this research is to identify the role of feature selection in optimizing machine learning models in a supply chain 4.0 context, starting from a dataset of past clients orders all over the world in order to get a forecasting for next orders delivery status. In the proposed method, a python model using random forest classifier was elaborated and then combined with ANOVA's selectKbest method by the mean of f_classif() function. The experimental results validate that feature selection increase the model's efficiency on prediction late delivery orders and outperformed the primary Random forest model with an accuracy over 99%. Our further work will extend the study on other Supply chain processes as we combine machine learning and deep learning models to achieve the higher level of performance.

REFERENCES

[1] Sarah El Hamdi. Mustapha Oudani. Abdellah Abouabdellah , Morocco's Readiness to Industry 4.0,International conference on the Sciences of Electronics, Technologies of Information and Telecommunications 2019.

[2] Castelo-Branco, I.; Cruz-Jesus, F.; Oliveira, T. Assessing Industry 4.0 readiness in manufacturing: Evidence for the European Union. Comput. Ind. 2019, 107, 22–32.

[3] Shahana AH, Preeja V. Survey on feature subset selection for high dimensional data. In: Circuit, power and computing technologies (ICCPCT), 2016 international conference on. IEEE; 2016. p. 1–4.

[4] Kumar, Vipin & Minz, Sonajharia. (2014). Feature selection: A literature review. Smart Computing Review. 4. 211-229. 10.1145/2740070.2626320

[5] Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Adv Bioinformatics. 2015;2015:198363. doi: 10.1155/2015/198363.

[6] Balakrishnan S, Narayanaswamy R, Savarimuthu N, Samikannu R. SVM ranking with backward search for feature selection in type II diabetes databases. In: Systems, man and cybernetics, 2008. SMC 2008. IEEE international conference on. IEEE; 2008. p. 2628–33.

[7] Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation- Based Filter Solution. In T. Fawcett, & N. Mishra (Eds.), Proceedings, Twentieth International Conference on Machine Learning (pp. 856-863).

[8] M. Dash [M. Dash, H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, Elsevier, pp. 131-156, 1997.

[9] Biswas, S. (2020). Measuring performance of healthcare supply chains in India: A comparative analysis of multi-criteria decision making methods. Decision Making: Applications in Management and Engineering, 3(2), 162–189.

[10] P. Ravisankar, V. Ravi, G. Raghava Rao, I. Bose, Detection of financial statement fraud and feature selection using data mining techniques, Decision Support Systems, Volume 50, Issue 2, 2011, Pages 491-500, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2010.11.006.

[11] Li, L., Ma, S., Han, X., Zheng, C. and Wang, D. (2021), "Data-driven online service supply chain: a demand-side and supply-side perspective", Journal of Enterprise Information Management, Vol. 34 No. 1, pp. 365-381. https://doi.org/10.1108/JEIM-11-2019-0352.

[12] Seong, Teh & Ponnusamy, Vasaki & Zaman, Noor & Annur, Robithoh & Talib, Maria. (2021). A comparative analysis on traditional wired datasets and the need for wireless datasets for IoT wireless intrusion detection. Indonesian Journal of Electrical Engineering and Computer Science. 22. 1165. 10.11591/ijeecs.v22.i2.pp1165-1176.

[13] Akbar Ghanadian, S., & Ghanbartehrani, S. (2021). Evaluating Supply Chain Network Designs: An Approach Based on SNA Metrics and Random Forest Feature Selection. Journal of Operations and Management Research, 1(1), 15-35.

[14] Frederico, G.F., Garza-Reyes, J.A., Anosike, A. and Kumar, V. (2020), "Supply Chain 4.0: concepts, maturity and research agenda", Supply Chain Management, Vol. 25 No. 2, pp. 262-282. https://doi.org/10.1108/SCM-09-2018-0339.

[15] Ghobakhloo, M. The future of manufacturing industry: A strategic roadmap toward Industry 4.0. J. Manuf. Technol. Manag. 2018, 29, 910–936.

[16] Frederico, G.F. From Supply Chain 4.0 to Supply Chain 5.0: Findings from a Systematic Literature Review and Research Directions. Logistics 2021, 5, 49. https://doi.org/10.3390/logistics5030049

[17] Why Traditional Supply-Chain Management Systems Are Dying Building Up from Supply- Chain 4.0 SupplyBloc Technology Jul 18, 2018.

[18] Cañas, Héctor, Josefa Mula, and Francisco Campuzano-Bolarín. 2020. "A General Outline of a Sustainable Supply Chain 4.0" Sustainability 12, no. 19: 7978. https://doi.org/10.3390/su12197978.

[19] Mariana Belgiu, Lucian Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 114, 2016, Pages 24-31, ISSN 0924-2716, https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[20] Divya Jain, Vijendra Singh, Feature selection and classification systems for chronic disease prediction: A review, Egyptian Informatics Journal, Volume 19, Issue 3, 2018, Pages 179- 189, ISSN 1110-8665, https://doi.org/10.1016/j.eij.2018.03.002.

[21] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3(Mar):1157–82.

[22] Tang J, Alelyani S, Liu H. Feature selection for classification: a review. Data Classif: Algor Appl 2014;37.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.