Quality prediction through machine learning for the inspection and manufacturing process of blood glucose test strips

Ching-Shih Tsou\textsuperscript{1,2,*}, Christine Liou\textsuperscript{2}, Longsheng Cheng\textsuperscript{3} and Hanting Zhou\textsuperscript{3}

**Abstract:** Although machine learning for quality prediction of manufacturing processes has attracted attention in the literature, there is a significant lack of case studies from industry, especially in medical sector. This paper proposes a data-driven approach to infer the batch quality of blood glucose test strips. Once the low quality of work in process is detected, unnecessary process waste could be eliminated. Starting from data pre-processing, which consists of Synthetic Minority Over-sampling TEnchnique and Random Over-Sampling Example, this project tries to balance the ill-distributed data first. Followed by machine learning aims to classify and predict the quality of blood glucose test strips. Different models are evaluated by the Receiver Operating Characteristic curve and Area Under Curve. Computational results show that the decision tree and random forest after SMOTE perform better than the counterpart of ROSE method. Ensemble learning, such as random forest, out-wins base learner decision tree. To sum up, random forest with SMOTE is the suggested model for accurately predicting the quality of blood glucose test strips. There is a 30\% improvement in error rate under random forest and SMOTE for NG class that could be of top concern for prognosis. Several factors, including the direction of applying test reagent onto test strips and the position where the strips are located, that affect quality of test strips have been identified. Explanations in terms of inspection and manufacturing are discussed subsequently. Finally, the prognosis of quality can be attained through big data and statistical machine learning.

**Subjects:** Artificial Intelligence; Industrial Engineering & Manufacturing; Machine Learning

**Keywords:** quality prediction; imbalanced data; over-sampling; decision trees; ensemble learning

1. Introduction

About 442 million people worldwide have diabetes according to World Health Organization (WHO). The WHO also estimates that diabetes is the ninth leading cause of death. Diabetes is a metabolic disease characterized by hyperglycaemia. It is broadly classified into three types by etiology and clinical presentation, including type 1, type 2, and gestational diabetes. Some other less common types of diabetes are monogenic diabetes and secondary diabetes (Goyal & Jialal, 2021). The type 1 and type 2 diabetes can lead to serious health problems that affect many organs including the nerve, the eye, and the kidney. These health problems can be disabling or even life-threatening, and to date, there are no effective therapies that can slow or prevent their development (Eid et al., 2020).
However, there is no curable treatment for diabetes till now. Self-Monitoring Blood Glucose (SMBG) is a common and important tool for health keeping of patients with diabetes. SMBG is necessary in all patients taking insulin, with increased frequency of monitoring recommended for patients taking meal-time insulin. Glycemic control targets included fasting and post-prandial glucose as determined by SMBG (McGill & Ahmann, 2017). Intensive treatment of type 1 diabetes is monitored at least four times a day at fasting state, and eight times when blood sugar is unstable including fasting and post-prandial state (Ludvigsson & Hanas, 2003).

The method of SMBG has been used in clinical practice for nearly 70 years. Blood glucose test strips are the first semi-quantitative tests to monitor blood glucose, which is put on the market in the late seventies and has been used up to now (Dufaitre-Patouraux et al., 2003). Accuracy and precision concern are in the heart of considerations at the beginning of self-blood glucose monitoring. Clarke et al. (1987) develop an Error Grid (EG) analysis, which describes clinical accuracy of SMBG systems. Taken into account the values of system-generated glucose and those of reference blood glucose, EG analysis draws the relation of these two values on two-dimensional space. Parkes et al. (2000) construct new EG to evaluate the accuracy of SMBG, which also consists of five ranks of categories and the accuracy of SMBG rated 98.6% of their measurements compared with 95% for traditional EG. Ginsberg (2009) assumes that accuracy may be limited due to strip manufacturing variances, strip storage, aging, as well as environments such as temperature or altitude. Patient factors such as improper coding, incorrect hand washing, altered haematocrit, or naturally occurring interfering substances could also be the reasons led to inaccuracy (Ginsberg, 2009).

Initially, the use of blood glucose test strips is intended for medical staff. With the handiness improvement, test strips become much more adequate for patients and are now a necessary tool for SBGM (Dufaitre-Patouraux et al., 2003). However, studies mentioned above focus on the accuracy of the strips rather than quality control of manufacturing process, which has been rarely studied. Assessing the quality of medical equipment or materials production by data-driven methods is also not emphasized enough in the past. Kuo and Kusiak (2019) reviewed data-driven approach for production research and proposed future trends. They mentioned that thanks to the versatile open-source software, data analytics in production research has been rapidly increasing after 2005. Production research papers associated with the applications of data mining have higher citations because of the overwhelming Big Data. This indicates that combining Big Data Analytics with quality information systems has been a popular and valuable topic nowadays.

Quality improvement of industrial products and processes require collection and analysis of data to solve quality-related manufacturing problems. Related tasks are product/process quality issues description, quality level prediction, quality grade classification, and quality parameter optimization (Köksal et al., 2011). Machine learning recently has been employed in various domains to develop intelligent systems. Generally speaking, scholars use statistics or machine-learning algorithms to construct data-driven models, such as principal component regression (PCR), support vector machine (SVM) and so on. These models help to monitor the process of product manufacturing and to predict the quality of product (Yao & Ge, 2018). Malhotra and Jain (2012) apply one statistical method and six machine learning methods (random forest, adaboost, BAGGING, multilayer perceptron, support vector machine and genetic programming) to predict the quality of software. Comparisons are made in terms of the Area Under the Curve (AUC) obtained from the Receiver Operating Characteristics (ROC) curve. Their results show that machine learning approach has a competitive performance with statistical methods. Among the machine-learning methods, random forest and BAGGING methods outperformed the other models.

Our study focuses on the inspection data from the manufacturing process of blood glucose test strips. The aim is to propose a data-driven approach to estimate the quality of the strips in process.
Once an alarm of low quality is issued, unnecessary process inputs could be eliminated. Various predictive models are developed based on in-process inspection data from a manufacturer in Taiwan. As expected, class distribution is ill-balanced for the strips under inspection. The efficacy of classifiers will be jeopardized by the skewness among different classes. Moreover, evaluation of the imbalanced learning is a difficult but critical part of predictive modelling process.

Learning from an imbalanced dataset arises in many practical applications, especially in quality inspection, biomedical diagnostic tests, etc. (Huda et al., 2016; Mazurowski et al., 2008; Seiffert et al., 2014). One of the popular approaches for imbalanced learning is the Synthetic Minority Oversampling Technique (SMOTE) since its publication in 2002. SMOTE has been considered as the “de facto” standard of learning from imbalanced data (Fernández et al., 2018). Moreover, Menardi and Torelli (2014) proposed a unified and systematic framework for dealing with the imbalanced classification problem. The framework is based on a smoothed bootstrap re-sampling technique and was named as the Random Over Sampling Examples (ROSE). Empirical studies showed that it outperforms the other remedies for imbalanced learning problems. Generally speaking, to increase the learning performance some data pre-processing, such as class distribution balancing, feature extraction and data reduction, are imperative. Therefore, our approach was divided into two parts. The first step is to use SMOTE and ROSE to balance the inspection data. Next, quality estimation models are built and evaluated for gaining insights from the manufacturing and inspection processes.

To sum up, the aim of this paper is to develop pre-processing and predictive modelling for inferencing the quality of blood glucose test strips. Insights elicited from the modelling results could highlight some operations issues, such as the unevenness of the carbon plated electrodes. If the quality of medical material is reliable, then they can help to accurately and precisely detect the pathological condition of patients.

The rest of this paper is organized as follows. Section 2 presents a brief review of the related literature, including manufacturing process, imbalanced learning and classification models. To learn more about quality control of high-tech productions, research on quality improvement of the wafer or electronic products are reviewed in this section. Section 3 focuses on how to employ the In-Process Quality Control (IPQC) data to estimate the quality of strips. Computation results are subsequently expounded in Section 4. Conclusions and recommendations for future research are presented finally.

2. Literature review

2.1. Manufacturing process

Gomes et al. (2010) mention that SMBG has the potential to play an important role in control and management of diabetes. Manufacturing of test strips is a complex process involving various steps. It is unreasonable to assume that all test strips in use will provide identical or similar measurement results, even they are produced by a same manufacturer (Erbach et al., 2016). Standards for blood glucose measurement, such as accuracy, are specified in the internationally accepted standard EN ISO 15197 (Freckmann, Schmidl et al., 2015). According to the authorization of European and American regulators, three different test strip lots must be included in the future for accurate evaluation of blood glucose system under the requirements of ISO 15197:2013 (Freckmann, Link et al., 2015).

The test strips of blood glucose are stored in a vial for future use. The variation of glucose monitoring systems comes from four sources: strip factors, physical factors, patient factors, pharmacological factors. Due to the manufacturing process, strip-to-strip and vial-to-vial variation may occur. For example, small changes in enzyme coverage may affect the accuracy of blood glucose system (Ginsberg, 2009). Research on manufacturing of blood glucose test strips or portable meters are not fruitful. However, the meticulous processing of some semiconductors or
electronic products is similar to blood glucose test strips. In particular, the manufacturing processes of both wafer and blood glucose strips are to cut circuits piece in piece after etching a whole circuit board (Lin et al., 2004). As a result, some researches on related manufacturing domains are reviewed subsequently to understand how the quality of the products is analysed.

Blue and Chen (2010) developed a novel Spatial Variance Spectrum (SVS) to analyse the systematic variations over the surface of wafers. The SVS is a series of spatial variations over a range of spatial moving window sizes from the smallest one consisting of only two metrology sites to the largest one covering all sites of the entire wafer. The SVS is used to characterize the wafer spatial variations and to detect possible systematic anomaly. Our approach to estimate the strip-to-strip variation on a panel is similar to theirs. Variations are calculated by the difference percentage of each strip with respect to corresponding row average (Equation (1)).

Choi et al. (2017) build decision trees to analyze the causes of defective automobile electronic parts. The data used in their study including information from the Manufacturing Execution System (MES), the Point of Production (POP), equipment sensor data, in-process/external air-conditioning sensors and static electricity. Similarly, data in this paper are collected during in-process quality inspection on test strips. They are utilized to estimate the quality of Work In Process (WIP) batch and to find possible causes of the defects observed.

2.2. Imbalanced learning
Building a useful model is not an easy task in case of ill-distributed classes. A naive classifier tends to classify all examples as the majority class, usually the non-defective class, because the defective samples are usually rare. Accuracy of such kind of model would be quite high, but its predictions are useless. To mitigate this issue, the class distribution should be rectified before building a model.

Haixiang et al. (2017) reviewed 527 papers involving imbalanced learning and summarized existing methods by a taxonomy. Resampling is usually the first one to the rescue. It is further divided into three categories, including over-sampling, under-sampling, and hybrid resampling. Over-sampling artificially creates new minority class samples to mitigate the skewness. Under-sampling, on the other side, discards samples in the majority class to reach a more balanced dataset. One weakness of under-sampling is that it is possible to lose important information within the majority class. Haixiang et al. (2017) mentioned that a large datasets would be a preferred condition for under-sampling, whereas hybrid resampling methods compromise the above twos in some ways.

In addition to resampling, feature selection and extraction might help us to learn a descent model. The former one tries to select a subset of \( k \) features (\( k \leq m \), \( m \) is the number of features and is usually large) from entire variable set to increase the performance of classifiers. The latter one composes original features into surrogate ones before building models (Chawla, 2009). The inspection data in this study is not a high-dimensional one. And the models employed here are quite robust against noisy or irrelevant features. Therefore, feature selection and extraction will not be covered in the following sections.

Lin and Chen (2013) discussed the imbalanced learning for high-dimensional data. Strategies of mitigating class-imbalanced effects are categorized into the data-based and algorithm-based approach. Data-based approach, also known as resampling, rectifies the imbalanced dataset before submitting to any algorithm. Algorithm-based approach leaves the dataset intact and adjusts classification algorithm, such as neural network, decision tree, k-nearest neighbours, random forest, and so on. However, it is more difficult for an algorithm to identify rare patterns than to identify common patterns. Hence, over-sampling approach is more suitable than others,
especially the challenge of imbalanced learning is absolute rarity instead of relative rarity (He & Ma, 2013). Only 3.7% of blood glucose test strips are defective in our dataset.

Chawla et al. (2002) developed the famous Synthetic Minority Over-sampling TEst Technique (SMOTE). It generates new examples of the minority class using its nearest neighbours. By taking each minority sample into consideration, synthetic examples are produced along the line segments connecting to all k nearest neighbours. On the other side, majority examples are also down-sampled to have a more balanced dataset. Compared with over-sampling with replacement, the “synthetic” samples of SMOTE had better performance under the Receiver Operating Characteristic (ROC) curve (Chawla et al., 2002).

Random Over-Sampling Example (ROSE) also produces a synthetic data, but the sampling is directed by a smoothed-bootstrap approach. It enhanced the efficiency and evaluation during the estimation of a binary classifier in the presence of a rare class (Lunardon et al., 2014). New examples are drawn from an estimated conditional kernel density of the two classes (Menardi & Torelli, 2014). The R packages DMwR and ROSE both provide functions to deal with imbalanced data (Lunardon et al., 2014; Torgo, 2017).

2.3. Classification models

Machine learning is a viable tool to develop intelligent quality systems for a variety of manufacturing processes. It encompasses models of pre-processing, feature extraction and selection, dimensionality reduction, clustering, regression and classification. For our WIP quality prediction problem, classification algorithms are chosen to estimate the quality grading of the test strips, in order not to waste resource in subsequent manufacturing process.

Common classification models include logistic regression, linear discriminant analysis, decision tree algorithms (ID3, C4.5, CART), k-Nearest Neighbours classifier (KNN), Naive Bayesian model (NB), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) (Nikam, 2015). Furthermore, ensemble of learning machines have been shown to outperform single classifiers in many cases by uniting a variety of base learners to tackle difficult problems (Chandra & Yao, 2006). Bootstrap AGGregatING (BAGGING) generates an ensemble of independent classifiers by resampling the training data given. However, boosting combines correlated and simple classifiers to out win a clever and complex model. Dietterich (2000) compares the effectiveness of randomization, BAGGING and boosting for improving the performance of C4.5 algorithm. The result shows that BAGGING is much better than boosting, and sometimes better than randomization.

Random forest is another famous ensemble learning algorithm. Based on BAGGING integration, random forest goes one step further toward building more independent trees by introducing random feature set in the training process (Breiman, 2001). Naghibi et al. (2016) use Boosted Regression Trees (BRT), Classification And Regression Tree (CART), and Random Forest (RF) to produce groundwater spring potential maps. Predicted results from these models are validated using the ROC curve. Kocev et al. (2007) mentioned that ensemble methods are able to improve the predictive performance of many base classifiers, yet the learning results are mostly uninterpretable.

3. Problem and solution

3.1. Production, sampling and calculation

As mentioned before, the manufacturing of blood glucose test strips is similar to that of wafers. Intermediate products are printed and plated from a single piece of panel. They are cut into small slices and ready for filling in a vial. One panel full of blood glucose strips is shown in Figure 1. Each panel is further cut into three hundred slices, comprising six rows and fifty blood glucose test strips
each row. After several testing, including carbon paste electrode test, IPQC, and final quality test, the strips are filled into a vial ready for shipping. All testing are executed under a stringent environment control including temperature and humidity to ensure a repeatable and accurate results.

IPQC is our focus here, which was done on four panels randomly selected from production line to estimate the batch quality. Because it is a destructive test, not all 1,200 strips in four panels are applied by test reagent. The row sampling plan for selected panels is shown in Table 1. Despite there are 50 strips in each row, ten strips are randomly tested row by row. Test liquid is made from blood or Glucose Control Solution (GCS) under high (hyperglycaemia, Level II) or low level (hypoglycaemia, Level I).

The range of test values depends on if the reagent is hypoglycaemia or hyperglycaemia. Stringent manufacturing practices can stabilize the variation of blood glucose readings from test strips under high and low levels. Variations are estimated by the difference in percentage of each value with respect to the corresponding row average.

\[
\text{ave}_{\text{diffpc}} = \frac{X_{RC} - \bar{X}_R}{\bar{X}_R} \times 100\%, \text{ for } R = 1, \ldots, 6, C = 1, \ldots, 10.
\] (1)

| Table 1. Row sampling plan for selected panels |
|-----------------------------------------------|
| Row 1  | Panel 1 | Panel 2 | Panel 3 | Panel 4 |
|--------|---------|---------|---------|---------|
| Row 2  |         |         |         |         |
| Row 3  | GCS     | Hyperglycaemia | Blood  | Hyperglycaemia |
| Row 4  | Hypoglycaemia |       |         | Blood  |
| Row 5  |         |         |         | Hyperglycaemia |
| Row 6  |         |         |         |         |

The slashes indicate that the area had not been tested.
where $X_{RC}$ is the blood glucose reading from strip C in row $R$, and $\bar{X}_R$ is the average value for each row. The percentage calculation in Equation (1) scales well for Level I and Level II blood glucose under different order of magnitude.

### 3.2. Pre-processing, balancing and modelling

Suppose there are 27 batches of blood glucose strips on hand. All sampled IPQC data of each batch are recorded on a separate Excel file consisting of several sheets. A lot of time is spent on data pre-processing, including sheets integration, feature extraction, and incomplete data handling. Thousands of lines of R scripts were used to organise data from Excel files. The sample size is totally 4,320 tests on selected panels. Once a quality prediction model has been validated, this could be a precaution system built on big data to help us estimate the quality level of WIP. After aggregating valuable information from separate Excel files, data are ready for exploration and analysis as shown in **Table 2**. Four columns are listed in our IPQC dataset. The first column ("rnd_obs") is the ten tests for each row. Test results, Pass or not pass (NG), of each strip from the same batch are shown in the second column ("output"). The third and fourth variables ("row" and "direction") record the position of the strips which represent the row number and the direction to apply the reagent, respectively. The last column is the $ave_{rdiffpc}$ for each strip computed by equation (1).

| rnd_obs | output | row | direction | $ave_{rdiffpc}$ |
|---------|--------|-----|-----------|-----------------|
| 1       | Pass   | 3   | left      | 1.26582278     |
| 2       | Pass   | 3   | left      | -3.79746835    |
| 3       | Pass   | 3   | left      | -1.26582278    |
| 4       | Pass   | 3   | left      | 1.26582278     |
| 5       | Pass   | 3   | left      | 1.26582278     |
| 6       | Pass   | 3   | right     | 1.26582278     |
| 7       | Pass   | 3   | right     | -3.79746835    |
| 8       | Pass   | 3   | right     | 3.79746835     |
| 9       | Pass   | 3   | right     | -1.26582278    |
| 10      | Pass   | 3   | right     | 1.26582278     |
| 1       | NG     | 6   | left      | -2.5913206     |
| 2       | NG     | 6   | left      | 3.3406182      |
| 3       | NG     | 6   | left      | 3.3406182      |
| 4       | NG     | 6   | left      | -1.6546987     |
| 5       | NG     | 6   | left      | -5.7133937     |
| 6       | NG     | 6   | right     | 0.8429597      |
| 7       | NG     | 6   | right     | 3.0284109      |
| 8       | NG     | 6   | right     | 2.7162036      |
| 9       | NG     | 6   | right     | -0.4058695     |
| 10      | NG     | 6   | right     | -2.9035279     |

The criteria to determine if the batch can pass or not (Pass or NG) depend on several conditions. The first one is the number of Standard Deviations (SD) or Coefficients of Variation (CV) calculated from each row cannot exceed certain limit. Note that CV (%) is used for blood glucose concentration greater or equal to 100 mg/dl and SD (mg/dl) is employed for concentration less than 100 mg/
Next, the number of out-of-range strips under hypoglycaemia and hyperglycaemia should not be over a pre-specified limit. Finally, the relative bias to some reference results (possible the best record ever in the past) should be less than a certain percentage. All criteria above are already used in the existing IPQC process. Under these criteria an educated estimate of the quality level on each batch of test strips can be reckoned in order to decide whether let the batch go on for subsequent processing.

The aim of IPQC is to infer the batch quality from the variation of readings of blood glucose test strips. The lower the variation of the testing result is, the more stable the quality of the batch is. However, the amount of test strips shipped outbound is decreased because IPQC test is destructive. If a model is capable to reasonably infer the batch quality, alerts on work in process (WIP) of bad quality will be issued to stop wasting time and resources of further processing. Heatmaps in Figure 2 summarize the ave_rdiffpc by Row and rnd.obs across Pass and NG samples. Divergent colours are used in this plot where positive is coloured from white to purple and negative is from white to red. Generally speaking, the NG one is more colourful than the Pass one. The lower part are more positive than the upper part.

Because the difference between the sample size of Pass and NG is huge, SOMTE and ROSE are employed to balance the dataset first. The former up-samples the minority data by linear interpolation instead of duplicating the same one. The latter generates synthetic minority by approximating its probability distribution through bootstrapping, although the approximation error could be large.

After balancing IPQC data, decision trees and random forest classifiers are built for interpretation and prediction, respectively (Liaw & Wiener, 2002). Different models are evaluated by the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC). The general framework of this study are summarized in Figure 3.
4. Results and analysis
This paper intends to propose a prognostic model of quality degradation. Factors that potentially affect the quality of test strips are identified. The inspection and manufacturing processes could be improved accordingly.

4.1. Data balancing
Among 27 batches of IPQC data one has been classified as defective and finally been scrapped. Therefore, there are 4160 Pass samples and only one batch is NG data. The number of negative samples (Pass) is much greater than that of positive samples (NG, or not pass). The prevalence of positive example is 3.7%. Any classifier trained could be biased towards the majority class Pass. Thus accuracy on the minority class NG is poor which is our focus here. Data sampling is a commonly used method to mitigate the issues of class imbalance described above (Seiffert et al., 2008).

SMOTE artificially generate artifact samples of the minority class by linearly interpolating between the randomly selected NG and its nearest neighbours. Function SMOTE of R package DMwR is used to get a near-balanced dataset (Torgo, 2017). The first two columns shown in Table 3 is basically balanced under two balancing methods. Function ROSE from package ROSE in R produces a more balanced samples than SMOTE by smoothed-bootstrap resampling approach mentioned in Section 2.2 (Lunardon et al., 2014). However, the resampling results from ROSE could

---

Figure 3. General framework of the proposed solution.
be more volatile because it involves three times random sampling. Table 3 also summarizes data splitting results under both methods.

To visualize the data distribution before and after balancing, aggregating medians and Inter-Quartile Ranges (IQRs) against six rows and ten observations are shown in Figures 4–6. Among 120 points one half is Pass and the other half is NG. The Pass data are relatively concentrated in a confined area and NG data are scattered around with greater scope in terms of medians and IQRs in Figure 4. The variations of Pass samples after balancing spread much wider than those before balancing. Take a closer look at SMOTE and ROSE in Figures 5 and 6, the medians and IQRs under SMOTE are smaller than those under ROSE. This phenomenon is probably due to the linear interpolation used by SMOTE to synthesize new minority data only. While the ROSE generates new minority and majority samples through the probability density function estimated by resampling.

### 4.2. Decision trees
Classification trees are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost (Loh, 2011). The *rpart* package in R is used to build the decision tree (Therneau & Atkinson, 2018). Models are trained

|       | Pass | Not pass | Total | Training | Test |
|-------|------|----------|-------|----------|------|
| SMOTE | 2640 | 1920     | 4560  | 4104     | 456  |
| ROSE  | 2151 | 2002     | 4153  | 3737     | 416  |
under cross validation as shown in Table 4. The last three columns are used to find the best complexity parameter (CP) of post pruning for avoiding overfitting. For each row the mean (Xerror) and standard deviation (Xstd) of error rate under cross (X) validation corresponds to a specific CP. As the tree size (Nsplit, number of split) increases, the Xerror and Xstd decrease too due to a more complex tree. Furthermore, the error rate relative to the root error rate (Rel error) also decreases as the tree grows. The best CP is 0.01 whose corresponding number of splits is 3 as shown in Table 4.

Figure 5. Scatter plot of IQR and median after balancing by SMOTE.
Table 4. Error rates under different complex parameters

| CP       | Nsplit | Rel error | Xerror | Xstd  |
|----------|--------|-----------|--------|-------|
| 0.2658593| 0      | 1.000000  | 1.000000 | 0.0182493 |
| 0.0158593| 1      | 0.7341407 | 0.7341407 | 0.0170896  |
| 0.0100000| 3      | 0.7024221 | 0.7081892 | 0.0169177  |

Figure 7. Decision tree built by the balanced dataset from SMOTE.

Figures 7 and 8 illustrate binary decision trees after data balanced by the SMOTE and ROSE, respectively. Label on the top of each node is the majority class of the corresponding subset. Decimal value on the center of each node is the percentage of NG samples within the subset. The proportion of positive observations is represented as a percentage at the bottom.

From Table 3, the tree in Figure 7 starts growing with 4104 training samples. NG data in the root node take 42.3% (0.423, 1736 samples). On the right-most leave node, 71.5% of row 3 and mostly left-soaked (i.e., 2-left, 3-left, 4-left, 5-left, 6-left) training samples are NG. To the left-most terminal node, rows 1, 2, 4, 5, and 6 training samples of which 71.4% (i.e., 1–0.286) are negative Pass class. In the middle part among row 3 samples, there are 12.9% right-soaked (1-left, 7-right, 8-right, 9-right, 10-right). These mostly right-soaked samples are further split into two leave nodes according to whether the average row difference percentage (ave_rdiffpc) is greater than or equal to –1.81. There are 58.3% (1–0.417) Pass samples among the 9.2% samples when ave_rdiffpc is greater than or equal to –1.81, whereas over half of the samples with ave_rdiffpc less than –1.81 are NG products (0.616 or 61.6%).

Figure 7 reveals that the row where the test strips locate shows differences in variation. Applying direction of test reagent was another significant factor for the quality prediction of test strips. Over
70% of the third-row data with left-soaked direction are classified as failed. It is concluded that the position of test strips and liquid applying direction would affect the test results.

The remaining 12.9% of test strips, which belong to the third row and are tested from the right, takes into account of another critical factor ave_rdiffpc. If ave_rdiffpc was greater than or equal to −1.81, the possibility of passing would be nearly 60%. On the contrary, there is a 61.6% proportion that the test strips won’t pass the test.

Figure 8 illustrates the decision tree using ROSE to balance the dataset. The tree starts from the root node with 3737 training samples, among which NG examples take nearly 48.2% (1801 samples). When the ave_rdiffpc is beyond 3.63 or lower than −4.57, percentage of NG product is more than 70% (i.e., 72.1%) or even more than 80% (i.e., 85.0%), respectively. To the left-most terminal node, 85.6% training samples whose ave_rdiffpc are within half-open interval [−4.57, 3.63) and more than half (i.e., 56.7%) are Pass class.

4.3. Random forest
In addition to decision trees, ensemble learning using bootstrap resampling can be utilized to build a forest of uncorrelated classification trees for the balanced datasets. Figure 9 illustrates the tuning results of the number of trees in the random forest. Prediction error decreases as the number of trees increases. No matter what kind of balancing mechanisms employed, forest size around 100 have been confirmed as the best hyperparameter under repeated bootstrapping. Table 5 summarized three kinds of error rate from both mechanisms. It shows that SMOTE could be a better choice in terms of out-of-bag (OOB) and minority (NG) error rates. Especially, there is
a 30% improvement in error rate under SMOTE for NG class that could be of top concern for prognosis.

4.4. Model evaluation
Confusion matrix is usually used to explore the relationship between prediction results and ground truth. There are four possible situations in a confusion matrix shown in Table 6. True positives (TP) are positive examples correctly predicted as positive. False positives (FP) express those negative

**Table 5. Comparison of error rates**

|              | OOB error rate | NG error rate | Pass error rate |
|--------------|----------------|---------------|-----------------|
| SMOTE        | 32%            | 26%           | 38%             |
| ROSE         | 38%            | 56%           | 22%             |

**Table 6. Confusion matrix**

|                | Actual positive | Actual negative |
|----------------|-----------------|-----------------|
| Predicted positive | 210(TP)         | 63(FP)          |
| Predicted negative     | 64(FN)         | 119(TN)         |
examples incorrectly labelled as positives. True negatives (TN) refer to negative examples correctly classified as negatives. False negatives (FN) correspond to positive examples incorrectly considered as negatives. A variety of classifiers metrics can be evaluated from the entries of confusion matrix, such as Sensitivity (True Positive Rate, TPR), Specificity, Precision, and False Positive Rate (FPR, aka. 1—Specificity). FPR and TPR can be used to construct the Receiver Operating Characteristic (ROC) curve. Both coordinates are the FPR and TPR from equations (7) and (8), respectively. The larger the value of TPR and the smaller the value of FPR indicate a better prediction of the model under examination. It is desired that the ROC curve displays closer to the upper left part and, hence, the Area Under Curve (AUC) gets larger.

\[
TPR = \frac{TP}{TP + FN} \tag{2}
\]

\[
FPR = \frac{FP}{FP + TN} \tag{3}
\]

Figures 10 and 11 demonstrate the ROC curves of tree classifiers and random forest under the SMOTE and ROSE, respectively. Apparently SMOTE has a better result in light of TPR against FPR. To sum up, SMOTE plays an important role in data pre-processing and random forest has better prediction than that of decision trees.
5. Conclusion and further work

A quality predictive framework using machine learning for the inspection and manufacturing process of blood glucose test strips has been proposed. In-Process Quality Control (IPQC) data is an important information asset in the manufacturing process of panel and test strips. In order to analyse the IPQC data, different data balancing and classification models are manipulated for quality prediction. Experimental results have shown that predictive models trained by IPQC data predict the quality level with acceptable accuracy.

For imbalanced learning, data balanced by the SMOTE can accomplish a better performance of binary classifier than that by the ROSE. Mechanism behind the SMOTE uses linear interpolation to synthesize new minority sample, while the ROSE generates new sample through estimated probability density function by bootstrap resampling. Although the resampling idea in the ROSE is novel, its estimation for minority class could be unstable because of insufficient evidence. Therefore, the variance of SMOTE result shrinks closer than that of ROSE, which finally influences predictive performance. Moreover, ensemble learning outperforms the base learner decision trees, hence, verifying the superiority of multiple models’ approach. Practically speaking, the predicted result of blood glucose test strips still have room to improve by other important methods such as cost-sensitive learning and boosting.

Several factors that affect quality of blood glucose test strips have been identified after inspecting the modelling results. It is interesting that the direction of applying test reagent onto the strips possibly influences the result of blood glucose strips test. Left-soaked direction would have higher probability of failed tests. The reason in terms of inspection could be that most Taiwanese are right-handed, so the left-handed results are more variable. From the manufacturing point of view, the strips studied here are based on electrochemical glucose biosensing. To ensure accurate glucose readings when dispensing glucose oxidase enzyme solution from a dispenser onto test strips fabricated from a printed circuit board, every drop of the enzyme solution needs to have nearly the same weight and to be evenly dispensed on the reaction zone of the test strips. It is assumed that the uniformity of dispensing should reflect a consistent variations of readings from
both directions of applying test reagent. The stability of spray droplets can be improved by filling pressure, nozzle aperture, striker stroke, baking temperature and formulation of enzyme compound (Kim et al., 2018). To reduce this variation more data need to be collected from production site for future study.

Another factor affecting quality is the row where the test strips locate. The quality of the test strips taken from the middle part of the panel (cf., Figure 1) is more unstable than other parts. Possible causes could be the easily sunken blobs tend to be clustered around the middle part instead of peripheral area during etching or baking process. More examination on production process and related equipment is necessary for spotting potential weakness in manufacturing. For example, one process named baking, using high-temperature or vacuum, is to eliminating moisture from the PolyEthylene Terephthalate (PET) or Carbon Plated Electrodes (CPE) panels. This process is essential for test strips with high precision. If the moisture is not removed properly, then it may lead to the dysfunction of the panel while implementing it for further processing. It is recommended to place the panels in such a way that the air can circulate freely around them during the baking process. Since the moisture (and other solvents, possibly) need to be removed, the best way to do this is to have the panels in a rack, vertically oriented with some space in between them. If the boards are stacked on top of one another or flat on the base of the oven, etc., then it can be more difficult for the moisture to escape.

Furthermore, variation of blood glucose readings from test strips are estimated by the difference in percentage of each value with respect to corresponding row average. The row difference percentage between 4.57 and 3.63 will be more stable than those out of range. Once the row difference in percentage is too small (less than ~4.57) or too large (greater than 3.63), whole batch of strips deserves much stricter inspection because of the warning issued by wider variation of quality tests.

For future studies more data from dispensing or baking process can help to make more accurate prediction. Through integrating data from different processes and updating models periodically, big data and statistical machine learning can be synergised to achieve the principles of quality engineering, i.e. continuous improvement of processes.

Acknowledgements
We are grateful to a local blood glucose test strips maker who provides part of funding to develop and validate methodology reported here.

Funding
The authors received no direct funding for this research.

Author details
Ching-Shih Tsou12
E-mail: ctsou@mail.mcut.edu.tw
ORCID ID: http://orcid.org/0000-0001-8937-1729
Christine Liu1
Longsheng Cheng1
Hanting Zhou1
1 Department of Mechanical Engineering, Ming Chi University of Technology, New Taipei City, Taiwan.
2 Institute of Information and Decision Sciences, National Taiwan University of Business, Taipei City, Taiwan.
3 School of Economics & Management, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, China.

Disclosure statement
No potential conflict of interest was reported by the author(s).

References
Blue, J., & Chen, A. (2010). Spatial variance spectrum analysis and its application to unsupervised detection of systematic wafer spatial variations. IEEE Transactions on Automation Science and Engineering, 8(1), 56–66. https://doi.org/10.1109/TASE.2010.2041775
Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324
Chandra, A., & Yao, X. (2006). Ensemble learning using multi-objective evolutionary algorithms. Journal of Mathematical Modelling and Algorithms, 5(4), 417–445. https://doi.org/10.1007/s10852-005-9020-3
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. Data Mining and Knowledge
Discovery Handbook, 875–886. https://doi.org/10.1007/978-3-642-25465-5_40

Choi, S., Battulga, L., Nasridinov, A., & Yoo, K. H. (2017). A decision tree approach for identifying defective products in the manufacturing process. International Journal of Contents, 13(2), 57–65. https://doi.org/10.5392/IJoc.2017.13.2.057

Clarke, W. L., Cox, D., Gonder-Frederick, L. A., Carter, W., & Pohl, S. L. (1987). Evaluating clinical accuracy of systems for self-monitoring of blood glucose. Diabetes Care, 10(5), 622–628. https://doi.org/10.2337/diacare.10.5.622

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40(2), 139–157. https://doi.org/10.1023/A:1007607519361

DuaFatre-Patouraux, L., Vague, P., & Lassmann-Vague, V. (2003). History, accuracy and precision of SMBG devices. Diabetes & Metabolism, 29(2 Pt 2), 57–14. https://pubmed.ncbi.nlm.nih.gov/12746620/

Eid, S. A., O'Brien, P. D., Hinder, L. M., Hayes, J. M., Mendelson, F. E., Zhang, H., Zeng, L., Kretzler, K., Narayanan, S., Abcouwer, S. F., Brosius, I. F. C. 3., Pennathur, S., Savelief, M. G. M., & Feldman, E. L. (2020). Differential effects of empagliflozin on microvascular complications in murine models of type 1 and type 2 diabetes. Biology, 9(11), 347. https://doi.org/10.3390/biology9110347

Erbach, M., Freckmann, G., Hinzmann, R., Kulzer, B., Ziegler, R., Heinemann, L., & Schnell, O. (2016). Interferences and limitations in blood glucose self-testing: An overview of the current knowledge. Journal of Diabetes Science and Technology, 10(5), 1161–1168. https://doi.org/10.1177/193229681664133

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research, 61, 863–905. https://doi.org/10.1613/jair.11192

Freckmann, G., Schmid, C., Baumstark, A., Rutschmann, M., Haug, C., & Heinemann, L. (2015). Analytical performance requirements for systems for self-monitoring of blood glucose with focus on system accuracy: Relevant differences among ISO 15197: 2003, ISO 15197: 2013, and current FDA recommendations. Journal of Diabetes Science and Technology, 9(4), 885–894. https://doi.org/10.1177/1932296815580160

Freckmann, G., Link, M., Schmid, C., Pleus, S., Baumstark, A., & Haug, C. (2015). System accuracy evaluation of different blood glucose monitoring systems following ISO 15197: 2013 by using two different comparison methods. Diabetes Technology & Therapeutics, 17(9), 635–648. https://doi.org/10.1089/dia.2015.0085

Ginsberg, B. H. (2009). Factors affecting blood glucose monitoring: Sources of errors in measurement. Journal of Diabetes Science and Technology, 3(4), 903–913. https://doi.org/10.1177/193229680900300438

Gomes, T., Juurlink, D. N., Shah, B. R., Paterson, J. M., & Mamdani, M. M. (2010). Blood glucose test strips: Options to reduce usage. CMAJ, 182(1), 35–38. https://doi.org/10.1503/cmaj.091017

Goyal, R., & Jialal, I. (2021). Diabetes mellitus type 2. StatPears Publishing.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

He, H., & Ma, Y. (2013). Imbalanced learning: Foundations, Algorithms, and Applications (Wiley).

Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2018). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE Access, 4, 9145–9154. https://doi.org/10.1109/ACCESS.2016.2647238

Kim, K.-Y., Chang, H., Lee, Y.-C., Lee, W.-D., Lee, J.-F., & Lee, J.-F. (2018). Effect of filling pressure in jetting dispenser on the performance of blood glucose test strips using immersion gold-plated printed circuit board. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 232(8), 779–786. https://doi.org/10.1177/095411971876321

Kocev, D., Vens, C., Struyf, J., & Dzeroski, S. (2007). September 2007: Estimators of multi-objective decision trees. In European conference on machine learning (pp. 624–631). Springer, Berlin, Heidelberg.

Köksal, G., Batmaz, I., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. Expert Systems with Applications, 38(10), 13468–13476. https://doi.org/10.1016/j.eswa.2011.04.063

Kuo, Y. H., & Kusiak, A. (2019). From data to big data in production research: The past and future trends. International Journal of Production Research, 57(15–16), 4828–4853. https://doi.org/10.1080/00207543.2018.1443230

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22. https://cran.r-project.org/doc/Rnews/2002-3.pdf

Lin, C. T., Chang, C. W., & Chen, C. B. (2004). Focus groups: Impact of quality and process capability factors on the silicon wafer slicing process. International Journal of Manufacturing Technology and Management, 61(1–2), 171–184. https://doi.org/10.1504/IJMTM.2004.004513

Lin, W. J., & Chen, J. J. (2013). Class-imbalanced classifiers for high-dimensional data. Briefings in Bioinformatics, 14(1), 13–26. https://doi.org/10.1093/bib/bbs006

Loh, W. Y. (2011). Classification and regression trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14–23. https://doi.org/10.1002/widm.8

Ludvigsson, J., & Hanss, R. (2003). Continuous subcutaneous glucose monitoring improved metabolic control in pediatric patients with type 1 diabetes: A controlled crossover study. Pediatrics, 111(5), 933–938. https://doi.org/10.1542/peds.111.5.933

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A package for binary imbalanced learning. R Journal, 6(1), 79–89. https://doi.org/10.22626/rj-2014-008

Malhotra, R., & Jain, A. (2012). Fault prediction using statistical and machine learning methods for improving software quality. Journal of Information Processing Systems, 8(2), 241–262. https://doi.org/10.3745/JIPS.2012.8.2.241

Mazurowski, M. A., Habas, P. A., Zuroda, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural
network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Networks, 21(2–3), 427–436. https://doi.org/10.1016/j.neunet.2007.12.031

McGill, J. B., & Ahmann, A. (2017). Continuous glucose monitoring with multiple daily insulin treatment: Outcome studies. Diabetes Technology & Therapeutics, 19(S3), S3–S12. https://doi.org/10.1089/dia.2017.0090

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery, 28(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

Naghibi, S. A., Pourghasemi, H. R., & Dixon, B. (2016). GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. Environmental Monitoring and Assessment, 188(1), 1–27. https://doi.org/10.1007/s10661-015-5049-6

Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. Oriental Journal of Computer Science & Technology, 8(1), 13–19 http://www.computerscijournal.org/?p=1592

Parkes, J. L., Slatin, S. L., Pardo, S., & Ginsberg, B. H. (2000). A new consensus error grid to evaluate the clinical significance of inaccuracies in the measurement of blood glucose. Diabetes Care, 23(8), 1143–1148. https://doi.org/10.2337/diacare.23.8.1143

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2008, December). A comparative study of data sampling and cost sensitive learning. In 2008 IEEE international conference on data mining workshops (pp.46–52), IEEE.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Folleco, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Information Sciences, 259, 571–595. https://doi.org/10.1016/j.ins.2010.12.016

Therneau, T., & Atkinson, B. (2018). rpart: Recursive partitioning and regression trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart

Torgo, L. (2017). Data mining with R: Learning with case studies (2nd ed.). Chapman and Hall/CRC.

Yao, L., & Ge, Z. (2018). Scalable semisupervised GMM for big data quality prediction in multimode processes. IEEE Transactions on Industrial Electronics, 66(5), 3681–3692. https://doi.org/10.1109/TIE.2018.2856200
