

Quantifying key factors for optimised manufacturing of Li-ion battery anode and cathode via artificial intelligence

Mona Faraji Niri^{a,b,*}, Kailong Liu^a, Geanina Apachitei^{a,b}, Luis A.A Román-Ramírez^{a,b}, Michael Lain^{a,b}, Dhammika Widanage^{a,b}, James Marco^{a,b}

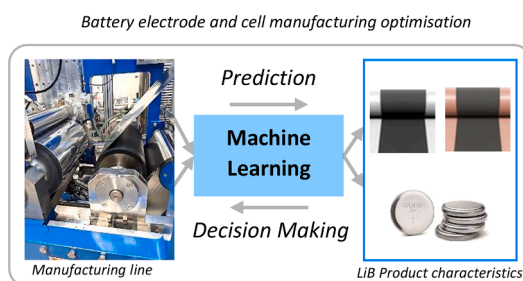
^a Warwick Manufacturing Group, University of Warwick, CV4 7AL Coventry, UK

^b The Faraday Institution, Quad One, Harwell Science and Innovation Campus, Didcot, UK

HIGHLIGHTS

- A data driven approach to predict the quality of Li-ion electrodes (anode and cathode).
- Identifying the key parameters and control variables affecting the intermediate and final products of battery electrode manufacturing process.
- Quantifying the effect of key control variables on the electrode and cell characteristics.
- Systematic design of experiments to generate representative data for ML activities with reduced waste and resource demand.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Li-ion battery electrodes
Anode
Cathode
Battery manufacturing
Machine learning
Artificial intelligence

ABSTRACT

Li-ion battery is one of the key players in energy storage technology empowering electrified and clean transportation systems. However, it is still associated with high costs due to the expensive material as well as high fluctuations of the manufacturing process. Complicated production processes involving mechanical, chemical, and electrical operations makes the predictability of the manufacturing process a challenge, hence the process is optimised through trial and error rather systematic simulation. To establish an in-depth understanding of the interconnected processes and manufacturing parameters, this paper combines data-mining techniques and real production to offer a method for the systematic analysis, understanding and improving the Li-ion battery electrode manufacturing chain. The novelty of this research is that unlike most of the existing research that are focused on cathode manufacturing only, it covers both of the cathode and anode case studies. Furthermore, it is based on real manufacturing data, proposes a systematic design of experiment method for generating high quality and representative data, and leverages the artificial intelligence techniques to identify the dependencies in between the manufacturing parameters and the key quality factors of the electrode. Through this study, machine learning models are developed to quantify the predictability of electrode and cell properties given the coating process control parameters. Moreover, the manufacturing parameters are ranked and their contribution to the electrode and cell characteristics are quantified by models. The systematic data acquisition approach as well as the quantified interdependencies are expected to assist the manufacturer when moving towards an improved battery production chain.

* Corresponding author.

E-mail address: mona.faraji-niri@warwick.ac.uk (M.F. Niri).

<https://doi.org/10.1016/j.egyai.2021.100129>

1. Introduction

Li-ion batteries (LiBs) have received a significant interest from various industries, especially those focused on electric transportation systems as they are enabling the transition of energy from traditional fossil fuel-based sources to the clean and renewables. In the last thirty years, the LiB cell energy density has increased by more than 3 times (from 80 to 250 Wh/kg) with the prices dropped by up to 45 times from \$4500 to \$100–250 per kWh [1]. The LiB total capacity is expected to exceed 1 TWh by 2028 to meet the need of the electric transportation system which are only on the ground. This is achievable only if the cost drops from \$100–250 to about \$70–80 per kWh [2, 3]. This price reduction is not easy to attain, that's why investigations are still required to suggest methods and approaches for optimisation of battery manufacturing process, increase cell performance and lifetime and decrease the cost. According to the reviews, cathode, and anode are the most expensive components of the Li-ion battery cells, followed by separator, electrolyte, and current collector [4, 5]. While almost 75% of the total LiB manufacturing costs are related to the raw material for cell components [6], the scrape rates have also an impact on the costs. During a conventional manufacturing run, scrape rates can vary from 6% to 15%, and up to 40% of the final manufactured cells can either be defective or in need of post-production adjustments [7]. Part of this scrapes come from the research and development (R&D) phase of the LiB manufacturing in search for an optimum combination of material and settings for increased performance and energy density. This activities when heavily based on trial-and-error, increase the production cost significantly [8]. Obviously, a systematic approach for a deep understanding of the LiB cell material and production processes would help to achieve a successful design. It would help identifying low quality products, specially during the R&D or the early stages of the manufacturing chain and reduce the rate of the failure.

While studies on the best combination of material started as early as the introduction of LiBs [9, 10], the correlation between the manufacturing control parameters and methods with the quality of LiB intermediate products such as slurry density, coating thickness, active material mass loading, as well as the final product characteristics such as energy capacity and internal resistance has attracted researcher more recently. In this direction, experiments have been designed to find the best combination of manufacturing parameters such as slurry descriptors [11], coating properties, and calendaring step settings [12, 13]. However, most of these designs have been conducted via try-and-error and the systematic production chain optimisation via advanced computational and experimental methods has been introduced only recently. In this concept machine learning (ML) models have had promising contributions due to their flexibility in dealing with large number of control and response variables that cannot be correlated via physics-based models. In fact, physics-based models face challenges around the effort required for their parameterisation and when compounded with the inherent variability of cells they result in local representations and reduced prediction accuracy [14]. Machine learning models allow researchers to extract patterns and trends from data and enable them to generalise the finding to the novel data with affordable computations.

A wide variety of ML algorithms are well investigated for the representation and performance improvement of the LiBs. For example, Gaussian classifiers and Markov models are combined for load prediction and state of energy estimation for batteries in electric vehicles [15]. Data-driven models are developed for knee point identification and state of health characterisation of cells [16]. Remaining useful life of cells is predicted in the early stages of ageing via hybrid ML models [17]. Gradient boosted trees are trained for state of health estimation in [18], and capacity fade is characterised by data-driven and experimental approaches in [19, 20]. Opportunities for improving the battery management systems (BMS) are reviewed in [21, 22] and algorithms such as wavelet-Markov models [23] or neural networks [24] are developed

for increasing the accuracy of state of power prediction for BMS. While the impact of ML algorithms on ready-to-use LiB cells has been clearly highlighted [25, 26, 14], their contribution to the battery manufacturing optimisation processes is rather new and still under investigation [26, 27, 28].

In the area of data mining and ML methods for battery manufacturing optimisation, [29] has proposed a method to indicate the failed products via an intelligent quality gate concept. In this approach the factors to be measured during the quality control process are determined such that the minimum effort is required for summarising the information on battery products. Here, the general quality control studies consider the quality measures and focus on classifying a product as OK, or Not-OK. Classification by models such as shallow and deep learning artificial neural networks are one of the examples [30]. Beside product classification by quality, various investigations have proposed data-driven methods to directly predict the final or intermediate product characteristics. In [31], a cross industry standard process for data mining (CRISP-DM) methodology is used and different regression techniques, neural networks, and ensemble models are compared to predict the energy capacity of the manufactured batteries given a combination of the manufacturing data. The effect of specific manufacturing factors such as wafer speed is used for quality prediction by [32], while a similar study via decision trees (DTs) and random forest (RF) is conducted in [33] to determine the importance of factors related to the laser cutting, assembly, dispersing and calendaring steps and their effects on the final cell's maximum capacity. Regression via Gaussian process models is proposed in [34] to predict the electrode mass loading considering the effect of mixing stage parameters, such as the slurry viscosity, amount of active material, liquid to solid ratio and coating parameter of comma bar gap. Models considering the aforementioned control variables are proposed in [35] to distinguish samples with low/medium/high mass loading and porosity. These models are further utilised to perform feature analysis through indices such as predictive measure of association and Gini index [36]. A combination of supervised and unsupervised approaches is proposed for electrode characteristics prediction in [37], with K-means clustering to putting data into groups of manufacturing conditions, and the Gaussian Naive Bayes model to predicting the heterogeneity properties such as active material mass loading and thickness. The focus of the study has been on the mixing process parameters of active material percentage, liquid to solid ratio and coating process parameters. Linear regression is utilised in [38] to clarify the relationship between the cathode physical characteristics and the cells gravimetric/volumetric capacity and its performance rate. A complementary and comparative study is report in [39], which utilises support vector machines and decision trees to predict the performance of LiB cells given the same physical characteristics.

While most of the aforementioned studies have been dedicated to the electrode mixing process, there are research conducted to address other LiB manufacturing steps such as calendaring and electrolyte assembly. Electrolyte mass effect on cell energy capacity is studied via CRISP-DM methodology in [40] and investigated via decision trees, support vector machines and neural networks in [41]. In [42] a combination of experimental results, in silico generation of electrode mesostructures and machine learning are used to link the calendaring step parameters and electrode properties. [43] suggests statistical methods such as principal component analysis to quantify the importance of the parameters of cathode calendaring on the electrode conductivity. Similar approach is addressed by [37] to select the best variables for heterogeneity prediction.

Although the above-mentioned research has adopted the data-driven strategies to address the issues of battery manufacturing optimisation and forecast product properties, several limitations are not yet discussed. In fact, most of the previous studies on the coating processes are only dedicated to the cathode of LiBs [34–41, 44] and a comprehensive analysis of both electrodes, cathode and anode is still required. Furthermore, the literature reviews confirm that the previous studies

have been focusing on either the intermediate [42, 41, 11] or the final products [44, 22, 23] and a thorough analysis from the manufacturing parameters to intermediate and final products characteristics is rare. Covering this research gap has been the main motivation for the present study.

The objective of this paper is to study the relationship between the manufacturing control parameters, the intermediate product characteristics, and the final cell performance in relation to both cathode and anode. In order to elevate the existing challenges of investigations raised due to try-and-error approaches, this research is taking the advantage of systematic design of experiments approach and ML techniques as a branch of artificial intelligence (AI).

This study pursues two goals via the application of DoE and ML techniques. First quantifying the predictability of products' characteristics given the manufacturing and control variables and second, performing correlation, feature ranking, feature contribution and model interpretation analysis. This study is considered novel compared to the previous ones as it (1) addresses both electrodes, (2) considers interest factors from initial, intermediate, and final manufacturing steps, (3) quantifies interconnections both in between control parameters and in between control-response parameters and (4) offers comprehensive feature analysis and ML model interpretations. In particular the contributions of the study are:

- proposing a systematic design of experiments for obtaining representative data for modelling purposes
- quantifying the effect of the coating process parameters on the quality of both electrodes and explicitly on the half-cell performance
- revealing the interconnections between manufacturing control parameters and variables
- interpretation of the ML models performance when correlating the control and response variables

The following steps have been taken to achieve the objectives:

1) A systematic design of experiments is conducted based on screening tests, correlation analysis, interconnection evaluations and fractional factorial saturated approach. Here, manufacturing parameters of coating speed, coating gap (comma bar gap), coating ratio, drying air speed, and drying temperature are considered as the interested feature factors, while cathode and anode thickness, active material mass loading, porosity, and the associated half-cell energy capacity are taken as the interest characteristics.

2) Machine learning models are trained, validated, and tested to forecast both positive and negative electrode and half-cell characteristics. Two advanced ML models of gradient boosted decision trees (GBT) and random forest are developed and compared via various indices.

3) The importance and contribution of each interest factor is calculated based on the final product's quality via representative indices of mean decrease in impurity (MDI) [45] and Shapely values [46] in

connection with the ML models.

The framework of this study is given in the Fig. 1. The flow chart shows the structure of the research and where the DoE and ML algorithms stand in the concept of battery manufacturing optimisation. According to this framework the purpose is to provide feedbacks and improvements for either the manufacturing line for battery production, or to the design of experiment experts for analysis of the production processes during research and development. This framework is following the CHAIN (Cyber Hierarchy and Interactional Network) framework [47]. This framework is enabling digital solutions for design, optimisation, and management of Li-ion battery manufacturing processes. It could accelerate the industry focusing on building a high performance, and low-cost LiBs.

The structure of this paper is as follows. In Section 2 the experimental studies are given, the electrode manufacturing processes, and the experiments design are explained. Data generation and collection procedures are briefed in Section 3 and machine learning models are developed to predict the electrode properties. Results, and analysis such as performance comparisons in between various models, feature importance, and feature contributions are given in Section 4. Finally, Section 5 summarises and concludes the findings.

2. Experimental studies

2.1. Electrode manufacture, anode and cathode

Fig. 2 visualises the main processes of the battery manufacturing chain. The process starts with selecting the material and continues with mixing those in a controlled environment to form a slurry. Then the slurry is placed on the electrode foils and goes for drying and calendaring. At the next step the electrodes are cut and then the cells are assembled and filled with the electrolyte. The final step is formation, testing, and characterisation of manufactured cells in various loading, charging, and discharging conditions. Further details of the LiB production processes are given in [8, 26].

The anode formulation used in this study was: 95.25% active material-graphite (S360-E3 graphite, BTR), 1% conductive additive-C45 (Imerys), binders- 1.5% carboxy methyl cellulose (BVH8 CMC, Ashland) and 2.25% styrene-butadiene rubber (BM-451B SBR, Zeon). The dry components (graphite, C45, CMC) were first mixed together, then water was added to reach 58% solid content during the kneading stage, then added again in the dilution stage for reaching 46% solid content in the final slurry. The SBR solution was added last and mixed in at a low speed.

For cathode, 96% of NMC (nickel manganese cobalt oxide) was used in addition to a 2% conductive additive of carbon black (C65) and 2% of polyvinylidene binder. The process was similar to anode but with N-methyl pyrrolidinone as solvent to achieve the 77% solid content in the first and then to 67% at the second step.

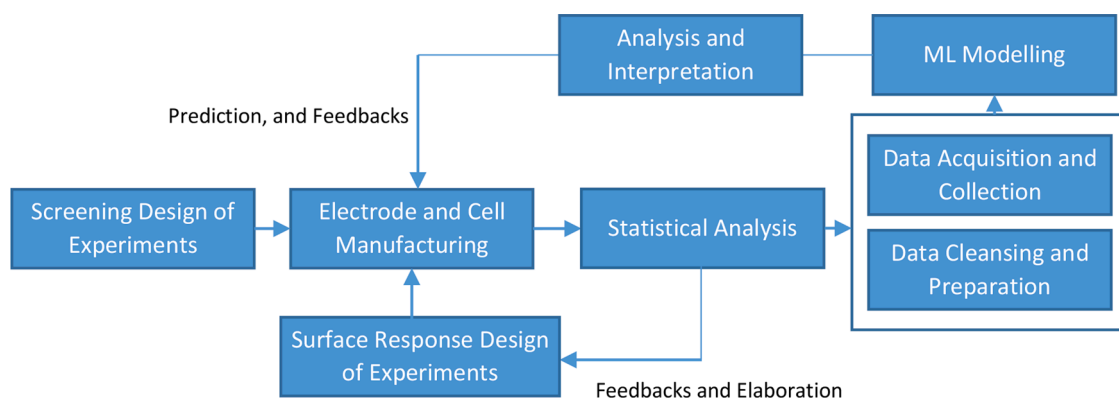


Fig. 1. Framework of this study.

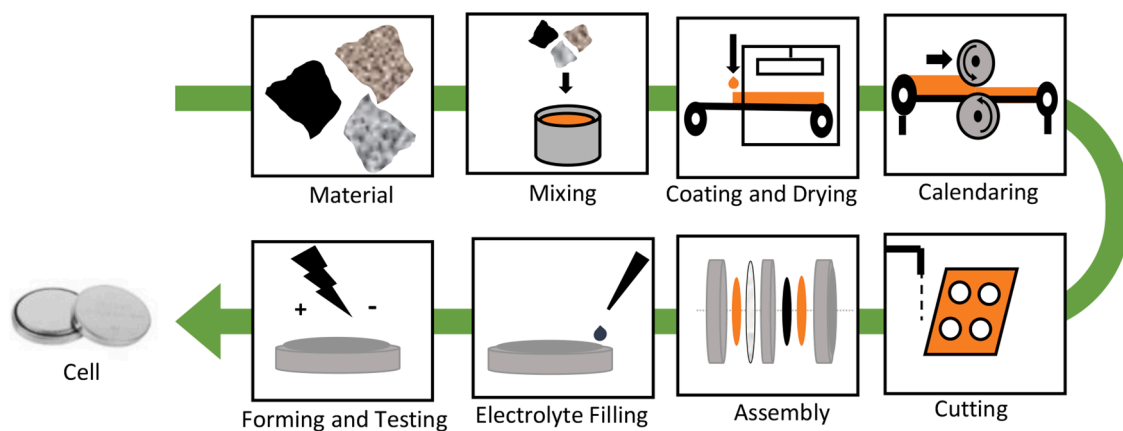


Fig. 2. Li-ion battery production processes.

The slurry was used to coat a 10 (μm) Copper foil for anode and a 15 (μm) Aluminium foil for cathode. This was performed via DÜRR Magtec pilot-line-scale coating machine which uses the roll-to-roll technology for material deposition. The material fills a reservoir that is created by the precision chrome roll and the comma bar of the machine and then released onto the foils. The drying was conducted in a controlled environment inside a roll supported 3-zone thermal dryer where the temperature and air speed were both set manually by an operator. Calendaring was performed inside a R&D calendaring machine from Innovative Machine Corporation with consistent settings for all experiments to keep focus on coating process and avoid introducing further uncertainties. Also, in order to minimise the residual moisture, the electrode rolls were further dried in a Binder vacuum furnace at 120 °C for another 12 h.

The comma bar coating technology and slurry transfer is shown for anode and cathode in Fig. 3. Based on each electrode obtained after the calendaring process, three half-coin cells (size 2023) were assembled. During cathode manufacturing the half-cells include a lithium disc as negative electrode, and similarly, during anode half-cell manufacturing the half-cells include a lithium plate as the positive electrode. It is worth mentioning that half-cells were preferred over full cells in order to keep the focus on each electrode individually for a comprehensive analysis.

The cells went for formation cycles at C/5 and then get tested in the room temperature, 25 °C via a commercial cyler, Maccor™, under C/20 discharge current. This particular Crate was preferred as it discharges the cell quite slowly and helps the cell reveals its electrochemical characteristics in the long term.

2.2. Systematic design of experiments

In order to perform a comprehensive analysis for the optimisation of the cathode and anode manufacturing, a rich data set covering the practical ranges and limits of each control parameter was necessary. Generally, the equipment for cathode and anode production are designed such that a minimum length of the coating (usually 1 m) is generated at each run to ensure that sections with desired quality are available for cell manufacturing. This requires large amount of material and resources and usually is followed by a large material and energy waste. Therefore, here a design of experiments (DoEs) approach is used to minimise the cost and waste and extract the maximum information from each experiment. The systematic DoE of this research starts with a selection of control parameters, and then continues with the parameters that are recognised to be more significant via correlation analysis.

In order to design the experiments first a screening design was promoted based on a fractional factorial saturated approach [48]. The control parameters as well as their associated minimum and maximum levels are reported in Table 1. The limits were decided based on the experts' recommendations and literature [41, 49]. For both of the electrodes the control parameters were similar but with different ranges.

Through the screening run of the experiments, 12 combinations of the levels mentioned in the Table 1 were used to perform the coating of cathode and anode. For each electrode and its associated key parameters (KPs), (active material mass loading, thickness, and porosity), the correlation analysis was performed. The purpose was to quantify the impact of each control variable on the one other as well as on the desired KPs. The correlation coefficients $r_{X,Y}$ for two variables of X and Y, are calculated via the Eq. (1) and recognised as Pearson product-moment correlation coefficients [50]. Where, μ is the mean value of each

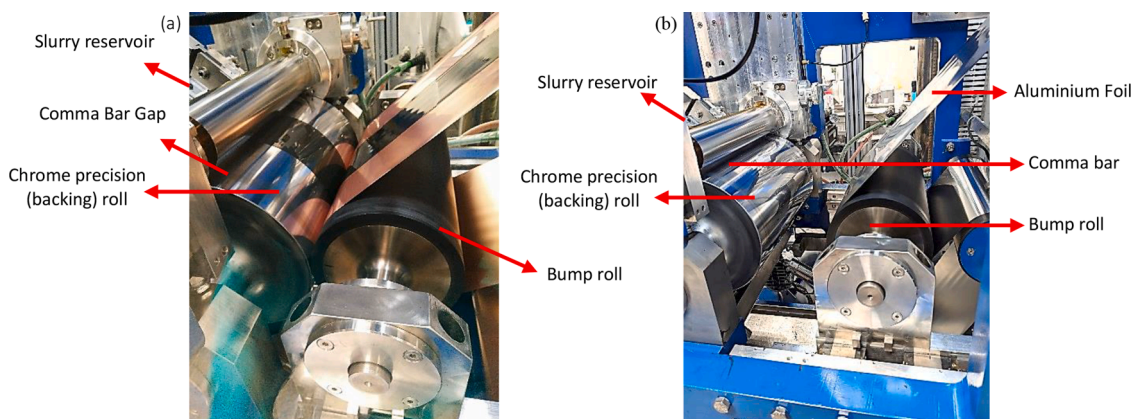


Fig. 3. a) comma bar coating technology, (a) Anode, (b) Cathode.

Table 1
Experiments for the screening DoE.

Parameters and levels	Comma bar gap (um)	Web speed (m/min)	Temperature (°C)	Air speed (m/s)	Coating ratio (%)
Cathode (min)	80	0.5	85	5	110
Cathode (max)	140	1.5	110	15	150
Anode (min)	100	0.5	45	5	110
Anode (max)	190	1.5	60	15	150

variable, σ is the standard deviation and E denotes the expectation.

$$r_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

The coefficients are a value within [0, 1] and could be positive or negative. A positive (negative) correlation, r , means that the two variables, X and Y are directly (inversely) connected, i.e., an increase in one variable is associated with an increase (decrease) of the other one. While a value of the coefficient significantly different with zero confirms a strong relationship between the two variables, the values close to zero mean that there is not sufficient evidence that there is a significant linear correlation between variables.

For each correlation coefficient a p -value is also calculated within a P matrix, the p -value is the probability that the ‘Null Hypothesis’ is true. Here the null hypothesis is no relationship between the two variables of coating process [51]. P -values are used to validate the correlation analysis and provide insights about the generalisation of the results from sample to population [52]. P -value is a number between 0 and 1. If a non-diagonal elements in P matrix is smaller than a threshold, α , then the correlation is considered significant.

Fig. 4 shows the correlation coefficient and p -value matrices for

cathode and anode manufacturing parameters.

According to Fig. 4, for both cathode and anode, the 5 control variables have correlation coefficients all below 0.1 with associated p -values all close to 1. This means that the variables are fairly independent and there is no evidence available for significant correlation. This implies that the five mentioned control variables are all are worth to be studied in an optimised manufacturing process and need to be measured individually during screening runs.

While Fig. 4 is about the correlation in between the control variables themselves, Fig. 5 shows the correlation coefficients and p -values between control and response variables. For cathode, Fig. 5(a) and (b), comma bar gap is showing high correlation coefficients with coating weight and thickness. Both coefficients are positive which confirm an increase in those variables will increase the mass loading and thickness. For porosity, comma bar gap and web speed are having larger correlation coefficients. While coating ratio is correlated positively, comma bar gap and web speed are negatively related.

Considering the p -values, the correlations of three response variables with comma bar gap, web speed and coating ratio are statistically significant with the small p -values. The correlation coefficients related to the temperature and airspeed are small and their related p -values are quite large. Therefore, in total there is not enough evidence that they are statistically significant.

The correlation coefficients of anode, Fig. 5(c) and (d), are smaller than those for cathode. Another difference between anode and cathode data is the correlation coefficient direction between temperature and electrode characteristics. For cathode, the correlation coefficients between temperature and all three KPs is negative, while for anode it is positive with the coating thickness and porosity.

Based on the findings from the abovementioned correlation analysis, a second DoE is performed to further focus on significant factors. The relation between the two Does is clarified in Fig. 1. For cathode the three control parameters of comma bar gap, coating ratio and web speed were recognised to be important based on the higher values of the correlation

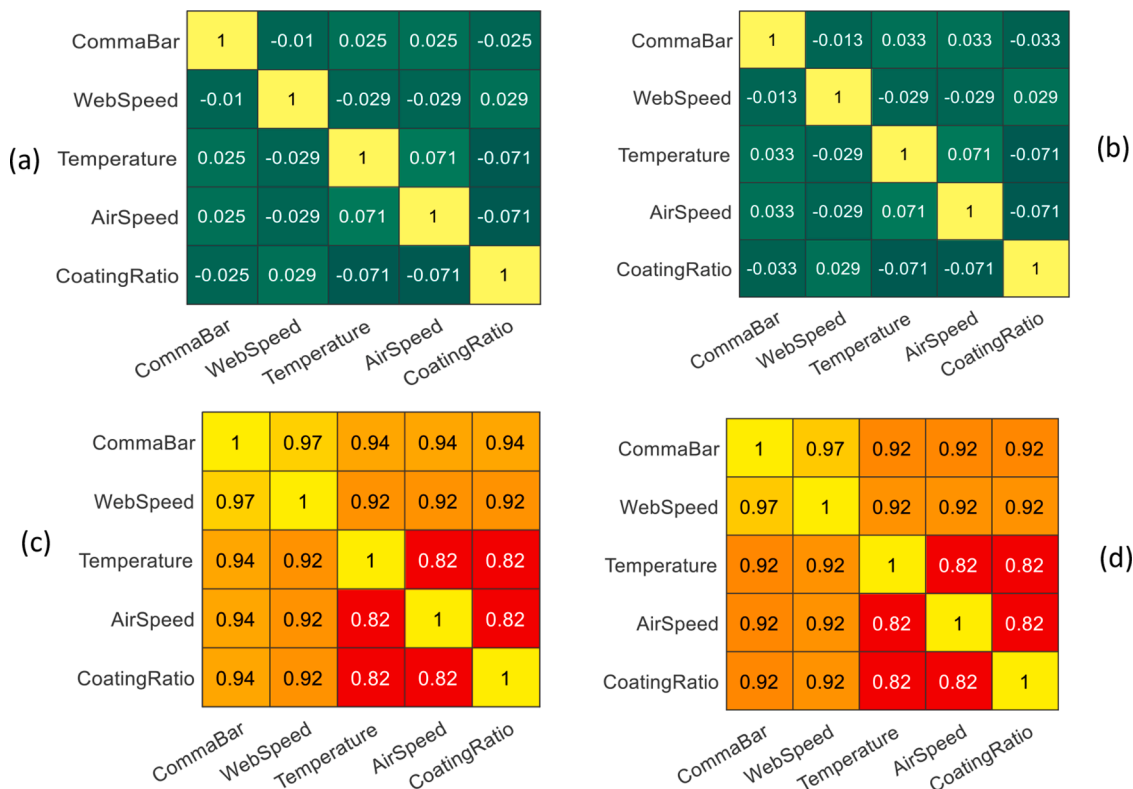


Fig. 4. Correlation analysis in between control variables. (a) Cathode correlation coefficients, (b) Anode correlation coefficients, (c) Cathode p -values, (d) Anode p -values.

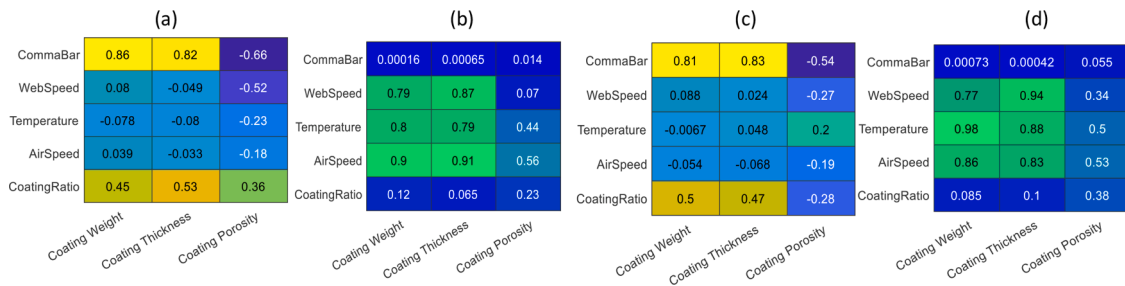


Fig. 5. Correlation analysis between control variables and KPs, correlation coefficients cathode (a), p-values cathode (b), correlation coefficient anode (c), p-values anode (d).

coefficients and p-values smaller than the threshold $\alpha = 0.2$. This threshold is selected as a compromise between the number of variables with statistically significant correlations and the availability of the resources to run experiments and measure the variables. For anode the important parameters are then limited to comma bar gap and coating ratio as other p-values are larger than the threshold of 0.2. Unlike cathode, the web speed is not taken to the second DoE as the correlation coefficient is small and p-value is rather large compared to the threshold. For further details about the correlation analysis for the data of this study refer to [38].

The second DoE was designed based on a response surface methodology (RSM) [53]. This RSM consisted of a Box-Wilson composite design (CCD) with 5 levels for each of the main control variables [54]. Although there are different available experimental designs [55], Box-Wilson CCD has been preferred due to a number of reasons. It presents a good balance between the number of experimental runs and the statistical power, it suggests a more stable variance with 5 levels, and thus is more suited to this manufacturing study due to the intrinsic variability of the experiments and non-repeatability of those. The levels of the DoE are given in the Tables 2 and 3 for the cathode and anode respectively. The DoE space volume and surface is depicted in Fig. 6 which show how DoE points fill the parameter space uniformly and comprehensively.

3. Data preparation and analysis

3.1. Quantifying the interest factors

This study considers the coating step parameters and the electrode physical characteristics as interest factors. The interest factors are selected based on the literature [41, 38, 39] and the experts view about their impact on electrode and cell characteristics. The manufacturing factors are, (1) Comma bar gap (um), which is the gap between the comma bar and the substrates of the coating machine. (2) Coating speed (m/min) also called the web and line speed, which is the speed of the coating line and defines the dragging speed of the foil. (3) Coating ratio (%), which is the ratio of the chrome roll speed to the backing roll speed of the machine. Backing roll is the same as bump roll supporting the foil when the material was transferred to it. These three factors affect the amount of slurry that is deposited onto the electrode foils (current collectors) that subsequently defines the electrode thickness, mass loading and eventually the final cells' characteristics. Other interest factors are related to the drying step of the electrode manufacturing, they are (4) Air speed (m/s) which is the air flow of the dryer and (5) Temperature (°C) which is the temperature of the dryer. Although, there are other

factors such as the ambient or drying equipment humidity but considering the limitations of the resources and the experts view, they were identified to stand at a lower priority compared to the ones already considered.

For each piece of the coated foils and the electrodes obtained after calendaring, three characteristics were measured. (1) Coating weight, or so-called active material mass loading (g/m^2), which was measured in real-time using two ultrasonic weight/thickness gauges (MeSys GmbH) positioned so the first scanned across the wet coating before entering the dryer, while the second one scanned across the dry coating, before rewinding it on the roll. (2) Coating Thickness (um), which was measured by a digital micrometre, Mitutoyo, with the accuracy of about 1 um. (3) Coating Porosity (%), which was calculated as a ratio of coating weight (g) to the coating thickness and its density.

It is worth mentioning that, since the data of this study were generated from the pilot scale battery manufacturing lines at WMG University of Warwick, the volume of data was directly manageable within the conventional data storage units of the computers. The data were stored in the form of tables within Excel and MATLAB without further requirements for data storage facilities.

3.2. Data preparation

The active material mass loading, thickness and porosity values that are collected during experimental runs as well as the cell energy capacity data are not necessarily in their best format for analysis and modelling activities and data preparation and curation is necessary. The mass loading recorded by the equipment includes data points related to bare foil as well as other deflections such as missing points and outliers that might affect the predictability of models. Beside the missing values, the outliers are also an issue specially with the measured thickness and cell capacity that are performed by operators in the lab.

To prepare the data for in-depth analysis, first they are cleaned from outliers. Outliers are considered to be data points away from median of all the records at the same experiment for at least three times. Secondly, the data are trimmed to only include the sections that are considerably higher in uniformity and consistency with a chance to be used for the next step of the cell assembly. Finally, the missing values are managed by replacing those with the median of all data of the same record. The output of these steps is a clean and efficient data set with 32 full electrode samples for cathode and 25 full electrode samples for anode. Examples of pre- and post-processed electrode samples mass loading data are given in Fig. 7.

From each electrode sheet, the features of mass loading, thickness,

Table 2
Experiments for the CCD DoE for cathode.

Experiment number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Comma bar gap (um)	128	110	80	110	110	128	92	92	110	92	110	140	110	92	110	110	128	110	128	110
Web speed (m/min)	0.7	1.0	1.0	1.5	1.0	0.7	1.3	0.7	1.0	0.7	1.0	1.0	1.0	1.3	1.0	1.0	1.3	0.5	1.3	1.0
Coating ratio (%)	142	130	130	130	110	118	142	142	130	118	130	130	130	118	150	130	118	130	142	130

Table 3
Experiments for the CCD DoE for anode.

Experiment number	1	2	3	4	5	6	7	8	9	10	11	12	13
Comma bar gap (um)	145	145	113	145	145	190	145	177	113	145	177	145	100
Coating ratio (%)	130	150	144	110	130	130	130	116	116	130	144	130	130

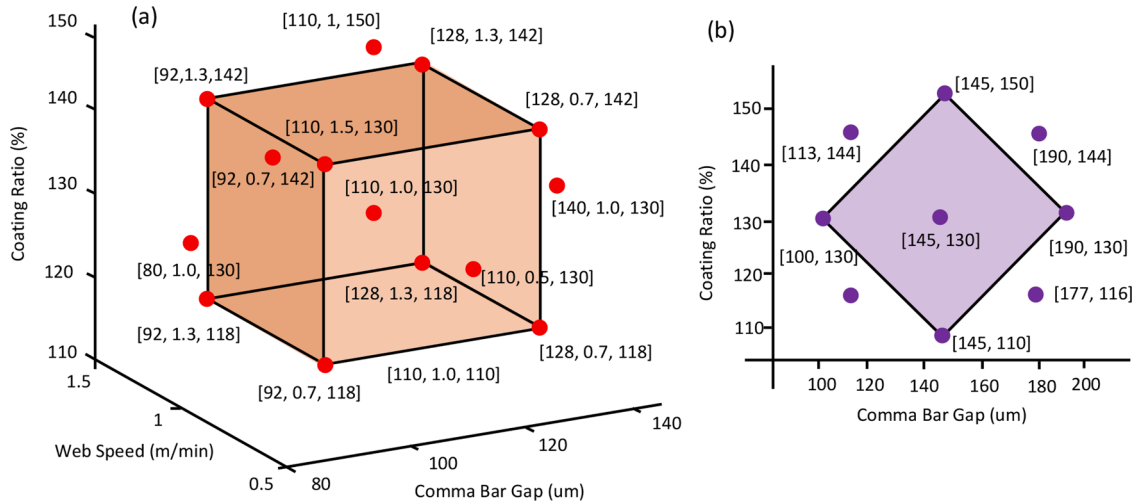


Fig. 6. Graphical representation of the CCD experimental design for (a) Cathode and (b) Anode.

porosity, and cell capacity are extracted for the next step analysis. Per sheet, 3 discs are cut and therefore 3 mass loading values, 3 thickness values, and 3 capacity values, are obtained by measurements. The total size of data is 96 cathode-related, and 75 anode-related electrode and half-cell data.

3.3. Modelling by machine learning

The three coating process factors of comma bar gap, coating speed and coating ratio are used as inputs to a machine learning model in order to predict the quality of the manufactured electrodes in the form of its active material mass loading, thickness, and porosity as well as the performance of the manufactured half-cell in terms of C/20 capacity. Considering the fact that porosity is a feature directly calculated based on the thickness and mass loading, the model has only three outputs of mass loading, thickness, and capacity.

Here the gradient boosted trees model has been the model of choice. The model is selected as it is based on a gradient boosting framework for machine learning and built upon the decision tree algorithm. It is a

powerful technique for prediction, ranking and decision making due to combining a large number of decision trees to reduce the risk of over-fitting. Successful implementation of this model for similar prediction problems of LiBs has been reported in [31, 18, 56].

GBT algorithm utilises gradient boosting method to combine the individual DTs. In fact, DTs or so-called weak learners are connected sequentially to form a strong learner, F.

Given the training samples of $\{y_i, x_i\}, i = 1, \dots, N$, where x_i is the input and y_i is the response or output, the purpose is to find a learner, $F^*(x)$ as in (2), which minimises a cost function of $\psi(y, F(x))$.

$$F^*(x) = \underset{F(x)}{\operatorname{argmin}} E_{y,x}[\psi(y, F(x))] \tag{2}$$

Boosting technique of GBT method helps to approximate the strong learner by an additive expansion of weak learners defined by (3).

$$h(x, \{R_{lm}\}_L, \bar{y}_{lm}) = \sum_{l=1}^L \bar{y}_{lm} 1(x \in R_{lm}) \tag{3}$$

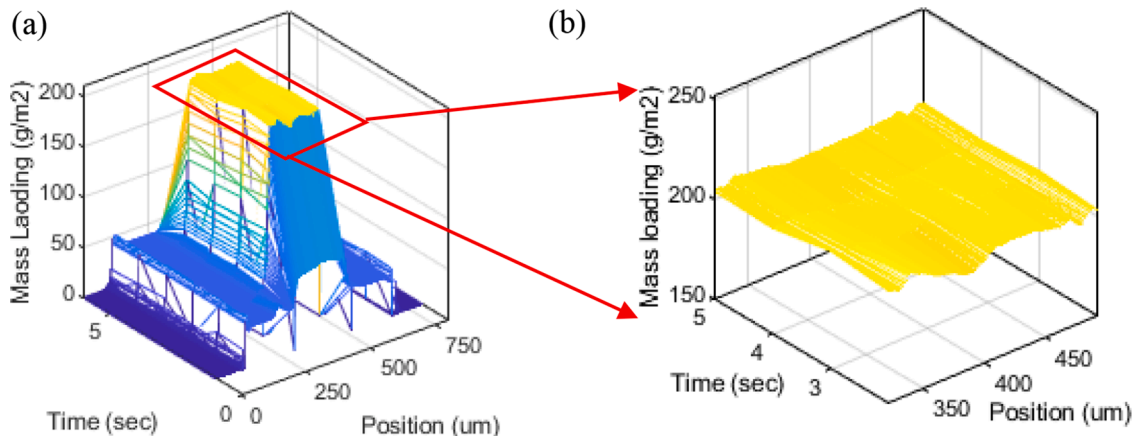


Fig. 7. Data Preparation, (a) pre-processed and (b) post-processed coating mass loading.

The learners are L-terminal node trees in this study, and form the strong learner in the following form:

$$F^*(x) = \sum_{m=0}^M \beta_m h(x, \{R_{lm}\}_L, \bar{y}_{lm}) = \sum_{m=0}^M \sum_{l=1}^L \beta_m \bar{y}_{lm} 1(x \in R_{lm}) \quad (4)$$

where each weak learner is indexed by m , and weighted by an expansion coefficient, β_m . $\{R_{lm}\}_L$ are L -disjoint regions and \bar{y}_{lm} is the corresponding split points determined by the m^{th} tree.

Each weak learner tries to minimize the error of the previous tree. Then the results are aggregated by the final model. In order to get the error or residuals of each weak learner a cost function is used which is the mean squared error function in here. To estimate the strong learner a stepwise method is necessary. The method starts with an initial condition $F_0(x)$, then the pseudo residuals for each tree is obtained by (5).

$$\tilde{y}_{lm} = - \left(\frac{\partial \psi(y_i, F(x_i))}{\partial F(x_i)} \right)_{F(x)=F_{m-1}(x)} \quad (5)$$

In order to get each region $\{R_{lm}\}_L$, the m th tree is estimated by the sample set of $\{\tilde{y}_i, x_i\}_N$, then the loss function optimisation is carried on by defining $\beta_m \bar{y}_{lm} = \gamma_{lm}$ according to:

$$\gamma_{lm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{lm}} \psi(y_i, F_{m-1}(x_i) + \gamma). \quad (6)$$

Finally, the strong learner is updated following (7),

$$F_m(x) = F_{m-1}(x) + \eta \cdot \gamma_{lm} 1(x \in R_{lm}) \quad (7)$$

where $\eta \in (0, 1)$ is a parameter that controls the amount of information used for estimating a new tree [57, 58].

In order to evaluate the capability of the designed GBT model in predicting the response for unseen data, the models need to be experimentally tested and validated. It is worth to highlight that generally there exists two ways to validate the developed models, Offline and Realtime.

The offline validation is where the data are separated to train, validation and test and then used to verify the predictability of response variables through the models. The second way is the Realtime validation, where models need to be implemented within the lab by appropriate hardware and processors. Via this implementation they can provide feedbacks in Realtime to the manufacturing run and therefore the experimental run output and the difference with the model predictions can provide a measure for accuracy and validity. This study considers the offline approach as the current battery manufacturing technology lacks the suitable actuators to provide commands to the line in Realtime.

Here to make most out of the available data for validation and test, the cross validation (CV) approach is utilised. CV is based on resampling procedure and very suitable for dealing with limited data samples, it results in less biased estimations and avoids the unnecessary optimism [59]. In CV the data are first split into K different, non-overlapping groups. K is considered as a design parameter of this procedure and the method is called K -fold CV as well. 1 out of K groups is taken as hold-out or test data, while the rest of the $K-1$ groups are used during the training of the model, during the training and test processes, the data stay in their groups. Then, to give all data the possibility of being considered once for training and once for testing, the data are then shuffled and regrouped into K new ones. This procedure is repeated for K times. For $K-1$ groups first the model is trained and then is tested on the last remaining group. The performance of the model on the test data is then recorded in the form of an accuracy score. All scores of the K runs are then summarised into an average and a standard deviation value. Obviously, as during the CV each sample goes for test at least once, all the samples have contributed to the final accuracy matrices but only

when they have had the test data role.

The value of design parameter K is selected such that the training and testing data are large enough to statistically represent the whole samples [60]. For this study, K is selected to be 5 which means that the data are split into 5 groups for 5 runs and during each run, 80% of data samples are used for training and the 20% of the samples are used for validation and testing.

4. Results and discussions

In order to visualise the data set of comma bar gap and coating ratio in relation with the structural features of the anode, Fig. 8 is depicted for an average of three electrode discs. Evidently an increase in the comma bar gap and coating ratio results in increased weight and thickness of the anode. Obviously, this is physically explainable as larger gaps allow a higher weight of slurry to be transferred onto the foil surface. The behaviour of porosity is almost the same as thickness at high and low comma bar gaps and coating ratios. However, it shows a slight reduction in the medium values of the two parameters.

Having comma bar gap, web speed and coating ratio, the distribution of cathode properties is depicted in the Fig. 9. As the data are 4 dimensional, a colour bar is used to visualise them.

As Fig. 9 suggests that a larger value of comma bar gap and coating ratio result higher cathode coating weights which is due to the larger amount of material that is deposited on the foil surface. Unlike the comma bar gap and coating ratio, web speed is a less affecting factor for active material mass loading.

Higher coating ratio and comma bar gap levels along with the slow speed of coating result thicker electrode as all let larger amount of material to sit on the coating surface. The relationship between the comma bar gap, coating ratio and web speed with porosity are different to that of coating weight or thickness. Higher comma bar gap, faster coating and lower coating ratio all reduce porosity as more open pores are left on the surface. Although Figs. 8 and 9 partly help to understand the effect of manufacturing parameters on the electrode and cell characteristics thanks to the DoEs, but still quantifying the impact of those on the electrode and cell quality is necessary and only achievable via advanced ML techniques in the next upcoming sections.

4.1. Predictability

In what follows, the predictability of the electrode features given the manufacturing variables is investigated. Porosity is not considered as a target for prediction as it has been calculated based on the physical measurements of mass loading and thickness. The main model of this section is the gradient boosted trees, and the results are also compared with a second machine learning algorithm, random forest. Random forest is an algorithm also based on decision trees and utilises a method called bagging, which is combining various decision trees to create an ensemble. It combines trees in parallel and involves bootstrap sampling [61, 62] as well. Although the two methods are using similar weak learners, the method of aggregating them is quite different which make them worth to be investigated for this data set. To obtain conclusive results all models are evaluated for a minimum of 10 independent runs and the indices are averaged.

Fig. 10 and Table 4, summarise the prediction accuracies for the two methods. To quantify the accuracy, four indices are used. R^2 , referred as the coefficient of determination, mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE). R^2 shows the proportionate amount of variation in the electrode or cell characteristic that has been explained by the manufacturing control variables. The closer the R^2 to 1, and the smaller the RMSE the more representative the models. All the accuracy metrics are calculated only over the test data fold during the CV. Obviously, as during the CV each sample goes for test at least once, all the samples have contributed to the final accuracy matrices. In other words, during each run of CV, the accuracy metrics

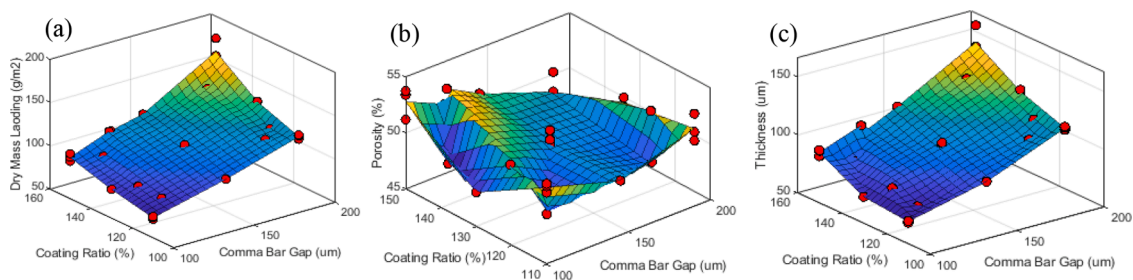


Fig. 8. distribution of manufacturing parameters with relation to (a) dry mass loading, (b) coating thickness, (c) Coating porosity of Anode.

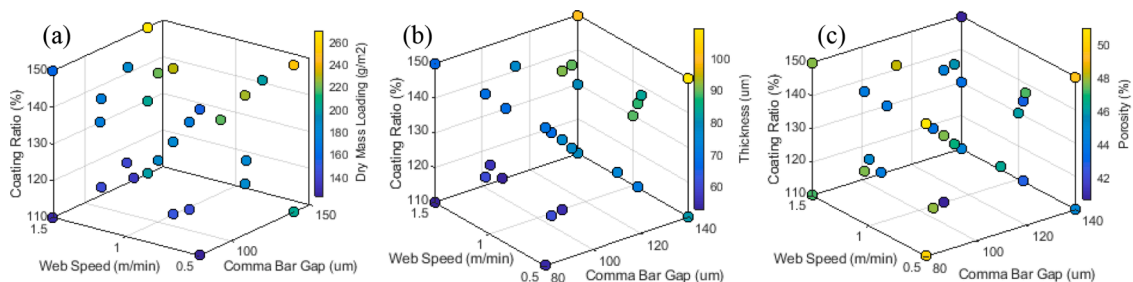


Fig. 9. distribution of manufacturing parameters with relation to (a) dry mass loading, (b) coating thickness, (c) Coating porosity of Cathode.

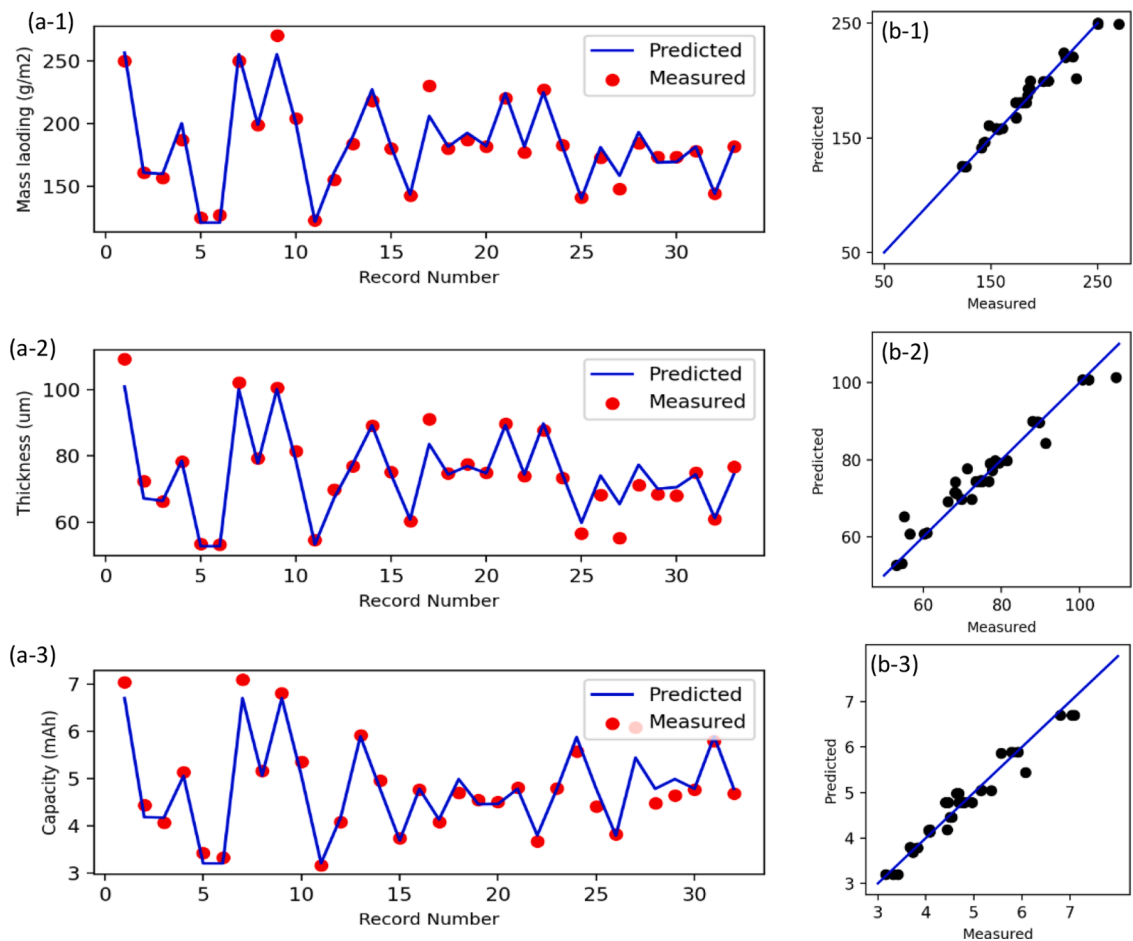


Fig. 10. Prediction results for the Cathode characteristics.

Table 4
Prediction indices for random forest and light gradient boosting tree for Cathode and Anode characteristics.

Cathode		GBT			RF			
Model Index (Mean, std)	R ²	RMSE	MAE	MSE	R ²	RMSE	MAE	MSE
Mass loading (g/m ²)	0.915 (0.070)	7.716 (2.822)	5.37 (1.951)	63.39 (54.72)	0.856 (0.102)	12.72 (5.163)	9.61(2.482)	150.32 (93.02)
Thickness (um)	0.786 (0.168)	4.922 (1.917)	4.035 (1.02)	34.453 (22.13)	0.771 (0.157)	6.418 (2.113)	4.793 (1.53)	48.282 (38.23)
Capacity (mAh)	0.767 (0.176)	0.387 (0.140)	0.283 (0.09)	0.139 (0.082)	0.756 (0.160)	0.403 (0.162)	0.345 (0.113)	0.187 (0.161)
Anode		GBT			RF			
Model Index (Mean, std)	R ²	RMSE	MAE	MSE	R ²	RMSE	MAE	MSE
Mass loading (g/m ²)	0.835 (0.161)	8.863 (4.069)	5.912 (3.05)	101.48 (86.71)	0.819 (0.186)	9.186 (4.564)	9.076 (4.730)	120.83 (130.84)
Thickness (um)	0.783 (0.152)	8.293 (2.851)	6.042 (2.11)	67.66 (55.34)	0.751 (0.184)	8.990 (3.579)	7.168 (3.89)	97.09 (84.03)
Capacity (mAh)	0.735 (0.257)	0.469 (0.143)	0.291 (0.12)	0.204 (0.107)	0.726 (0.225)	0.448 (0.135)	0.421 (0.148)	0.212 (0.133)

are calculated for the test fold, and then the results of all runs are collected to report a mean and std of the metrics.

According to them, via GBT, the mass loading of the electrode is predictable with an accuracy of 0.915 in terms of R2, 7.716 (g/m2) in terms of the average root mean squared error (RMSE) and 2.822 (g/m2) as the standard deviation of the RMSE. Fig. 10 (a-1) and (b-1) show the distribution of the measured and predicted active material mass loading. Considering the average value of cathode mass loading, the mean and standard deviation of the prediction RMSE is 4.23% and 1.55% respectively.

Similar results are depicted for thickness in Fig. 10 (a-2) and (b-2) with 0.786 R2 and 4.922 (um) RMSE. In this case the RMSE mean and standard deviation are 6.59% and 2.57% respectively.

Fig. 10 (a-3) and (b-3) depict the distribution of the cathode half-cell capacity data and the predicted values. The capacity is predicted by 0.737 R2, and 0.419 (mAh) of average RMSE. The mean and standard deviation of the prediction error are 8.72% and 4.10% respectively.

For the random forest model, the accuracy of prediction for all three outputs is slightly behind the GBTs. Furthermore, the standard deviation of multiple runs is much higher in RF compared to GBTs. This is believed to be due to the capability of GBTs in better dealing with unbalanced data intrinsically [58].

The results of the predictive models for anode are plotted on Fig. 11, and right section of Table 4. Compared to the cathode, the anode mass loading, thickness, and capacity are showing slightly lower predictability. The distribution of predicted and real anode active material mass

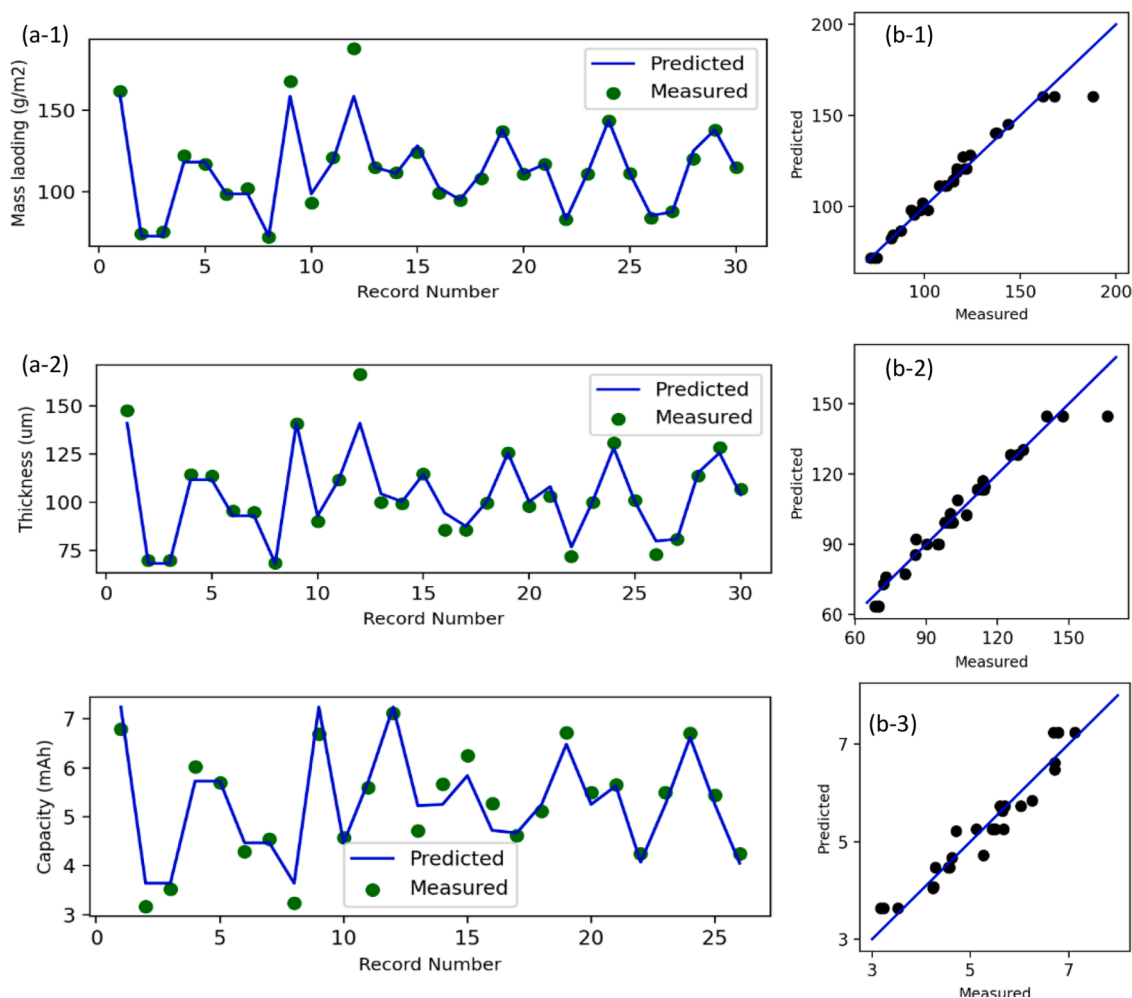


Fig. 11. Prediction results for the Anode characteristics.

loading data is given in Fig. 11 (a-1) and (b-1), the mass loading is predictable with 0.835 R2, and 8.863 (g/m2) RMSE. This error is translated to be 7.82% with a standard deviation of 3.59% to the average mass loading.

The R2 and RMSE for anode thickness are 0.783 and 8.293 (um), the predictability is visualized in Figs. 11 (a-2) and (b-2) and quantified to be 8.02% mean RMSE with 2.76% standard deviation. For Anode half-cell capacity the GBT provides results with an R2 of 0.745 and 0.399 (mAh). The RMSE mean and standard deviation are 8.91% and 2.72% respectively. Similar to the cathode, RF is also slightly behind the GBT in terms of performance indices and their standard deviation.

The RMSE values normalised to the average data of the same type are reported in Table 5.

For all cases cathode is higher predictable than anode, and the standard deviation of the indices are higher for electrode characteristics and lower for capacity of anode. Considering all three characteristics the prediction accuracy for cathode is 93.48% and for anode is 91.75% by average.

4.2. Factor analysis

To obtain a deep understanding of the effect of control factors on the intermediate and the final products' quality, a factor (feature) contribution and factor importance analysis is necessary.

For a conventional and traditional factor contribution and importance analysis, it is necessary to only have a single factor changing from one experiment to another, so that its effect on the response variables can be determined. This approach is not applicable to data sets with having more than one factor changing in each run. The experiments in this study have been designed such that the most representative data points are driven with minimum number of experiments to reduce the material waste, experiment time, required energy as well as resources. This has resulted a combination of levels for control factors of comma bar gap, web speed, and coating ratio where in a single experiment more than one factor is different with the other experiments. In order to perform the contribution and importance analysis for the data of this DoE approach, machine learning models are utilized in connection with two representative indices of shapely [63, 46] and mean decrease in impurity [45, 64].

The contribution of each of the control variables on the physical characteristics of the electrodes, or electrochemical features of Li-ion half-cells are very exclusive. It was previously shown in Fig. 4 that the factors correlate with the characteristics quite differently. In order to quantify this contribution, the GBT model is trained to obtain Shapely values [63, 46].

Shapely values indicate how each feature is contributing to the overall prediction of the response variables, here electrode and half-cell characteristic. Shapely equation is as (8).

$$\varphi_i(v) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \left(\frac{|N| - 1}{|S|} \right)^{-1} (v(S \cup \{i\}) - v(S)) \quad (8)$$

In this equation, N is the total number of factors and S is a subset of all factors. v indicates the value of each subset and $\varphi_i(v)$ is the Shapely value.

In order to calculate the Shapely values for factor i (comma bar gap, web speed, coating ratio), first that factor is excluded from the group of factors, then all the possible subsets of factors without factor i is for-

Table 5

RMSE Normalised to the average data of the same type.

Electrode RMSE normalised	Cathode		Anode	
	Mean	Std	Mean	Std
Mass loading (%)	4.23	1.55	7.82	3.59
Thickness (%)	6.59	2.57	8.02	2.76
Capacity (%)	8.72	4.10	8.91	2.72

med, $S \subseteq N \setminus \{i\}$ Then next step, the marginal value of adding a factor to the prediction problem is calculated via the term at the right-hand side of (8). The summation of marginal values considering the permutations of each subset size that can be generated out of all remaining factors excluding the factor i is then obtained. This summation when scaled to the number of all factors provides the final shapely value of that factor.

The Figs. 12 and 13 are the contribution of each factor for each individual characteristics of mass loading and thickness of the anode and cathode respectively.

For completeness, the normal distribution of the contributions is also included locally in the group of data with the same value. The distribution graphs of the values perform as a confidence bound for the calculated contributions. Each distribution graph is accompanied with a mean value of the shapely values, the dashed lines, the colour bar is used to distinguish the predicted characteristic's data points based on their value for a better overview, as well. The graphs are in fact the first order contribution of factors to the characteristics of the electrode and half-cells. Within this context, a direct (inverse) contribution means that an increase in the factor leads to an increase (decrease) in the response variable.

According to Fig. 12 and considering the dashed lines of the average contribution values, it is evident that for anode case study, the contribution of comma bar gap to active material mass loading and thickness is a plateau at lower values and then becomes fairly linear for higher values, Fig. 12 (a-1) and (b-1). In fact, lower comma bar gap values have a negative shapely value for the predicted mass loading and thickness, and the distribution of samples with the same measured value is fairly in the negative region as well. For comma bar gap values larger than 110 (um) the plateau shape turns into a linear contribution and after 177 (um) the shapely values become positive for both predicted characteristics.

The contribution of coating ratio to mass loading and thickness starts with a plateau at lower coating ratio levels and changes to an almost exponential behaviour after 120 (%). Considering the trend of the shapely values, comma bar gap and coating ratio both contribute directly to the thickness and mass loading of anode which means an increase in the mentioned factors would be associated with an increase in the response and the shape of this relation is either plateau, linear, or exponential with range of control variables.

For cathode the contribution of comma bar gap to the mass loading and thickness is direct and fairly linear for the whole range, Fig. 13 (a-1), (b-1), (a-2) and (b-2). For the comma bar gaps larger than 128 (um) and the shapely value sign changes from negative to positive. For coating ratio, Fig. 13 (a-3) and (b-3), the contribution is direct but with exponential behaviour, which the positive shapely values after 130 (%).

Unlike the first two factors, for web speed the contribution values are spread in a wider range and show higher variance, Fig. 13 (a-2) and (b-2). The contribution of web speed to the value of mass loading and thickness of the cathode is inverse which means that an increase in web speed reduces the mass loading and thickness of the cathode. According to the figures, the lower web speed levels have positive shapely values for both thickness and mass loading, and the opposite is the case for higher web speed values.

The contribution graphs for anode and cathode half-cell capacity are given in Figs. 14 and 15.

According to Figs. 14 and 15, the comma bar gap and coating ratio are contributing linearly and directly to the cathode half-cell capacity values. The contribution of web speed, Fig. 15 (a-2) is also fairly linear but inverse. This means that an increase in the comma bar gap or coating ratio, or a decrease in web speed would lead to cathode half-cells with higher energy capacity. For anode half-cells, while the contribution of comma bar gap and coating ratio are both direct, the comma bar gap is linearly contributing to capacity levels compared to coating ratio, which is showing plateaus at very low or very high values, Fig. 14 (a-1), (a-2).

In what follows, feature (factor) importance analysis is performed. This analysis helps to quantify the strength of the relationship between

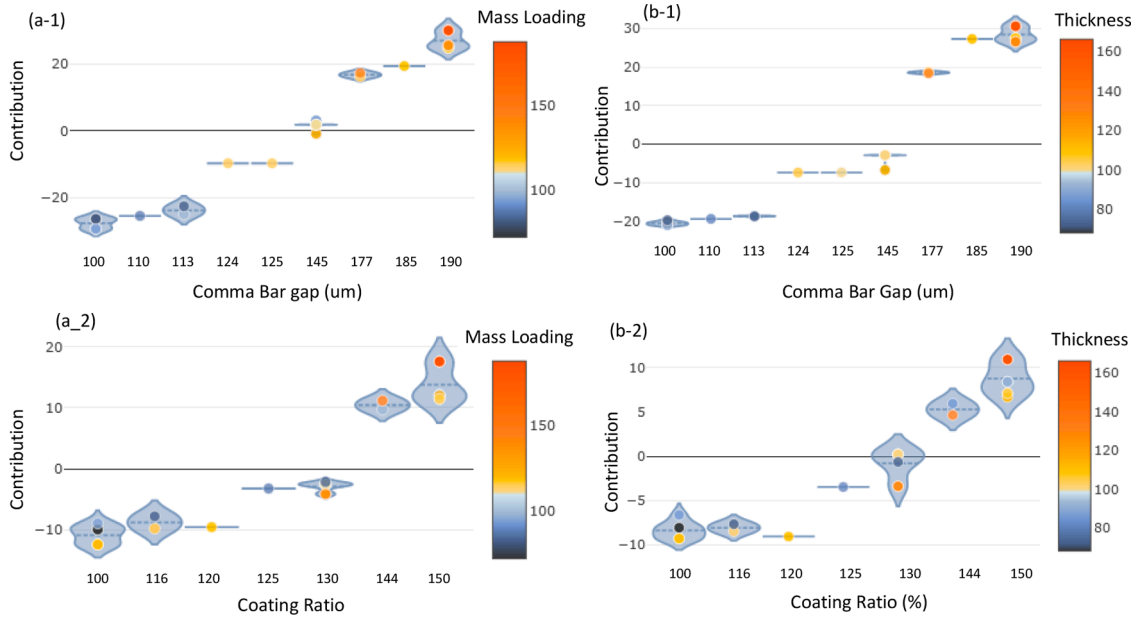


Fig. 12. Contribution results for Anode, (a-1) comma bar gap for mass loading, (b-1) Comma bar gap for thickness, (a-2) Coating Ratio for mass loading, (b-2) Coating ratio for Thickness.

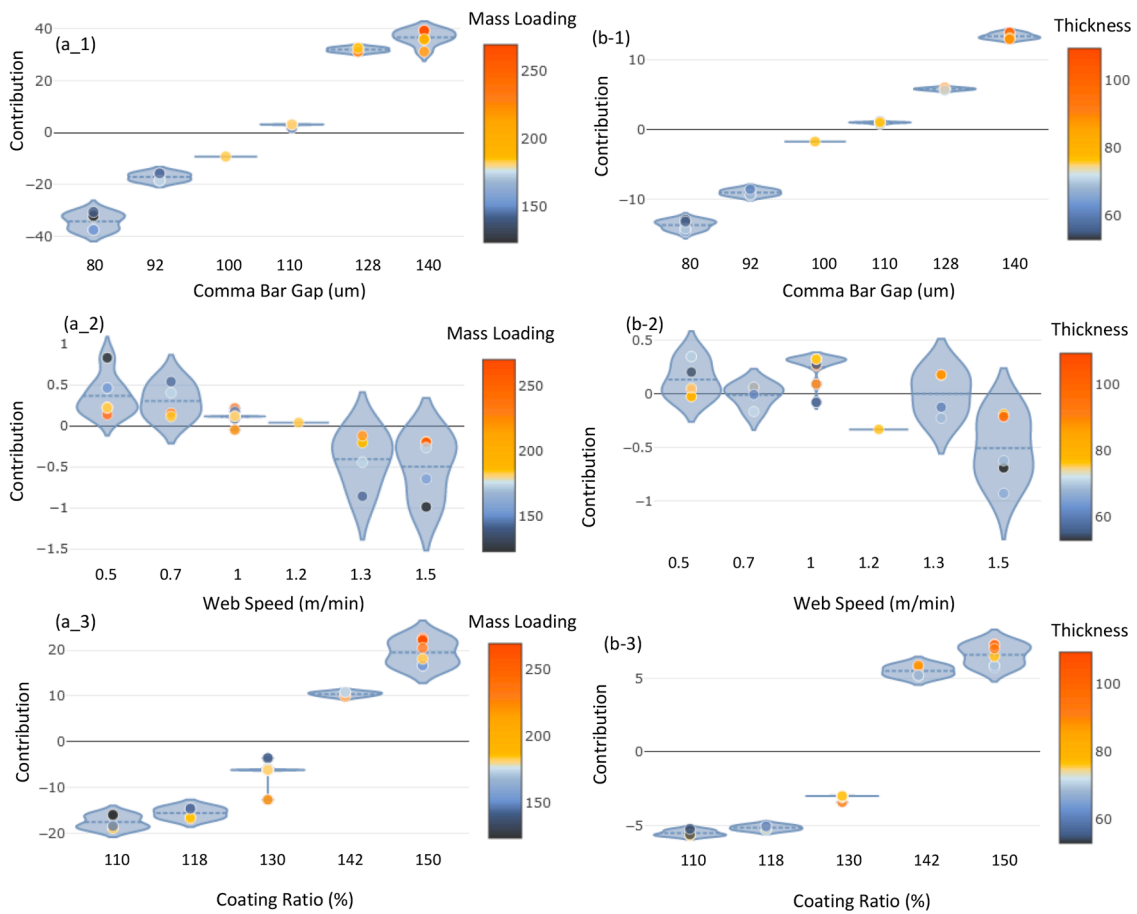


Fig. 13. Contribution results for Cathode, (a-1) comma bar gap for mass loading, (a-2) web speed for mass loading, (a-3) coating Ratio for mass Loading, (b-1) comma bar gap for thickness, (b-2) web speed for thickness (b-3) Coating Ratio for Thickness.

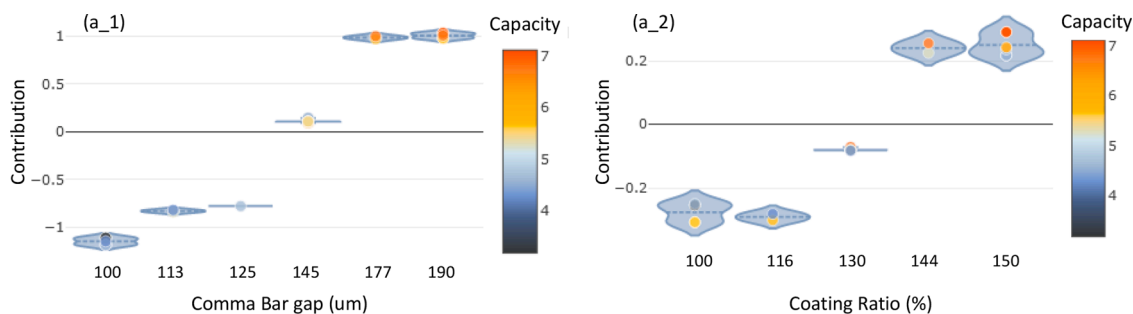


Fig. 14. Contribution results for anode half-cell capacity, (a-1) Comma bar gap, (a-2) Coating ratio.

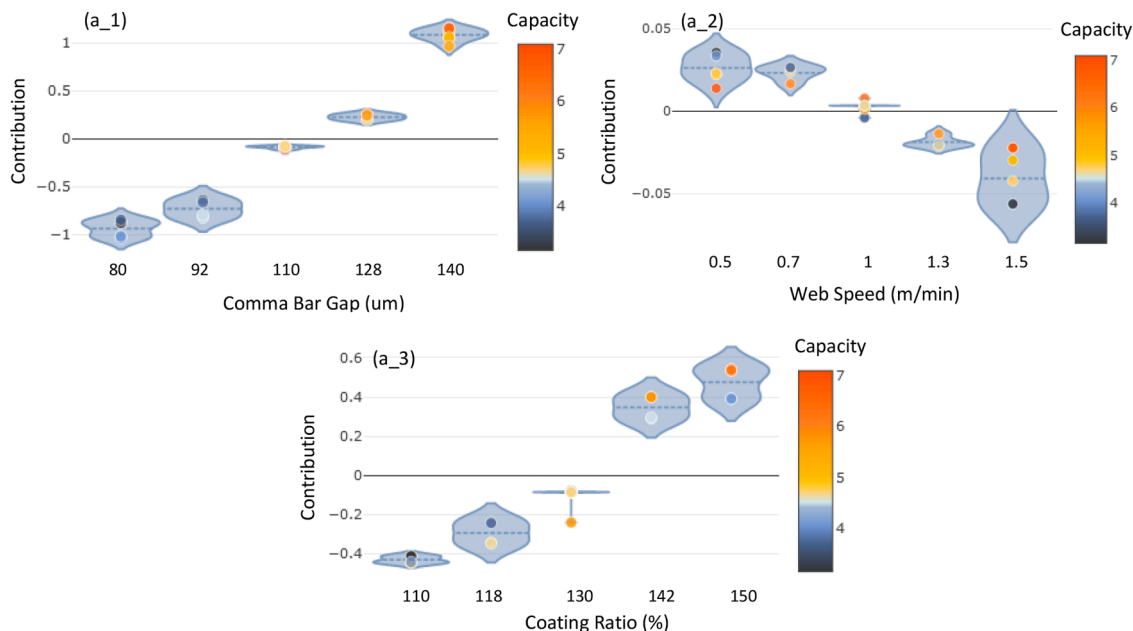


Fig. 15. Contribution results for cathode half-cell capacity, (a-1) comma bar gap, (a-2) web speed, (a-3) coating Ratio.

the manufacturing control parameters and the electrode characteristics as well as their associated half-cells. It helps the control parameters to be ranked and compared in order to make the key decisions.

In order to perform the feature importance analysis, the mean decrease in impurity index [45, 64] has been used in association with the GBT model. In each individual tree of the GBT model, each node splits the data from its parent node based on the information provided by a feature that improves the MDI to the highest amount. The final feature importance is obtained as the sum of the impurity improvements of the nodes using that specific feature [45, 64]. The feature importance charts are plotted as in Fig. 16 for the cathode, Fig. 17 for anode electrode and in Fig. 18 for both half-cell capacity.

This importance charts are highlighting the average of the scores for 10 runs of the GBT model and are reported as normalized values for a

better overview. As the Fig. 16 suggests, comma bar gap and coating ratio have much higher impact compared to the web speed for both mass loading and thickness of the cathode. Web speed has slightly higher score for thickness than mass loading in this case. Similarly, for anode the score of comma bar gap is a more important feature than coating ratio for both characteristics, Fig. 17.

For the capacity of cathode half-cells, Fig. 18, the comma bar gap is the dominant feature, and similar case applies to anode half-cell. Furthermore, for anode the score of comma bar gap is much higher than that for cathode half-cell. The importance scores for characteristics are summarised in Tables 6 and 7 for the cathode and anode respectively.

The analysis reveals the relative scores which highlight the features with the most, or least relevancy to the target variable. For the electrode

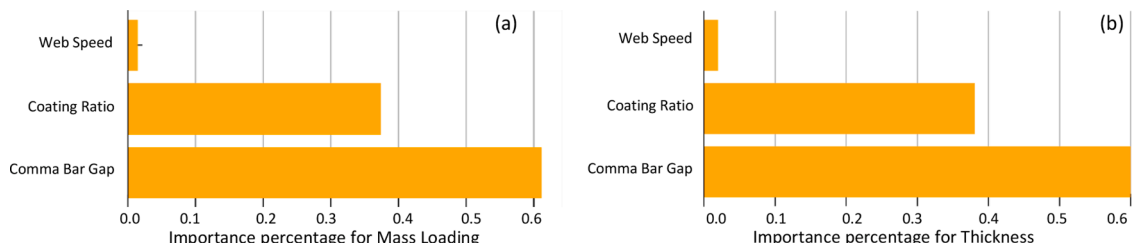


Fig. 16. Feature importance for Cathode, (a) Mass Loading, (b) Thickness.

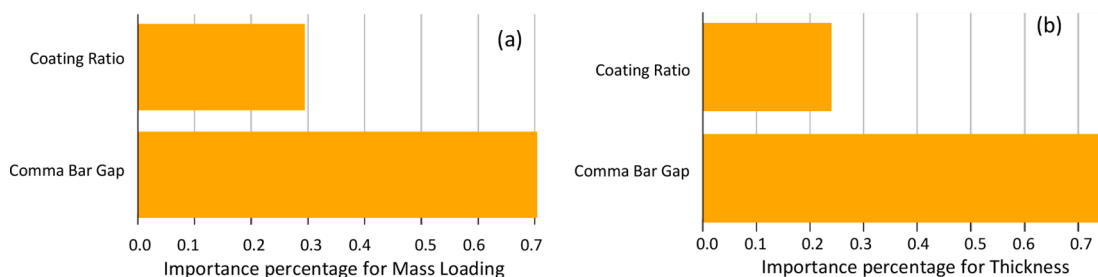


Fig. 17. Feature importance for Anode, (a) Mass Loading, (b) Thickness.

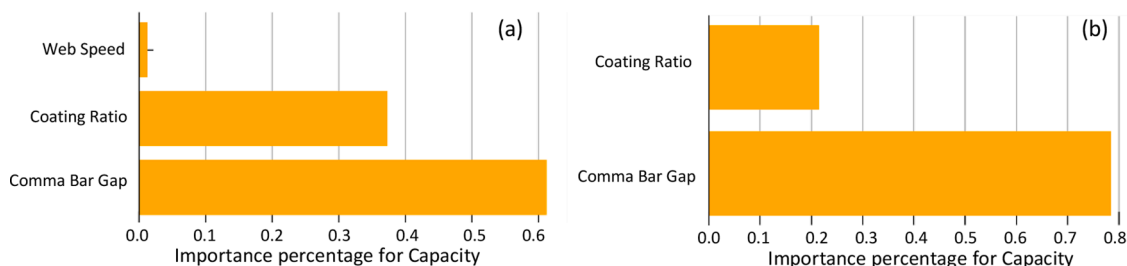


Fig. 18. Feature importance for (a) Cathode and (b) Anode half-cell Capacity.

Table 6

Factor importance (%) for cathode.

Characteristics	Cathode factors		
	Comma bar	Coating ratio	Web speed
Mass loading	0.6114	0.3741	0.0144
Thickness	0.5998	0.3804	0.0198
capacity	0.6129	0.3736	0.0134

Table 7

Factor importance (%) for anode .

Characteristics	Anode factors	
	Comma bar	Coating ratio
Mass loading	0.7048	0.2952
Thickness	0.7606	0.2394
capacity	0.7846	0.2154

manufacturing while the importance of comma bar gap is quite obvious, the importance of the other control parameters could be used as a basis to refine the levels for an efficient design of experiments in order to gather representative data for further modelling activities.

5. Conclusions

This study has focused on improving battery manufacturing processes via systematic design, analysis, and advanced machine learning approaches. It starts by proposing an efficient design of experiments for generating representative data during the manufacturing process of the electrodes and Li-ion cells. The 2-stage DoE helps to cover the space of the variables without extensive efforts for running a large number of tests. The study then continues with developing machine learning models, including gradient boosted trees and random forests to quantify the effect of key control parameters on the intermediate and final products. Unlike the traditional methods the importance and contribution analysis do not require only a single factor to change within the experiments but are applicable to more representative DoEs with more than one factor changing in each run.

Via the models, not only the most important factors are identified, but their relative scores are also calculated and quantified. The

characteristics of the electrodes including, mass loading, and thickness are concluded to be predictable with an average accuracy of 93.48%. The results also show that the cell capacity can be forecasted given the manufacturing parameters of comma bar gap, mass loading and coating ratio, with an average accuracy of 91.75%. The bootstrapping method as well as the cross-validation approach used in this study reduces the risk of overfitting and makes the models highly reliable and transformable to the similar data sets at volume scale. Therefore, this study is expected to provide a methodology for a systematic DoE and data analysis for high volume manufacturing. It is also expected to pave the way towards the in-situ and online control of battery manufacturing processes.

Future works in this direction are still in order and amongst those, performing an extensive analysis on the other battery manufacturing processes such as slurry mixing, and calendaring are the most crucial ones. Extending the models and the DoE approach from half-cell to full cell manufacturing is another direction worth to be addressed in future works.

CRedit authorship contribution statement

Mona Faraji Niri: Conceptualization, Methodology, Software, Investigation, Data curation, Formal analysis, Validation, Visualization, Writing – original draft. **Kailong Liu:** Investigation. **Geanina Apachitei:** Methodology. **Luis A.A Román-Ramírez:** Methodology. **Michael Lain:** Methodology. **Dharmika Widanage:** Funding acquisition. **James Marco:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was undertaken as part of the NEXTRUDE project, funded by The Faraday Institution, UK. [Grant Number: FIRG015].

Appendix

The hyperparameters of all models have been set via a grid search method and optimisation in the ranges provided as the following. The list is provided in order to increase the reproducibility of the results. The models are implemented in Python 3.0.

- Random Forest: Number of estimators = [100 1000], Minimum samples leaf = [2 5], minimum samples split = 2, Criterion = [Squared error Absolute error].
- Gradient boosted tree: Number of leaves = [2 50], Maximum depth = -1, Learning rate = [0.005 0.5], Number of model estimators = [50 10,000], Sample bin size = [20 100], Minimum child samples = [1 10], Min child weight = 0.01.

The data of this study could also be shared upon request and by the approval of the funding body.

References

- [1] Turcheniuk K, Bondarev D, Amatucci GG, Yushin G. Battery materials for low-cost electric transportation. *Mater Today* 2021;42:57–72.
- [2] Schmidt O, Hawkes A, Gambhir A, Staffell I. The future cost of electrical energy storage based on experience rates. *Nature Energy* 2017;2(8):1–8.
- [3] K. Edström, “ Battery 2030+ roadmap,” 2020.
- [4] Wentker M, Greenwood M, Leker J. A bottom-up approach to lithium-ion battery cost modeling with a focus on cathode active materials. *Energies* 2019;12(3): 504–22.
- [5] Berckmans G, Messagie M, Smekens J, Omar N. Cost projection of state of the art lithium-ion batteries for electric vehicles up to 2030. *Energies* 2017;10(9): 1314–34.
- [6] Kwade A, Haselrieder W, Leithoff R, Modlinger A, Dietrich F, Droeder K. Current status and challenges for automotive battery production technologies. *Nature Energy* 2018;3(4):290–300.
- [7] Stangel L. Tesla wasted \$150 M on scrap materials making cars this year. *Silicon Valley Business Journal* 2018.
- [8] Turetskyy A, Wessel J, Herrmann C, Thiede S. Battery production design using multi-output machine learning models. *Energy Stor Mater* 2021;38:93–112.
- [9] Goodenough J. Energy storage materials: a perspective. *Energy Stor Mater* 2015;1: 158–61.
- [10] Liu Y, Zhou G, Liu K, Cui Y. Design of complex nanomaterials for energy storage: past success and future opportunity. *Acc Chem Res* 2017;50(12):2895–905.
- [11] Kraysberg A, Ein-Eli Y. Conveying Advanced Li-ion Battery Materials into Practice The Impact of Electrode Slurry Preparation Skills. *Adv Energy Mater* 2016;6(21): 1600655.
- [12] Haselrieder W, Ivanov S, Christen D, Bockholt H, Kwade A. Impact of the Calendaring Process on the Interfacial Structure and the Related Electrochemical Performance of Secondary Lithium-Ion Batteries. *ECS Trans* 2013;50(29):59.
- [13] Meyer C, Kosfeld M, Haselrieder W, Kwade A. Process modeling of the electrode calendaring of lithium-ion batteries regarding variation of cathode active materials and mass loadings. *J Energy Storage* 2018;18:371–9.
- [14] Wu B, Widanage W, Yang S, Liu X. Battery digital twins: perspectives on the fusion of models, data and artificial intelligence for smart battery management systems. *Energy AI* 2020;1:100016.
- [15] Faraji Niri M, Bui T, Dinh T, Hosseinzadeh E, Yu T, Marco J. Remaining energy estimation for lithium-ion batteries via Gaussian mixture and Markov models for future load prediction. *J Energy Storage* 2020;8:101271.
- [16] Fermín-Cueto P, McTurk E, Allerhand M, Medina-Lopez E, Anjos M, Sylvester J, Dos Reis G. Identification and machine learning prediction of knee-point and knee-onset in capacity degradation curves of lithium-ion cells. *Energy AI* 2020;1: 100006.
- [17] Tong Z, Miao J, Tong S, Lu Y. Early prediction of remaining useful life for Lithium-ion batteries based on a hybrid machine learning method. *J Clean Prod* 2021;317: 128265.
- [18] Khaleghi S, Firouz Y, Berecibar M, Mierlo J. Ensemble gradient boosted tree for SoH estimation based on diagnostic features. *Energies* 2020;13(5):1262.
- [19] Carnovale A, Li X. A modeling and experimental study of capacity fade for lithium-ion batteries. *Energy AI* 2020;2:100032.
- [20] Nagulapati V, Lee H, Jung D, Paramanatham S, Brigljevic B, Choi Y, Lim H. A novel combined multi-battery dataset based approach for enhanced prediction accuracy of data driven prognostic models in capacity estimation of lithium ion batteries. *Energy AI* 2021;5:100089.
- [21] Lipu M, Hannan M, Karim T, Hussain A, Saad M, Ayob A, Miah M, Mahlia T. Intelligent algorithms and control strategies for battery management system in electric vehicles: progress, challenges and future outlook. *J Clean Prod* 2021;292: 126044. 126044.
- [22] Lipu M, Hannan M, Hussain A, Ayob A, Saad M, Karim T, How D. Data-driven state of charge estimation of lithium-ion batteries: algorithms, implementation factors, limitations and future trends. *J Clean Prod* 2020;277:124110.
- [23] Niri M, Dinh T, Yu T, Marco J, Bui T. State of Power Prediction for Lithium-Ion Batteries in Electric Vehicles via Wavelet-Markov Load Analysis. *IEEE Trans Intell Transp Syst* 2020 b:1–16.
- [24] Chemali E, Kollmeyer P, Preindl M, Emad A. State-of-charge estimation of Li-ion batteries using deep neural networks: a machine learning approach. *J Power Sources* 2018;400:242–55.
- [25] Aykol M, Herring P, Anapolsky A. Machine learning for continuous innovation in battery technologies. *Nat Rev Mater* 2020;5(10):725–7.
- [26] Lombardo T, Duquesnoy M, El-Bouysidy H, Årén F, Gallo-Bueno A, Jørgensen PB. Artificial Intelligence Applied to Battery Research: hype or Reality? *Chem Rev* 2021.
- [27] Liu Y, Esan O, Pan Z, An L. Machine learning for advanced energy materials. *Energy AI* 2021;3:100049.
- [28] Thiede S, Turetskyy A, Kwade A, Kara S, Herrmann C. Data mining in battery production chains towards multi-criterial quality prediction. *CIRP Annals - Manufacturing Technology* 2019;68:463–6.
- [29] Schnell J, Reinhart G. Quality management for battery production: a quality gate concept. *Procedia CIRP* 2016;57:568–73.
- [30] Lee J, Do Noh S, Kim H, Kang Y. Implementation of cyber-physical production systems for quality prediction and operation control in metal casting. *Sensors* 2013;18(5):35–40.
- [31] Schnell J, Nentwich C, Endres F, Kollenda A, Distel F, Knoche T, Reinhart G. Data mining in lithium-ion battery cell production. *J Power Sources* 2019;413:360–6.
- [32] Lieber D, Stolpe M, Konrad B, Deuse J, Mori K. Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. *Procedia CIRP* 2013;7:193–8.
- [33] Turetskyy A, Thiede S, Thomitzek M, von Drachenfels N, Pape T, Herrmann C. Toward data-driven applications in lithium-ion battery cell manufacturing. *Energy Technology* 2020;8(2). 1900136.
- [34] Liu K, Wei Z, Yang Z, Li K. Mass load prediction for lithium-ion battery electrode clean production: a machine learning approach. *J Clean Prod* 2021 a;289:125159.
- [35] Liu K, Hu X, Zhou H, Tong L, Widanage D, Marco J. Feature analyses and modelling of lithium-ion batteries manufacturing based on random forest classification. *IEEE/ASME Trans Mechatron* 2021 b. Early Access.
- [36] Farris F. The Gini index and measures of inequality. *Am Math Mon* 2010;117(10): 851–64.
- [37] Duquesnoy M, Boyano I, Larraitz G, Cereijo P, Elixabete A, Franco A. Machine Learning-Based on Assessment of the Impact of the Manufacturing Process on Battery Electrode Heterogeneity. *Energy AI* 2021;5. 100090.
- [38] Román-Ramírez L, Apachitei G, Faraji-Niri M, Lain M, Widanage W, Marco J. Understanding the effect of coating-drying operating variables on electrode physical and electrochemical properties of lithium-ion batteries. *J Power Sources* 2021;516. 230689.
- [39] Niri M, Liu K, Apachitei G, Ramirez L. Machine learning for optimised and clean Li-ion battery manufacturing: revealing the dependency between electrode and cell characteristics. *J Clean Prod* 2021;324. 129272.
- [40] Kornas T, Daub R, Kara MZ, Thiede S, Herrmann C. Data-and Expert-Driven Analysis of Cause-Effect Relationships in the Production of Lithium-Ion Batteries. In: *IEEE 15th International Conference on Automation Science and Engineering*. BC, Canada: Vancouver; 2019.
- [41] Cunha RP, Lombardo T, Primo EN, Alejandro AF. Artificial intelligence investigation of NMC cathode manufacturing parameters interdependencies. *Batteries Supercaps* 2020;3(1):60–7.
- [42] Duquesnoy M, Lombardo T, Chouchane M, Primo M, Franco A. Data-driven assessment of electrode calendaring process by combining experimental results, in silico mesostructures generation and machine learning. *J Power Sources* 2020;480: 229103.
- [43] Primo E, Touzin M, Franco A. Calendaring of Li (Ni0.33Mn0.33Co0.33) O2-based cathodes: analyzing the link between process parameters and electrode properties by advanced statistics. *Batteries Supercaps* 2021;4(5):834–44.
- [44] Hong J, Wang Z, Chen W, Wang L, Lin P, Qu C. Online accurate state of health estimation for battery systems on real-world electric vehicles with variable driving conditions considered. *J Clean Prod* 2021;294:125814.
- [45] Sandri M, Zuccolotto P. Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. *Stat Comput* 2010;20(4):393–407.
- [46] Merrick L, Taly A. The Explanation Game: explaining Machine Learning Models Using Shapley Values. *Machine learning and knowledge extraction*. cd-make 2020. lecture notes in computer science, vol 12279. Cham: Springer; 2020.
- [47] Yang S, He R, Zhang Z, Cao Y, Gao X, Liu X. CHAIN: cyber Hierarchy and Interactional Network Enabling Digital Solution for Battery Full-Lifespan Management. *Matter* 2020;3(1):27–41.
- [48] Steinberg D. Robust design: experiments for improving quality. *7 Robust design: Experiments for improving quality* 1996;13:199–240.
- [49] Mishra A, Mehta A, Basu S, Malode SJ, Shetti NP, Shukla SS, Nadagouda MN, Aminabhavi TM. Electrode materials for lithium-ion batteries. *Materials Science for Energy Technologies* 2018;1(2):182–7.
- [50] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient,” in noise reduction in speech processing. springer topics in signal processing. Berlin, Heidelberg: Springer; 2009.
- [51] Gogtay N, Thattai U. Principles of correlation analysis. *J Assoc Physicians India* 2017;65(3):78–81.
- [52] Komaroff E. Relationships Between p-values and Pearson Correlation Coefficients, Type 1 Errors and Effect Size Errors, Under a True Null Hypothesis. *J Stat Theory Pract* 2020;14(3):1–13.
- [53] Oehlert GW. A first course in design and analysis of experiments. USA: W. H. Freeman and Company; 2010.

- [54] Myers RH, Montgomery DC, Anderson-Cook CM. Response surface methodology: process and product optimization using designed experiments. New Jersey: John Wiley & Sons; 2016.
- [55] Adams BM. Dakota a multilevel parallel object-oriented framework for design optimization parameter estimation uncertainty quantification and sensitivity analysis: version 6.14 user's manual. et al. Albuquerque, USA: U.S. Department of Energy; 2021.
- [56] Wang F, Mamo T. Gradient boosted regression model for the degradation analysis of prismatic cells. *Comput Ind Eng* 2020;144:106494.
- [57] Friedman JH. Greedy Function Approximation: a Gradient Boosting Machine. *Ann Stat* 2001;29(5):1189–232.
- [58] Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002;38(4): 367–78.
- [59] Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.
- [60] Berrar D. Cross-Validation. *Encyclopedia Bioinformatics Computational Biology* 2019;1:542–5.
- [61] Efron B, Tibshirani R. An introduction to the bootstrap. Boca Raton, Florida: CRC Press; 1993.
- [62] Lee T, Ullah A, Wang R. Bootstrap aggregating and random forest. *Macroeconomic forecasting in the era of big data*. Springer; 2020. p. 389–429.
- [63] Hart S. Shapley value. *Game theory*. London, The New Palgrave: Palgrave Macmillan; 1989.
- [64] Xu Z, Huang G, Weinberger K, Zheng A. Gradient boosted feature selection. 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York USA 2014.