

Automated machine learning: the new data science challenge

Ilham Slimani¹, Nadia Slimani², Said Achchab³, Mohammed Saber¹, Ihame El Farissi¹, Nawal Sbiti²,
Mustapha Amghar²

¹SmartICT Laboratory, ENSAO, Mohammed First University, Oujda, Morocco

²Computer Systems and Productivity Team, EMI, Mohammed V University, Rabat, Morocco

³ADMIR Laboratory, ENSIAS, Mohammed V University, Rabat, Morocco

Article Info

Article history:

Received Jun 19, 2021

Revised Mar 23, 2022

Accepted Apr 15, 2022

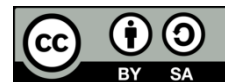
Keywords:

Artificial intelligence
Artificial neural networks
Automated machine learning
Convolutional neural network
Data science
Long short term memory
Machine learning

ABSTRACT

The world is changing quite rapidly while increasingly tuning into digitalization. However, it is important to note that data science is what most technology is evolving around and data is definitely the future of everything. For industries, adopting a “data science approach” is no longer an option, it becomes an obligation in order to enhance their business rather than survive. This paper offers a roadmap for anyone interested in this research field or getting started with “machine learning” learning while enabling the reader to easily comprehend the key concepts behind. Indeed, it examines the benefits of automated machine learning systems, starting with defining machine learning vocabulary and basic concepts. Then, explaining how to, concretely, build up a machine learning model by highlighting the challenges related to data and algorithms. Finally, exposing a summary of two studies applying machine learning in two different fields, namely transportation for road traffic forecasting and supply chain management for demand prediction where the predictive performance of various models is compared based on different metrics.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Ilham Slimani
SmartICT Laboratory, Mohammed First University
Oujda, Morocco
Email: slimani.ilham@gmail.com

1. INTRODUCTION

We have all probably heard the expression «data is the new oil»; well it turns out that data now worth so much more than oil. Indeed, it is true that both data and oil generate value, however oil is “used up” but data is not in a way that it can be renewable and reused in so many different ways while increasing value [1]. Certainly, the future is already here and big data is everywhere since we are living in a constantly changing world that creates huge amount of data every single day. The term “data” includes everything from words and ideas, to sounds, pictures or videos, to personal data (name, age, gender, height) and of course, to anything that can be collected from industrial processes or internet of things (IoT) sensors [2]. In other words, data is anything operated by computers and can be converted to binary numbers (0s and 1s).

Data science is definitely the future of everything”. It is an interdisciplinary field based on scientific methods, algorithms and processes with the aim of extracting knowledge and value from data in both structured and unstructured forms, analogous to data mining. In order to gain a competitive advantage, companies should start leveraging data by using it for their profit maximization. Nevertheless, data is not an instantly valuable resource; data must be collected, cleaned, improved, before it can be analyzed; and this process is accomplished using artificial intelligence (AI) techniques.

In a typical business problem, where a bank of three million costumers, for instance, introduces a new card. The bank's sales marketing team executed a campaign, offering it to thousands of their existing costumers. Unfortunately, only 3% of them signed up for it and that is much fewer than expected. Therefore, costumers are divided to two groups: i) the first group: 3% are delighted to use this new card maybe for online shopping and so on and ii) the second group: 97% did not want the new credit card, maybe because they prefer to shop locally and pay cash.

With the purpose of profit maximization, the main question for the bank is how to target more people from the first group and fewer people from the second one in the next marketing campaign? Well the answer is in the data! Usually, it takes months to get a project like that off the ground and deploy a predictive targeting model. However, it goes faster with the use of AI techniques, where data resides in a database and the next marketing action is planned based on information from the last campaign: historical data with a detailed record of all activities on costumers' bank account transactions. Consequently, it becomes an easy task separating costumers who are interesting in the new product from those who are not.

The present paper offers a roadmap for anyone interested in this research field or getting started with "machine learning" learning while enabling the reader to easily comprehend the key concepts behind. Indeed, this paper is structured: As an introduction, the importance and power of data in a digitalization global trend are highlighted in the first part. The second part is dedicated to the fundamentals of machine learning (ML) starting with basic concepts such as AI, automation and automated ML [3], classification and regression, deep learning (DL). Afterwards, in the forth section of this paper, the "how to" of building a ML model is explained in details. Followed by ML applications for forecasting purposes in different fields namely, transportation for road traffic forecasting [4] and supply chain management (SCM) [5] for demand prediction [6] using real datasets from Morocco. In those studies, the predictive performance of various models including auto regressive moving average (ARIMA), multi layer perceptron (MLP), support vector regression (SVR or SMOReg), long short term memory (LSTM), convolutional neural network (CNN) are compared based on different metrics. Finally, the conclusion and perspectives.

2. THE FUNDAMENTALS OF MACHINE LEARNING: LITERATURE REVIEW

2.1. AI vs ML vs DL: Biggest confusion

AI has become the latest "buzzword" in the industry today. As displayed in the following Figure 1, it has many applications in many subdomains such as robotics, planning, speech or picture regognition, cyber security. However, we often confuse terms like AI, ML [7] or DL [8]: i) AI enables computers to perform tasks that normally require human intelligence by allowing machines to mimic human behavior, ii) ML is a sub-field of AI that uses statistical learning algorithms and enables machines to learn and improve from experience to learn from their own [9], iii) DL is a ML subset, also defined as a multilayered neural network architecture containing a large number of parameters and layers. Naming CNN or recurrent neural network (RNN). DL is a technique that processes information in a similar way to how a human brain processes it. It is used in various industries such as transportation, finance, and advertising. Most of DL's models are based on neural networks structures this is why they are often denoted as deep neural networks (DNN).

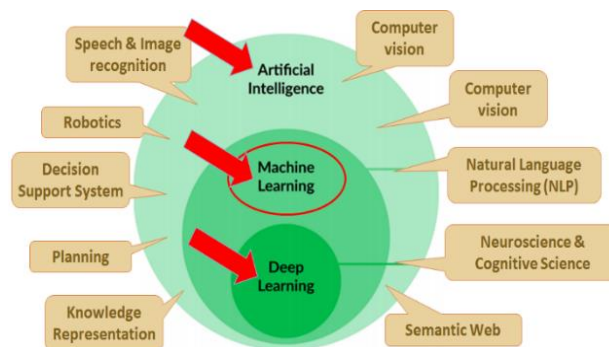


Figure 1. AI vs ML vs DL, AI applications

2.2. Difference between automation, ML and automated ML concepts

The terms "automation" and "Machine learning" are two very different concepts that perfectly complement each other, but often confused when used interchangeably. In short, Automation is more about

“doing” without any further thought or guidance and ML is all about “thinking” by mimicking human intelligence. The concept of automation has been around since ancient times, its practice was focused on controlling machines and devices. More precisely, it is the idea of using a machine to do repetitive tasks without human involvement in order to speed workflows. Whereas, ML is a process that enables a system to make itself smarter over time by taking advantage of the various data and experiences it has. There are different levels of automation in business: starting from desktop application and manual intervention, to robotic and process automation with self-service, then AI and ML based on data-driven process. A new related trend is gaining momentum called “Automated ML” [10]. Essentially, this concept offers ML expert tools to apply ML to ML itself in order to automate repetitive tasks related to ML like the choice of the convenient dataset, data preparation or even feature selection [11].

2.3. Traditional vs ML programming approach

If the problem is complex and you opt for the traditional approach as a resolution method, you might end up with a long list of rules that are hard to maintain. As illustrated in the following Figure 2, after studying the problem and writing the appropriate program, the proposed solution is evaluated in order to launch it if it works or analyze and correct the errors if it does not. This approach is usually not very accurate and can be hard to implement. The ML programming approach, on the other hand, is usually the best solution for complex problems and is mostly based on data. Indeed, after studying the problem, the resolution process starts from an initial database where data is majorly treated and pre-processed before it can be loaded into ML algorithms for feature engineering, and then we go through training and test phases with the aim of searching the best classification or predictive performance with the possibility to iterate if needed. To get started with ML learning, we should be familiar with ML vocabulary and of course with other disciplines starting from maths to programming, to databases and ML algorithms and tools.

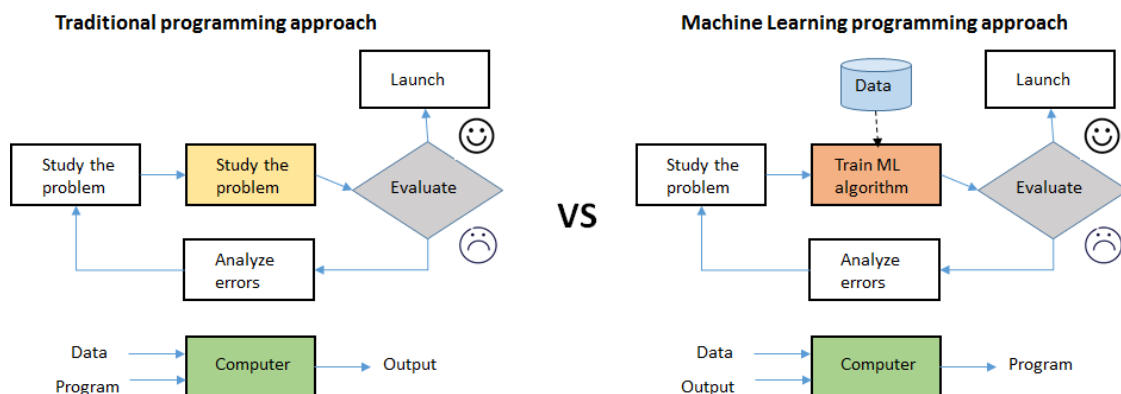


Figure 2. Traditional vs ML approach [7]

2.4. Taxonomy of ML systems

Based on the amount of data and type of supervision gotten during the training process, ML systems can be divided into four main categories: i) supervised learning: in this type of learning algorithm, the problem can be either classification (logistic regression, k-nearest neighbour (KNN), support vector machine (SVM), naïve Bayes) or regression (decision tree (DT), linear regression, random forest (RF), support vector regression (SVR) [12]). It is used with labeled data where the mapping patterns from the training set of a given model is learned from input to output. Then, the proposed model is trained to predict the response of a new dataset called test set; ii) unsupervised learning: this learning algorithm is performed when data is not labeled. Moreover, unlike supervised learning, this algorithm is left on his own to group data by finding differences or resemblances in the input patterns. It is generally used for clustering problems (k-means, mean-shift, apriori) or dimensionality reduction problems (principal component analysis (PCA), feature selection, linear discriminant analysis (LDA)); iii) semisupervised learning: with a mixture of some labeled and many unlabeled training data, semisupervised algorithms are a combination of supervised and unsupervised techniques, such as heuristic approaches, generative models, and iv) Reinforcement learning: Is an environment where an agent learns how to find the best strategy by continuously interacting with it in a trial and error method. The agent receives feedback for its earlier actions and experiences and is rewarded for a correct performance or punished for incorrect actions.

3. BUILDING A MACHINE LEARNING MODEL

3.1. Challenges and limitations of machine learning

In general, the aim of machine learning is to select the best algorithm and train it on an adequate dataset. It follows that the relevance of the algorithm and the training database play a crucial role in the accuracy of the results. In this section, we will focus on the different challenges and limitations that can hinder the performance of the chosen model, namely: bad data (insufficient training data, irrelevant features) and bad algorithm (topologie, overfitting or underfitting of the training dataset).

3.1.1. Insufficient training data

The first hurdles faced by the machine learning user is the availability, completeness, and relevance of the learning and testing database. According to the widely cited article [13], Google researchers Halevy, Norvig, and Pereira argue that for complex problems such as linguistic expressions and learning from texts, the performance of the machine learning process is of paramount importance dependent on the data regardless of the algorithm used, so “we may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus development”. The size of the training database is very important for the model’s accuracy, which gives good results depending on the size of the database; and regardless of the model used [14].

3.1.2. Nonrepresentative and poor quality training data

By utilizing a nonrepresentative training set, we prepare a model that is probably not going to make precise forecasts; it is harder for the system to detect the underlying patterns, so the system is less likely to perform well. Chen *et al.* [15] deduces that the results of the experiments showed a strong correlation between the quality of the datasets and the performance of the machine learning system. The results also demonstrated that a rigorous evaluation of data quality is necessary for guiding the quality improvement of machine learning.

3.1.3. Irrelevant features

The success of machine learning project depends on the detection of the key characteristics that drive the output. It consists on detecting the good set of features to train on, this process is called feature engineering and involves [16]: i) feature selection: selecting the most useful features to train the model by analysing the historical data and ii) feature extraction: combining and producing more useful features and introduction to the test set.

3.1.4. Overfitting/underfitting the training data

Overfitting means that the model performs well on the training data [17], but it does not generalize well as shown in Figure 3. It happens when the model is too complex relative to the amount and noisiness of the training data. The possible solutions are [18]: i) to simplify the model by reducing parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of attributes in the training data or by constraining the model, ii) to expand the training data, iii) to clean training data (e.g., fix data errors and remove outliers), and iv) constraining a model to make it simpler and reduce the risk of overfitting is called regularization.

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> • High training error • Training error close to test error • High bias 	<ul style="list-style-type: none"> • Training error slightly lower than test error 	<ul style="list-style-type: none"> • Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			

Figure 3. Illustration of overfitting and underfitting in ML [7]

Besides, underfitting is something contrary to overfitting: it happens when the model is too easy to even consider learning the basic construction of the information. The fundamental choices to fix this issue are: i) selecting an all the more remarkable model, with more boundaries, ii) redoing the historical analysis of the data to incorporate more relevant characteristics for the model, and iii) reducing the requirements on the model.

3.2. End to end machine learning operating mode

This section is dedicated to explain in detail how to design and implement a machine learning model from end to end. Indeed, starting from understanding the business problem and identifying related data, then preparing it by determining the training and test sets, followed by an evaluation of the model's performance. The main steps to follow are listed below [19], [20] and are shown in Figure 4.

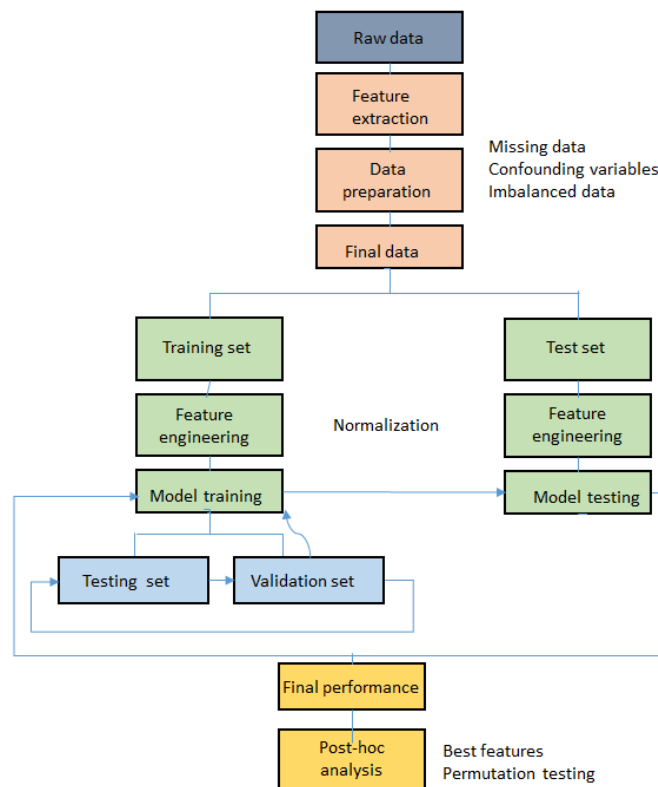


Figure 4. Steps to build a ML model

3.2.1. Prepare the data

The first step of implementing a machine learning model is preparing the data. This phase consists of framing the problem by doing requirement analysis, then characterizing and developing the problem being addressed (objectives, inputs and outputs). Based on this analysis, sources of information is identified in order to do the cleaning and set up, then to be filled up with the features' values.

3.2.2. Handling missing data

There are some techniques to manage the missing data in machine learning algorithm. The obvious method is to ignore instances with unknown feature values [21]. The other ones are to fill up the missing feature with most common feature value, or the mean value computed from available cases to fill in missing data values on the remaining cases. Another method is treating missing feature values as special values. These would be classified as outliers in the laterstage.

3.2.3. Discretization

The discretization process refers to the conversion of continuous attributes, variables or features to discretized ones. It aims to fundamentally reduce the quantity of potential upsides of the constant component since largenumber of conceivable element esteems to slowand ineffective process. However, the problem of

choosing the interval borders for the discretization of a numerical value range remains an open problem in numerical feature handling [22].

3.2.4. Feature selection

Feature selection is the process of identifying and removing possible irrelevant and redundant features by getting rid of noise in data. Generally, features are characterized as relevant if they have an influence on the output and their role cannot be assumed by the rest feature. This is assured by choosing relevant features based on the type of problem studied.

3.2.5. Normalization

This step consists on “scaling down” transformation of the features. Within a feature there is often a large difference between the maximum and minimum values. When normalization is performed the value magnitudes are scaled to appreciably low values. The two-most-common methods for this scope are min-max normalization and z-score normalization.

3.2.6. Select and train the model

To ensure the proper functioning, the model needs an input data and a desired output data in order to compare it with the obtained output then calculate the total model error. To reduce the model’s error, a learning phase is adopted. During this phase, a model adjustment is performed in order to determine the appropriate value of the connections. Select performance measurement indicators, with which to measure the accuracy of the results and eventually compare the results provided by the designed system with other models.

4. MACHINE LEARNING APPLICATIONS: TRANSPORTATION AND SUPPLY CHAIN MANAGEMENT

4.1. Transportation: Road traffic prediction

4.1.1. Research method

With the rapid development of the cities, and the lack of public transport with low availability coupled with long waiting times, households encouraged by credit facilities own at least one car. That is why every country in the world is experiencing a rapid development of motorized transport. However, the development of road infrastructure has not kept pace, which implies several bottlenecks [23]. This phenomenon has a direct impact on the quality of life, especially in urban areas, which are still characterized by almost permanent traffic jams and high levels of air pollution [24].

Prediction is definitely a key component of road traffic management. Considering that, this study tackles the use of various methods (parametric and non-parametric [25]) to forecast traffic based on real Moroccan dataset of a toll station with high traffic volume. The approach consists on comparing, according to predefined criteria, the predictive performances of three different methods, namely: i) neural networks structure MLP [4] as a non parametric model, ii) mathematical modeling method seasonal ARIMA (SARIMA) [26] as a parametric model, and iii) SMOREg inspired from the SVM algorithm [12] as a non parametric model.

4.1.2. Results and discussion

The dataset contains 30792 recordings of hourly traffic flow. This dataset is divided into 42 months of traffic flow information; it is separated into two parts: the daily traffic of three years (i.e. 85% of the data) for training, and the daily traffic of the following six months (i.e. 15% of the remaining data) for testing. Several networks topologies and combinations were tested and the best results were offered with a neural network with the following characteristics:

- a) An input layer provided by 18 values: i) 8 calendar information: working day, weekend, national holiday, religious holiday, school holiday, Ramadan, strike and chronological order of the day in the holidays and ii) 10 previous daily traffic flows: the flow for a given day d is predicted using the historical flows of the days: $d-1$, $d-2$, $d-3$, $d-4$, $d-5$, $d-6$, $d-7$, $d-14$, $d-21$, $d-365$.
- b) The output represents the expected traffic flow d of the day in the next order.
- c) Multi-layer perceptron (MLP) of three hidden layers, the first hidden layer is composed of five neurons, the second is composed of eight neurons and the third one is composed of two neurons. The transfer function is Sigmoid and Backpropagation as learning algorithm with the existence of the Bias neuron.

The best MLP result is obtained with a total mean square error (MSE) of 0.00927 in the train set and 0.01321 in the test set. The MLP recorded the best forecasting performance with 0.57% absolute error, as illustrated in the following Tables 1 and 2. The obtained results, presented in Table 2, prove that the neural

network gives a satisfactory accuracy of the forecast. The comparison of the observed and modelled traffic flows in Figure 5 shows that over 80% of the forecasts were made with less than $\pm 5.0\%$ error. A more intensive gander at the fundamentally various conjectures showed that the observed traffic peaks that recorded a worse forecast performance are days when other exogenous factors were not introduced as features: mainly bad weather and pavement maintenance operations.

Table 1. Daily traffic forecasting with MLP 5-8-2

Day	Actual traffic	Forecasted traffic	Difference
08/01/2018	23147	22902	245
17/01/2018	27739	28176	-437
28/02/2018	27471	28347	-876
24/05/2018	25080	24587	493

Table 2. Absolute error comparison on June 2018

Model	Total traffic	Absolute error	Relative error
Actual traffic	827913	-	-
MLP	832837	4924	0,57%
SARIMA	841929	14016	1,69%
SMOreg	807560	20353	2,45%

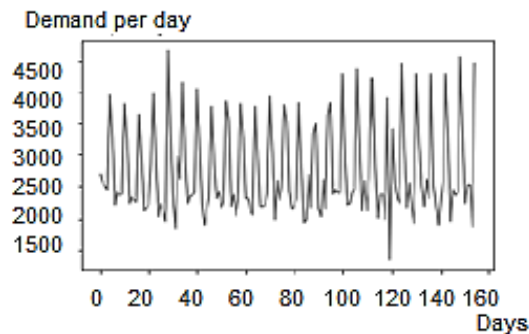


Figure 5. Data visualization per day

4.2. SCM: Demand forecasting

4.2.1. Research method

Prediction is a process that uses historical data to forecast future trends. It is performed using various time series models. Demand forecasting is a key component to improve the supply chain's (SC) performance, since having a clear vision of future demand gives each SC's member the opportunity to optimize its performance by minimizing costs, related to inventory, manufacturing, transportation or distribution; and maximizing the profit. In this study [6], different statistical and deep learning models are studied in a comparison analysis aiming to demonstrate which one performs better in terms of providing accurate forecasts, namely: i) Autoregressive integrated moving average (ARIMA) as a statistical model, ii) Multi layer perceptron (MLP) as a feedforward neural network, iii) Long short term memory (LSTM) [27] as a recurrent neural network: is capable of learning long-term dependencies, its architecture is composed by a collection of subnets or also called blocks of memory where the memory cell stores the state, the front door controls what to learn, the door of oblivion controls what to forget and the exit door controls the amount of data to modify, and iv) Convolutional neural networks (CNN or ConvNet) [27]: Like the MLP structure, CNN has three type of layers: the input layer, the output layer and the hidden layer that is categorized in two types the feature learning layers (convolution, pooling, and rectified linear) and the classification layers (fully connected layers and normalization layers). CNN shows its performance in many fields such as image classification and segmentation, face recognition, object detection, traffic sign recognition, speech processing.

4.2.2. Results and discussion

In the proposed neural network, the demand of a given day d is predicted using demand quantities of the days: $d-6$, $d-12$, and $d-18$. Therefore, we model the problem as a three time-steps prediction on one input corresponding to the quantity of the day, and this for each output. This instead of considering that the data is

composed of three inputs as we did in the MLP study. The LSTM network is composed of 50 neurons in the hidden layer and 1 neuron in the output layer. The root mean squared error and the efficient Adam version of the stochastic gradient is used. This yielded an RMSE after normalization of 0.138 and an root-mean-square error (RMSE) before normalization of 457.958.

The proposed convolutional neural network is composed of the following layers: i) a one-dimensional convolutional layer with 64 filters, a kernel size of 2 and an input shape of (3,1), the activation function consists of the relu function; ii) A one-dimensional max pooling layer with a filter of size 2; iii) a flatten layer; and iv) a fully connected layer composed of 50 neurons with relu being the activation function. The model was trained using the adam version of stochastic descent and the MSE loss function. The training after 200 epochs results in a RMSE of 0.138 after normalization, and 457.079 before normalization.” [6]. Daily demand tendency with an overview of comparing actual and forecasted demand using LSTM model are illustrated in the following Figure 5.

The following Table 3 illustrates the various metrics used to measure the performance of each proposed model including ARIMA, MLP, LSTM and CNN. The numerical experimentations are approved using a real dataset provided by a recognized supermarket in Morocco. The results clearly show that the CNN gives slightly better forecasting results than the LSTM network [6].

Table 3. ARIMA vs MLP vs LSTM vs CNN: performance metrics

Performance metrics	ARIMA	MLP	LSTM	CNN
RMSE (before normalization)	485 690	464 261	457 958	457 079
RMSE (after normalization)	-	0.14	0.138	0.138
Training time	1 129 201	2 341 361	2 541 292	2 720 236
Consumed energy (Joule)	16 938 010	35 120 412	38 119 376	40 803 534

5. CONCLUSION AND FUTURE WORK

Machine learning algorithms automatically extract knowledge from input datasets duly accompanied by the identified features. Unfortunately, their success is usually dependant on the quality of data that they operate on. Our society is creating vast amounts of data every day, besides; programming relieves people by managing routine tasks that is why the real future lies within DL. Although not fully developed, this branch of data science is the closest we have gotten to any time of applied AI. The strength of ML lies on computers' capacity to learn how to best perform such tasks. Automated ML helps computers to learn how to optimize the outcome of learning and how to perform the routine actions (applying ML to ML itself). In other words, Automated ML offers ML experts tools to automate repetitive tasks by applying ML to ML itself. Yet, to improve business performances, there is a new data science challenge called “dark data” which is the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes (for example, analytics, business relationships and direct monetizing). It often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only. Storing and securing data typically incurs more expense (and sometimes greater risk) than value.

As future work, dark data is an interesting research field that still needs to be exploited using of course automated ML techniques. Overall, it is worth thinking about dark data as unfulfilled value. The key to monetising dark data lies not only in gathering it, but also in analyzing it to discover patterns and putting the insights to use. By utilising new technologies around ML, specifically deep learning, businesses can join structured and unstructured data sets together to provide high-value results and enhance business performances by treating: Neglected information: unmanaged, unclassified or unknown data; web log files, audio/video, email archive, visitor tracking data; storage costs, hidden risks.




REFERENCES

- [1] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, and H. Oza, “Application of machine learning and artificial intelligence in oil and gas industry,” *Petroleum Research*, vol. 6, no. 4, pp. 379–391, Dec. 2021, doi: 10.1016/j.ptlrs.2021.05.009.
- [2] A. Ghasempour, “Internet of things in smart grid: architecture, applications, services, key technologies, and challenges,” *Inventions*, vol. 4, no. 1, pp. 22–33, Mar. 2019, doi: 10.3390/inventions4010022.
- [3] L. Vaccaro, G. Sansonetti, and A. Micarelli, “An empirical review of automated machine learning,” *Computers*, vol. 10, no. 1, p. 11, Jan. 2021, doi: 10.3390/computers10010011.
- [4] N. Slimani, I. Slimani, N. Sbiti, and M. Amghar, “Traffic forecasting in Morocco using artificial neural networks,” *Procedia Computer Science*, vol. 151, pp. 471–476, 2019, doi: 10.1016/j.procs.2019.04.064.
- [5] H. Bousqaoui, I. Slimani, and S. Achchab, “Information sharing as a coordination tool in supply chain using multi-agent system and neural networks,” in *Advances in Intelligent Systems and Computing*, Springer International Publishing, 2018, pp. 626–632.
- [6] H. Bousqaoui, I. Slimani, and S. Achchab, “Comparative analysis of short-term demand predicting models using ARIMA and




- deep learning,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 4, pp. 3319–3328, Aug. 2021, doi: 10.11591/ijece.v11i4.pp3319-3328.
- [7] B. Boehmke and B. Greenwell, *Hands-on machine learning with R*. Chapman and Hall/CRC, 2019.
- [8] C. C. Aggarwal, *Neural networks and deep learning*. Springer International Publishing, 2018.
- [9] M. Saber *et al.*, “A comparative performance analysis of the intrusion detection systems,” 2020, pp. 192–200.
- [10] J. Waring, C. Lindvall, and R. Umeton, “Automated machine learning: Review of the state-of-the-art and opportunities for healthcare,” *Artificial Intelligence in Medicine*, vol. 104, Apr. 2020, doi: 10.1016/j.artmed.2020.101822.
- [11] R. Budjač, M. Nikmon, P. Schreiber, B. Zahradníková, and D. Janáčková, “Automated machine learning overview,” *Research Papers Faculty of Materials Science and Technology Slovak University of Technology*, vol. 27, no. 45, pp. 107–112, Sep. 2019, doi: 10.2478/rput-2019-0033.
- [12] N. Slimani, I. Slimani, N. Sbiti, and M. Amghar, “Machine learning and statistic predictive modeling for road traffic flow,” *International Journal of Traffic and Transportation Management*, vol. 03, no. 01, pp. 17–24, Mar. 2021, doi: 10.5383/JTTM.03.01.003.
- [13] A. Halevy, P. Norvig, and F. Pereira, “The unreasonable effectiveness of data,” *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009, doi: 10.1109/MIS.2009.36.
- [14] A. Géron, *Hands-on machine learning with scikit-learn, keras, and tensorflow: concepts, tools, and techniques to build intelligent systems*. O’Reilly Media, Inc., 2019.
- [15] H. Chen, J. Chen, and J. Ding, “Data evaluation and enhancement for quality improvement of machine learning,” *IEEE Transactions on Reliability*, vol. 70, no. 2, pp. 831–847, Jun. 2021, doi: 10.1109/TR.2021.3070863.
- [16] Y. Li, R. Liu, C. Wang, L. Yangning, N. Ding, and H.-T. Zheng, “Learning purified feature representations from task-irrelevant labels,” *arXiv preprint arXiv:2102.10955*, Feb. 2021.
- [17] X. Ying, “An overview of overfitting and its solutions,” *Journal of Physics: Conference Series*, vol. 1168, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
- [18] D. Bashir, G. D. Montañez, S. Sehra, P. S. Segura, and J. Lauw, “An information-theoretic perspective on overfitting and underfitting,” in *AI 2020: Advances in Artificial Intelligence*, Springer International Publishing, 2020, pp. 347–358.
- [19] S. Vieira, R. Garcia-Dias, and W. H. Lopez Pinaya, “A step-by-step tutorial on how to build a machine learning model,” in *Machine Learning*, Elsevier, 2020, pp. 343–370.
- [20] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, “‘What is relevant in a text document?’: An interpretable machine learning approach,” *Plos One*, vol. 12, no. 8, Aug. 2017, doi: 10.1371/journal.pone.0181142.
- [21] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data preprocessing for supervised learning,” *International Journal of Computer Science*, vol. 1, pp. 111–117, 2006.
- [22] H. Shrivastava and S. Sridharan, “Conception of data preprocessing and partitioning procedure for machine learning algorithm,” *International Journal of Recent Advances in Engineering and Technology (IJRAET)*, vol. 1, no. 3, 2013.
- [23] T. Nagatani, “Traffic flow stabilized by matching speed on network with a bottleneck,” *Physica A: Statistical Mechanics and its Applications*, vol. 538, Jan. 2020, doi: 10.1016/j.physa.2019.122838.
- [24] J. Lu, B. Li, H. Li, and A. Al-Barakani, “Expansion of city scale, traffic modes, traffic congestion, and air pollution,” *Cities*, vol. 108, Jan. 2021, doi: 10.1016/j.cities.2020.102974.
- [25] B. L. Smith, B. M. Williams, and R. Keith Oswald, “Comparison of parametric and nonparametric models for traffic flow forecasting,” *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, Aug. 2002, doi: 10.1016/S0968-090X(02)00009-8.
- [26] N. Slimani, I. Slimani, M. Amghar, and N. Sbiti, “Road traffic forecasting using a real data set in Morocco,” *Procedia Computer Science*, vol. 177, pp. 128–135, 2020, doi: 10.1016/j.procs.2020.10.020.
- [27] H. Eskandari, M. Imani, and M. P. Moghaddam, “Convolutional and recurrent neural network based model for short-term load forecasting,” *Electric Power Systems Research*, vol. 195, Jun. 2021, doi: 10.1016/j.epsr.2021.107173.

BIOGRAPHIES OF AUTHORS






Ilham Slimani    Engineer graduated from National School of Applied Sciences Oujda (ENSAO) and PhD from the National School of Computer Science and System Analysis (ENSIAS) in Rabat. She is currently Professor at the Department of Computer Science at Mohammed First University, Faculty of Sciences Oujda, Morocco. Her research and publication interests include machine learning, supply chain management and demand forecasting, transportation and road traffic forecasting, intrusion detection in computer networks, IoT and Finance. She can be contacted at email: slimani.ilham@gmail.com.






Nadia Slimani    Engineer graduated from the Mohammedia School of Engineering (EMI) in Rabat and currently PhD student in Mohammedia School of Engineering. She has a long experience with a national transport operator in Morocco. She also attended the higher cycle of management at the ISCAE of Rabat. Her research and publication interests include machine learning, transportation and road traffic forecasting. She can be contacted at email: slimani.nadia@gmail.com.






Said Achchab    holder of a PHD in Applied Mathematics from the Mohammedia School of Engineering in Rabat as well as a university habilitation in Business Intelligence from the same institution, he also attended the higher cycle of management at the ENCG of Settat. He is currently Professor of Quantitative Finance, Artificial Intelligence and Risk Management at ENSIAS, and Head of the Department “Computer Science and Decision Support” since January 2018. He is coordinator of the Specialized Master “Engineering for Sustainable Finance and Risk Management” and of the engineering program “Digital Engineering for Finance”. He is the founding President of the African Institute of Fintechs. For more than 15 years, he has been carrying out consulting missions for public and private organizations. He can be contacted at email: sachchab@gmail.com.






Mohammed Saber    Associate Professor at ENSAO (Ecole Nationale des Sciences Appliquées Oujda), Mohammed First University, Morocco. He is currently Director of Smart Information, Communication and Technologies Laboratory (SmartICT). His interests include Network Security (Intrusion Detection System, Evaluation of security components, Security in Mobile Ad Hoc Networks (MANET), Security IoT), Robotics and Embedded Systems. He can be contacted at email: m.saber@ump.ac.ma.






Ihame El Farissi    Engineer and Associate at ENSAO (Ecole Nationale des Sciences Appliquées Oujda). Her research interests involve Machine learning, classification of smart card attacks and intrusion detection systems. She can be contacted at email: i.elfarissi@ump.ac.ma.



Nawal Sbiti    holder of a PHD in automation and industrial computing from the Mohammedia School of Engineering in Morocco as well as a university habilitation in research from the same institution. She is currently Professor in Electrical Department at Mohammedia School of Engineering. For more than 30 years, she has been carrying out consulting missions for public and private organizations. She can be contacted at email: sbiti@emi.ac.ma.



Mustapha Amghar    holder of a PHD in industrial computing from Liege University in Belgium as well as a university habilitation in research from the Mohammedia School of Engineering in Morocco. He is currently Professor in Electrical Department at Mohammedia School of Engineering. For more than 30 years, he has been carrying out consulting missions for public and private organizations. He can be contacted at email: amghar@emi.ac.ma.