In-process Estimation of Ascorbic Acid Assay Using Machine Learning Algorithms

M A Felipe and R Baldovino
Department of Manufacturing Engineering and Management De La Salle University
2405 Taft Avenue Manila Philippines
Email: renannbaldovino@dlsueduph

Abstract. Demand for greater volume of pediatric vitamin C syrups become higher every year whilst maintaining an expected high quality. Current laboratory techniques employ costly time-consuming and destructive testing which inhibit manufacturers to attain higher process efficiency. In this study, machine learning methods: random forest RF, support vector machine SVM, artificial neural network ANN, and adaptive neuro-fuzzy inference system ANFIS are used for the estimation of percentage ascorbic acid %AA assay. Likewise, these methods are compared to conventional linear regression LR model and to each other. pH, specific gravity, viscosity, and %AA assay measurements were used for the training of the network. Preprocessing technique involving data smoothing was employed on each nonlinear main effect relationship to reduce the noise and achieve better prediction accuracy. Upon training, it was found the ANN coupled with Bayesian Regularization exhibited 0.2314 MSE, a higher accuracy among other algorithms. Furthermore, it generally pushed ANN to be relatively more accurate than ANFIS with a minimal MSE of 0.30810.

1. Introduction
Batch processing and laboratory sampling/testing currently state the entirety of a pharmaceutical manufacturing process. Though still providing decent quality pharmaceuticals to the patients, laboratory testing in particular limit manufacturers to achieve higher efficiency whilst maintaining atmost drug efficacy due to its cost and complexity [1].

Ascorbic acid AA or commonly known as vitamin C is an important factor to prevent accumulation of abnormal collagen, the most abundant protein essential in the formation of teeth, bones, and skin [2]. Throughout the years, demand for AA increases particularly in pediatrics. To cope with the situation, machine learning techniques provide the necessary tools to analyze in-process parameters of the product and estimate its AA content without solely relying on laboratory methods.

Machine learning ML has resurfaced to distinction as a prevalent tool in performing tasks based on rules learned from data instead of rules explicitly described by humans [3]. Over the last few years, regression learners such as linear regression LR, random forest RF, support vector machine SVM, artificial neural network ANN, and adaptive neuro-fuzzy inference system ANFIS have become the staple algorithm when it comes to estimating the nonlinear relationships of input-output variables. Likewise, they become common algorithms due to their high flexibility and proven prediction abilities.

Linear regression – the rudimentary approach for predictive analysis that relates the dependent to the independent variable/s by the form of \( y = ax + b \).
**Random forest** – an ensemble of decision trees trained through the bagging method. It runs by generating multiple trees then outputting an average prediction. Lebanov et al [4] comparatively used RF on chromatogram mass spectra of 20 different essential oils to classify each of them as well as determine its quality.

**Support vector machine** – a learning algorithm that maps the nonlinear input into the feature space through various kernel functions to avoid high dimensional computation and then constructed with a linear function with $\varepsilon$-insensitive loss function in the feature space. Zhao et al [5] used least squares SVM to classify samples of bamboo cotton and hemp fiber based on its NIR spectra; and compared it to soft independent modelling of class analogy SIMCA with the latter performing poorly.

**Artificial neural network** – Commonly known as a multilayer perceptron MLP uses processing points known as nodes in three different succeeding layers – input hidden and output. Similar to a biological neuron each node in a layer is linked to a node of a neighboring layer harboring a corresponding weight. Each of these weights undergo iterative adjustments during network training until the difference between the predicted output and the target output is within a desired minimum error. Wang et al [6] used ANN with Back Propagation BP coupled with Partial Least Squares PLS to rapidly and concurrently determine the amount of preservatives in beverages sodium benzoate and potassium sorbate from their corresponding UV-Vis spectra.

**Adaptive neuro-fuzzy inference system** – a product of the combination of the learning abilities of the ANN and the reasoning capabilities of fuzzy logic in order to provide reinforced prediction capabilities and efficient technique for modelling. Bouhedda et al [7] used ANFIS to estimate the adsorption profile of Cephalexin on OSA starch as well as how parameters such as pH, contact time, and temperature affect the adsorption process.

In this study pH specific gravity and viscosity as well as the %AA assay of a high-volume pediatric Vitamin C syrup shall be used in five different machine learning regressors – LR, RF, SVM, ANN, and ANFIS to develop and choose an optimal model. The optimal model chosen shall have the highest accuracy among the others of estimating %AA assay.

### 2. Materials and Methods

#### 2.1. Data Acquisition and Pre-Processing

The studied product is a nutraceutical syrup containing 100mg/5mL AA has an appearance of a slightly hazy yellow-green thickened syrup and commercially manufactured using a 10000L compounding tank. In this study data containing input pH, specific gravity, and viscosity and output assay concentration physicochemical properties from 264 batches of the nutraceutical product were utilized to construct a predictive model.

As per manufacturing procedure each of the input properties were determined from samples taken at final volume. Each of the properties were determined using the following instruments: 1 pH – taken via ThermoFisher© Benchtop pH meter at 27°C; 2 SG – taken via Durac© hydrometer at 27°C; Viscosity – taken using a Brookfield© Digital Viscometer at 27°C Spindle 2 with a speed of 20 rpm; and Ascorbic Acid AA concentration – taken using Mettler Toledo© UV/Vis Spectrophotometer.

To ensure accuracy of analysis raw data was smoothened using the moving average technique. Data smoothing refers to the removal of random variation noise and linear trends from data. Using the “smoothdata” function in Matlab© raw data was processed. Figure 1 shows the plot for the smoothed data against the raw data of the pH, SG, and viscosity versus ascorbic acid concentrations.

### Table 1: Linear and quadratic main effects regression models for assessing %AA assay

| Relation  | Fitted Equation | $R^2$ |
|-----------|-----------------|-------|
| pH - %AA  | $pH = -1159 + 1334[AA] - 0003743[AA]^2$ | 0.61  |
| SG - %AA  | $SG = 1752 - 01896[AA] + 000551[AA]^2$ | 0.65  |
| Visco - %AA | $Visco = -4873 + 3382[AA]$ | 0.022 |

2
Based on the smoothed data of the input and output parameters, simple regression analysis was employed through Minitab® 1710. Relations between % AA assay with other input parameters were tested using linear and quadratic functions. Statistically significant p<0.005 relationships between the input parameters and AA concentration were determined to be linear. Viscosity vs %AA and quadratic pH, SG vs %AA regressions. However, despite the significance percent variation from each regression model only shows that the relationship is truly nonlinear and cannot be simply fitted using simple linear or quadratic relations. Table 1 lists all of the obtained binary relationships together with their fitted equations and percentage of variation.

2.2. Regression Modelling
Matlab® was used for the training and creation of the regression models. The regression learner app was used for LR, RF, and SVM learning. Main effects linear linear regression was chosen for the model type of LR Bagged trees ensemble corresponding to RF was also selected. All configurations, i.e., cubic quadratic Gaussian, etc., for SVM were run and the minimum MSE was chosen to be the optimal model. Neural net fitting tool nftool employing 10 neurons encompassing a single hidden layer were used to construct the model. Likewise, all training algorithms, i.e., Levenberg-Marquardt Bayesian Regularization and Scaled Conjugate Gradient, were employed to train the network. Lastly, Neuro-fuzzy designer was utilized for the creation of the fuzzy inference system. Furthermore, 9 membership functions utilizing each of the possible types were used for the creation of the fuzzy inference system.

3. Results and Discussion
Training of the input data employing five different learning algorithms at different configurations yielded the following performance stated in Table 2. As shown in the comparison of different training algorithms, generally ANNs outperform the other learning algorithms with BRANN see Figure 2 as the best fitting regression with a mean square error MSE of 0.23143. In addition, as expected among regularization techniques both BRANN and LMANN exhibited higher functioning approximation than SCGANN and among other training algorithms.
Table 2 Performance index for training algorithms

| Regression Model | Training Algorithm          | Mean Square Error |
|------------------|-----------------------------|-------------------|
| ANN              | Bayesian Regularization BRANN | 0.23143           |
| ANN              | Levenberg-Marquardt LMANN    | 0.24063           |
| ANFIS            | Triangular MF trimf          | 0.29520           |
| ANFIS            | Trapezoidal MF trapmf        | 0.30220           |
| ANFIS            | Pi MF pimf                  | 0.30810           |
| ANFIS            | Gaussian MF gaussmf          | 0.31050           |
| ANFIS            | Sigmoid difference MF dsigmf | 0.31080           |
| ANFIS            | Sigmoid product MF psigmf    | 0.31080           |
| ANFIS            | Gaussian Bell MF gbellmf    | 0.31140           |
| ANFIS            | Combined Gaussian gauss2mf  | 0.32550           |
| ANN              | Scaled Conjugate Gradient SCGANN | 0.36070      |
| SVM              | Cubic C-SVM                 | 0.41125           |
| SVM              | Quadratic Q-SVM             | 0.45343           |
| SVM              | Linear regression           | 0.45374           |
| SVM              | Course Gaussian CG-SVM      | 0.45622           |
| SVM              | Linear L-SVM                | 0.45851           |
| SVM              | Medium Gaussian MG-SVM      | 0.41617           |
| SVM              | Fine Gaussian FG-SVM        | 0.75096           |
| RF               | Bagged Trees RF             | 0.87937           |

A focus on ANN regression shows that despite BRANN’s higher correlation than LMANN the observed not so significant difference of 0.0092 between their regression values may be attributed to BRANN’s algorithm BRANN also known as Bayesian regularization backpropagation is a network training function that uses Bayesian regularization to update the weight and bias values and use the Gauss-Newton approximation to the Hessian matrix available inside the Levenberg-Marquardt algorithm.

In this process Bayesian regularization minimizes a linear combination of squared errors and weights and modifies the linear combination so that at the end of training the resulting network has good generalization qualities In addition even though BRANN achieved a good generalized model than LMANN see figure 2 the later exhibited far faster convergence as seen in figure 3 Levenberg-Marquardt function achieved faster convergence due to the fact that it approaches second-order training speed without having to compute the Hessian matrix.
Figure 3 Training performance of LMANN versus BRANN

Though less accurate ANFIS performed comparatively with ANN as depicted by the 0.05 difference between the ANFIS trimf and ANN LMANN. An increase in the number of membership functions without compromising overfitting would probably make the model more accurate. Of all the training algorithms, RF exhibited the least accurate followed by FG-SVM which unexpectedly outperformed by LR.

4. Conclusion

It is shown that the constructed ANN models using Bayesian Regularization and Levenberg-Marquardt exhibit a high performance than ANFIS, SVM, LR, and RF in decreasing accuracy for predicting %AA assay. The performance comparison showed that the machine learning system is a good tool for estimating API assays even at in-process without solely relying on laboratory tests. Higher performance of machine learning models were sourced from greater degree of robustness and fault tolerance than traditional statistical models because there are more processing neurons each with primarily local connections.

References

[1] Sacher S et al 2020 Towards A Novel Continuous HME-Tableting Line: Process Development And Control Concept Eur J Pharm Sci 142 105097
[2] Griffiths J K 2020 Vitamin Deficiencies Tenth Edit Elsevier Inc
[3] Alsharif M H, Kelechi A H and Yahya K 2020 Machine Learning Algorithms for Smart Data Analysis in Internet of Things Environment: Taxonomies and Research Trends Symmetry Basel 12(1) 1–17
[4] Lebanov L, Tedone L, Ghiasvand A Paull and B 2019 Random Forests Machine Learning Applied To Gas Chromatography – Mass Spectrometry Derived Average Mass Spectrum Data Sets For Classification And Characterisation Of Essential Oils Elsevier BV
[5] Zhao Z, Li B, Liu W, Yuan H and Song C 2019 A Qualitative Recognition Method Based On Karhunen-Loeve Decomposition For Near-Infrared NIR And Mid Infrared MIR Spectroscopy Analysis Anal Methods 12(1) 112–121
[6] Wang M et al 2020 Simultaneous Ultraviolet Spectrophotometric Determination Of Sodium Benzoate And Potassium Sorbate By BP-Neural Network Algorithm And Partial Least Squares Optik Stuttgart 201 163529
[7] Bouhiedda M, Lefnaoui S, Rebour S and Yahoum M M 2019 Predictive Model Based On Adaptive Neuro-Fuzzy Inference System For Estimation Of Cephalexin Adsorption On The Octenyl Succinic Anhydride Starch Chemom Intell Lab Syst 193 103843