WINTER 2015/2016 • Volume 14/Issue 4 • ISSN 1538-8786

BioProcessing JOURNAL

Trends & Developments in BioProcess Technology

A Production of BioProcess Technology Network

BioProcessing JOURNAL
TRENDS & DEVELOPMENTS IN BIOPROCESS TECHNOLOGY

Improving Biopharmaceutical Manufacturing Yield Using Neural Network Classification

By Will Fahey and Paula Carroll

Abstract

Traditionally, the Six Sigma framework has underpinned quality improvement and assurance in biopharmaceutical manufacturing process management. This paper proposes a neural network (NN) approach to vaccine yield classification and compares it to an existing multiple linear regression approach. As part of the Six Sigma process, this paper shows how a data mining framework can be used to extract further value and insight from the data gathered during the manufacturing process, and how insights into yield classification can be used in the quality improvement process.

1. Introduction

The World Health Organization states that pneumococcal disease is the world's number one vaccine-preventable cause of death among infants and children under the age of five.[1] Vaccines are a crucial resource in the fight to lower infant mortality rates for developing countries[2], coming second only to clean drinking water. However, the vaccine manufacturing sector is quite fragile due to strict regulatory licensing and cost concerns.[3] Approaches to public health policies that could contribute to the sustainability of the vaccine manufacturing sector are outlined by Robbins and Jacobson.[3] The focus of Proano *et al*.[2] is on price-bundle determination for combination vaccines, which maximise social good by ensuring sufficient vaccine is produced while also ensuring minimum profit levels to guarantee the long-term viability of manufacturers.

This article focuses on a different idea that may contribute to ensuring the long-term viability of the sector by exploiting value from the data gathered for regulatory compliance and operations management.

1.1. The Vaccine Manufacturing Process

A vaccine is typically made up of a number of individual polysaccharide components, called serotypes, which immunize the recipient against a particular strain of the targeted disease.[4] Manufacturing a pneumococcal vaccine is a complicated procedure, involving the use of bioreactors to manage cells so they produce the various active biological substances.[5] A bioreactor is a vessel used to replicate the conditions found in a mammalian body that are needed to promote the creation of the biological components that combine to form the vaccine product. These components are passed through various ultrafiltration steps which purify the product, and diafiltration steps which concentrate the product to desired levels. The process is categorised by long lead times of up to 30 days. Manufacturing one batch may involve between 40 and 50 process steps which may be characterised by explanatory variables (EVs) such as temperature, pressure, and both nutrient and gas flow rates at various points during the production process.

A vaccine manufacturing process involves the combination of saccharide components which elicit the desired immune response from a recipient. In this paper, the success of this combination is reported as the "yield" and is defined as the amount of vaccine created in a production batch as a percentage of the expected amount (based on the quantity of raw materials used). The dynamic nature of biological components used in vaccine manufacturing renders static methods of measurement as only indicative, and adds to the complexity of identifying root causes of yield fluctuation.

Traditionally, Six Sigma (6σ) approaches such as design of experiments (DOE) and statistical process control (SPC) techniques have been used to improve yield and decrease variability.[6] The method utilised by 6σ to achieve quality improvement (QI) is the use of brainstorming to generate hypotheses about which EVs have an impact on quality. Then, statistical methods such as regression, or hypothesis testing, are used to confirm or disprove these hypotheses. Such analyses provide the necessary feedback for product/process design (or redesign), as well as other corrective QI actions.[7]

To identify the root cause of poor yield, 6σ is the incumbent approach used by example company "Z." The vaccine manufacturing process has multiple biological inputs, each with multiple quality characteristics that may potentially explain yield fluctuation. There are many combinations of control equipment settings and possible EVs, so it is often unclear how the inputs interact and how the multiple measurement settings affect process outputs. However, if the 6σ process is taken to its theoretical conclusion, a potentially exponential number of hypotheses could be generated during the measurement of a complex issue such as yield variability. In brainstorming sessions, there are also considerations of inherent human bias that might influence the identification of a possible root cause during the 6σ measurement phase.

Large amounts of data are generated and collected by automated manufacturing processes, most of which are used for process control rather than process improvement.[8] This paper proposes a novel approach for extracting real business value from the wealth of data that has already been gathered. The method focuses on an evaluation of neural networks (NNs) to generate and test hypotheses about which process parameters, or combinations of parameters, lead to a high or low yield. In this case, a hypothesis is that a process parameter setting contributes to yield fluctuation. Two serotypes, referred to as serotype "X" and serotype "Y", are the subject of this study which was undertaken using manufacturing data from company Z.

2. Vaccine Manufacturing Challenges and Opportunities

Biopharmaceutical manufacturing is one of the most heavily regulated industries in the world today. Regulatory bodies such as the US Food and Drug Administration (FDA) have been relentless in driving higher levels of process control and understanding in the biopharmaceutical sector. These bodies recognise the significance and untapped potential of data mining (DM) methods to enable more robust biological manufacturing processes through increased process knowledge. Analysis of manufacturing data using multivariate data analysis (MVDA) was stimulated by the FDA's landmark guidance on process analytical technology (PAT) in 2004.[9] Regulatory authorities are demanding a greater level of process characterisation and robustness in the biopharmaceutical industry as a means of ensuring a consistent supply of safe, efficacious products for patients. However, there remains a gap between the huge quantities of manufacturing data available and how much knowledge the industry derives from it.[10]

Regular changes to production processes are inevitable in a manufacturing industry—particularly when a strong culture of continuous improvement in 6σ exists. However, every change involves risk. Quantified risk assessment can only be effective in mitigating this risk when the process is sufficiently understood. DM then becomes an essential tool in assessing the impacts of changes to critical process parameters, such as those in downstream operations.[11]

Some of the challenges faced by the vaccine manufacturing industry are outlined next. These challenges can also be interpreted as DM opportunities[12] and include:

- A high number of possible EVs is needed to ensure an adequate description of the process, including many statistical measures associated with input components and process-stage metrics, such as temperature and pressure. This is especially true for biological manufacturing processes.
- A high number of dependencies must be modelled when several components are integrated into one system. However, it is not only the high number of statistically proven dependencies that require significant resources to be modelled, but there are other potential dependencies that have to be accepted or rejected as contributing to an improved model. This calls for an efficient way of pruning hypothesised relations. Inherent yield variability, which is referred to in 6σ as a common cause issue, is very rarely attributable to a single input value or process setting. It is much more likely that interdependencies among EVs conspire to produce a low yield.
- Uncertainty of measurement data, such as the proportion of manually recorded data, is getting smaller, but it is still present and indicates the possibility that transcription error still exists. Methods to capture the uncertainty associated with the auto-capture of other manufacturing data also aim to quantify doubt about the validity of sampling, precision, and possible calibration errors.
- Incomplete information is a common problem when using raw manufacturing data. Values are sometimes deemed unimportant to the process outputs, and due to resource constraints, are not fully gathered. DM has an advantage over traditional statistical methods as it offers intelligent ways to replace missing values. One such method is *k*-means clustering.[13]

With missing data, statistical tests can lose power, results can be biased, and analysis may not be feasible at all. Missing values are replaced with estimated values according to an imputation method or model. In the *k*-nearest neighbor (*k*-NN) method, a case is imputed using values from the *k* most similar cases. *k*-NN is a non-parametric, lazy learning, algorithm where "non-parametric" means that the method does not make any assumptions about the underlying data distribution. This property is useful in this case study since the data does not necessarily allow typical theoretical assumptions like following a typical distribution such as normal or exponential.

"Lazy" refers to the fact that the algorithm does not use the training data points to do any generalization. In other words, there is no explicit training phase. This speeds up the algorithm, making it practical for use in one of the nested operators in the DM process, and allows a stronger

model than simply replacing each value with the mean of the other values. This approach to missing data has been used successfully[14] and illustrates another advantage that DM techniques have over 6σ regression, which cannot be performed with missing values.

2.1. The 6σ and Cross-Industry Standard Process (CRISP) for DM Methodologies

QI programmes aim for improvements in manufacturing yield by using the *define-measure-analyse-improve-control* (DMAIC) approach to reach 6σ quality levels with less than 3.4 defects per million opportunities. Each project in the 6σ methodology has five phases represented by the initials in DMAIC. An overview of each phase is as follows:

- ***Define*** the nature of the problem and frame the problem statement. Make sure this statement aligns with the project sponsor's outlook on the issue, and then map the process to ensure consensus.
- ***Measure*** key aspects of the current process and collect relevant data. This involves visualising and investigating the data to provide insight and potential root causes of the issue. Use these as a benchmark for brainstorming all potential root causes of the issue.
- ***Analyse*** the data to investigate and verify cause-and-effect relationships. Use statistical techniques to rule-in or rule-out the potential root causes. Techniques include regression and hypothesis testing.
- ***Improve*** the confirmed root causes by error-proofing the issue, and set up pilot runs to establish process capability.
- ***Control*** by piloting the future-state process to ensure that any deviations from the target are corrected before they result in defects. Implement control systems such as statistical process control, and monitor the process to make sure the improvements are effective.

The 6σ process has many advantages, providing structure to the problem-solving effort so that the goals are clear and well-defined. The structure and sequential nature provide a common language so that stakeholders from every level can understand the problem and how it will be solved.

DMAIC also provides a data-driven structure to a diverse team of subject matter experts (SMEs) who each bring an expert, but possibly biased, understanding of the root cause in the process problem. In the absence of the DMAIC structure, SMEs may jump to premature conclusions based on their own process experiences.

Wu *et al.*[15] point out that classical methods, such as control charts, aim to monitor the process and not imply the relationship between the EVs and the highly important outputs. Büchner *et al.*[16] elaborate on the shortcomings of retrospective statistical methods and state that they considerably limit the potential for continuous process improvement.

CRISP is the *de facto* industry standard process methodology for DM. The process was inspired by the 6σ DMAIC methodology and must be identified by practitioners to allow the adoption of DM as a key part of business processes.[17] It is an iterative and adaptive hierarchical process based on real-world experience of how people conduct DM projects, and provides an overview of the lifecycle of a DM project. The CRISP DM process framework defines six phases of a DM project, their respective tasks, the relationships between these tasks, and the deliverables of each phase. A brief outline of the phases is given next:

1. ***Business understanding.*** This initial phase focuses on understanding the project objectives and requirements from a business perspective. This phase is comparable to the *define* phase of a 6σ project, where a plan is formed and the project goals are reviewed by the project sponsor.
2. ***Data understanding.*** The data understanding phase starts with initial data collection and proceeds with identification of data quality problems. Some early exploratory data analysis is also carried out in order to gain an initial impression of the possible relationships present in the data. This can be compared to the preliminary stage of the "measure" phase of a 6σ project.

3. ***Data preparation.*** The data preparation phase covers all activities needed to construct the final data set from the initial raw data. This includes dimensionality reduction, dealing with missing values, data normalisation, and dealing with outliers. This phase is not usually required during a 6σ project.
4. ***Modelling.*** In this phase, modelling techniques are selected and applied, and their parameters are calibrated to optimal values.
5. ***Evaluation.*** The practical applications of the model are evaluated. But before proceeding to the final deployment of the model, it is important to thoroughly evaluate and review the steps executed to create it to be certain the model properly achieves the business objectives. Any risk in applying the model must also be assessed.
6. ***Deployment.*** Creation of the model is not the end of the project. The knowledge gained will need to be translated into a format that the customer can use and understand.

Figure 1 shows the structure of the team required to complete an analytics manufacturing project, adapted from the findings of Büchner *et al*.[16] Three skillsets are essential in building a team for a DM project in the manufacturing domain: a data expert, a DM expert, and a domain expert.

Ideally the data expert should belong to the IT department and have experience with relational databases. This proved to be the case in this study. The data expert was an automation engineer with experience in querying databases using structured query language (SQL) programming. This point is expanded upon in section 5.

The domain expert in this study belonged to the technical operations group and had significant experience with the manufacturing process and 6σ statistical techniques.

DM is usually carried out in large organisations; however, a domain expert who is also an expert in the data stored by the organisation is rare. Often the DM expert is a consultant with no knowledge of the manufacturing process (which is a distinct disadvantage). The team for this project was fortunate in that the DM expert had experience in 6σ techniques and data retrieval using SQL, and was also familiar with the manufacturing process at a high level.

2.2. DM and Analytics Opportunities

Machine learning (ML), which is closely related to computational statistics, extends the use of DM through the use of algorithms that learn patterns from the data. ML approaches, such as NNs, are well-equipped to deal with the range of problems outlined in section 2. In many cases, NNs are used for modelling complex non-linear relations with a large number of EVs, as described by Hickey *et al*.[18] Chien *et al*.[19] illustrate how NNs can also adapt dynamically to changes occurring in the modelling system in real-time. This is essential for manufacturing applications. Even though the initial training results may not be accurate, the NN performance improves with time as more training data samples are provided.

One advantage of using a NN is that it can be fitted to any kind of data set and does not require the relationships in the model to be explicitly stated. NNs are particularly useful when data may be noisy and relationships may be

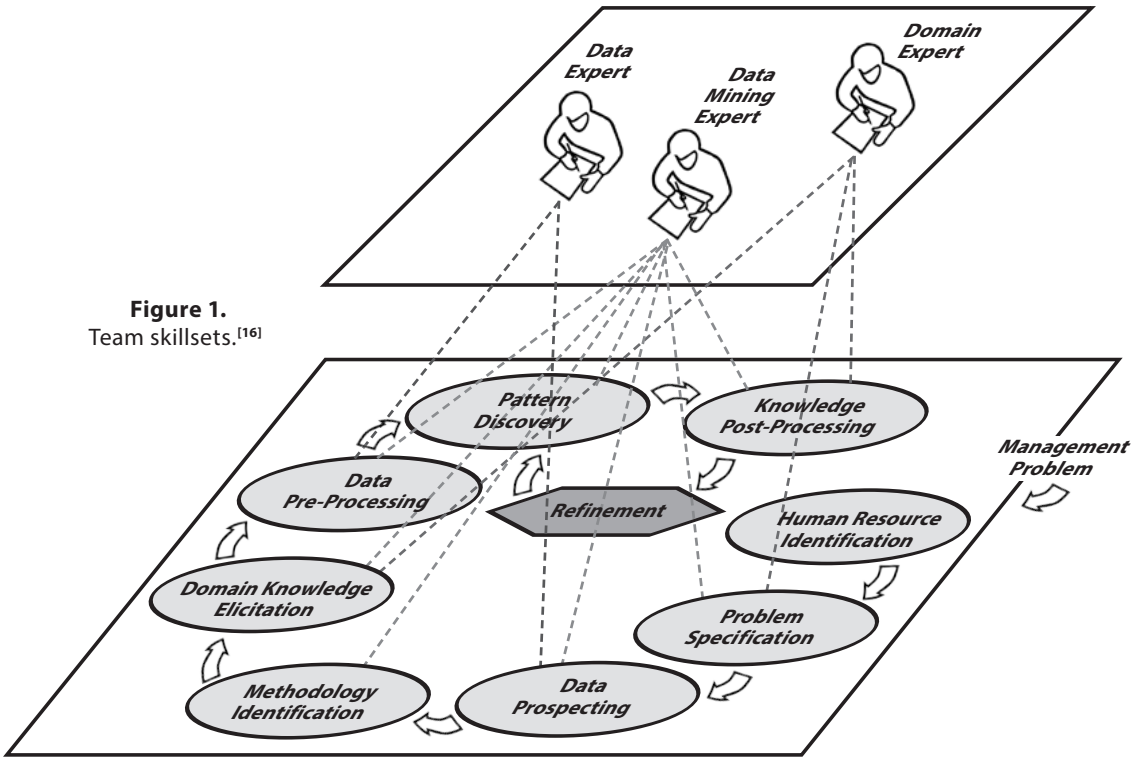

Figure 1.
Team skillsets.[16]

non-linear, such as the data set in this study. Because of the complexity and non-linearity involved in vaccine manufacturing systems, such systems lend themselves well to the use of NNs, where they benefit from the NN online learning and adaptive abilities. NNs are criticised for being a black box but have demonstrated their usefulness in many practical applications within the manufacturing sector.[7]

NNs are a supervised learning approach designed to model the method by which human brains accomplish a certain task. Tetko *et al.* give some characteristics of NNs that have led to their widespread use.[20] A NN can learn by adjusting the topology (also called architecture or structure) and edge weights of a network connecting certain input signals to a desired output response. Such a training process is an iterative one which is run until no further adjustment is required. Once a NN design has been based on a training data set, it can then be tested and evaluated on a test data set. NNs can be used for classification or prediction tasks.

In this paper, a NN is used to classify a production batch as high or low yield, depending on the values of the manufacturing production process EVs.

2.2.1. *Cross-Validation of the Model*

The cross-validation method involves repeated training of the neural network using a number of partially-overlapping and arbitrarily large portions of data as the training sets, with the remainder of the data in each case being used as the independent test set. In this way, all data will eventually be used in the test set, and errors due to the inclusion of non-representative data in either set are avoided. This is effective but computationally expensive.

A validation set is used either to refine the topology of the network or to serve as a stopping criterion. NN topology design parameters, such as the number of units in a hidden layer, or the number of hidden layers, determine the structure of the network.

In the first methodology, the network designer assesses the performance of different trained networks by evaluating an objective function with the validation set. The network with the smallest error is selected. In the second approach, training and validating take place concurrently. The network stops learning once the sum of residuals, based on the validation set, starts to increase beyond a user-specified number of iterations. A testing set is later used to avoid overfitting where the network has learned the noise in the training data and is no longer useful to generalise unseen data.[21]

The accuracy measure for evaluating the performance of classifiers is defined as:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

In this study which predicts high or low yield, a true positive is a high yield production batch that is correctly identified by the classifier as high yield. Accuracy is then defined as the number of correct (or true) high and correct low yield predictions divided by the total number of tests.

Precision is another measure used in ML quality assessment to measure the ability of the classifier to make positive predictions correctly. Precision of the classifier to identify high yield is defined for this study as the number of correct high yield predictions divided by the total number of high yield predictions (both correct and incorrect). A similar low yield precision measure is also defined.

Recall is a measure used to quantify the sensitivity of an ML classification system. In the case of a high yield classification, it is defined as the number of correct (true) high yield predictions divided by the sum of the correct high and incorrect low yield predictions. A similar low yield recall measure is also defined.

The design of a NN is more of an art than a science. There is no unified approach for setting the design parameters of a NN. Zobel and Cook give a good overview of selecting the design parameters of NNs.[22] The general approach is one of trial and error to change the design parameters and note if it has an effect on the performance of the model. The NN design parameters include:

- ***Hidden Layers.*** This parameter describes the number and size of all hidden layers. The user can define the structure (network topology) of the NN with this parameter. The hidden layer links the input layer to the output layer. Within each node in the hidden layer, a weighted sum calculation is carried out relating the input layer to the output using a predefined function.
- ***Training Cycles.*** This parameter specifies the number of training cycles used for the NN training. In a back-propagation approach, the output values are compared with the correct answer to compute the value of some predefined error function. The error is then fed back through the network. Using this information, the back-propagation algorithm adjusts the weights of each connection in order to reduce the value of the error function by some small amount. This training process is repeated a number of times.
- ***Learning Rate.*** This parameter determines how much the weights are changed at each step.
- ***Momentum.*** The momentum adds a fraction of the previous weight update to the current one. This prevents local maxima and smooths optimisation directions.

2.2.2. *Principal Component Analysis and Data Reduction*

Bellman's "curse of dimensionality" indicates that a large number of EVs and a small number of batches (or samples) can lead to a poor model. In this study, this problem occurs and many possible EVs are available for a number of production runs. A time series plot (TSP) shows a sequence of observations of the variables of interest such as temperature, pressure, and flow rate during the manufacturing process steps. TSP data can have extremely high

dimensionality because each time point can be viewed as a single dimension (giving a tuple of values for the EVs). High dimensionality can lead to an over-fitted ML model, and the raw time series may be too expensive computationally to process during the NN training stage, so the number of dimensions must be reduced.[23]

One of the challenges faced early in the data preparation phase of this study was deciding how to deal with time series (TS) data from the pivotal filtration/concentration steps of the manufacturing process. It was essential to add this data to the model for consideration; however, distributions of the values over time were erratic and did not fall into a recognised pattern (*e.g.*, binomial, log, normal). To achieve data dimension reduction, the distribution of each EV (*e.g.*, temperature) was represented by a number of descriptive statistical values such as the "moments" of the EV (for example, the mean is the first moment). These statistics were then passed as the inputs to the NN. The descriptive statistics method of data reduction was motivated by Bickel and Lehmann[24] as a means to summarise a non-parametric model. They recommend the following group of statistics:

- ***Mean.*** The average of the values.
- ***Standard Error.*** The standard deviation of the sampling distribution of a distribution. Standard error of the mean (SEM) is calculated by dividing the standard deviation by the square root of the number of observations.
- ***Median.*** The value in the middle of a set of numbers: half the numbers have values that are greater than the median, and half have values that are less.
- ***Mode.*** The most frequently occurring or repetitive value in an array or range of data.
- ***Standard Deviation.*** The standard deviation tells us how much variation is present in a distribution.
- ***Trimmed Mean (20%).*** Trimmed mean discards the top 10% and lowest 10% of values. This was included to account for a large number of outliers. A significant number of outliers can be identified by comparing this value to the mean.
- ***Kurtosis.*** The relative peakedness (positive kurtosis) or flatness (negative kurtosis) of a distribution compared with the normal distribution.
- ***Skewness.*** The degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.
- ***Maximum.*** The maximum value recorded.
- ***Minimum.*** The minimum value recorded.
- ***Quartile 1 (Q_1).*** The middle value in the first half of the rank-ordered data set.
- ***Quartile 3 (Q_3).*** The middle value in the second half of the rank-ordered data set.
- ***Interquartile Range (IQR).*** A measure of variability based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles, and are denoted by Q1, Q2, and Q3 respectively.

Principal component analysis (PCA)[25] is a multivariate dimension reduction technique applicable to large data sets. The set of possibly correlated EVs is reduced to a set of linearly uncorrelated principal components values. PCA identifies a vector similar to a basis that uncovers the underlying structure in the data. They are mathematical constructs that point in the direction where there is the most variance. PCA takes into account the combined contribution to the variation of a number of vectors as opposed to the univariate point of view represented by a correlation matrix.

3. Methodology

The incumbent baseline multiple linear regression (MLR) used by company Z to identify yield improvement models the yield based on a correlation factor for univariate relationships within the data set. The NN model is compared with the incumbent MLR model using root mean squared as the error metric.

3.1. CRISP-DM Implementation

For the NN method, TSP data is first summarised using the measures suggested by Bickel and Lehmann.[24] PCA is then used for dimension reduction to identify a minimal set of prioritised EVs. A NN model is then created to relate the identified EVs to manufacturing yield output. The CRISP-DM framework is used to support the implementation of the study, and a summary of the CRISP-DM phases of this study follows:

1. ***Business Understanding.*** As this methodology was a new approach for company Z, significant groundwork had to be completed to ensure management buy-in. This involved presentation of the methods as win-win since there was little or no capital required, and no new data was required since historical data was re-used. Manpower resources whose skill sets were complementary to the project's requirements were identified, and in this case, a mix of process and statistical knowledge, with an interest in process modelling, were required. Due to company Z's unfamiliarity with the CRISP-DM methods, it was difficult to assign a realistic goal for yield improvement. Company Z agreed to an initial study as proof of concept of the CRISP-DM process under the banner of a company-wide innovation initiative.
2. ***Data Understanding.*** The data was extracted from multiple data sources and assembled into a format that could be read by the DM software. This was the most

time-consuming part of the process as most of the data preparation tasks were completed manually. Missing values were imputed using the *k*-means clustering algorithm.

3. ***Data Preparation.*** The data from each database was combined into one data set, the TSP data was summarised, and PCA was applied to reduce the data to the most significant EVs. The software platform used for these tasks was RapidMiner, which is an integrated software tool for ML and DM applications.

   Each principal component (PC) explains a certain proportion of the variation in the target variable (yield), and is a mathematical construct for the dimension reduction filtering effect. It is useful, from a practical point of view, to understand which EVs are significant.[26] Each PC is correlated with a number of EVs which the RapidMiner software ranks in order from highest to lowest. Grading for the study data showed there was a natural drop-off in correlation at 40 EVs. This process produces a prioritised data set which is then passed to the NN for the modelling phase. For example, in this study, a PC that explains 50% of the yield variation contributes 20 EVs to the pool of 40. This is discussed further in section 4.
4. ***Modelling.*** The 40 EVs identified during the data preparation step are used to create a NN model.
5. ***Evaluation.*** The output of the model shows an equation that gives each variable a correlation coefficient to illustrate how it relates to the other variables when yield is at the optimum. These findings are validated by the domain expert.
6. ***Deployment.*** Creation of the model is not the end of the project. The settings for EVs are implemented on the manufacturing floor, normally under protocol to validate the findings before they are committed to standard operating procedures.

Data Preparation Phase:
Time Series Representation and Data Reduction

The data preparation phase of the CRISP-DM methodology required careful consideration. As noted in section 2.2.2, large volumes of data had to be reduced to a manageable representation to allow a tractable model. Due to the high number of variables, 800–900 for each vaccine batch in this study, it was necessary to distill the number down to a more manageable size before it was passed to the NN model. In the data preparation phase of the CRISP-DM process, each of the statistics proposed by Bickel and Lehmann[24], as described in section 2.2.2, was calculated for each EV time series, creating a set of values that represent the TS, like the components of a fingerprint.

Due to the high number of EVs, or *p*, in comparison to *n* (the number of batches), it was necessary to further reduce the number of possible EVs that are presented to the NN in order to obtain a tractable model. The dimension reduction was necessary as the software either could not handle the number of EVs, or in the cases where it could, the NN classification model was poor. This ratio of *n*:*p* was 24:180 for serotype X, and 21:344 for serotype Y. A number of NN training runs were attempted without reducing the dimension of the data, and the NN training process was stopped after 60 hours without having yielded a result.

Multi-group modelling is based on the assumption that a common eigenvector subspace exists for the individual variance/covariance matrix. Through the pooled sample variance/covariance matrix of the batches relating to different yields, the principal component loading is calculated. The EVs that are most strongly correlated to yield in isolation are identified, and these 15–20 EVs (as there is a natural drop-off in correlation coefficient after this point) are used in the MLR model. **Table 1** shows a comparison of the data reduction techniques for the MLR and NN approaches.

3.2. Modelling Using NNs

3.2.1. *NN Design Parameter Optimisation*

Having prepared the data, the next phase of the CRISP-DM method focused on building an appropriate model. The RapidMiner software platform was used to develop the NN model and find good settings for the NN design parameters described in section 2.2.1. A summary of the impact of changes in the NN design parameters for serotype X are shown in **Table 2**. This information shows that the number of hidden layers, and the momentum, are significant factors. These NN design parameters were fed to the "optimise parameters" operator in RapidMiner to ensure high accuracy was reached in the final NN model.

TABLE 1. Comparison of approaches.

Technique	Incumbent MLR Model	Proposed NN Model
Dimension reduction	Correlation matrix	Principal component analysis
Modelling	Multiple linear regression	Neural network

TABLE 2. NN model parameters for serotype X.

Model Parameter Adjusted	Values	Effect On Accuracy
Training cycles	100, 200, 300, 500, 1000	None
Learning rates	0.1, 0.3, 0.5, 0.8, 1.0	None
Hidden layers	1, 2, 3	Two hidden layers increased accuracy
Momentum	0.1, 0.3, 0.5, 1.0	1.0 decreased accuracy

Table 3 shows the impact on accuracy for changes in the serotype Y NN design parameters.

The number of hidden layers and momentum are also significant in the serotype Y NN model.

TABLE 3. NN model parameters for serotype Y.

Model Parameter Adjusted	Values	Effect On Accuracy
Training cycles	100, 200, 300, 500, 1000	None
Learning rates	0.1, 0.3, 0.5, 0.8, 1.0	None
Hidden layers	1, 2, 3	More than one hidden layer reduced accuracy
Momentum	0.1, 0.3, 0.5, 1.0	1.0 decreased accuracy

4. Results and Analysis

4.1. Model Performance Comparisons

Table 4 compares root mean squared errors (RMSEs) for each model. The NN model offers a significant improvement over the MLR model for both serotype X and serotype Y. Both NN models have a better RMSE, and the variance is also considerably smaller. This indicates that the distance from the residuals to the fitted model does not vary significantly from point-to-point.

In the serotype X MLR model, there is a possible additional error of ±5.751 on top of the already large RMSE. The *p*-values indicate that the MLR model is unsuitable. The residuals are large and not normally distributed, so the resulting outputs would be susceptible to misinterpretation. The serotype Y MLR model shows similar findings, but on a smaller scale.

TABLE 4. Performance measure for model comparison.

Model	Root Mean Squared Error	*p*-Value
Serotype X NN	0.244 ± 0.279	NA
Serotype X MLR	11.892 ± 5.751	0.05
Serotype Y NN	0.464 ± 0.301	NA
Serotype Y MLR	3.724 ± 2.827	0.06

The cumulative proportion column of **Table 5** shows that the top 12 PCs are responsible for almost 90% of the variability of the target yield variable for serotype X. The standard deviation column indicates how far the variables are dispersed from the principal component vector. As noted in section 3, PCA is used to identify a set of significant EVs. The PCs themselves are not passed to the NN; rather the set of prioritised EVs are identified by PCA. The number of prioritised EVs passed to the NN model is 40 for each of the serotypes. This number was selected as there is a natural drop-off in the correlation of the EVs to the principal components from this point forward. From this pool of 40, the number taken from each PC vector is proportional to its cumulative contribution, as shown in **Table 5**. For example, if a PC contributes 40% to the variability of the target yield variable, then the top 16 EVs constituting that PC are passed to the NN model. **Table 5** shows that PC1 explains 20% of the yield variation so it identifies eight of the 40 EVs, and this is shown in the EV entitlement column.

TABLE 5. PCA for serotype X.

PCA Component	Standard Deviation	Proportion	Cumulative Proportion	EV Entitlement
PC 1	3.492	0.200	0.200	8
PC 2	3.315	0.180	0.380	7
PC 3	2.697	0.119	0.499	5
PC 4	2.273	0.085	0.584	3
PC 5	1.839	0.055	0.639	2
PC 6	1.810	0.054	0.693	2
PC 7	1.653	0.045	0.738	2
PC 8	1.623	0.043	0.781	2
PC 9	1.448	0.034	0.815	1
PC 10	1.300	0.028	0.843	1
PC 11	1.250	0.026	0.868	1
PC 12	1.168	0.022	0.891	1
PC 13	1.093	0.020	0.910	1

TABLE 6. PCA for serotype Y.

PCA Component	Standard Deviation	Proportion	Cumulative Proportion	EV Entitlement
PC 1	5.928	0.302	0.302	12
PC 2	3.965	0.135	0.437	5
PC 3	3.669	0.116	0.553	5
PC 4	2.769	0.066	0.619	3
PC 5	2.555	0.056	0.675	2
PC 6	2.433	0.051	0.726	2
PC 7	2.387	0.049	0.775	2
PC 8	1.990	0.034	0.809	1
PC 9	1.850	0.029	0.838	1
PC 10	1.732	0.026	0.864	1
PC 11	1.674	0.024	0.888	1
PC 12	1.479	0.019	0.907	1

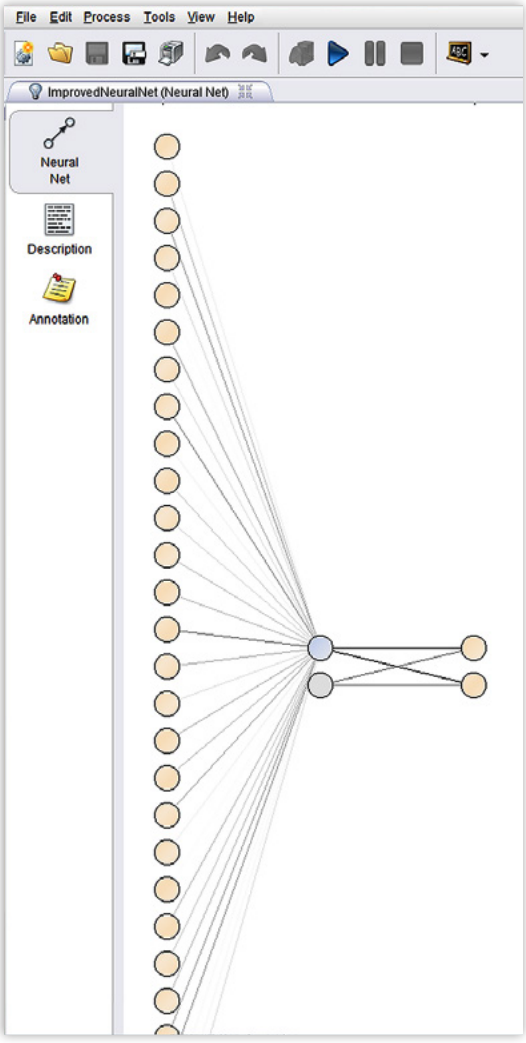

Figure 2. Serotype X NN.

Table 6 shows a similar summary for serotype Y where the EV's contribution to each principal component is based on how much it contributes to yield variation. The top 12 PCs explain 90% of the variability in serotype Y, as shown in **Table 6**. PC1 explains 30.2% of serotype Y yield variation and so identifies 12 of the 40 EVs.

4.2. Serotype X NN Results

The RapidMiner software tool was used to create a NN for serotype X by using the prioritised EVs from the PCA as inputs. An image of the best NN is shown in **Figure 2**. The 40 EVs are seen as inputs on the left feeding into a small number of hidden nodes. The outputs on the right are the yield classification (high or low).

A sample output of the serotype X NN model is shown in **Table 7**. The model performs very well and predicts a serotype X yield with an accuracy of 87.5%. To calculate accuracy, the confusion table shows the true values in the columns, as compared with the predicted values in the rows. For example, in this case the model correctly predicted 16 batches as having a high yield but predicted one batch as having a high yield when in fact it was low. The model has extremely high class precision in predicting high yield (94%).

The precision results show that the NN has a high capability of correctly identifying serotype X high yield batches (94.12%). Precision and recall measure often have an inverse relationship, but we see that the NN also has a high recall value (88.89%). The NN shows good performance in predicting low yield production batches.

TABLE 7. Confusion table for serotype X NN.

Yield	True High	True Low	Class Precision
Predicted high	16	1	94.12%
Predicted low	2	5	71.43%
Class recall	88.89%	83.33%	—

4.3. Serotype Y NN Results

Figure 3 shows the topology of the NN created using the RapidMiner software for serotype Y. The 40 prioritised EVs identified by the PCA are fed in as inputs to produce the yield output classification.

The accuracy for the serotype Y model is given as 66.7% (**Table 8**). This is not as good as the serotype X model, and the reasons for this are explored in section 5. Again, the NN is better able to predict high yield production runs than correctly predicting low yield production runs for serotype Y. Unbalanced data sets are a common problem when performing DM on manufacturing data.[27] Failure rates tend to be so low that the data are unbalanced with only a low percentage of failed items. It is difficult for the ML technique to distinguish failures, as they occur so infrequently.

4.4. MLR Results

The incumbent MLR method is used to identify EVs with high correlation to yield in a univariate manner. These EVs are then investigated under the statistical process control framework. However, as noted in section 4.1, the MLR model results in a poor fit to the data, and direct comparison to the multivariate NN results is not possible. Whereas, the NN approach allows a combination of parameters to be identified and adjusted in unison to improve yield.

4.5. SME Interpretation of Results

Throughout the CRISP-DM process, it was necessary to work closely with the domain expert to validate the modelling process. At each stage of the process, the model outputs were reviewed by the domain expert to make sure that redundant variables that could not affect the yield of the batch were removed. In addition, batches that were not representative, due to process changes, were excluded from the data set.

Perhaps the most crucial considerations, while at the same time being the most nebulous, were "effects" that were removed from the data set at the domain expert's recommendation. Correlation is not to be confused with causality, and the domain process expert reviewed the data to remove what were perceived as effects rather than causes of yield fluctuations. With the high number of variables, there was a chance that some of these effects were missed and used as input data for the model. This would mean that they would be correlated with yield and be prioritised as significant during data pre-processing.

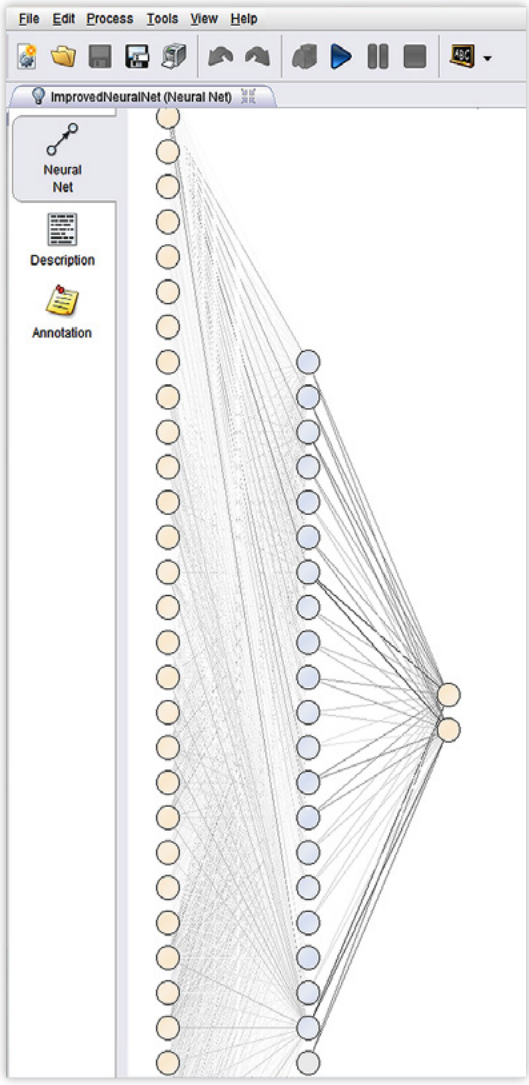

Figure 3. Serotype Y NN topography.

5. Recommendations and Discussion

5.1. Challenges Identified During the Study

A number of challenges remain in applying an analytics model to a complex manufacturing system such as conjugating vaccines. It is true to say that there is a wealth of data accumulated from modern-day manufacturing, but it is also true to say that it is not stored with ease of access or extraction of value in mind. A large amount of time for this project was spent gathering data from disparate locations and converting them to a usable format for the RapidMiner program. There are significant challenges to aggregating and cleaning data from several different sources. The number of batches available and eligible for this study was smaller than the ideal (<25), and was a result of changes in the ongoing production process during the study period. Changes to the process were, for example, procuring a raw material from a different vendor. Because the components of the system are biological, these changes may not just affect the subsystem they are applied to, but could also have unforeseen effects on consequences further downstream. Changes

TABLE 8. Confusion table for serotype Y NN.

Yield	True High	True Low	Class Precision
Predicted high	12	4	75.00%
Predicted low	3	2	40.00%
Class recall	80.00%	33.33%	—

in the production configuration of the system and subsystems constantly occur over time, in contrast to traditional DOE-controlled settings. The control charts used in the 6σ approach identified some batches that were excluded from the study, as they were deemed to be too different due to incremental process improvement changes.

The approach to the data in this paper aligns with the 6σ philosophy of continuous process improvement. As changes to the process occur in small incremental steps, sufficient data is available for the NN approach on a rolling basis in line with incremental process improvement changes, but care needs to be taken when identifying batch data that is representative of the process as it currently stands.

Although an expansive data set was gathered, it was not exhaustive. The problem of a lurking or hidden variable is ever-present, such as one that is significant to yield but has not been analysed. It is important to recognise that there are limits to what we can capture and explain due to the sheer number of possible permutations. In the words of George E. P. Box: "All models are wrong, but some are useful."

With so much data produced by a modern manufacturing process, analytics has a distinct advantage in that it is exploratory rather than ruling in or out a particular hypothesis. This is a very important quality when analysing manufacturing data, as investigators may not always know what they are looking for. In the case of this study, the results were a significant improvement on the incumbent 6σ method, and the NN approach has been adopted on a trial basis by company Z for other serotypes.

Currently, statistics are viewed as the domain of experts, but analytics has the potential to be a more widely accessible toolkit because of the availability of DM tools with graphical user interfaces (GUIs). Importantly, coaching on the statistical significance of results, and a grounding in the limitations of the models, are a prerequisite for the appropriate application of analytics.

With the advent of electronic batch records and manufacturing execution systems, the raw materials required for the application of analytics are readily available. There is an abundance of real-time, shop-floor data. However, the skill sets using analytics to translate this into knowledge are scarce. It is an opportune time to start combining the two most valuable resources a manufacturing company has—its data and its people. The challenge is to invest in the systems and skill sets that will allow companies to optimise their use of existing process information. The first step is the commissioning of a dedicated analytics server which combines all the disparate pockets of data into a format that is easily and quickly analysed by a DM package. The true power of these techniques lies in their accessibility, with an ideal scenario being that the domain expert becomes proficient in the use of these tools.

With very little outlay, analytic techniques have the potential to significantly increase profit margins—particularly in the fragile vaccine manufacturing domain.[3] The success of this project has led company Z to extend the methods to the remaining vaccine serotypes that make up the product. But this is a secondary consideration compared with the effect these vaccine products have on the patients who receive them. The vaccine that is the subject of this study is predicted to save 1.5 million lives by 2020.

While promising, the NN model results have limitations. NNs are a heuristic technique, so the results are empirical evidence only. Much of this document describes measures taken in order to secure management buy-in, but measures must also be taken to manage management expectations. This NN classifies production settings that produce a high and low yield. It is important that management understands the model outputs and limitations of the NN and ML approaches.

The analytics era is in its infancy, from a manufacturing standpoint, but the practice of advanced analytics is grounded in years of mathematical research with successful applications in the equally volatile and complex banking and finance industries. While these powerful tools are easy to use, a good understanding of their statistical foundations is crucial to the valid interpretation of results, and to ensure that assumptions underlying the statistical techniques are not violated. This is why the company-wide initiative, and the use of 6σ at all levels of the company, should provide a fertile ground for making the case for DM and facilitating its acceptance. The 6σ mindset of measuring process performance and analysing data promotes data-based decision-making, and therefore makes DM a natural extension of this methodology.

Acknowledgements

We would like to thank Gareth Thornton for his help in proofreading this document. We would also like to thank the staff of the Smurfit Business School for sharing their expertise and insights in the course of this work. Thanks also to Dermot O'Malley and Paraic Fahey for their endless patience and help. We are thankful to Tony Walsh for taking a keen interest in the methodology, and for sponsoring the project. Lastly, many thanks to Eamonn Nixon for his flexibility and support.

References

[1] Rudan I, Tomaskovic L, Boschi-Pinto C, Campbell H. Global estimate of the incidence of clinical pneumonia among children under five years of age. *Bull World Health Organ*, 2004; 82: 895–903.

[2] Proano RA, Jacobson SH, Zhang W. Making combination vaccines more accessible to low-income countries: the antigen bundle pricing problem. *Omega*, 2012; 40(1): 53–64. http://dx.doi.org/10.1016/j.omega.2011.03.006.

[3] Robbins MJ, Jacobson SH. Pediatric vaccine procurement policy: the monopsonist's problem. *Omega*, 2011; 39(6): 589–597. http://dx.doi.org/10.1016/j.omega.2010.12.004.

[4] Black S, Shinefield H, Fireman B, Lewis E, Ray P, Hansen JR *et. al*. Efficacy, safety and immunogenicity of heptavalent pneumococcal conjugate vaccine in children. *Pediatr Infect Dis J*, 2000; 19(3): 187–195. http://dx.doi.org/10.1097/00006454-200003000-00003.

[5] Gambillara V. The conception and production of conjugate vaccines using recombinant DNA technology. *BioPharm Int*, 2012; 25(1): 28–32.

[6] Raisinghani MS, Ette H, Pierce R, Cannon G, Daripaly P. Six Sigma: concepts, tools, and applications. *Ind Manage Data Syst*, 2005; 105(4): 491–505. http://dx.doi.org/10.1108/02635570510592389.

[7] Köksal G, Batmaz I, Testik MC. A review of data mining applications for quality improvement in manufacturing industry. *Expert Syst Appl*, 2011; 38(10): 13448–13467. http://dx.doi.org/10.1016/j.eswa.2011.04.063.

[8] Chakravorty SS. Six Sigma failures: an escalation model. *Oper Manage Res*, 2009; 2(1): 44–55. http://dx.doi.org/10.1007/s12063-009-0020-8.

[9] FDA. Guidance for Industry: PAT — a framework for innovative pharmaceutical development, manufacturing, and quality assurance. Sept 2004. http://www.fda.gov/downloads/Drugs/Guidances/ucm070305.pdf.

[10] Molcho G, Zipori Y, Schneor R, Rosen O, Goldstein D, Shpitalni M. Computer aided manufacturability analysis: closing the knowledge gap between the designer and the manufacturer. *CIRP Annals — Manuf Technol*, 2008; 57(1): 153–158. http://dx.doi.org/10.1016/j.cirp.2008.03.046.

[11] Thomassen YE, van Sprang ENM, van der Pol LA, Bakker WAM. Multivariate data analysis on historical IPV production data for better process understanding and future improvements. *Biotechnol Bioeng*, 2010; 107(1): 96–104. http://dx.doi.org/10.1002/bit.22788. PMid:20506395.

[12] Rokach L. Decomposition methodology for classification tasks: a meta decomposer framework. *Pattern Anal Appl*, 2006; 9(2): 257–271. http://dx.doi.org/10.1007/s10044-006-0041-y.

[13] Arteaga F, Ferrer A. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *J Chemometr*, 2002; 16 (8–10): 408–418. http://dx.doi.org/10.1002/cem.750.

[14] Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *J Artif Intel Med*, 2010; 50(2): 105–15. http://dx.doi.org/10.1016/j.artmed.2010.05.002.

[15] Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H. Top 10 algorithms in data mining. *Knowl Inform Syst*, 2008; 14(1): 1–37. http://dx.doi.org/10.1007/s10115-007-0114-2.

[16] Büchner AG, Mulvenna MD. Discovering internet marketing intelligence through online analytical web usage mining. *ACM Sigmod Record*, 1998; 27(4): 54–61. http://dx.doi.org/10.1145/306101.306124.

[17] Wirth R, Hipp J. *CRISP-DM: towards a standard process model for data mining*. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000; pp 29–39.

[18] Hickey C, Kelly S, Carroll P, O'Connor J. Prediction of forestry planned end products using Dirichlet regression and neural networks. *Forest Sci*, 2015; 61(2): 289–297. http://dx.doi.org/10.5849/forsci.14-023.

[19] Chien C-F, Wang W-C, Cheng J-C. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Syst Appl*, 2007; 33(1): 192–198. http://dx.doi.org/10.1016/j.eswa.2006.04.014.

[20] Tetko IV, J Livingstone DJ, Luik AI. Neural network studies. 1. Comparison of overfitting and overtraining. *J Chem Inform Comput Sci*, 1995; 35(5): 826–33. http://dx.doi.org/10.1021/ci00027a006.

[21] Feng C-XJ, Wang XF. Data mining techniques applied to predictive modeling of the knurling process. *IIE Trans*, 2004; 36(3): 253–263. http://dx.doi.org/10.1080/07408170490274214.

[22] Zobel CW, Cook DF. Evaluation of neural network variable influence measures for process control. *Eng Appl Artif Intel*, 2011; 24(5): 803–812. http://dx.doi.org/10.1016/j.engappai.2011.03.001.

[23] Wang X, Smith K, Hyndman R. Characteristic-based clustering for time series data. *Data Min Knowl Disc*, 2006; 13(3): 335–364. http://dx.doi.org/10.1007/s10618-005-0039-x.

[24] Bickel PJ, Lehmann EL. Descriptive statistics for nonparametric models. *Ann Stat*, 1975; 3(5): 1038–1044 (I. Introduction), http://dx.doi.org/10.1214/aos/1176343239; and 1045–1069 (II. Location), http://dx.doi.org/10.1214/aos/1176343240.

[25] Wold S, Kim E, Paul G. *Principal component analysis, chemometrics and intelligent laboratory systems*. In: Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, 1987; 2(1–3) pp 37–52.

[26] Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometr Intel Lab*, 1995; 28(1): 3–21. http://dx.doi.org/10.1016/0169-7439(95)80036-9.

[27] Provost F. *Machine learning from imbalanced data sets 101*. In: Proceedings of the AAAI workshop on imbalanced data sets, 2000; pp 1–3.

About the Authors

Will Fahey and **Paula Carroll***

Centre for Business Analytics, Smurfit School of Business, University College Dublin (UCD), Ireland

Dr. Carroll is the corresponding author:

Phone: 00 353 1 716 4776 | Email: paula.carroll@ucd.ie | Website: www.ucd.ie

Will is a Lean Six Sigma Black Belt with ten years of experience in biopharmaceutical manufacturing. He recently completed a MSc in Business Analytics at the Smurfit School of Business, UCD, Ireland. He is interested in the application of machine learning techniques to better understand complex manufacturing processes.

Paula is a lecturer at UCD. She is a member of the UCD Centre for Business Analytics and the UCD Electricity Research Centre. Her research interests are in the application of business analytics and operations research optimisation techniques to solve real world business problems.