



CMAC

Machine Learning Models for the Prediction of Pharmaceutical Powder Properties

EPRSC – CMAC Future Manufacturing Research Hub

UNIVERSITY OF STRATHCLYDE

Strathclyde Institute of Pharmacy and Biomedical Sciences

Glasgow, UK

Laura Pereira Diaz

2022



**Engineering and
Physical Sciences
Research Council**

Declaration of Authenticity and Author's Rights

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree. The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed:

A handwritten signature in black ink, appearing to be 'S. J. ...', written over a horizontal line.

Date: 23/12/2022



To my grandmother Lita

Abstract

Understanding how particle attributes affect the pharmaceutical manufacturing process performance remains a significant challenge for the industry, adding cost and time to the development of robust products and production routes. Tablet formation can be achieved by several techniques however, direct compression (DC) and granulation are the most widely used in industrial operations. DC is of particular interest as it offers lower-cost manufacturing and a streamlined process with fewer steps compared with other unit operations. However, to achieve the full potential benefits of DC for tablet manufacture, this places strict demands on material flow properties, blend uniformity, compactability, and lubrication, which need to be satisfied. DC is increasingly the preferred technique for pharmaceutical companies for oral solid dose manufacture, consequently making the flow prediction of pharmaceutical materials of increasing importance. Bulk properties are influenced by particle attributes, such as particle size and shape, which are defined during crystallization and/or milling processes. Currently, the suitability of raw materials and/or formulated blends for DC requires detailed characterization of the bulk properties. A key goal of digital design and Industry 4.0 concepts is through digital transformation of existing development steps be able to better predict properties whilst minimizing the amount of material and resources required to inform process selection during early-stage development.

The work presented in Chapter 4 focuses on developing machine learning (ML) models to predict powder flow behaviour of routine, widely available pharmaceutical materials. Several datasets comprising powder attributes (particle size, shape, surface area, surface energy, and bulk density) and flow properties (flow function coefficient) have been built, for pure compounds, binary mixtures, and multicomponent formulations. Using these datasets, different ML models, including traditional ML (random forest, support vector machines, k-nearest neighbour, gradient boosting, AdaBoost, Naïve Bayes, and logistic regression) classification and regression approaches, have been explored for the prediction of flow properties, via flow function coefficient. The models have been evaluated using multiple sampling methods and validated using external datasets, showing a performance over 80%, which is sufficiently high for their implementation to improve manufacturing efficiency. Finally, interpretability methods, namely SHAP (SHapley Additive exPlanations), have been used to understand the predictions of the machine learning models by determining how much each variable included in the training dataset has contributed to each final prediction.

Chapter 5 expanded on the work presented in Chapter 4 by demonstrating the applicability of ML models for the classification of the viability of pharmaceutical formulations for continuous DC via flow

function coefficient on their powder flow. More than 100 formulations were included in this model and the particle size and particle shape of the active pharmaceutical ingredients (APIs), the flow function coefficient of the APIs, and the concentration of the components of the formulations were used to build the training dataset. The ML models were evaluated using different sampling techniques, such as bootstrap sampling and 10-fold cross-validation, achieving a precision of 90%.

Furthermore, Chapter 6 presents the comparison of two data-driven model approaches to predict powder flow: a Random Forest (RF) model and a Convolutional Neural Network (CNN) model. A total of 98 powders covering a wide range of particle sizes and shapes were assessed using static image analysis. The RF model was trained on the tabular data (particle size, aspect ratio, and circularity descriptors), and the CNN model was trained on the composite images. Both datasets were extracted from the same characterisation instrument. The data were split into training, testing, and validation sets. The results of the validation were used to compare the performance of the two approaches. The results revealed that both algorithms achieved a similar performance since the RF model and the CNN model achieved the same accuracy of 55%.

Finally, other particle and bulk properties, i.e., bulk density, surface area, and surface energy, and their impact on the manufacturability and bioavailability of the drug product are explored in Chapter 7. The bulk density models achieved a high performance of 82%, the surface area models achieved a performance of 80%, and finally, the surface-energy models achieved a performance of 60%. The results of the models presented in this chapter pave the way to unified guidelines moving towards end-to-end continuous manufacturing by linking the manufacturability requirements and the bioavailability requirements.

Acknowledgments

I would like to first thank my supervisors, Prof. Alastair J. Florence and Dr Cameron J. Brown for their support and guidance throughout this project. I would like to thank Dr Chantal Mustoe, for her wonderful input and advice. I would like to thank Dr Antony Vassileiou, Dr John Armstrong and Matthew Wilkinson for their help and support. I would like to thank all my colleagues at EPSRC CMAC Future Manufacturing Research Hub for all their help, encouragement, and advice.

Finally, I would like to thank my family for their unconditional support, inspiration, and motivation. Thank you, Nuria.

Contents

Declaration of Authenticity and Author’s Rights	ii
Abstract.....	iv
Acknowledgments.....	vi
Contents.....	vii
Table of Figures.....	xi
Table of Tables	xxii
1. Introduction	1
1.1. Secondary Processing in Pharmaceutical Manufacturing	1
1.2. Manufacturing classification system: overcoming formulation challenges.....	2
1.3. The importance of powder flowability in pharmaceutical manufacturing	4
1.3.1. Powder flow measurement.....	5
1.3.2. Prediction of powder flowability.....	6
1.4. Artificial Intelligence and Machine Learning.....	8
1.5. Summary	9
1.6. References	10
2. Aims and objectives	13
2.1. Aims	13
2.2. Objectives	13
2.2.1. Prediction of powder flow of pharmaceutical materials using ML models (Chapter 4).	13
2.2.2. Prediction of powder flow of pharmaceutical blends from physical particle properties of the API using ML models (Chapter 5).....	14
2.2.3. Deep learning (DL) approaches for the investigation of the robustness of particle size and shape descriptors (Chapter 6).....	14
2.2.4. Interrogation of particle and bulk property descriptors in the context of machine learning and prediction of pharmaceutical materials (Chapter 7).....	15
2.3. References	16
3. Materials and methods	17

Contents

3.1.	Materials	17
3.2.	Particle size and shape measurements.....	19
3.2.1.	Dynamic image analysis: QICPIC® Sympatec.....	19
3.2.2.	Static image analysis: Morphologi® G3, Malvern	22
3.2.3.	Laser diffraction: Mastersizer® 3000, Malvern	27
3.3.	Surface area and surface energy: Surface Energy Analyzer®	28
3.4.	Bulk measurements	31
3.4.1.	Bulk density: FT4® Powder Rheometer, Freeman Technology	31
3.4.2.	Shear Cell Measurements: FT4® Powder Rheometer, Freeman Technology	32
3.5.	Machine Learning (ML) in pharmaceutical manufacturing.....	36
3.5.1.	Unsupervised learning	37
3.5.2.	Supervised learning.....	39
3.5.3.	Model evaluation: metrics for classification and regression	44
3.5.4.	Model interpretability: Shapley Additive exPlanations (SHAP) values	49
3.6.	Deep Learning (DL) in pharmaceutical manufacturing	50
3.7.	References	52
4.	Prediction of powder flow of pharmaceutical powders using machine learning	55
4.1.	Introduction	55
4.2.	Materials and methods.....	57
4.2.1.	Materials	57
4.2.2.	Experimental methods.....	58
4.2.3.	Data analytics: correlation between features and correlation filtering	59
4.2.4.	Machine learning methods	59
4.3.	Results and discussion	61
4.3.1.	Experimental results	61
4.3.2.	Machine learning results.....	70
4.3.3.	Exploration of the sensitivity of machine learning models to the instrument used to generate the training dataset: understanding the impact of changes in input variables.....	88

Contents

4.3.4.	Reverse engineering: guidelines to design a direct-compressible material.	95
4.4.	Conclusions	100
4.5.	References	103
5.	Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.	106
5.1.	Introduction	106
5.2.	Materials and methods.....	109
5.2.1.	Experimental methods.....	109
5.2.2.	Machine learning methods: data curation, unsupervised (PCA) and supervised (classification and regression) learning algorithms.....	110
5.2.3.	Monte Carlo Simulation considering the analytical error to estimate the probability of classification.....	111
5.3.	Results and discussion	113
5.3.1.	Powder flow: flow function coefficient (FFc) models	113
5.3.2.	Powder flow: global wall friction angle (WFA) regression models	142
5.4.	Conclusions	145
5.5.	References	147
6.	Deep learning (DL) approaches for the investigation of particle size and shape descriptors.....	150
6.1.	Introduction	150
6.2.	Materials and methods.....	154
6.2.1.	Materials	154
6.2.2.	Particle size and shape analysis: Morphologi G3, Malvern.....	155
6.2.3.	Powder flow analysis: FT4 Powder Rheometer, Freeman Technology Ltd.....	155
6.2.4.	Machine learning methods: RF classification models.....	156
6.2.5.	Machine learning methods: CNN classification model.	158
6.2.6.	External Validation.....	159
6.3.	Results.....	160
6.3.1.	Experimental methods.....	160
6.3.2.	Machine learning results: comparison between the RF and CNN models for the	

Contents

prediction of powder flow of pharmaceutical materials	165
6.4. Conclusions	168
6.5. References	169
6.6. Appendix	172
6.6.1. Machine learning results: benchmark of the prediction of powder flow classes from particle size and shape using data-driven models.	172
7. Interrogation of particle and bulk property descriptors in the context of machine learning and prediction of pharmaceutical materials.....	177
7.1. Introduction: the importance of the prediction of bulk density, surface area and surface energy on tablet manufacturability	177
7.1.1. Bulk density.....	178
7.1.2. Surface properties.....	179
7.2. Materials and methods	181
7.2.1. Experimental methods for data generation.....	181
7.2.2. Machine learning methods	182
7.3. Results and discussion	183
7.3.1. Bulk density measurements and associated ML models	183
7.3.2. Surface area measurements and associated ML models.....	201
7.3.3. Surface energy measurement and associated ML models	209
7.4. Conclusions	217
7.5. References	219
8. Conclusions and future work	221
8.1. Future work.....	226
8.2. References	228

Table of Figures

Figure 1-1: MCS based on drug loading (%w/w) and API flow function coefficient (M. Leane et al., 2015).	3
Figure 1-2: The particle properties required for each class of the Manufacturing Classification System (MCS): Class 1: direct compression (DC), Class 2: dry granulation (DG), Class 3 (Wet granulation), Class 4 (Other technologies).	4
Figure 3-1: The QICPIC® – RODOS dynamic image analysis instrument.	20
Figure 3-2: The Feret diameter, the distance between the two parallel planes that restrict the object.	21
Figure 3-3: The convexity of a particle calculated as the ratio of the projection area (A) to the area of the convex hull (A+B)	22
Figure 3-4: The Morphologi® G3 - Malvern G3	23
Figure 3-5: The calculation of the CE diameter of a three-dimensional particle captured as a two-dimensional image. Taken from the Morphologi® G3 User Manual.	24
Figure 3-6: The circularity values for different shapes.....	25
Figure 3-7: The convexity values for particles with different shapes.	26
Figure 3-8: The elongation of four particles with different shapes and sizes.....	26
Figure 3-9: The Mastersizer® 3000 and the Aero S dispersion unit.....	28
Figure 3-10: Wettability, cohesion within the powder, and adhesion increase as the surface energy increases.	30
Figure 3-11: The inverse Gas Chromatography-Surface Energy Analyzer (iGC-SEA).	31
Figure 3-12: The 10 ml x 25 mm split vessel used to carry out the powder flow measurement.	33
Figure 3-13: The Freeman FT4® Powder Rheometer conditional cycle: a) vented piston; b) normal stress applied to consolidate the powder; c) removing the excess of powder from the vessel.	33
Figure 3-14: The 24 mm diameter blade used to perform the shear cell test.....	34
Figure 3-15: Normal stress plotted against shear stress. The FFc is defined by the ratio of the major principal stress (σ_1) to the unconfined yield strength (f_c) (Mehos et al., 2017).....	35
Figure 3-16: Clustering algorithm for the identification high-density areas of data points.	38
Figure 3-17: The optimum hyperplane to find the maximum distance between the classes in binary classification tasks.....	40
Figure 3-18: a) Non-linear dataset b) The hyperplane of a non-linear dataset.	40
Figure 3-19: A RF decision tree. The length of the decision trees is decided by the availability of training data whereas the nodes are decided based on a randomly selected number of features that provide the best split at each node.....	41

Figure 3-20: Diagram of the multi-layer perceptron (MLP) neural network interconnected layers. 42

Figure 3-21: Sigmoid function used in LR to classify observations in class 0 or class 1 43

Figure 3-22: AUC – ROC curve, calculated from the TNR plotted against the FPR to study the performance of a supervised classification algorithm 47

Figure 4-1: The distribution of the PSD D10. The values of PSD D10 ranged between 0 and 300 μm ; and most of the powders had a PSD D10 smaller than 100 μm 62

Figure 4-2: The distribution of the PSD D10 values smaller than 100 μm . The histogram shows that most of the powders had a PSD D10 between 20 and 60 μm (60% of the total of the pharmaceutical powders included for this study). 62

Figure 4-3: The distribution of the PSD D50. The PSD D50 values ranged from 0 to 600 μm , and most of the powders analysed had a PSD D50 smaller than 300 μm 63

Figure 4-4: The distribution of the PSD D50 of the powders that exhibited a PSD D50 smaller than 300 μm . Of these powders, most of them had a PSD D50 between 50 and 150 μm 63

Figure 4-5: The distribution of the PSD D90 ranged between 0 and 1000 μm , and most of the materials had a PSD D90 value smaller than 400 μm 64

Figure 4-6: The distribution of the PSD D90 values smaller than 400 μm and most of the materials had a PSD D90 between 200 and 250 μm 64

Figure 4-7: The distribution of the SMD values ranged from 0 to 500 μm . Most of the powders had a SMD smaller than 200 μm 65

Figure 4-8: Sauter Mean Diameter (SMD) values smaller than 200 μm . The histogram shows that most of the powders had a SMD between 40 and 120 μm 65

Figure 4-9: The distribution of the aspect ratio D50 values across the pharmaceutical powder included in the training dataset..... 66

Figure 4-10: The distribution of the sphericity D50 values of the materials included in the training dataset 67

Figure 4-11: The distribution of the sphericity D50 values greater than 0.5. Most of the materials had a sphericity D50 between 0.7 and 0.8..... 67

Figure 4-12: The distribution of bulk density of the pharmaceutical powders of study..... 69

Figure 4-13: 4-stage data-driven model workflow: data, model training, model evaluation, and model interpretability 70

Figure 4-14: The PCC heatmap. High values of the Pearson correlation coefficient are presented in light orange, whereas low values of the PCC are presented in dark red. 71

Figure 4-15: The PCC heatmap after filtering out highly correlated variables (PPC > 0.9). 72

Figure 4-16: Louvain clustering analysis 73

Figure 4-17: The diagram of the single-step classification model..... 73

Figure 4-18: The comparison of the performance of the algorithms selected for the single-step classification model, from left to right: kNN, RF, MLP NN, NB, AB, and GB. Even though almost all of them exhibited a similar performance, MLP was selected for further analysis. 75

Figure 4-19: MLP confusion matrix. 75

Figure 4-20: The diagram of the two-step classification model. 76

Figure 4-21: The performance of the classification algorithms included in Step 1 of the two-step model evaluated using 10-fold cross-validation 78

Figure 4-22: The MLP neural network confusion matrix of Step 1 of the two-step classification model evaluated using 10-fold cross-validation 78

Figure 4-23: The results of the external validation of Step 1 of the two-step classification model of the MLP neural network model. Only 62.5% of the materials were correctly classified 79

Figure 4-24: The results of the external validation of Step 1 of the two-step classification model of the RF model. 79

Figure 4-25: The comparison of the performance of the algorithms selected to train Step 2. RF was slightly higher than the other algorithms. 80

Figure 4-26: The combined confusion matrices for Step 1 and Step 2 for the RF models, evaluated by 10-fold cross-validation 81

Figure 4-27: The combined confusion matrices for Step 1 and Step 2 for the RF models, evaluated by the external dataset..... 81

Figure 4-28: Feature importance analysis for the RF model in a) Step 1 and b) Step 2. The features are ranked based on their absolute mean SHAP score. The impact of each feature on each class is represented using colours and hence, the impact of the prediction on the output non-free-flowing is represented in blue and the impact of the output free-flowing is represented in red for Step 1. The impact of the prediction on the output cohesive is represented in blue and the impact of the output non-cohesive is represented in red for Step 2 82

Figure 4-29: Comparison of the algorithms trained on particle size, particle shape, bulk density, surface area, specific surface energy, surface energy (com), and dispersive surface energy at 0, 3, 5 and 10% of coverage for Step 1 of the RF two-step classification model. 83

Figure 4-30: Feature importance analysis for the Step 1 of the two-step RF classification model, including surface area and surface energy as inputs of the model. 84

Figure 4-31: Feature importance analysis for the RF regression model. The features are ranked based on their absolute mean SHAP score..... 86

Figure 4-32: SHAP dependence plot of the RF regression model for the prediction of the reciprocal of the FFC with interaction visualisation between the bulk density and the PSD D10 87

Figure 4-33: Combined results of the RF model Step 1 and Step 2 for the sensitivity analysis carried out with Morphologi G3 data..... 89

Figure 4-34: Combined results of the Step 1 and Step 2 RF models for the sensitivity analysis carried out with Mastersizer® 3000 PSD data combined with the particle shape Morphologi G3 data. 89

Figure 4-35: Scatter plot of PSD D10 measured using a) QICPIC® (x-axis) and Morphologi® G3 (y-axis), achieving a R² of 0.903; b) QICPIC® (x-axis) and Mastersizer® 3000 (y-axis), achieving a R² of 0.975; c) Morphologi® G3 represented (x-axis), and Mastersizer® 3000 (y-axis), achieving a R² of 0.906 91

Figure 4-36: The PSD D10 measured by the QICPIC® (light blue), Morphologi® G3 (cobalt blue), and Mastersizer® (indigo blue) of the 12 powders included in this analysis..... 91

Figure 4-37: Scatter plot of PSD D50 measured using a) QICPIC® (x-axis) and Morphologi® G3 (y-axis), achieving a R² of 0.905; b) QICPIC® (x-axis) and Mastersizer® 3000 (y-axis), achieving a R² of 0.899; c) Morphologi® G3 represented (x-axis), and Mastersizer® 3000 (y-axis), achieving a R² of 0.867 92

Figure 4-38: a) Scatter plot of PSD D90 measured using a) QICPIC® (x-axis) and Morphologi® G3 (y-axis), achieving a R² of 0.937; b) QICPIC® (x-axis) and Mastersizer® 3000 (y-axis), achieving a R² of 0.565; c) Morphologi® G3 represented (x-axis), and Mastersizer® 3000 (y-axis), achieving a R² of 0.645. 92

Figure 4-39: The PSD D50 measured by the QICPIC® (light green), Morphologi® G3 (medium green), and Mastersizer® (dark green) of the 12 powders included in this analysis. 93

Figure 4-40: The PSD D9010 measured by the QICPIC® (light pink), Morphologi® G3 (dark pink), and Mastersizer® (purple) of the 12 powders included in this analysis..... 93

Figure 4-41: Scatter plot of sphericity results measured using QICPIC® (x-axis) and Morphologi® G3 (y-axis) of a) sphericity D10, achieving a R² of -0.065; b) sphericity D50, achieving a R² of -0.315; c) sphericity D90, achieving a R² of -0.504..... 94

Figure 4-42: Scatter plot of aspect ratio results measured using QICPIC® (x-axis) and Morphologi® G3 (y-axis) of a) aspect ratio D10, achieving a R² of -0.347; b) aspect ratio D50, achieving a R² of -0.499; c) aspect ratio D90, achieving a R² of -0.721 95

Figure 4-43: RF Step 1 classification model SHAP dependence plot of PSD D10. The colour of the data points corresponds with the x-axis values in that data points are presented in blue for low values of PSD D10, and pink for high values of PSD D10..... 97

Figure 4-44: RF Step 1 model SHAP dependence plot of Sphericity D90. The colour of the data points corresponds with the x-axis values in that data points are presented in blue for low values of Sphericity D90, and pink for high values of Sphericity D90 98

Figure 4-45: Flowability predictions for *Materials 1 (a), 2 (b), and 3 (c)*. Variables that increased the probability of being free-flowing were presented in pink, and variables that decreased the probability of being free-flowing were presented in blue. The length of the bars of the variables represented how much the variables impact the model outcome. 100

Figure 5-1: Number of PCs plotted against the explained variance ratio. The columns represent the individual explained variance by each component and the line represents the cumulative variance. The first PC explains most of the variance (approximately 20%), and when 12 PCs are used, 82% of the variance is explained. 115

Figure 5-2: The PCA plot of PC1 vs PC2 for the whole dataset, without differentiation between classes. 116

Figure 5-3: Loadings of the negative correlated variables with PC1: a) HS Circularity D[n, x] D50 ($R^2 = -0.92$), b) Circularity D[x, n] D10 ($R^2 = -0.81$), c) HS Circularity D[n, x] D90 ($R^2 = -0.80$), and d) Aspect Ratio D[n, x] D50 ($R^2 = -0.79$). 118

Figure 5-4: Loadings of the positive correlated variables with PC2: a) Concentration of FlowLac 90 ($R^2 = -0.685$), b) Concentration of Aerosil 200 ($R^2 = 0.67$), c) CE Diameter D[n, x] (μm) D10 ($R^2 = 0.54$), and d) Length D[n, x] D10 ($R^2 = 0.54$). 119

Figure 5-5: Distribution of the FFc of the formulations used for training, showing that the majority of the formulations had a low FFc (viable for cDC greater than 5). 120

Figure 5-6: Confusion matrix of the RF classification model using randomly splitting between training and testing (75:25). In the testing set, 16 non-viable formulations (FFc < 5, non-viable) and 4 viable formulations (FFc > 5, viable) were included. 121

Figure 5-7: Precision distribution calculated using a bootstrapping method of the RF model. 122

Figure 5-8: Recall distribution calculated using a bootstrapping method of the RF model. 122

Figure 5-9: The SHAP values plot. The features are ranked based on their SHAP score: the larger the value, the more important the feature. 125

Figure 5-10: The SHAP bee swarm plot. The direction in which the variables contribute to the prediction: the more scatter the data points, the more important the variable. High variables of the variables are represented in pink and low values are represented in blue. 126

Figure 5-11: SHAP dependence plot to analyse the impact of the FFc of the API included in the formulation on the flowability of the blend 127

Figure 5-12: SHAP dependence plot of the impact of the Area D[n,x] (μm^2) D90 of the API on the flowability of the formulation 127

Figure 5-13: SHAP dependence plot of the impact of the HS Circularity D[n,x] D10 of the API on the flowability of the formulation 128

Figure 5-14: Histogram of the distribution of the FFC values of the API included in test formulation 3 across the 1000 simulated formulations (mean value = 1.87, standard deviation = 0.55)..... 129

Figure 5-15: SHAP individual plot of one of the simulated formulations from the formulation test 1. 131

Figure 5-16: SHAP individual plot of one of the simulated formulations from the formulation test 2. 131

Figure 5-17: SHAP individual plot of one of the simulated formulations from the formulation test 3. 132

Figure 5-18: SHAP individual plot of one of the simulated formulations from the formulation test 4. 132

Figure 5-19: Scatter plot of the actual values of the FFC of the test formulations and the predicted value of the FFC of the formulations. The R² value achieved was 0.81..... 133

Figure 5-20: Measured values of the FFC of the formulations used for external validation plotted against the predicted values by the RF regressor. 134

Figure 5-21: Ranking of the most important variables using SHAP values for the regression models. 135

Figure 5-22: SHAP dependence plot of the impact of the value of the FFC of the API included in the formulation on the FFC of the formulation 136

Figure 5-23: SHAP individual plot of the prediction of test formulation 1. The main drivers that increased the predicted value of the FFC of the formulation were the concentration of Avicel PH102 of the formulation and the value of Aspect ratio D[n,x] D10 of the API. On the other hand, the drivers that decreased the predicted value were the HS Circularity D[n,x] D10 of the API. 136

Figure 5-24: Distribution of the values of the reciprocal of the FFC of the formulations included in the training dataset..... 137

Figure 5-25: Scatter plot of the measured reciprocal of FFC against the predicted reciprocal value of FFC of the formulations of the internal validation set (R²=0.856)..... 138

Figure 5-26: SHAP feature importance analysis of the reciprocal of FFC model..... 139

Figure 5-27: The individual SHAP analysis for the first formulation of the external dataset. The predicted value of the reciprocal of FFC was 0.48 (FFC = 2.07). The main drivers of this prediction were the API FFC (1.5), and the API concentration (50%), which increased the predicted value of reciprocal of FFC (decreased the FFC). 140

Figure 5-28: The SHAP feature importance analysis of the RF regression model when the reciprocal of the FFC of the API was an input variable..... 141

Figure 5-29: SHAP bee swarm plot for the PHIW regression model..... 143

Figure 5-30: The scatter plot actual values of WFA plotted against the predicted values of WFA. The 95% confidence intervals and the best-fit line are represented 144

Figure 5-31: The SHAP individual plot of the testing formulation that was predicted with the largest error. The actual value of the WFA was 9.9° and the predicted value was 13.79° 144

Figure 5-32: The SHAP individual plot the testing formulations that was predicted with the smallest error. The actual value of the WFA was 16.7° and the predicted value was 16.5° 145

Figure 6-1: The difference in the methodology of ML and DL models. For ML models, the particle information is extracted from the images generated by the particle size characterisation instrument by the built-in software, whereas for the DL models, this first feature extraction step is already performed by the DL model and therefore, the images can be fed directly to the model. Therefore, the inputs of these models would be different, i.e., the inputs of the ML models are particle size and shape descriptors, and the inputs for the DL model are pixel information 152

Figure 6-2: The diagrams of the RF models: a) multi-step classification model: the images are classified into cohesive, easy flowing or free flowing, and b) two-step classification model: the images are initially classified into free flowing or non-free flowing, and if they are classified as non-free-flowing, a second step is applied to classify them into cohesive or non-cohesive. 157

Figure 6-3: Particle size distribution (D50 values) across the powders included in the training dataset. Approximately 45% of the materials had a D50 value greater than 100 µm and smaller than 200 µm. 163

Figure 6-4: Distribution of the aspect ratio D50 values across the materials included in the training dataset. Approximately 25% of the materials had an aspect ratio D50 value between 0.6 and 0.65. 164

Figure 6-5: The area composite images taken using the Morphologi G3 of a) cohesive material, b) easy-flowing material, c) easy-flowing material, and d) free-flowing material. 165

Figure 6-6: The results of the external validation of a) the RF multiclass classification model and b) the CNN multiclass classification model..... 166

Figure 6-7: The performance of a) Step 1 of the two-step RF classification model, b) Step 1 of the two-step CNN classification model, c) Step 2 of the two-step RF classification model, and d) Step 2 of the two-step CNN classification model. 167

Figure 6-8: The results of the combination of the performance of Step 1 and Step 2 for a) the two-step RF classification model and b) the two-step CNN classification model. 167

Figure 6-9: Classification metrics of the multi-step classification RF model, including accuracy (0.526 ± 0.11), AUC (0.676 ± 0.12), precision (0.591 ± 0.11), and recall (0.558 ± 0.08)..... 172

Figure 6-10: RF confusion matrix for multi-class classification RF model (cohesive vs. easy-flowing vs. free-flowing materials) a) obtained from the validation and b) obtained from the external test. 173

Figure 6-11: Classification metrics of the Step 1 of the RF model, including accuracy (0.668 ± 0.1), AUC-ROC (0.794 ± 0.09), Precision (0.703 ± 0.08) and Recall (0.525 ± 0.22). 174

Figure 6-12: RF confusion matrix for Step 1 (free-flowing vs. non-free-flowing materials) a) obtained from the validation and b) obtained from the external test..... 174

Figure 6-13: Classification metrics of the Step 2 of the RF model, including accuracy (0.753 ± 0.09), AUC (0.855 ± 0.08), precision (0.769 ± 0.07), and recall (0.892 ± 0.06). 175

Figure 6-14: RF confusion matrix for Step 2 (cohesive vs. non-cohesive materials) a) obtained from the validation and b) obtained from the external test..... 176

Figure 6-15: SHAP analysis for the misclassification of cetyl alcohol. The plot shows that the probability of being class 1 (non-cohesive) was 0.54, and therefore, the model was not confident in this prediction. The main drivers that increase the probability of being non-cohesive where the particle size distribution D10 ($43.02 \mu\text{m}$), and the Sauter Mean Diameter ($92.94 \mu\text{m}$). 176

Figure 7-1: Bulk density (g/ml) experimental results..... 184

Figure 7-2: Pearson correlation coefficient heatmap of the features used for training bulk density models..... 186

Figure 7-3: Pairplot of the particle size (D10, D50, D90), particle shape (sphericity D10 and D90, and aspect ratio D10) and bulk density. This figure shows the scatter plot of each pair of variables to explore correlations. 187

Figure 7-4: A plot of the top 2-components in the PCA analysis which can account for 48% of the variance of the dataset. Low-density powders (class 0) are represented in red, whereas high-density powders (class 1) are represented in green..... 188

Figure 7-5: Comparison of the performance of the classification algorithms trained for the prediction of bulk density, calculated using 10-fold cross-validation. The metric analysed was AUC-ROC. 189

Figure 7-6: External validation results using the RF model. 5 of the powders included in the dataset had low density ($BD > 0.5 \text{ g/ml}$), and 3 powders had high density ($BD > 0.5 \text{ g/ml}$)..... 190

Figure 7-7: SHAP values plot (model output: class 1) for the RF model. Aspect ratio D10 and PSD D50 were the most important variables for the classification of bulk density. Moreover, high values of aspect ratio D10 and high values of PSD D50 led to higher values of bulk density 191

Figure 7-8: SHAP dependence plot for the RF model. The values of the aspect ratio D10 are represented on the x-axis, and the SHAP values of the aspect ratio D10 (“importance score”) are represented on the y-axis. The colour of the data points indicates the value of aspect ratio D10, with low values represented in blue and high values represented in pink. 192

Figure 7-9: SHAP dependence plot for the RF model. The values of FFC are represented on the x-axis, and the SHAP values of the FFC (“importance score”) are represented on the y-axis. The colour of the data points depends on the value of FFC, with low values represented in blue and high values represented in pink..... 193

Figure 7-10: The comparison of the results of the classification models only including particle size, sphericity and aspect ratio, and the composition of the blends in the training dataset..... 194

Figure 7-11: External validation results using the RF model. 2 of the powders included in the dataset had low density (BD > 0.5 g/ml), and 6 powders had high density (BD > 0.5 g/ml)..... 194

Figure 7-12: The comparison of the MAE achieved by supervised learning (RF, GB, AB), and unsupervised learning (PLS)..... 196

Figure 7-13: RF SHAP values plot. Aspect ratio D10 and PSD D10 were the most important variables for the prediction of bulk density. Moreover, high values of aspect ratio D10 and high values of PSD D10 led to higher values of bulk density 197

Figure 7-14: SHAP dependence plot for the RF model. The values of the PSD D10 are represented on the x-axis, and the SHAP values of the PSD (“importance score”) are represented on the y-axis. The colour of the data points indicates the value of PSD D10, with low values represented in blue and high values represented in pink..... 197

Figure 7-15: Scatter plot of the actual values of bulk density against the predicted values by the gradient boosting model..... 199

Figure 7-16: Force plot of the Span 60 bulk density prediction for the RF model. As seen in this plot, the main driver of the prediction is the PSD, which takes the prediction from the base value to the predicted value (0.74)..... 199

Figure 7-17: Force plot of Calcium carbonate. The main drivers for the underprediction were the particle size distribution (D90), the aspect ratio D10, and the particle size distribution D10..... 200

Figure 7-18: Force plot of ibuprofen (20%) and Plasdne povidone blend. The main drivers for the prediction were the particle size distribution D50, the sphericity D10, and the PSD D10 200

Figure 7-19: The distribution of the specific surface area (surface area per unit of mass or volume) of the powders included in the training dataset..... 201

Figure 7-20: The heatmap of the PCC of the variables used to train surface area models..... 203

Figure 7-21: PCA of surface area. Low surface area powders (class 0) are represented in red and high surface area powders (class 1) are represented in green..... 205

Figure 7-22: Comparison of the performance of the classification models for the prediction of the surface area, using a binary classification system 206

Figure 7-23: a) RF model SHAP summary plot of the ranking of the most important variables for the classification of surface area; b) Pairplot of surface area and bulk density to study the correlation between the actual data of the two variables. 207

Figure 7-24: a) RF model SHAP values summary plot that shows the direction in which each variable contributes to the prediction. The further to the right the data points appear, the more their contribution to a higher surface area; b) RF model SHAP dependence plot. Bulk density was plotted on the x-axis and the SHAP values for bulk density (the contribution of the variable to classification of surface area) were plotted on the y-axis. 207

Figure 7-25: RF model SHAP dependence plot shows the impact of PSD D90 on the surface area class. As the PSD D90 increases, the surface area decreased, and therefore, a PSD D90 greater than 300 μm would have a negative impact on the surface area class. The red rectangle shows the target PSD D90 to ensure suitability for direct compression based on achieving desired powder flow. 209

Figure 7-26: The distribution of surface energy values across the powders included in the training dataset. Most of the materials had a surface energy between 4 and 8 mJ/m^2 , and the mean value was 6.97 mJ/m^2 210

Figure 7-27: PCA for the classification of surface energy..... 212

Figure 7-28: The comparison of the results of the performance of the classification algorithms trained for the prediction of surface energy 213

Figure 7-29: a) Feature importance analysis using the RF SHAP methods for model interpretability. The SHAP values are assigned to the variables depending on their impact on the model target (surface energy); b) RF model SHAP summary plot of the direction in which each variable contributes towards the prediction of surface energy. High values of the variables are represented in pink and low values in blue. 214

Figure 7-30: a) RF model SHAP dependence plot between sphericity D90 and its impact on the surface energy; b) RF SHAP dependence plot of PSD D50..... 215

Figure 7-31: The comparison of the performance of classification learning algorithms for the prediction of surface energy from particle size, particle shape and surface area. 216

Figure 7-32: RF model SHAP dependence plot of surface area values, plotted on the x-axis, and its SHAP values, plotted on the y-axis, indicating the impact of surface area on surface energy. 217

Figure 8-1: A diagram of the reverse engineering approach to obtain the “ideal” API for the desired pharmaceutical manufacturing route, including the flow map showing the FFC of the API and the drug loading for direct compression, wet granulation and dry granulation presented by the MCS (Leane et al., 2018). 223

Figure 8-2: The interface of the app implemented on-site. The model can predict the viability of a pharmaceutical formulation for cDC based on the physical properties of the API and the concentration of the components..... 224

Table of Tables

Table 3-1: The materials included in the development of data-driven models.....	18
Table 3-2: The Jenike’s classification for powder flow showing the correlation between the FFC and the powder behaviour. Powders that have a FFC smaller than 4 are classified as cohesive, powders that have a FFC between 4 and 10 are classified as easy-flowing, and powders that have a FFC greater than 10 are classified as free-flowing.....	36
Table 3-3: Confusion matrix.....	45
Table 4-1: The composition of binary blends. All binary blends included FastFlo 316 and one of the following APIs: Ibuprofen 50, Paracetamol Granular Special, Paracetamol Powder, Mefenamic Acid, Calcium Carbonate.....	57
Table 4-2: The composition of multi-component blends. All multicomponent blends included FastFlo 316, one of the following APIs: Ibuprofen 50, Paracetamol Granular Special, Paracetamol Powder, Mefenamic Acid, Calcium Carbonate, and the remaining 25% of a combination of 20% Avicel PH-102, 3.5% Croscarmellose Sodium, and 1.5% Magnesium Stearate.....	58
Table 4-3: PSD results, including the range of values and the median value for each parameter.	66
Table 4-4: The surface area and surface energy measurements for the 35 powders analysed.....	68
Table 4-5: The number of pharmaceutical powders in each range of interest of FFC.....	69
Table 4-6: The performance of the algorithm involved in model training. MLP achieved the highest performance (AUC-ROC = 0.823).....	74
Table 4-7: The performance of the classification algorithms included in Step 1 of the two-step model evaluated using 10-fold cross-validation.....	77
Table 4-8: The performance of the classification algorithms included in Step 2 of the two-step model evaluated using 10-fold cross-validation.....	80
Table 4-9: Regression metrics to evaluate the performance of the algorithms used to build the regression model, using FFC as the independent variable.....	85
Table 4-10: Regression metrics to evaluate the performance of the algorithms used to build the regression model, using 1/FFC as the independent variable.....	85
Table 4-11: Results of the external validation performed with the regression model, setting the target variable first as “FFC”, and then as “1/FFC”.....	86
Table 4-12: The powders selected for the sensitivity analysis (4 cohesive, 4 easy-flowing and 4 free-flowing powders).....	88
Table 4-13: Properties of a direct-compressible material.....	96

Table 5-1: The independent variables used to train the ML model measured with the Morphologi® G3 and the Mastersizer® 3000	109
Table 5-2: The independent variables included in the training dataset after filtering	110
Table 5-3: Standard deviation of the measurements performed by the three different instruments that were used to analyse the data required to build the model.	113
Table 5-4: Loadings of the highest correlated variables with PC1 and PC2 and their squared loading scores for comparison. The highest correlated variable with PC1 was HS Circularity D[n, x] D50, and the highest correlated variable with PC2 was the concentration of FlowLac 90.....	117
Table 5-5: Classification metrics of the RF classification model using randomly splitting between training and testing (75:25).....	121
Table 5-6: Classification metrics of the RF model using 10-fold cross-validation.....	123
Table 5-7: Results of the classification of the four formulations included in the external validation using the RF classification model.	124
Table 5-8: Classification using the RF model of the 4 formulations used for external validation after using MC methods to simulate formulations considering the analytical error of the measurements.	130
Table 5-9: Regression metrics used to evaluate the performance of the model.....	133
Table 5-10: Results of external validation of the FFC regression model.	134
Table 5-11: Results of the regression models for the prediction of the reciprocal value of the FFC of the formulations.....	138
Table 5-12: Results of the prediction of the FFC of the formulations included in the external set, after retransforming the data from the reciprocal of FFC to the FFC.	140
Table 5-13: The results of the performance of the RF regression model for the prediction of the reciprocal of the FFC of the formulation, considering the reciprocal of the FFC of the APIs as an input variable.....	141
Table 5-14: Results of the evaluation of the WFA regression model on the testing set.....	142
Table 6-1: Powders included in the training dataset for the RF and the CNN model.....	154
Table 6-2: The Morphologi G3 descriptors used to train the RF classifier model.....	158
Table 6-3: The powders included in the external validation set, classified into cohesive, easy-flowing or free-flowing based on their FFC value.	160
Table 6-4: Number of powders per class used to train the RF model.....	160
Table 6-5: Number of powders per class used to validate the RF model.	161
Table 6-6: Materials included in the external test set	161

Table 6-7: Descriptors used to train the RF classifier model. The range of values and the mean value are presented in this table. 162

Table 7-1: Powder Flowability based on Hausner ratio from excellent flow (HR smaller than 1.11) to very poor flow (HR greater than 1.46) (Gorle & Chopade, 2020). 179

Table 7-2: Bulk density experimental results of the external dataset used to validate the machine learning model. 184

Table 7-3: Variables included in the bulk density model. 185

Table 7-4: Variables included in the bulk density model after PCC 186

Table 7-5: Bulk density categories. Powders with a bulk density less than 0.5 g/ml belong to the low-density class, and powders with bulk density greater than 0.5 g/ml belong to the high-density class. 188

Table 7-6: The classification scores which predict whether or not a powder has a bulk density that is optimal for direct compression for each of the materials included in the external dataset. When the classification was correct, the classification score is reported in green, whereas when the classification was wrong, the classification score is reported in red. Classification scores above 0.5 indicate “high-density class” will be predicted class and classification scores below 0.5 indicate that “low-density class” will be the predicted class. 190

Table 7-7: Classification scores of predictions for each of the materials included in the external dataset. 195

Table 7-8: External validation results of the regression model. 198

Table 7-9: Number of powders in each class of surface area. 202

Table 7-10: The variables included in the surface area models. 202

Table 7-11: The variables considered to train the surface area model after removing highly correlated variables. 203

Table 7-12: number of powders that belong to each class in the training dataset. 211

Table 7-13: The five high-surface energy powders that were clustered in a different group using PCA. These five powders were pharmaceutical blends, as opposed to the rest of the powders grouped in the main cluster that were individual materials. 212

1. Introduction

Pharmaceutical manufacturing can be divided into two principal stages termed primary and secondary manufacturing. Primary manufacturing involves the synthesis, work-up and crystallisation of the active pharmaceutical ingredient (API). Secondary manufacturing refers to the transformation of API into a formulated dosage form that can be administered safely to the patient. Crystallisation is a crucial step of the pharmaceutical primary manufacturing stage as a widely used, effective method for purification (Chen, Sarma, Evans, & Myerson, 2011). During crystallisation key particle and bulk attributes that determine the manufacturability of the final product are defined, including particle size and particle shape. However, the integration of primary and secondary manufacturing remains challenging due to the lack of understanding of how the physical properties defined during primary manufacturing affect secondary manufacturing. An integrated holistic view of the control of attributes to achieve the required performance in downstream operations would better allow integration of primary and secondary manufacturing to enable end-to-end continuous pharmaceutical manufacturing (Hatcher, Burgess, Payne, & Wilson, 2020). Hence, in this thesis, a data-driven approaches are taken to explore and interrogate the correlation between particle attributes and process performance to be able to better inform the manufacturing process, and to enable formulation optimization, thereby removing the unnecessary trial-and-error design steps and reducing the amount of material required in early stages of pharmaceutical development.

1.1. Secondary Processing in Pharmaceutical Manufacturing

Secondary manufacturing of tablets involves the blending of the API with selected excipients, tablet formation and final dedusting, finishing and/or coating stages. Tablet formation can be achieved through several techniques, such as direct compression (DC), wet granulation, or dry granulation. Tablets are one of the most used oral solid dosage forms due to their functionality, safe administration, and long shelf life. Furthermore, oral dosage forms are the most convenient way to administer medicines because they are cost-efficient and easier for the patient to handle and take, improving compliance (Kottke & Rudnic, 2002). Tableting is a single unit operation, and thus allows manufacturers to mass-produce them at a low cost. Moreover, tablets are reliable due to their consistent content uniformity, physical and chemical stability, and ability to control quality (Gad, 2008).

Tablet formation is achieved by compressing and compacting the formulation powder blend into the predesigned shape. Particle and bulk properties impact the tablet formation, and therefore understanding these properties will allow a better tablet formulation development (Dai, S., Xu, B., Zhang, Z., Yu, J., Wang, F., Shi, X., & Qiao, Y., 2019). To achieve the desired characteristics of the tablet,

the API is blended with excipients that do not impact on therapeutic function but aid processing (e.g., glidant), drug release (e.g., disintegrant) or ease of handling (e.g., diluent). Excipients improve the manufacturing properties of the formulation and drug bioavailability, and the selection of the appropriate formulation is crucial to define the manufacturability, stability, and performance of the tablet.

Direct compression (DC) is the most cost-effective manufacturing technique since it requires fewer unit operations. Though cost-effective, DC still has strict requirements regarding bulk properties, i.e., flowability, and uniformity, to be successful. If the blend does not exhibit the appropriate powder properties, the formulation will need to be pre-treated. Dry granulation (DG) and wet granulation (WG) are the most commonly used techniques to improve powder properties of blends that are not initially viable for DC. However, these pre-treatment techniques result in more steps, and therefore more time, than DC. Additionally, materials that are sensitive to heat and/or moisture are not suitable for dry granulation and/or wet granulation. Both techniques require heat treatment, and wet granulation requires the addition of a liquid to increase cohesion between particles (Dürig & Karan, 2019).

As stated, DG and WG are two main types of granulation techniques. In WG, water or a binding liquid is added to the powder, which is agitated to create agglomerates. These agglomerates increase particle size, bulk density and when they are dried and sieved, the desired particle size distribution can be achieved. The granules can be blended with the excipients and then . DG, also called roller compaction, creates granules without liquid. DG also enhances powder flowability and it is suitable for materials that are sensitive to moisture. However, it is a time-consuming process and not suitable for heat-sensitive materials (Saddik, 2020). Therefore, DC has a number of benefits in requiring less time, less equipment, less space, and less power consumption to make the tablets. Nevertheless, it has stricter requirements on the formulation blend than dry or wet granulation with regard to powder flow, uniformity, compactability, and lubrication to assure process performance (Schaller et al., 2019; Shangraw, 1989; Trementozzi et al., 2017).

1.2. Manufacturing classification system: overcoming formulation challenges.

Leane *et al.* developed the Manufacturing Classification System (MCS) as a tool for scientists to rank the feasibility of different processing routes for the manufacture of oral solid dosage forms, based on the properties of the API and the drug loading (M. Leane, Pitt, Reynolds, & Group, 2015). Before the proposal of the MCS, pharmaceutical development was exclusively focused on the processes that affect the patient, as captured in the Biopharmaceutical Classification System (BCS). The BCS was

introduced in 1995 to classify APIs to reduce the number of *in vivo* studies. The BCS enables the prediction of the dissolution, solubility, and intestinal permeability of the API and hence helps in the discovery and early development of new APIs (Ku, 2008). However, the BCS is not useful for drug product development efforts. To overcome this limitation the MCS was created. Fig 1-1 shows the framework that helps decide the suitable manufacturing process for a given formulation, considering the drug loading and the flow function coefficient of the API included in the formulation (Yasir, Asif, Kumar, & Aggarval, 2010).

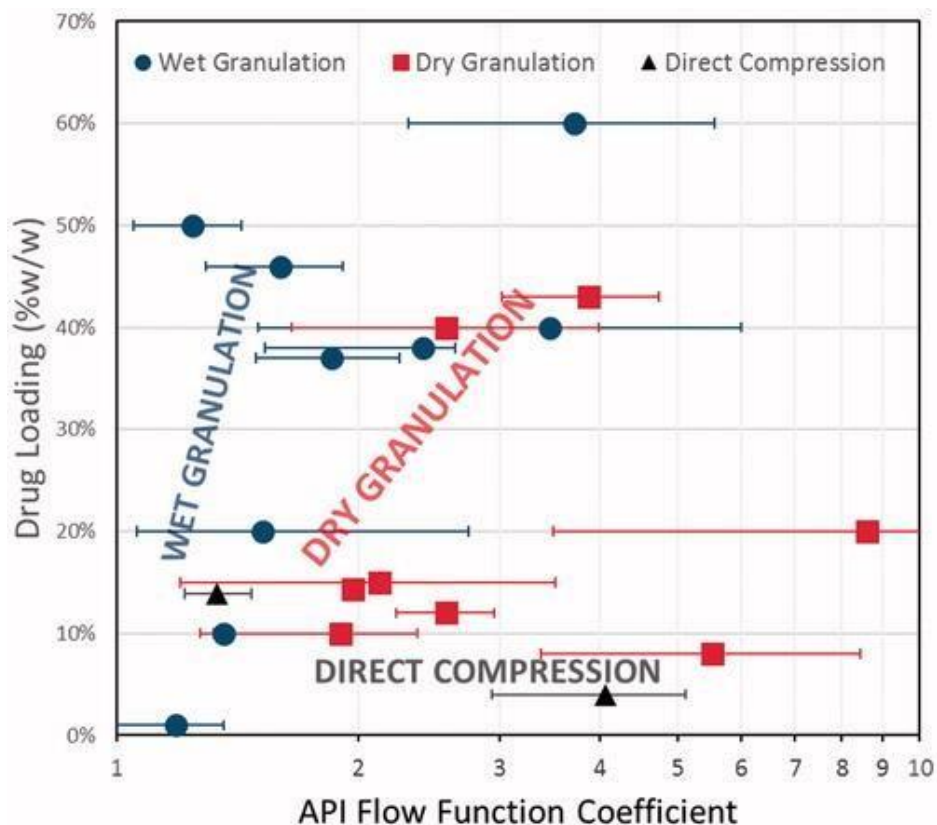


Figure 1-1: MCS based on drug loading (%w/w) and API flow function coefficient (M. Leane et al., 2015).

The MCS has identified four classes of manufacture depending on the materials and performance requirements, i.e., Class 1 is DC, Class 2 is DG, Class 3 is WG, and Class 4 is other technologies (OT), with increasing complexity and cost from class 1 to 3 (see Fig 1-2). The goal of the MCS is to improve the understanding of the relationship between particle properties and process performance to facilitate accelerated, rational drug development (M. Leane et al., 2015).

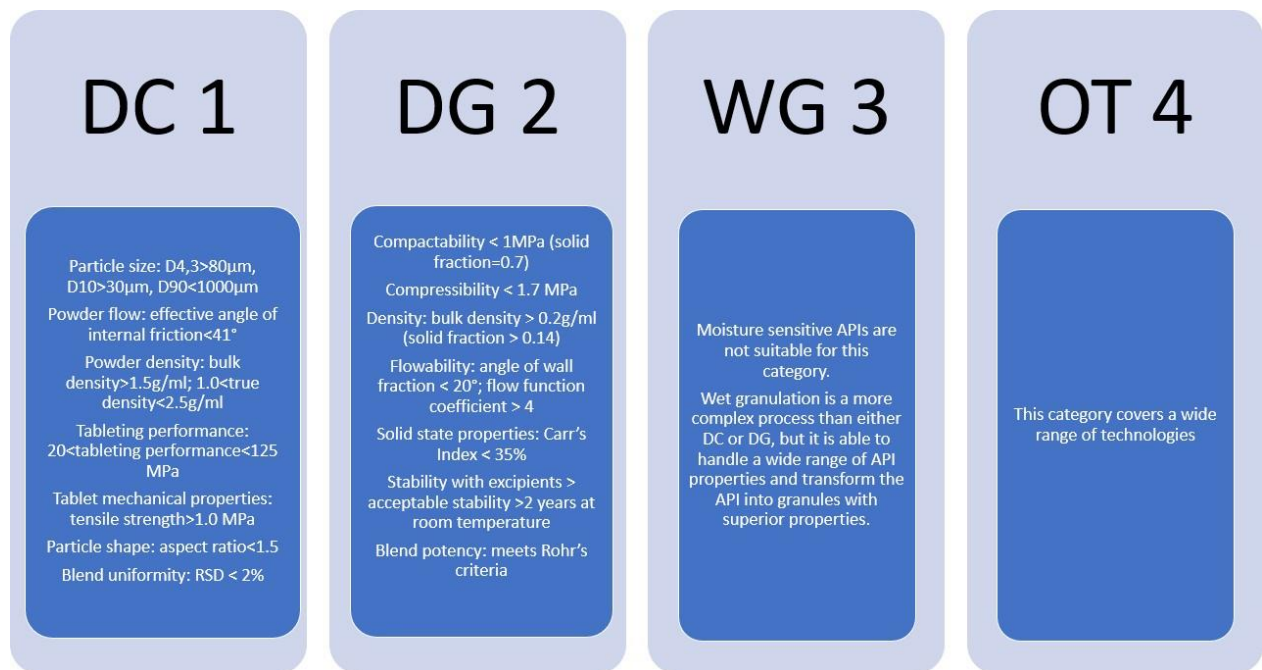


Figure 1-2: The particle properties required for each class of the Manufacturing Classification System (MCS): Class 1: direct compression (DC), Class 2: dry granulation (DG), Class 3 (Wet granulation), Class 4 (Other technologies).

The MCS highlights the importance of powder flow for deciding the appropriate tablet manufacturing process for a given formulation and dose and highlights the importance of powder flow properties to assuring efficient tableting operation. Powder flowability is impacted by the physical properties of the particles and by external variables. It is known that properties such as small particle size (Bellamy, Nordon, & Littlejohn, 2008) and needle-like particle shape (Kim, Wei, & Kiang, 2003) can lead to poorly flowable powders which may lead to difficulties regarding flow, feeding, blending and adhesion to surfaces in the powder path through the tableting line (Waknis et al., 2014) that can cause the failure of critical quality attributes e.g. from segregation, variable content uniformity or tablet properties.

1.3. The importance of powder flowability in pharmaceutical manufacturing.

Particle size distribution has been considered previously for the prediction of powder flow behaviour given the well-recognised impact of particle size on powder flow (Sun & Grant, 2001; Yu, Muteki, Zhang, & Kim, 2011). Kaerger *et al.* demonstrated the effect of particle shape of two different paracetamol blends with similar particle sizes on powder flowability, showing that blends containing spherical particles exhibited a 21% improvement in flowability compared to the blends containing needle-like particles (Kaerger, Edge, & Price, 2004), and Cleary proved that the strength of a given material depends on its particle shape (Cleary, 2008). External variables, such as humidity,

temperature, aeration, consolidation between particles, and storage time, affect powder behaviour too (Freeman, Brockbank, & Sabathier, 2017; Prescott & Barnum, 2000). Aeration is often used in industrial processes to enhance flow and Lloyd *et al.* concluded that aeration reduces shear stress up to 60% compared to the shear stress of non-aerated powders (Lloyd & Webb, 1986). However, using air to improve flowability can be particularly challenging in cohesive materials (Klein, Höhne, & Husemann, 2003). Lower levels of particle consolidations can also result in decreased powder flow which can, in turn, lead to manufacturing issues (Lupo, Schütz, Riedl, Barletta, & Poletto, 2019).

Storage time also impacts flowability which can be assessed by monitoring powder flowability after samples have been stored under compressive stress for a long time (Schulze, 2021; Teunou & Fitzpatrick, 2000). During storage particles are continuously in various degrees of contact with other particles and prolonged contact can lead to changes in the physical bonds between particles or in the particles themselves (Prescott & Barnum, 2000). Finally, equipment variables can also impact on the behaviour of the powder with factors including the hopper width, orifice diameters, tube lengths, materials of construction contributing to the overall flow of powder through feeding, blending, transfers, feed frame and compaction processes. Thus, overall, powder flow is a complex attribute of bulk powders that is impacted by various material properties and process and equipment parameters making it difficult to predict reliably but is of high importance for manufacturability.

1.3.1. Powder flow measurement

There are several experimental methods available to measure powder flow, such as angle of repose, bulk density, Carr's compressibility index (Carr, 1965), Hausner ratio (Hausner, 1967), and shear cell testers to measure the flow function coefficient (FFc). Carr's compressibility index is calculated by subtracting the final bulk density (ρ_{bulk}) from the initial bulk density (ρ_{tapped}) and divided by the initial bulk density (see Eq 1-1). Free-flowing powders have a Carr's index smaller than 15, and cohesive powders have a Carr's index greater than 32 (Agarwal, Goyal, & Vaishnav, 2018; Gorle & Chopade, 2020; Reddy et al., 2014).

$$\text{Carr's Index} = \frac{\rho_{\text{tapped}} - \rho_{\text{bulk}}}{\rho_{\text{tapped}}} \times 100 \quad (1-1)$$

The Hausner ratio is calculated by dividing the tapped density (ρ_{tapped}) by the poured density (ρ_{bulk}) (see Eq 1-2). Thus, powders that have a ratio of less than 1.2 are classified as free-flowing, and powders that have a ratio greater than 1.5 are classified as cohesive (Agarwal et al., 2018; Gorle & Chopade, 2020; Reddy et al., 2014). The FFC is calculated by shear cell testers, which will be described in detail in Chapter 3.

$$\text{Hausner ratio} = \frac{\rho_{tapped}}{\rho_{bulk}} \quad (1-2)$$

These experimental methods are relatively time and product-consuming and are not comparable between each other, since different instruments measure different aspects of powder flow, making it challenging to achieve reproducibility. Whilst these methods are widely used in practice, no one test can be considered as a standard, representative measurement of powder flow (Faqih et al., 2006). The US Pharmacopeia states that there is no one method to measure powder flow that will completely and adequately characterise all the powder flow properties seen in the pharmaceutical industry (Sheehan, 2013).

1.3.2. Prediction of powder flowability

Given the impact of powder flow on pharmaceutical manufacturing, different approaches to estimating powder flow have been published. Sandler and Wilson studied packing efficiency by measuring the particle size of granular intermediates using Principal Component Analysis (PCA) (Sandler & Wilson, 2010). In this study, their model used size and shape distributions to predict flowability and density for granular material using Partial Least Squares (PLS). Megarry *et al.* used a big-data approach using a shear cell test data to better understand the typical flow properties of pharmaceutical materials, finding that the APIs have usually a poorer flow than excipients and granulates (Megarry, Swainson, Roberts, & Reynolds, 2019). Yu *et al.* determined the relevance of particle shape on powder flow prediction using a PLS approach to predict the FFC of binary blends, resulting in a performance of 70% evaluated using cross-validation sampling (Yu et al., 2011). Capece *et al.* explored how the granular Bond number correlates to the flow function coefficient and illustrated the complexity involved in predicting powder behaviour (Capece, Silva, Sunkara, Strong, & Gao, 2016; Nalluri & Kuentz, 2010). The applicability of their model in downstream drug development is limited since the model requires prior assumptions, such as 200 nm for the diameter of asperities of the particles, i.e., natural surface roughness, or the sphericity of the particles. Barjat *et al.* investigated statistical modelling techniques to predict powder flow of APIs for LIW feeders and demonstrated the feasibility of the prediction of powder flow from particle size and shape, yielding classification models

with a performance between 0.79 and 0.84 (Barjat et al., 2021). The prediction of powder flow of APIs is particularly critical in the feeding of the individual materials of a formulation to the blending step of a continuous process, i.e., continuous direct compression. While the requirements and assumptions of these models restrict their use in an industrial setting, the studies detailed here nonetheless establish the possibility of predicting powder flow using digital design approaches.

The prediction of powder flowability from physical properties of the material, i.e., particle size and shape would allow streamlining experimental design, through reducing the number of experiments and the energy consumption in comparison with traditional methods, i.e., shear cells testing. Therefore, the prediction of powder properties offers the benefit of bringing the pharmaceutical industry closer to its Net-Zero goals to reduce carbon emissions during development and manufacturing. As one example of these streamlining efforts, Quality by Design (QbD) builds quality into manufacturing by identifying and understanding the impact of the critical attributes of materials on the critical parameters of the manufacturing process. QbD was developed by Dr Joseph M. Juran, based on the premise that quality should be designed rather than tested, guaranteeing the quality of the final product. QbD is currently considered one of the most important enablers to reduce the time to market, ensure product quality and reduce waste and cost while achieving regulatory compliance (Reklaitis, Khinast, & Muzzio, 2010). Adopting the QbD principal has been promoted by the FDA as the systematic approach for pharmaceutical development (Chatterjee, 2013), for which the companies need to identify their quality goals and develop appropriate processes. In this context, Machine Learning (ML) models are emerging as a tool that can help predict if a pharmaceutical material or process will meet the required standards of quality and safety whilst minimizing extensive destructive testing and trial and error experimentation (Grangeia, Silva, Simões, & Reis, 2020).

Currently, the average cost of bringing a new medicine to market is \$1.3 billion and takes over 10 years of development (Wouters, McKee, & Luyten, 2020). Thus, the interest of pharmaceutical companies and regulatory bodies in the application of Artificial Intelligence (AI) to optimize pharmaceutical manufacturing and thereby reduce the cost of bringing a new medicine to the market has increased in recent years. In April 2019, the “Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) – Discussion Paper and Request for Feedback” was published by the FDA and followed up by “Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan” in January 2021 to discuss the current approach of the application of AI and ML to medical devices. This framework recognizes the valuable application of data science, AI, and ML in earlier human disease detection, accurate diagnosis, increasing knowledge and understanding of human physiology, and development of personalized medicines and diagnostics. Accordingly, this framework highlights the

importance of the forthcoming digitalization of the pharmaceutical industry (Food & Administration, 2019; U. Food & Administration, 2021). In the UK, the Made Smarter review highlights significant benefits of digital technologies such as AI/ML on the development and manufacture of medicines by implementing industrial digital technologies (Maier, 2017).

1.4. Artificial Intelligence and Machine Learning

AI was first coined as “the science and engineering of making intelligent machines” in a proposal at the Dartmouth Summer Research Project in 1956 (McCarthy, 1989). The field of AI was started at this summer workshop by the “founding fathers” John McCarthy, Alan Turing, Marvin Minsky, Allen Newell, and Herbert A. Simon (Roberts, 2016). Since then, many definitions have been used to describe it, and it has become more widespread in society. ML, as a subset of AI, can learn from data and it can predict further outcomes. AI and ML have proven to deliver excellent results in several fields, such as speech recognition, robot control, natural language processing, and computer vision. In pharmaceutical manufacturing, ML can be a useful tool to identify correlations between attributes that play an important part in achieving process outcomes and therefore, opening the possibility of optimizing development and making manufacture more efficient and thereby reducing the time for new therapies to reach patients.

ML algorithms can be classified into supervised and unsupervised learning where the difference between them is the presence or absence respectively of responses or outcomes in the training dataset. Furthermore, there are two types of supervised learning: classification and regression. In classification, the output of the model is a class or category, whereas, for regression, the output of the model is a continuous variable. Unsupervised learning is mainly used for clustering (exclusive and overlapping clustering, hierarchical clustering, probabilistic clustering), association rules (apriori algorithms), or dimensionality reduction (Principal Component Analysis (PCA)). Deep learning (DL) is a subset of ML that was recognized as one of the ten breakthrough technologies of 2013 (Zhou et al., 2021) and has been gaining interest over the last decade. DL is formed by artificial neuron networks (ANN), which are designed to mimic the activity of brain neurons, enabling the development of models that can learn to recognise patterns in digital representations of data (Ionescu, 2020).

ML has already been implemented in different stages of the pharmaceutical industry, from drug discovery to drug development to pharmaceutical preformulation. ML was applied to QSAR, a technique used to correlate the physicochemical properties of a compound with its biological or chemical activity, in the early stages of drug discovery for molecular design and selection (Wu et al., 2021; Zhang, Tan, Han, & Zhu, 2017). To optimize the preformulation step, ML has been shown to

have application to predict the solubility enhancement effect of hydrotrope molecules for example (Damiati, Martini, Smith, Lawrence, & Barlow, 2017). ML has also become popular in pharmaceutical formulation development (Damiati, 2020) to predict the drug release profile of minitables (Barmpalexis, Grypioti, Eleftheriadis, & Fatouros, 2018; M. M. Leane, Cumming, & Corrigan, 2003; Valizadeh, Pourmahmood, Mojarrad, Nemati, & Zakeri-Milani, 2009), to predict the tensile strength (Onuki et al., 2012; Takagaki, Arai, & Takayama, 2010), and to predict physical stability (Han et al., 2019).

1.5. Summary

In this thesis, ML approaches will be investigated as digital design tools to predict the powder properties of a range of model pharmaceutical materials and blends to evaluate their potential to reduce the time required to make key decisions during the development of pharmaceuticals products and processes. The implementation of ML models requires access to suitable, well defined training data and considerable effort is required to develop such data sets as they are not currently widely available. However, based on previous successful applications of ML summarized above there is considerable potential to deploy ML to provide rapid, reliable, and interpretable predictions of bulk properties including flow function coefficient, bulk density and wall friction angle, from easy to measure, standard data types that describe powder physical properties. Such predictions would allow rapid decision-making regarding manufacturing route selection, saving time and effort in early-stage development. The performance of the models included in this thesis is sufficiently high in that the use of the model would be expected to improve manufacturing efficiency. Moreover, the models could be extended to inform formulation optimization or even to provide a performance target for particle engineering efforts to develop “ideal” materials for the intended manufacturing route while also ensuring drug bioavailability. The results presented in this thesis could pave the way toward a rapid digital screening tool that can reduce pharmaceutical manufacturing costs.

1.6. References

- Agarwal, P., Goyal, A., & Vaishnav, R. (2018). Comparative Quality Assessment of Three Different Marketed Brands of Indian Polyherbal Formulation-Triphala Churna. *Biomedical Journal*, 2, 9.
- Barjat, H., Checkley, S., Chitu, T., Dawson, N., Farshchi, A., Ferreira, A., . . . Tobyn, M. (2021). Demonstration of the Feasibility of Predicting the Flow of Pharmaceutically Relevant Powders from Particle and Bulk Physical Properties. *Journal of pharmaceutical innovation*, 16(1), 181-196. Doi:10.1007/s12247-020-09433-5
- Barmpalexis, P., Grypioti, A., Eleftheriadis, G. K., & Fatouros, D. G. (2018). Development of a new aprepitant liquisolid formulation with the aid of artificial neural networks and genetic programming. *Aaps Pharmscitech*, 19(2), 741-752.
- Bellamy, L. J., Nordon, A., & Littlejohn, D. (2008). Effects of particle size and cohesive properties on mixing studied by non-contact NIR. *International journal of pharmaceuticals*, 361(1-2), 87-91.
- Capece, M., Silva, K. R., Sunkara, D., Strong, J., & Gao, P. (2016). On the relationship of inter-particle cohesiveness and bulk powder behavior: Flowability of pharmaceutical powders. *Int J Pharm*, 511(1), 178-189. Doi:10.1016/j.ijpharm.2016.06.059
- Carr, R. L. (1965). Evaluating flow properties of solids. *Chem. Eng.*, 18, 163-168.
- Chatterjee, S. (2013). *QbD considerations for analytical methods—FDA perspective*. Paper presented at the US IFPAC annual meeting.
- Chen, J., Sarma, B., Evans, J. M., & Myerson, A. S. (2011). Pharmaceutical crystallization. *Crystal growth & design*, 11(4), 887-895.
- Cleary, P. W. (2008). The effect of particle shape on simple shear flows. *Powder Technology*, 179(3), 144-163.
- Dai, S., Xu, B., Zhang, Z., Yu, J., Wang, F., Shi, X., & Qiao, Y. (2019). A compression behavior classification system of pharmaceutical powders for accelerating direct compression tablet formulation design. *International Journal of Pharmaceutics*, 572, 118742.
- Damiati, S. A. (2020). Digital pharmaceutical sciences. *Aaps Pharmscitech*, 21(6), 1-12.
- Damiati, S. A., Martini, L. G., Smith, N. W., Lawrence, M. J., & Barlow, D. J. (2017). Application of machine learning in prediction of hydrotrope-enhanced solubilisation of indomethacin. *International journal of pharmaceuticals*, 530(1-2), 99-106.
- Dürig, T., & Karan, K. (2019). Binders in wet granulation. In *Handbook of pharmaceutical wet granulation* (pp. 317-349): Elsevier.
- Faqih, A., Chaudhuri, B., Alexander, A. W., Davies, C., Muzzio, F. J., & Tomassone, M. S. (2006). An experimental/computational approach for examining unconfined cohesive powder flow. *International journal of pharmaceuticals*, 324(2), 116-127.
- Food, & Administration, D. (2019). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD).
- Food, U., & Administration, D. (2021). 'Artificial intelligence/machine learning (ai/ml)-based software as a medical device (SAMD) action plan. *US Food Drug Admin., White Oak, MD, USA, Tech. Rep*, 145022.
- Freeman, T., Brockbank, K., & Sabathier, J. (2017). *Characterising powder flow properties—the need for a multivariate approach*. Paper presented at the EPJ Web of Conferences.
- Gad, S. C. (2008). *Pharmaceutical manufacturing handbook: production and processes* (Vol. 5): John Wiley & Sons.
- Gorle, A. P., & Chopade, S. S. (2020). Liquisolid Technology: Preparation, Characterization and Applications. *Journal of Drug Delivery and Therapeutics*, 10(3-s), 295-307.
- Grangeia, H. B., Silva, C., Simões, S. P., & Reis, M. S. (2020). Quality by design in pharmaceutical manufacturing: A systematic review of current status, challenges and future perspectives. *European Journal of Pharmaceutics and Biopharmaceutics*, 147, 19-37.
- Han, R., Xiong, H., Ye, Z., Yang, Y., Huang, T., Jing, Q., . . . Ouyang, D. (2019). Predicting physical stability of solid dispersions by machine learning techniques. *Journal of Controlled Release*, 311, 16-25.

- Hatcher, L. E., Burgess, A. J., Payne, P., & Wilson, C. C. (2020). From structure to crystallisation and pharmaceutical manufacturing: the CSD in CMAC workflows. *CrystEngComm*, 22(43), 7475-7489.
- Hausner, H. H. (1967). *Friction conditions in a mass of metal powder*. Retrieved from
- Ionescu, D. (2020). Deep learning algorithms and big health care data in clinical natural language processing. *Linguistic and Philosophical Investigations*(19), 86-92.
- Kaerger, J. S., Edge, S., & Price, R. (2004). Influence of particle size and shape on flowability and compactibility of binary mixtures of paracetamol and microcrystalline cellulose. *European Journal of Pharmaceutical Sciences*, 22(2-3), 173-179.
- Kim, S., Wei, C., & Kiang, S. (2003). Crystallization process development of an active pharmaceutical ingredient and particle engineering via the use of ultrasonics and temperature cycling. *Organic process research & development*, 7(6), 997-1001.
- Klein, J., Höhne, D., & Husemann, K. (2003). The influence of air permeation on the flow properties of bulk solids. *Chemical Engineering & Technology: Industrial Chemistry-Plant Equipment-Process Engineering-Biotechnology*, 26(2), 139-146.
- Kottke, M. J., & Rudnic, E. M. (2002). Tablet dosage forms. In *Modern pharmaceuticals* (pp. 458-532): CRC Press.
- Ku, M. S. (2008). Use of the biopharmaceutical classification system in early drug development. *The AAPS journal*, 10(1), 208-212.
- Leane, M., Pitt, K., Reynolds, G., & Group, M. C. S. W. (2015). A proposal for a drug product Manufacturing Classification System (MCS) for oral solid dosage forms. *Pharmaceutical development and technology*, 20(1), 12-21.
- Leane, M. M., Cumming, I., & Corrigan, O. I. (2003). The use of artificial neural networks for the selection of the most appropriate formulation and processing variables in order to predict the in vitro dissolution of sustained release minitables. *Aaps Pharmscitech*, 4(2), 129-140.
- Lloyd, P. J., & Webb, P. J. (1986). The characterisation of the flow of aerated powders. *Particle & Particle Systems Characterization*, 3(4), 174-178.
- Lupo, M., Schütz, D., Riedl, E., Barletta, D., & Poletto, M. (2019). Assessment of a powder rheometer equipped with a cylindrical impeller for the measurement of powder flow properties at low consolidation. *Powder Technology*, 357, 281-290.
- Maier, J. (2017). Made smarter review. *UK Industrial Digitalisation Review*.
- McCarthy, J. (1989). Artificial intelligence, logic and formalizing common sense. In *Philosophical logic and artificial intelligence* (pp. 161-190): Springer.
- Megarry, A. J., Swainson, S. M. E., Roberts, R. J., & Reynolds, G. K. (2019). A big data approach to pharmaceutical flow properties. *INT J PHARMACEUT*, 555, 337-345. Doi:10.1016/j.ijpharm.2018.11.059
- Nalluri, V. R., & Kuentz, M. (2010). Flowability characterisation of drug–excipient blends using a novel powder avalanching method. *European Journal of Pharmaceutics and Biopharmaceutics*, 74(2), 388-396.
- Onuki, Y., Kawai, S., Arai, H., Maeda, J., Takagaki, K., & Takayama, K. (2012). Contribution of the physicochemical properties of active pharmaceutical ingredients to tablet properties identified by ensemble artificial neural networks and kohonen's self-organizing maps. *Journal of pharmaceutical sciences*, 101(7), 2372-2381.
- Prescott, J. K., & Barnum, R. A. (2000). On powder flowability. *Pharmaceutical technology*, 24(10), 60-85.
- Reddy, R. S., Ramachandra, C., Hiregoudar, S., Nidoni, U., Ram, J., & Kammar, M. (2014). Influence of processing conditions on functional and reconstitution properties of milk powder made from Osmanabadi goat milk by spray drying. *Small Ruminant Research*, 119(1-3), 130-137.

- Reklaitis, G., Khinast, J., & Muzzio, F. (2010). Pharmaceutical engineering science—New approaches to pharmaceutical development and manufacturing. *Chemical Engineering Science*, 21(65), iv-vii.
- Roberts, J. (2016). Thinking machines: The search for artificial intelligence. *Distillations*, 2(2), 14-23.
- Saddik, J. (2020). *Investigating and Predicting Tablet Sticking Using Powder Rheology, Thermal Analysis. And Molecular Simulation*. Long Island University, The Brooklyn Center,
- Sandler, N., & Wilson, D. (2010). Prediction of granule packing and flow behavior based on particle size and shape analysis. *Journal of pharmaceutical sciences*, 99(2), 958-968.
- Schaller, B. E., Moroney, K. M., Castro-Dominguez, B., Cronin, P., Belen-Girona, J., Ruane, P., . . . Walker, G. M. (2019). Systematic development of a high dosage formulation to enable direct compression of a poorly flowing API: A case study. *INT J PHARMACEUT*, 566, 615-630. Doi:10.1016/j.ijpharm.2019.05.073
- Schulze, D. (2021). Flow properties of bulk solids. In *Powders and Bulk Solids* (pp. 57-100): Springer.
- Shangraw, R. F. (1989). Compressed tablets by direct compression. *Pharmaceutical dosage forms: Tablets, 1*, 195-246.
- Sheehan, C. (2013). General chapters:< 1174> Powder flow. *USP29-NF24, 3017*.
- Sun, C., & Grant, D. J. (2001). Effects of initial particle size on the tableting properties of L-lysine monohydrochloride dihydrate powder. *International journal of pharmaceuticals*, 215(1-2), 221-228.
- Takagaki, K., Arai, H., & Takayama, K. (2010). Creation of a tablet database containing several active ingredients and prediction of their pharmaceutical characteristics based on ensemble artificial neural networks. *Journal of pharmaceutical sciences*, 99(10), 4201-4214.
- Teunou, E., & Fitzpatrick, J. (2000). Effect of storage time and consolidation on food powder flowability. *Journal of Food Engineering*, 43(2), 97-101.
- Trementozzi, A. N., Leung, C.-Y., Osei-Yeboah, F., Irdam, E., Lin, Y., MacPhee, J. M., . . . Zawaneh, P. N. (2017). Engineered particles demonstrate improved flow properties at elevated drug loadings for direct compression manufacturing. *Int J Pharm*, 523(1), 133-141. Doi:10.1016/j.ijpharm.2017.03.011
- Valizadeh, H., Pourmahmood, M., Mojarrad, J. S., Nemati, M., & Zakeri-Milani, P. (2009). Application of artificial intelligent tools to modeling of glucosamine preparation from exoskeleton of shrimp. *Drug Development and Industrial Pharmacy*, 35(4), 396-407.
- Waknis, V., Chu, E., Schlam, R., Sidorenko, A., Badawy, S., Yin, S., & Narang, A. S. (2014). Molecular basis of crystal morphology-dependent adhesion behavior of mefenamic acid during tableting. *Pharmaceutical research*, 31(1), 160-172.
- Wouters, O. J., McKee, M., & Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama*, 323(9), 844-853.
- Wu, Z., Zhu, M., Kang, Y., Leung, E. L.-H., Lei, T., Shen, C., . . . Hou, T. (2021). Do we need different machine learning algorithms for QSAR modeling? A comprehensive assessment of 16 machine learning algorithms on 14 QSAR data sets. *Briefings in bioinformatics*, 22(4), bbaa321.
- Yasir, M., Asif, M., Kumar, A., & Aggarwal, A. (2010). Biopharmaceutical classification system: An account. *International Journal of PharmTech Research*, 2(3), 1681-1690.
- Yu, W., Muteki, K., Zhang, L., & Kim, G. (2011). Prediction of Bulk Powder Flow Performance Using Comprehensive Particle Size and Particle Shape Distributions. *J. Pharm. Sci*, 100(1), 284-293. Doi:10.1002/jps.22254
- Zhang, L., Tan, J., Han, D., & Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today*, 22(11), 1680-1685.
- Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A., . . . Summers, R. M. (2021). A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5), 820-838.

2. Aims and objectives

2.1. Aims

The purpose of the work described herein is to obtain a greater understanding of the relationship between particle and bulk properties of pharmaceutical powders to enable greater opportunity to couple primary and secondary manufacturing more closely. This, for example, could better enable the establishment of fully integrated end-to-end continuous processes, e.g., MIT-Novartis (Mascia et al., 2013), particularly focused on direct compression (DC). The research was performed using Artificial Intelligence (AI) modelling approaches, trained on a novel set of data covering a wide range of experimental data representing key attributes of pharmaceutical materials including particle size, particle shape, bulk density, surface area, surface energy, flow function coefficient, and wall friction angle. A number of different AI algorithms have been assessed to predict the relationships between these attributes and manufacturability, including machine learning (ML) and deep learning (DL). These models were employed to predict and to better understand the relationship between characteristic particle and bulk properties arising from upstream crystallisation processes to enable rapid decision-making regarding the suitability of the compound for DC. An additional aspect of this work was to assess the implementation of the models in an industrial context to demonstrate the applicability and build trust in these methods in real applications. Finally, the goal of enabling the reverse engineering the particle properties of an ideal API to ensure its suitability for DC by design was explored.

2.2. Objectives

To achieve these aims, the following objectives were sought:

2.2.1. Prediction of powder flow of pharmaceutical materials using ML models (Chapter 4).

- i. Establish a dataset of characteristic particle properties, including particle size distribution, particle shape, surface area, and surface energy; and bulk properties, including bulk density, and powder flow for an array of pharmaceutically relevant materials.
 - ii. Data curation, including creating, organising, and maintaining datasets, of the experimental measurement data.
 - iii. Data interrogation and data visualisation of the experimental data.
-

- iv. Develop a workflow for applying data-driven models spanning data generation through data visualisation, data curation, model training, model evaluation and model interpretability.
- v. Select the best classification and regression ML model to predict the powder flow of pharmaceutical materials through algorithm screening by comparing metrics of evaluation.
- vi. Tune all hyperparameters of the best-performing model.
- vii. Implement Shapley Additive exPlanations (SHAP) parameters to assist model interpretability to ensure an understanding of how the model makes the predictions and draw recommendations for designing direct-compressible APIs.

2.2.2. Prediction of powder flow of pharmaceutical blends from physical particle properties of the API using ML models (Chapter 5).

- i. Set up model matrix to predict powder flow of formulations from physical properties of the API.
- ii. Data interrogation of particle size and particle shape of APIs measured by static image analysis, dynamic image analysis, and laser diffraction analysis.
- iii. Comparison of the performance of different algorithms for flow function coefficient (FFc) and wall friction angle (WFA) prediction for classification and regression models.
- iv. Implement SHAP parameters to assist the interpretability of the models and to guide formulators in the design of viable formulations for continuous DC.
- v. Assess the real-world application of the modes in collaboration with Roche and deploy a user-friendly interface to implement the ML models on-site, accompanied by training material for formulators.

2.2.3. Deep learning (DL) approaches for the investigation of the robustness of particle size and shape descriptors (Chapter 6)

- i. Data generation using static image analysis to produce images to train DL models of small samples of pharmaceutical powders.
- ii. Assess image cropping as an appropriate method to increase image processability while maintaining the representation of bulk properties.

- iii. Decide the Neural Network DL model architecture and tune the hyperparameters.
- iv. Applying the workflow developed in Chapter 4 to train the DL model and evaluate its performance.
- v. Elucidate the feasibility of a DL model for the prediction of powder flow.

2.2.4. Interrogation of particle and bulk property descriptors in the context of machine learning and prediction of pharmaceutical materials (Chapter 7).

- i. Apply the workflow developed in Chapter 4 to build machine learning models for the classification or prediction of bulk density, surface area, and surface energy.
- ii. Select the best classification and regression ML model to predict the bulk density, surface area, and surface energy of pharmaceutical materials through algorithm screening by comparing metrics of evaluation.
- iii. Assess the manufacturability of pharmaceutical materials based on the classification of their bulk density.
- iv. Asses the bioavailability of pharmaceutical materials based on the classification of their surface area.

2.3. References

Mascia, S., Heider, P. L., Zhang, H., Lakerveld, R., Benyahia, B., Barton, P. I., . . . Jamison, T. F. (2013). End-to-end continuous manufacturing of pharmaceuticals: integrated synthesis, purification, and final dosage formation. *Angewandte Chemie International Edition*, 52(47), 12359-12363.

3. Materials and methods

The materials and methods used to achieve the aims and objectives of this thesis are described herein. The particle and bulk properties of over one hundred pharmaceutical powders were analysed to assemble datasets. Machine Learning (ML) models were built using these particle and bulk properties datasets to predict powder flow properties, i.e., flow function coefficient, bulk density, wall friction angle, and surface properties, i.e., surface area and surface energy. The ML models were evaluated using different sampling methods to calculate classification and regression metrics. The predictions made by the ML models were interrogated using interpretability methods to gain a deeper understanding of the impact of particle attributes on bulk properties, and therefore, on downstream processability.

3.1. Materials

Table 3-1 shows the materials that have been included in the development of ML and DL models for the prediction of powder flow, bulk density, and surface properties. The materials include APIs and excipients that can be commonly found in pharmaceutical formulations. The properties of the materials presented in this thesis refer to the powder properties of the grade materials, and in any case, they refer to the intrinsic properties of the materials.

Table 3-1: The materials included in the development of data-driven models.

Material	Supplier	Material	Supplier
4-aminobenzoic acid	Sigma-Aldrich	Ibuprofen 70	Sigma-Aldrich
Ac-Di-Sol	Dupont	Lactose	Sigma-Aldrich
Acetazolamide	Sigma-Aldrich	Lidocaine	Sigma-Aldrich
Affinisol	Dupont	Lubritose AN	Kerry
Aspirin	Sigma-Aldrich	Lubritose Mannitol	Kerry
Avicel PH-101	Dupont	Lubritose MCC	Kerry
Avicel PH-102	Dupont	Lubritose PB	Kerry
Benecel K100M	Dupont	Lubritose SD	Kerry
Benzoic acid	Sigma-Aldrich	Magnesium Stearate	Roquette
Benzydamine hydrochloride	Sigma-Aldrich	Magnesium Stearate	Sigma-Aldrich
Bromhexine hydrochloride	Sigma-Aldrich	Mefenamic acid	Sigma-Aldrich
Caffeine	Sigma-Aldrich	Methocel MC2	Colorcon
Calcium carbonate	Sigma-Aldrich	Microcel MC-102	Roquette
Calcium phosphate dibasic	Sigma-Aldrich	Microcel MC-200	Roquette
Cellulose	Sigma-Aldrich	Nimesulide	Sigma-Aldrich
Croscarmellose Na	Dupont	Paracetamol Granular Special	Sigma-Aldrich
D-glucose	Sigma-Aldrich	Paracetamol Powder	Sigma-Aldrich
D-mannitol	Sigma-Aldrich	Pearlitol 300DC	Roquette
D-sorbitol	Sigma-Aldrich	Plasdone povidone	Ashland
Dropropizine	Sigma-Aldrich	Plasdone K29/32	Ashland
FastFlo 316	Dupont	Phenylephedrine	Sigma-Aldrich
FlowLac 90	Meggle Pharma	Roxithromycin	Sigma-Aldrich
Granulac 140	Meggle Pharma	S-carboxymethyl-L-cysteine	Sigma-Aldrich
Granulac 230	Meggle Pharma	Soluplus	BASF
HPMC	Sigma-Aldrich	Span 60	Sigma-Aldrich
Ibuprofen 50	BASF	Stearic acid	Sigma-Aldrich

3.2. Particle size and shape measurements

The impact of the critical quality attributes (CQA) on the drug substance must be understood to achieve a robust manufacturing process. The lack of understanding of such impact can jeopardise the quality and efficacy of the drug product. In oral dosage forms, particle size and particle shape are the most influential CQA (Chattoraj et al., 2018).

Particle size is defined by the radius or diameter of the particle of study. For spherical particles, the diameter is that characteristic of a sphere. The most widely used method of defining particle size is the equivalent diameter, which is the diameter of a sphere with equivalent physical properties or geometry (Augsburger & Hoag, 2016).

The shape of a particle is defined by its external morphology. Particle shape can influence powder properties such as powder flowability, compactability, content uniformity, dissolution, drug release, and even particle size analysis. Particles with different shapes can have the same size but very different properties. Commonly used parameters to describe particle shape are aspect ratio, sphericity, circularity, and convexity (Peck, Baley, McCurdy, & Banker, 1989).

3.2.1. Dynamic image analysis: QICPIC® Sympatec

Dynamic image analysis is a technique that captures the size and shape of the particles when they are in motion. To perform this measurement, a few grams of sample are required, enabling the analysis of a large number of particles. The measurement of this large number of particles is useful to obtain statistically significant results. To initiate the measurement, the powder sample must be dispersed and streamed in front of the camera. The multiple settings available to disperse the sample hinder the reproducibility of the measurement (Gamble, Tobyn, & Hamey, 2015). For this thesis, the QICPIC® RODOS/L VIBRI/L has been used. The RODOS/L dispersion unit of the image analytical sensor QICPIC® ensures the appropriate dispersion of fine and coarse particles. The dosing of the sample before it is streamed in front of the QICPIC® high-speed camera is performed by the vibratory feeder VIBRI/L (see Fig 3-1).

The QICPIC® characterises the particle size distribution (PSD) of the powder sample. The PSD is typically reported using volume-based percentile metrics, i.e., $D[v, 0.1]$ (μm) where D is distribution and v is volume, and $D[v, 0.1]$ (μm) indicates the diameter which 10% of the particles are smaller than. Similarly, $D[v, 0.5]$ (μm) indicates the diameter which 50% of the particles are smaller than, and $D[v, 0.9]$ (μm) indicates the diameter which 90% of the particles are smaller than. Another common metric

to report the PSD is the Sauter Mean Diameter (SMD or $D[3,2]$). The SMD is an average particle size, which can be defined as the diameter of the sphere that has the same volume-to-surface area ratio as the particle. The QICPIC[®] also characterises particle shape, i.e., sphericity, aspect ratio, and convexity. The shape descriptors are also reported using percentile metrics.



Figure 3-1: The QICPIC[®] – RODOS dynamic image analysis instrument.

The QICPIC[®] uses the equivalent circle of the equal projection area (EQPC), and the diameter of the EQPC (X_{EQPC}) is calculated following Eq 3-1, calculated by the squared root the area of the particle (A) divided by π .

$$X_{EQPC} = 2\sqrt{A/\pi} \quad (3-1)$$

The QICPIC® also provides the Feret diameters (X_{Feret}), which are a group of diameters derived from the distance of two tangents to the contour of the particle in a specific orientation, also defined as the distance between the two parallel planes restricting the object perpendicular to that direction (see Fig 3-2). The maximum Feret diameter ($X_{Feret,max}$) is always larger than the EQPC, whereas the minimum Feret diameter ($X_{Feret,min}$) is always smaller than the EQPC.

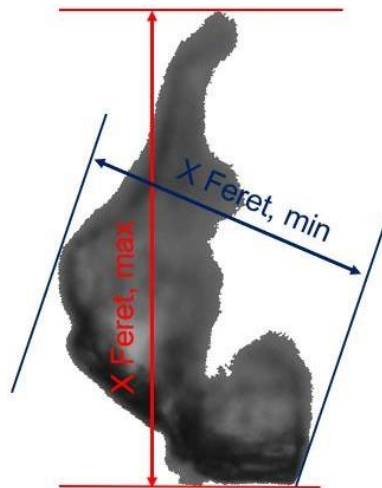


Figure 3-2: The Feret diameter, the distance between the two parallel planes that restrict the object.

The aspect ratio is calculated from the Feret diameter (see Eq 3-2). The aspect ratio takes values from 0 to 1. High values of aspect ratio indicate that the shape of the particles is close to a sphere, whereas small values of aspect ratio indicate that the particles are elongated.

$$Aspect\ ratio = \frac{X_{Feret,min}}{X_{Feret,max}} \quad (3-2)$$

Sphericity is calculated from the ratio of the perimeter of EQPC (P_{EQPC}) to the real perimeter of the particle (P_{real}) (see Eq 3-3). As the aspect ratio, sphericity values range from 0 to 1, with being 1 a sphere.

$$Sphericity = \frac{P_{EQPC}}{P_{real}} \quad (3-3)$$

Another descriptor that is commonly used to describe the shape of the particles is convexity. The convexity, defined as the ratio of the projection of the area of the particle to the area of the convex hull (see Eq 3-4), describes the compactness of the particle. The convexity values range from 0 to 1, being 1 when there are no concave regions. Figure 3-3 shows that the convexity is the ratio of the projection area (A) and the area of the convex hull (A+B).

$$\text{Convexity} = \frac{A}{A+B} \quad (3-4)$$

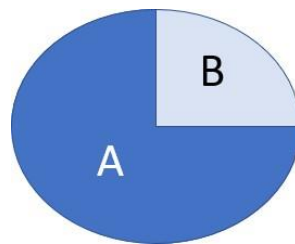


Figure 3-3: The convexity of a particle calculated as the ratio of the projection area (A) to the area of the convex hull (A+B)

To carry out the particle characterisation using the QICPIC[®], the powder sample was streamed through the measuring volume controlled by a high-speed camera, which leads to a random orientation of the particles that can be used to characterise the size and shape accurately. The measuring range selected was M7, which can accurately analyse particles with a range of size from 30 to 8,665 μm . The feeder VIBRI/L was modified to a feed rate of 40%, a pressure of 1 bar, and a frame of 100 Hz. The measurements were performed in triplicate. Further information about dynamic image analysis can be found in the ISO guidance 13322-2:2021, “Particle size analysis — Image analysis methods — Part 2: Dynamic image analysis methods”.

3.2.2. Static image analysis: Morphologi® G3, Malvern

Static image analysis involves the imaging of particles that have been pre-dispersed onto a glass slide. This approach provides several benefits. First, the particles are presented in their most mechanically stable position, which means that their dominant faces are always presented to the camera. Second, as the particles are static, instruments are better able to obtain higher-resolution morphological images and thereby increase the capacity for thresholding of the particles within the sample.

Static systems can deal with a wide range of particles, although highly three-dimensional samples can be challenging because of limitations in the depth of field. This challenge can be overcome by taking images at multiple focal places and combining the “in-focus” element of each plane into an amalgamated image, a process often referred to as vertical plane stacking or z-stacking. An important consideration of static image analysis is ensuring the adequate position of the particles on the plate once they have been dispersed. The preparation of a sample that contains elongated, fine particles can be particularly challenging. Additionally, static image analysis is more time-consuming than dynamic image analysis. A typical static image analysis measurement requires between 40 minutes and 3 hours, depending on certain specifications of the procedure, i.e., analysis area, and z-stacking. On the other hand, static image analysis enables a greater ability to conduct an in-depth analysis of the particles than dynamic image analysis (Clarke et al., 2019). Further information about static image analysis can be found in the ISO guidance document ISO 13322-1:2014 (Standard, 2004).

The Malvern Morphologi® G3 (Malvern Panalytical, Malvern, UK) was used in this thesis to perform static image analysis (see Fig 3-4). This instrument is a high-resolution analytical tool that can be used to measure the morphological characteristics of particles. It provides both number-based statistics, which are generated on individual particles and on the sample allowing the detection of fines; and volume-based statistics, which are calculated based on the importance of the particle in the sample in relation to its volume.

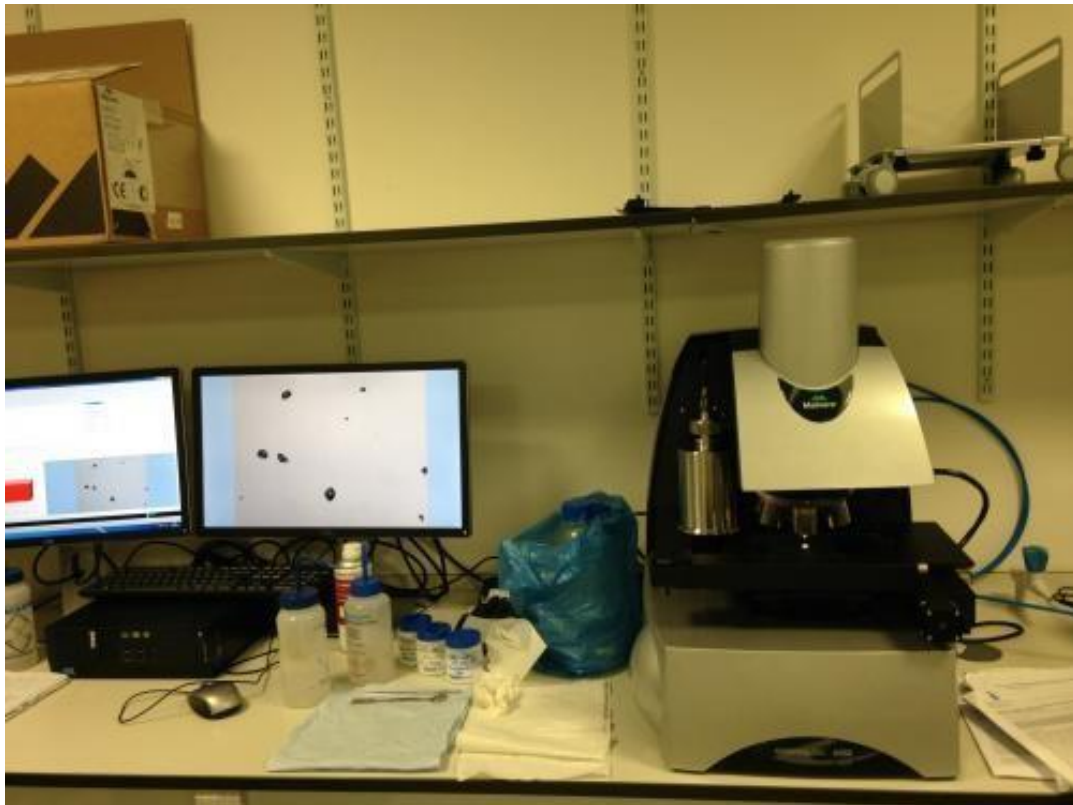


Figure 3-4: The Morphologi® G3 - Malvern G3.

Statistic image analysis captures a two-dimensional image of a particle and calculates size and shape descriptors. The Morphologi® G3 calculates the particle size diameter from the Circle Equivalent (CE) diameter (see Fig 3-5). The particle captured as a two-dimensional image is converted into a circle that has the same area as the particle. Then, the diameter of this circle is measured and reported as the diameter of the particle.

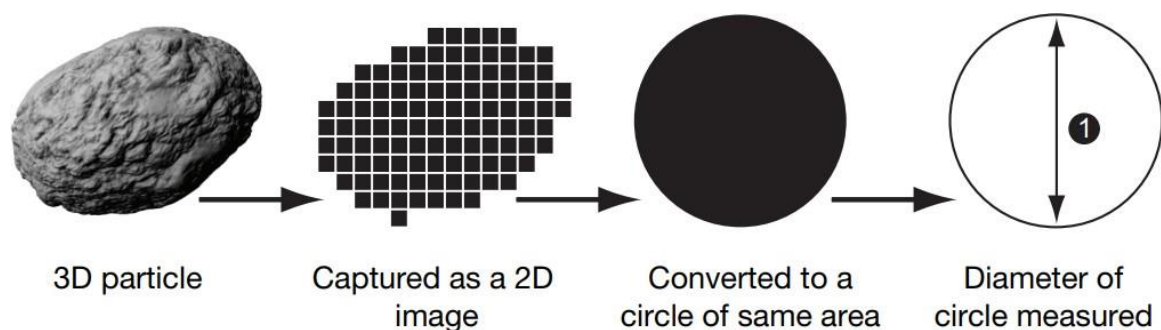


Figure 3-5: The calculation of the CE diameter of a three-dimensional particle captured as a two-dimensional image. Taken from the Morphologi® G3 User Manual.

The Morphologi[®] G3 software reports the statistics of the distribution of the descriptors calculated from the CE diameter. The arithmetic mean of the particle size distribution, usually denoted as D[1,0], and other means that take into account the volume weighting or area are calculated following Eq 3-5, where v_i represents the frequency of occurrence of particles in size class i , with a mean diameter of d_i shows the method to calculate the mean and other descriptors of the PSD.

$$D [m, n] = \left[\frac{\sum v_i d_i^{m-3}}{\sum v_i d_i^{m-3}} \right]^{\frac{1}{m-n}} \quad (3-5)$$

Volume-based particle size descriptors reported by the Morphologi[®] G3, i.e., D[v, 0.1], D[v, 0.5], D[v, 0.9], and D[3,2] are the default measurements reported by the software, similar to the QICPIC[®]. The Morphologi[®] G3 also provides shape data. Circularity quantifies how close to a circle the shape of the particle is. Circularity is calculated from the ratio of the perimeter of a circle with the same area as the particle to the perimeter of the actual particle. Therefore, the Morphologi[®] G3 circularity is equivalent to the QICPIC[®] sphericity. Fig 3-6 shows four particles with different circularity values to illustrate how the shape of these particles changes with the circularity value. Particles that have a high value of circularity (close to 1) exhibit a similar shape to a sphere. In contrast, particles that have a low value of circularity (close to 0) have an elongated shape.

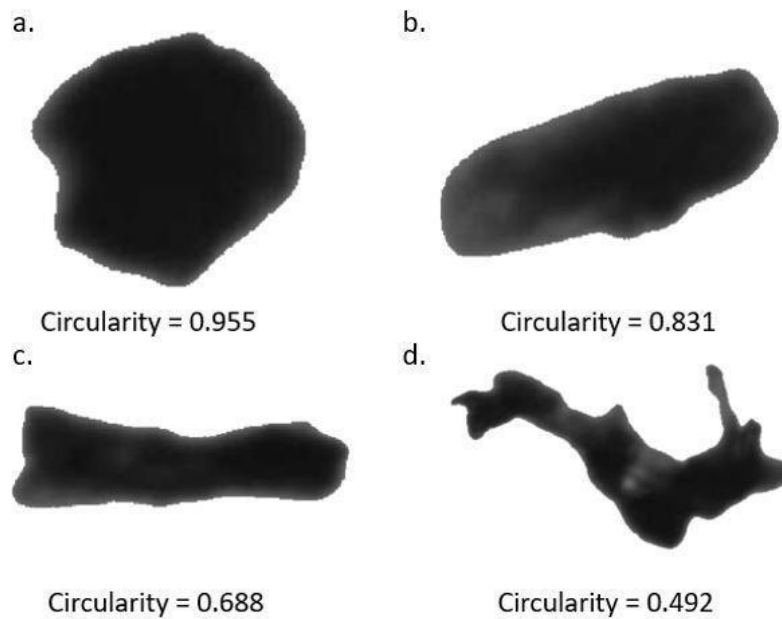


Figure 3-6: The circularity values for different shapes.

The Morphologi[®] G3 also reports High Sensitivity (HS) Circularity, which reduces the impact of the variations in the relationship between the area and the perimeter. Eq 3-6 shows that the HS Circularity is calculated by calculating the square of the circularity, where A is the particle area and P is the particle perimeter.

$$HS\ Circularity = \frac{4\pi A}{P^2} \quad (3-6)$$

Convexity is a measurement of the edge roughness of a particle. The Morphologi[®] G3 measures convexity following the same principle as the QICPIC[®]: the convex hull perimeter is divided by the actual perimeter of the particle. Likewise, the shape descriptors detailed previously, convexity takes values from 0 to 1. Fig 3-7 shows three particles with different convexity values to illustrate the differences in shape. Small variations in convexity result in major differences regarding the overall shape. For instance, two particles with a difference in convexity of 0.004 can have completely different shapes (see Figs 3-7(b) and 3-7(c)).

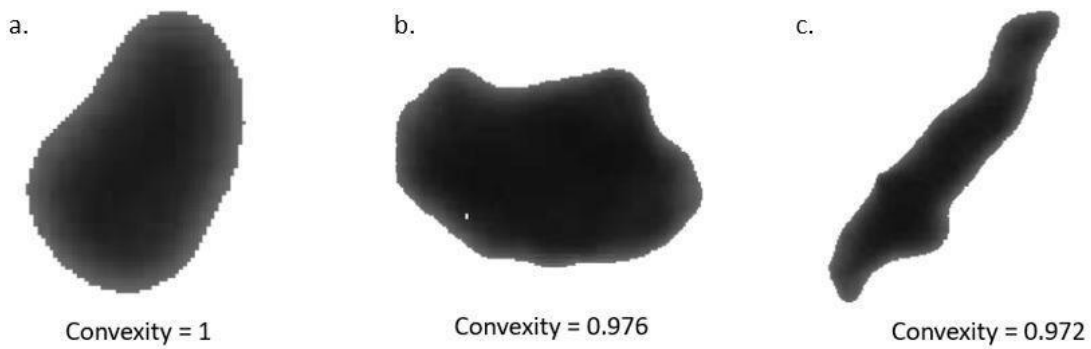


Figure 3-7: The convexity values for particles with different shapes.

Elongation is defined as the subtraction of the aspect ratio from 1. Fig 3-8 shows four particles with different elongation values to illustrate their differences in shape. Particles that have low values of elongation have a shape similar to a sphere, whereas particles with high values of elongation have an elongated shape.

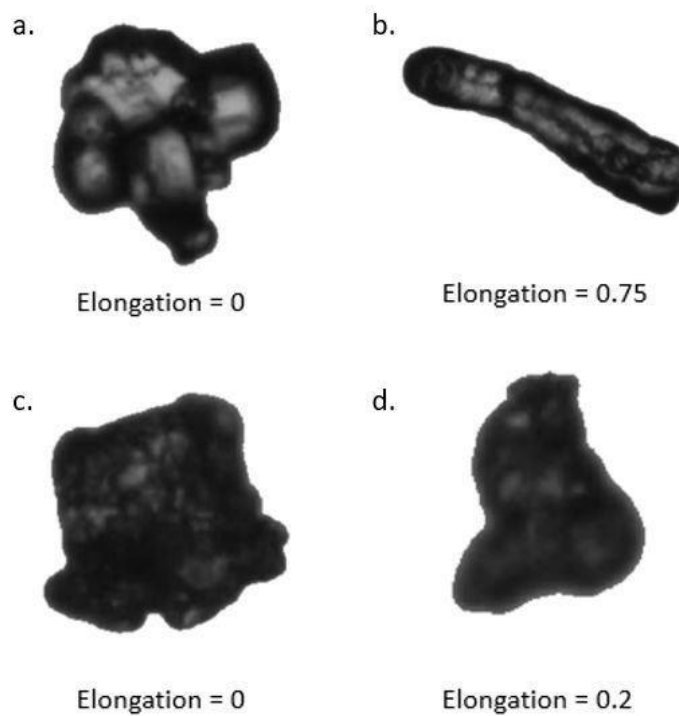


Figure 3-8: The elongation of four particles with different shapes and sizes.

The measurement parameters are set in the Standard Operating Procedure (SOP). For this work, the sample dispersion unit (SDU) is selected as the sample carrier for all the measurements. The powder will be dispersed from the SDU onto a glass plate at 1 bar for 20 milliseconds of injection time and a

settling time of 60 seconds. The light source selected is diastopic (bottom light). The magnification of the optics (2,5x, 5x, 10x, 20x, or 50x) was selected based on the estimated particle size. For this thesis, one of the following optics magnifications were selected:

- i. 2,5x, suitable for particle size between 13 and 1000 μm ,
- ii. 5x, suitable for particle size between 6.5 and 420 μm , or
- iii. 10x, suitable for particle size between 3.5 and 210 μm .

Finally, a 40% frame overlap was used to ensure that all particles are detected. Once the SOP was set up, the appropriate spatula was used to sample the powder from the bulk. The selection of the volume of the spatula depends on the powder characteristics. For a very fine powder, a small volume spatula (1 to 5 mm^3) would be more suitable than a bigger spatula (11 to 19 mm^3).

3.2.3. Laser diffraction: Mastersizer® 3000, Malvern

Laser diffraction is one of the most used particle characterization techniques as it can be validated for GMP testing (ISO, 2009). Laser diffraction determines particle size by measuring the variation of the light scatter angle as the laser beam passes through the sample. Small particles scatter light at a larger angle relative to the laser beam than large particles. The angular variation is analysed to calculate the size of the particle that created that scattering pattern. Laser diffraction used the Mie theory of light scattering. The Mie theory was proposed by Gustav Mie as a solution to Maxwell's equation describing the scattering of an electromagnetic plane wave by a sphere (Wriedt, 2012).

There are three key aspects of method development for laser diffraction analysis to ensure reliable results: sampling, dispersion, and measurement conditions. Usually, sampling is more significant for larger particles, whereas dispersion is more significant for smaller particles. The dispersion of the powder follows one of the following mechanisms:

- i. Velocity gradients caused by shear stress
- ii. Particle-to-particle collisions
- iii. Particle-to-wall collisions

Particle-to-wall collisions require more energy than particle-to-particle collisions, which require more energy than dispersion by velocity. With the increase in energy requirement, the effectiveness

increases too, and hence, the selection of the mechanism needs to be balanced between effectiveness and the risk of damaging a particle.

In this thesis, the laser diffraction particle size characterization was carried out using the Mastersizer® 3000 (Malvern Panalytical, Malvern, UK) with the Aero S dry dispersion unit (see Fig 3-9).



Figure 3-9: The Mastersizer® 3000 and the Aero S dispersion unit.

An SOP needs to be created to proceed with the measurement. For the Mastersizer® 3000, the feed rate required to be manually tested using high pressures to observe high obscurations. The obscuration range is set depending on particle size, and thus, for small particles, the obscuration can range from 0.5 to 3%, whereas for larger particles it will range from 0.5 to 6%. Once the feed rate was set, the powder is dispersed, and the degree of dispersion is controlled by the primary air pressure. Segregation might occur with very free-flowing powders that have wide particle size distribution. If segregation occurs, the measurement should be repeated.

3.3. Surface area and surface energy: Surface Energy Analyzer®

The surface area and surface impact the behaviour of pharmaceutical materials during manufacturing. Shah *et al.* reported that attributes such as cohesion, are affected by surface area and surface energy (Shah *et al.*, 2014), and many studies have been published regarding the relationship between the surface energy and compactability of powders (Alyami, Dahmash, Bowen, & Mohammed, 2017; Fichtner, Mahlin, Welch, Gaisford, & Alderborn, 2008).

Different techniques are available to measure the surface area and the surface energy of powders, such as sessile drop or dynamic contact angle measurements (Kwok, Gietzelt, Grundke, Jacobasch, &

Neumann, 1997). Inverse gas chromatography (iGC) was created as an alternative analysis to categorise the surface by measuring quantities of vapours that are injected into the glass column packed with the material of interest. It was first introduced in 1941 by the Nobel Prize winners Martin and Synge, and it was commercialised in the 1950s. However, iGC did not become popular until the 1970s when this technique started to be used to characterise the bulk properties of powders.

iGC is a useful technique for the characterization of surface properties within a wide range of temperature and humidity. The materials that are being analysed with the iGC are located in the stationary phase, and their properties influence the retention data (retention time, retention volume) of the test solutes that are passed through. The retention data will be transformed into parameters that describe the properties of interest, in this case, surface area and surface energy (Voelkel, 2021).

The surface energy measures the wettability of a powder. This property indicates the maximum surface tension of liquid that can wet a powder under ideal conditions. Materials that have high surface energy are easier to adhere to than materials that have low surface energy. Therefore, the surface energy indicates the cohesion within a powder, which is a critical descriptor to understanding compactability. Fig 3-10 shows how as the surface energy increases, the wettability, the cohesion, and the adhesion of the powder increase.

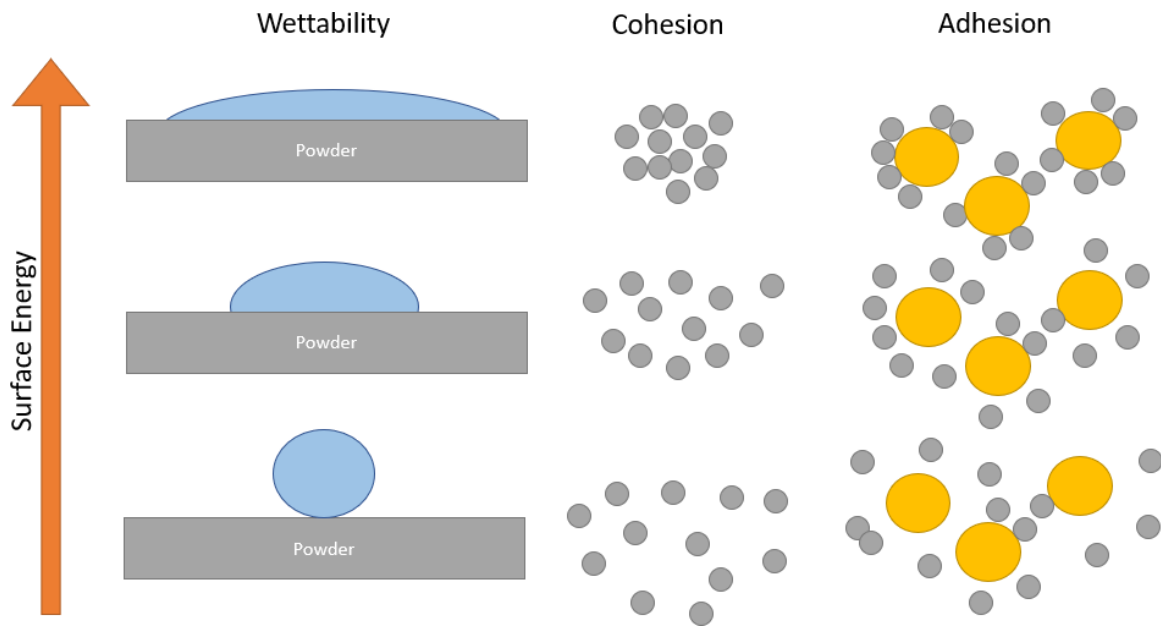


Figure 3-10: Wettability, cohesion within the powder, and adhesion increase as the surface energy increases.

Following the Owens-Wendt model, the surface energy can be split into two components: polar or specific (acid-base) (γ^p) and dispersive (γ^D) (see Eq 3-7) (Kozbial et al., 2014). There are two main methods to calculate the dispersive surface energy: Schultz (Schultz, Lavielle, & Martin, 1987) and Dorris-Gray (Dorris & Gray, 1980). Then, the specific surface energy is calculated from the dispersive energy using acid-base liquid probes.

$$\gamma = \gamma^p + \gamma^D \quad (3-7)$$

In this thesis, the Surface Area Analyzer® (SEA) has been used to measure the surface area and the surface energy (see Fig 3-11). To run this measurement, approximately 500 mg of sample were introduced into the silanised glass column, through which a probe will be injected using an inert carrier gas, normally helium, argon, and nitrogen. The probe used for dispersive interaction is a non-polar solvent, and the probe used for specific interactions is a polar solvent. In both cases, the properties of the solvent, i.e., acidity, molecular area, and polarity, are known. After the sample was introduced into the equipment, the Flame Ionization Detector (FID) was switched on. The FID measures the destruction of the organics and the formation of ions. The sensitivity of the FID to hydrocarbons is high; however, it is very low to the water. The interaction between the solvent and the sample resulted in a peak, and the time required to generate this peak was the retention time. From the retention

time of the solvent in the powder, the physicochemical properties (i.e., surface energy) are calculated. The peak was reported using two descriptors: peak max and peak com. Peak max is the maximum FID signal in the peak, whereas peak com is the centre-of-mass peak.



Figure 3-11: The inverse Gas Chromatography-Surface Energy Analyzer (iGC-SEA).

3.4. Bulk measurements

3.4.1. Bulk density: FT4® Powder Rheometer, Freeman Technology

Bulk density is calculated by the ratio of the mass of the powder to the volume that occupies, including interparticle void volume, and it is usually expressed in g/cm^3 , kg/cm^3 , or $\text{g}/100 \text{ ml}$. Bulk density is calculated using Eq 3-8, where the mass in grams (M) is divided by the volume in milliliters (V_o) (Pharmacopoeia, 2002).

$$\text{Bulk density} = \frac{M}{V_o} \quad (3-8)$$

Three different bulk densities are defined: aerated, poured, and tap density (López Córdoba & Goyanes, 2017). The bulk density of a powder is impacted by the way the particles fill the voids when they collapse with each other, which depends on the PSD, the texture of the surface, and the interparticle forces. Bulk density determines how much powder can fit into the feeder, the hopper, the capsule filler or the tablet press, influencing the manufacturing process and hence, the final product quality (Fitzpatrick, 2013). Therefore, bulk density should be considered an essential parameter when developing a pharmaceutical formulation (Shenoy et al., 2015).

For this work, the bulk density was measured with the FT4[®] Powder Rheometer (Freeman Technology Ltd.). The powder samples were poured into a 25 mm x 10 ml split vessel until achieving a volume above the split level. The first step of the measurement is the conditioning cycle, where the stress is removed from the sample and the weight of the sample is recorded. After conditioning, the vessel is split, and the excess of powder is removed to achieve the desired volume (10 ml). The weight is again recorded, which is used to calculate the bulk density (see Eq 3-9). The tests were done in triplicate and the average value was calculated.

$$\text{Conditionated bulk density} = \frac{\text{Split mass after conditioning}}{\text{Sample volume}} \quad (3-9)$$

3.4.2. Shear Cell Measurements: FT4[®] Powder Rheometer, Freeman Technology

Powder flowability plays a crucial part in the tablet manufacturing process. Good flow properties are essential to assuring efficient tableting operation. In this work, the powder flow was measured using the Freeman FT4[®] Powder rheometer (Freeman Technology, Malvern, UK). This instrument was designed to enable the analysis of pharmaceutical powders simulating industrial conditions by measuring the powder's resistance to flow whilst in motion.

As for the measurement of the bulk density, the first step of the measurement is to prepare the sample in a conditioning cycle, carried out by a 23.5 mm diameter x 6 mm wide blade that rotates upwards and downwards through the powder. The conditioning cycle eliminates the stress that might have been caused in storage or while handling the powder and achieves a homogeneous sample. To do the measurement, a 10 ml volume x 25 mm diameter split vessel is used. Once the vessel is assembled (see Fig 3-12), the powder is poured into the vessel and the blade starts rotating upwards and downwards.



Figure 3-12: The 10 ml x 25 mm split vessel used to carry out the powder flow measurement.

After the conditioning cycle, the blade is replaced by a vented piston to compress the powder to predetermined normal stress. For this thesis, a normal stress of 9 kPa was used. When the powder is consolidated, the vessel is split into two, and the excess of powder is removed. The vented piston is replaced by a 24 mm diameter blade that will perform the shear cell test (see Fig 3-14).

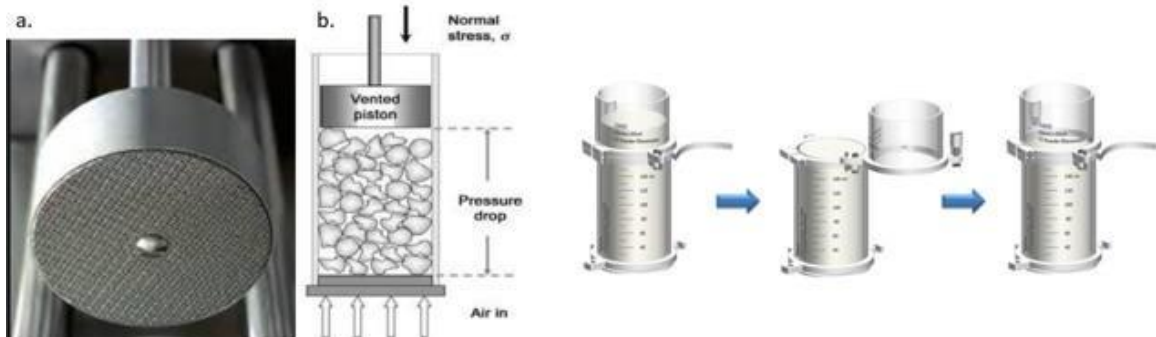


Figure 3-13: The Freeman FT4[®] Powder Rheometer conditional cycle: a) vented piston; b) normal stress applied to consolidate the powder; c) removing the excess of powder from the vessel.

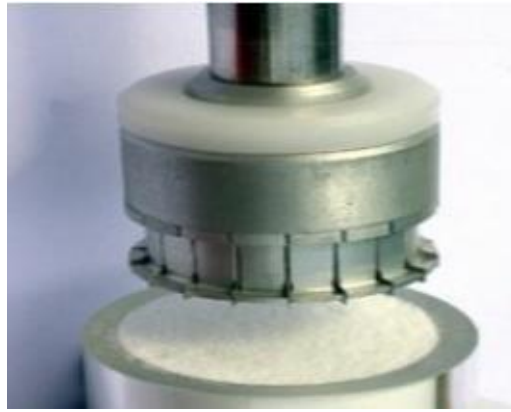


Figure 3-14: The 24 mm diameter blade used to perform the shear cell test.

For the shear cell test the 24 mm diameter blade moves at low speeds downwards into the powder inducing the predefined normal stress (see Fig 3-14). Then, the shear cell blade rotates on the first layer of the sample until it overcomes the shear strength. The sample is sheared at different normal and shear stresses to obtain five yield points from which the yield locus will be calculated. The Mohr's circles are fitted to the yield locus (shown in Fig 3-15). The major principal stress (σ_1) is calculated from the intersection of the outer point of the bigger Mohr's circle with the x-axis. The bigger Mohr's circle is calculated from the tangent to the yield locus when it crosses the steady-state flow. The unconfined yield locus (f_c) is calculated from Mohr's circle which is the tangent of the yield locus when it intercepts the origin. The ratio of σ_1 , which represents the normal stress-induced in pre-shear, to f_c , which is the cohesion of the powder gained during the consolidation step, is the flow function coefficient (FFc).

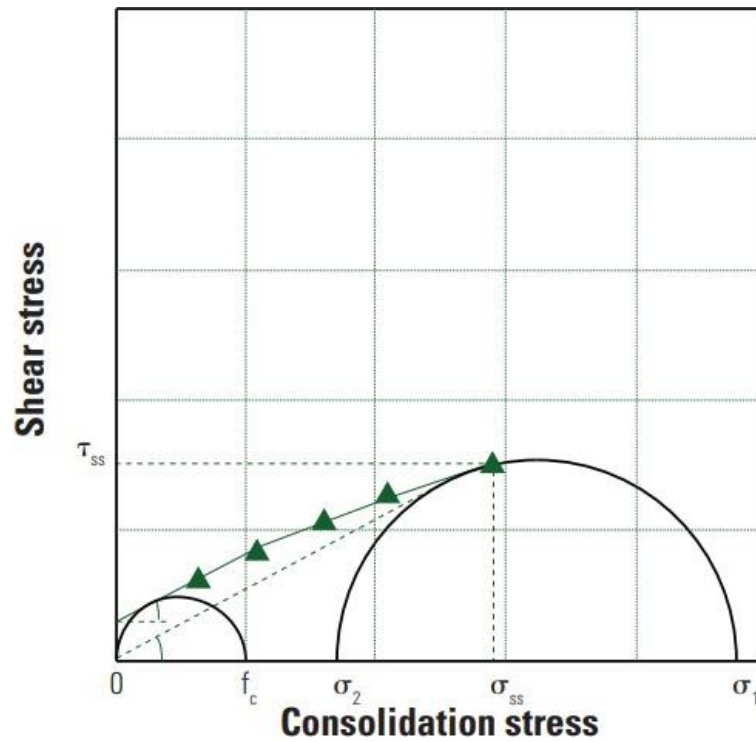


Figure 3-15: Normal stress plotted against shear stress. The FFC is defined by the ratio of the major principal stress (σ_1) to the unconfined yield strength (f_c) (Mehos et al., 2017).

Following Jenike's powder classification (Jenike, 1964), the FFC is used to rank powder flowability. Powders with a value of FFC smaller than 4 are grouped as cohesive; powders that have a value of FFC between 4 and 10 are classified as easy-flowing; and powder that have a FFC greater than 10 are classified as free-flowing (see Table 3-2).

Table 3-2: The Jenike's classification for powder flow showing the correlation between the FFC and the powder behaviour. Powders that have a FFC smaller than 4 are classified as cohesive, powders that have a FFC between 4 and 10 are classified as easy-flowing, and powders that have a FFC greater than 10 are classified as free-flowing.

FFc	Powder behaviour
$0 < \text{FFc} < 1$	Non-flowing
$1 < \text{FFc} < 2$	Very cohesive
$2 < \text{FFc} < 4$	Cohesive
$4 < \text{FFc} < 10$	Easy-flowing
$\text{FFc} > 10$	Free-flowing

3.5. Machine Learning (ML) in pharmaceutical manufacturing

ML, as a subset of AI, utilizes large sets of data to train models to make predictions. Many ML algorithms have been developed and applied in several programming languages, such as Python, R, or MATLAB, over the past decades. Python is perhaps the most widely used programming language because it is easy to access and supported by a large community. One of the main advantages that Python offers is the availability of many libraries of prewritten code that reduce the complexity of developing advanced models, e.g., Scikit-learn. Another useful technique to visualize data is the Orange Data Mining Software. The Orange Data Mining Software was first released in 1996. It is open-source software written in Python (mainly), Cython, C++, and C, and it can be used for ML, data mining, data visualisation, and data analysis. This software can also run in different operating systems (Windows, macOS, and Linux). The main advantage of the Orange Data Mining Software is that the user is not required to code. The software includes a broad range of data visualisation techniques that improve data analysis.

There are two different types of ML algorithms: supervised and unsupervised. Supervised algorithms need the input data ($x_1, x_2, x_3...$) also known as independent variables, and the desired outcome ($y_1, y_2, y_3...$) also known as the dependent variable. Supervised algorithms aim to output the correct value for a given input, either in classification (the output is a category) or regression (the output is a continuous variable). In contrast, unsupervised learning algorithms only have input data ($x_1, x_2, x_3...$), without the dependent variable. Unsupervised algorithms try to find a pattern within the data to either cluster or

reduce the dimensionality of the data (Ghahramani, 2003). For this thesis, unsupervised and supervised algorithms were used.

3.5.1. Unsupervised learning

The two main uses of unsupervised learning are dimensionality reduction and clustering, and one of the most widely used dimensionality reduction techniques is Principal Component Analysis (PCA). PCA is a multivariate statistical method suitable to describe the relationship between variables (Godoy, Vega, & Marchetti, 2014; Höskuldsson, 1995). In PCA, the variables of the dataset are reduced to “principal components” (PCs) without losing much information. The first step needed to apply PCA is to normalise the data, calculating the z-score, where x is the value of the datapoint, μ is the mean, and σ is the standard deviation (see Eq 3-10).

$$Z = \frac{x - \mu}{\sigma} \tag{3-10}$$

Then, the variance between the variables and the mean is explored to generate the covariance matrix. The covariance matrix enables the observation of the correlation between variables. This correlation could be either positive (if they increase together) or negative (one variable increases when the other variable decreases). Eigenvectors and eigenvalues of the covariance matrix are computed to calculate the PCs. The eigenvectors of the covariance matrix determine the direction of the axis where the maximum variance is, and the eigenvalues are the coefficients of the eigenvectors, and they indicate the variance of each PC.

As mentioned above, the PCs reduce the dimensionality of the data while explaining the maximum amount of the variance. However, the interpretation of the PCs is more difficult than the interpretation of the initial variables. The first PC (PC1) finds the largest possible variance in the data, and the second PC (PC2) will try to find the largest possible variance without being correlated to PC1. Then, the feature vector is calculated to decide whether to keep all PCs. This is the step when dimensionality reduction happens. Finally, the data is reorganised along the PCs axes using the eigenvectors of the covariance matrix (Holland, 2008).

Another widely used unsupervised algorithm is clustering. Clustering aims to group data points based on their similarities or by identifying high-density areas of similar data points (Massart, 2000) (see Fig 3-16). There are four main clustering algorithms: exclusive clustering, overlapping clustering, hierarchical clustering, and probabilistic clustering (Girra, Crucianu, & Boujemaa, 2004).

- i. Exclusive clustering is based on the idea that one data point can only belong to one cluster, e.g., K-means clustering (Max, 1960). K-means clustering classifies the data points into a “k” number clusters that have a centroid. The distance between a given data point and the centroid of the cluster will determine whether the datapoint belongs to said cluster (Davidson, 2002).
- ii. Overlapping clustering is based on the idea that one data point can belong to more than one cluster. One of the most efficient overlapping algorithms is overlapping K-means (OKM) (Khanmohammadi, Adibeig, & Shanehbandy, 2017).
- iii. Hierarchical clustering groups the data points into groups of trees. Hierarchical clustering can be either agglomerative or divisive. The difference between these two methods is that agglomerative clustering groups the data points based on their similarities, whereas decisive clustering groups the data points based on their differences (Nielsen, 2016).
- iv. Probabilistic clustering groups the data points based on their probability to belong to a specific cluster (Ben-Israel & Iyigun, 2008).

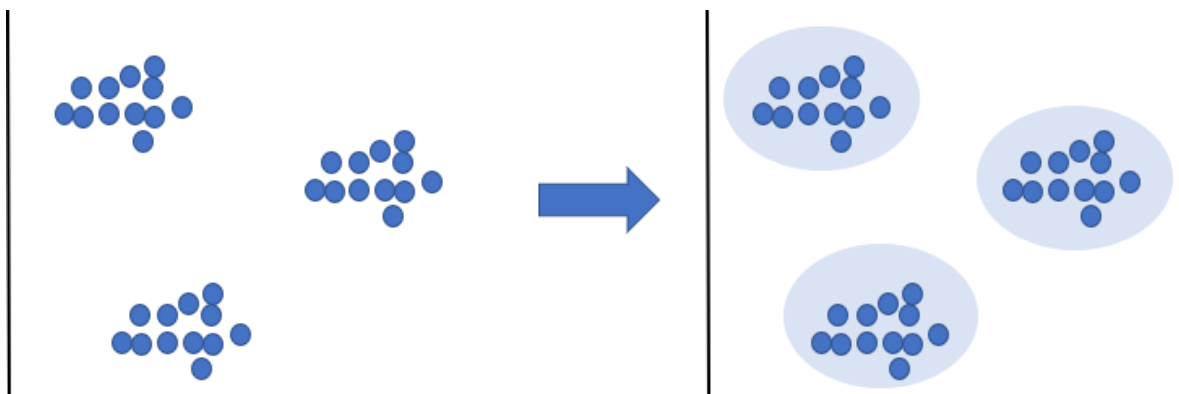


Figure 3-16: Clustering algorithm for the identification high-density areas of data points.

For this thesis, hierarchical clustering, specifically Louvain clustering, was used to find patterns and groups within the different datasets to understand the impact of particle attributes on the bulk substance.

3.5.2. Supervised learning

Supervised learning is one of the most frequently used ML methods. In contrast to unsupervised learning, supervised learning requires labelled data. In this section, a wide range of classification and regression algorithms used to build predictive models are described in detail.

3.5.1.1. Classification

Classification algorithms aim to categorise the input data. These models can either be binary classification models or multi-class classification models. Among commonly used ML methods, this thesis is mainly focused on k-Nearest Neighbours (kNN), Support Vector Machines (SVM), Random Forest (RF), multi-layer perceptron (MLP) neural network, Naïve Bayes (NB), Logistic Regression (LR), gradient boosting (GB), and AdaBoost (AB).

The kNN algorithm assumes that similar observations are close to each other and uses this proximity to classify new observations. The algorithm aims to identify the closest neighbour for a given data point. The distance between data points is calculated using different metrics, i.e., Euclidean distance (Hu, Huang, Ke, & Tsai, 2016), Manhattan distance (Suwanda, Syahputra, & Zamzami, 2020), Minkowski distance (X. Xie, 2018), and Hamming distance (Norouzi, Fleet, & Salakhutdinov, 2012), among others. The “k” of kNN refers to the number of neighbours that will be checked to classify a given data point. A common approach to calculate “k” is by calculating the square root of the total number of observations included in the training dataset. Then, the number of “k” will be rounded to an odd or even number depending on the number of classes including in the dataset. If the number of classes is even, “k” should be odd. In contrast, if the number of classes is odd, “k” should be even (Witten & Frank, 2002).

Support Vector Machine (SVM) finds a boundary line (hyperplane) to classify data points that belong to different classes. The optimum hyperplane is the one that maximises the distances between the two classes (see Fig 3-17). This technique for finding the hyperplane is relatively straightforward for linear data; however, the complexity of finding this hyperplane increases when the dataset is non-linear (see Fig 3-18). For non-linear models, SVM uses kernel functions, also known as the *kernel trick*, to find the optimum hyperplane. Using kernel functions instead of transforming the data to a higher-dimensional space speeds up the calculation that otherwise would be needed. The most popular kernel functions are linear kernel, polynomial kernel (usually less efficient), Gaussian Radial Basis (RBF)

function, generally selected for non-linear data; sigmoid kernel, mostly used for neural networks, and Gaussian kernel (Hofmann, 2006).

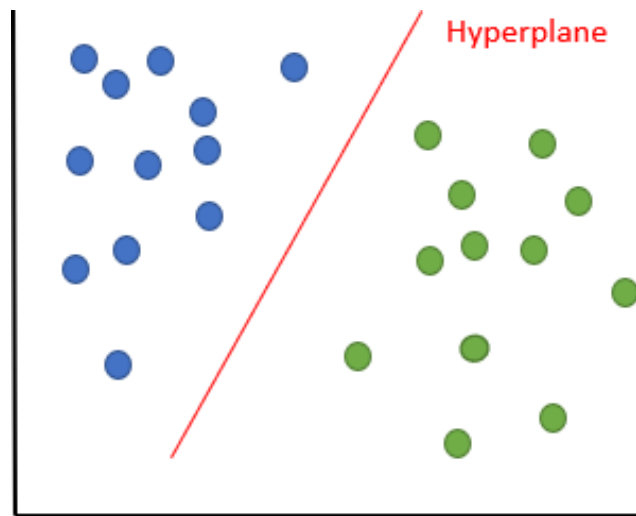


Figure 3-17: The optimum hyperplane to find the maximum distance between the classes in binary classification tasks.

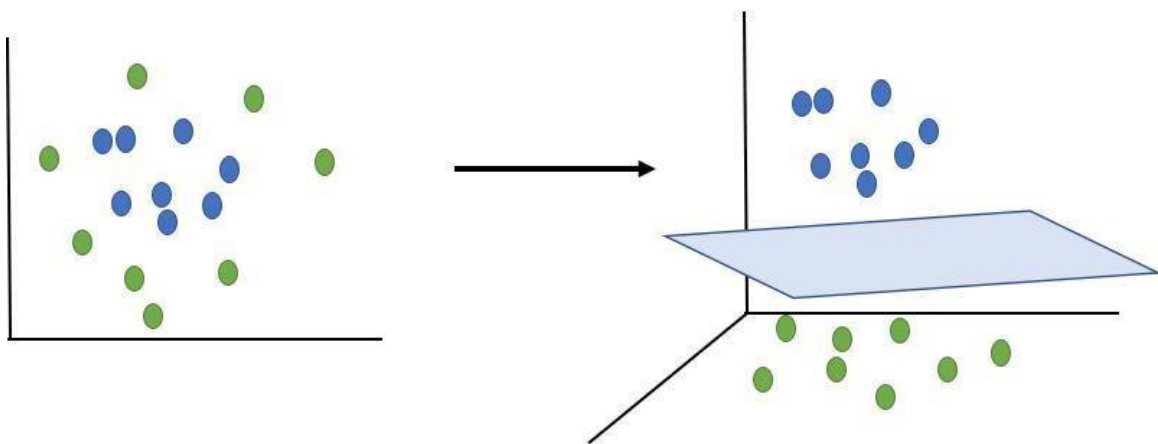


Figure 3-18: a) Non-linear dataset b) The hyperplane of a non-linear dataset.

Random Forest (RF) was first implemented by L. Breiman in 2001 (Breiman, 2001), and is applicable to solve both classification and regression tasks. RF is built on randomised independent decision trees (see Fig 3-19) that are ensembled to create the random forest. To classify a new instance or observation, the decision trees operate as an ensemble and the majority vote of all the trees is taken as the final output of the model (Breiman & Cutler, 2016).

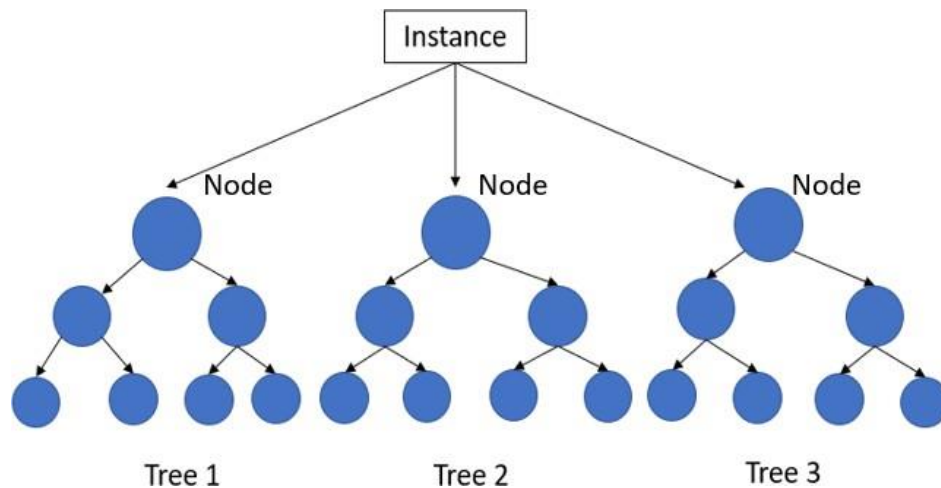


Figure 3-19: A RF decision tree. The length of the decision trees is decided by the availability of training data whereas the nodes are decided based on a randomly selected number of features that provide the best split at each node.

A Multi-layer perceptron (MLP) neural network is a feedforward artificial neural network, designed for binary classification tasks. It is a linear classifier that uses backpropagation to compute the training gradient, reducing the model error by adjusting the weight and bias. In a MLP neural network model, the information goes from the input layer through the hidden layers to the output layer, and the result of the output layer is compared to the true labels (Rosenblatt, 1958). Fig 3-20 shows that the MLP neural network layers are fully connected. Each node, except for the input layer nodes, has an activation function, generally, a logistic function, that activates the node so the node can pass the information forward (Gardner & Dorling, 1998).

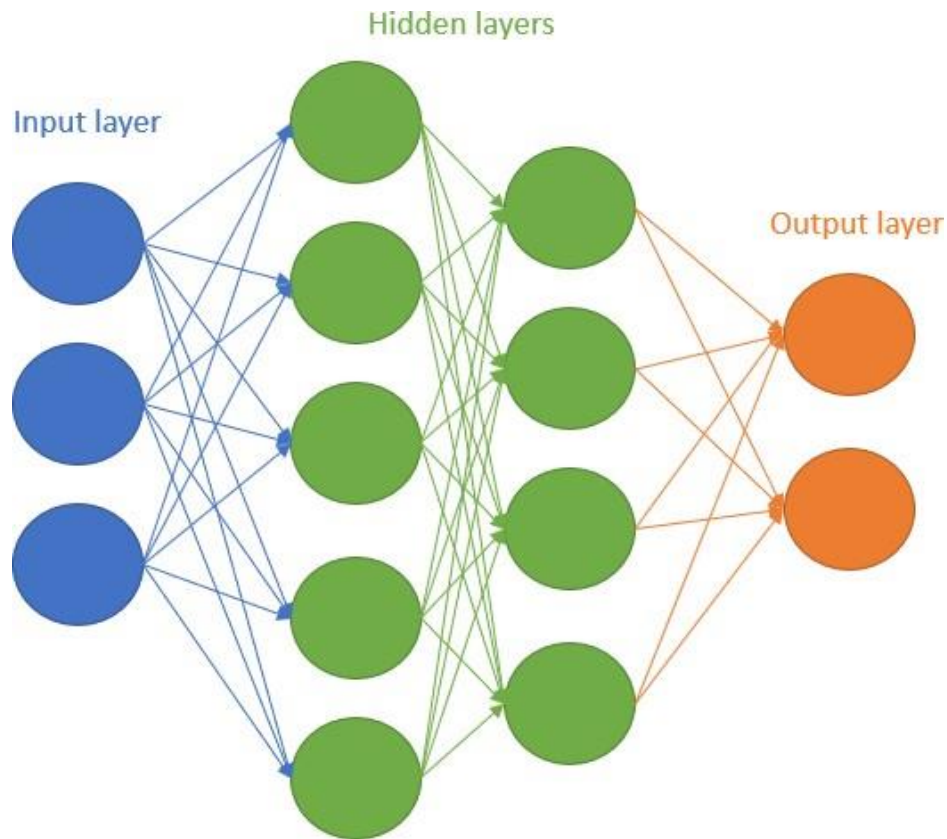


Figure 3-20: Diagram of the multi-layer perceptron (MLP) neural network interconnected layers.

Naïve Bayes (NB) classification algorithm is built on Bayesian classification methods relying on Bayes' theorem, which is used to calculate the probability of one event after another event had already occurred (Berrar, 2018). It is based on the premise that the variables are independent of one another, and that the data is Gaussian distributed. The marginal probability of an event irrespective of the outcomes of another probability is represented by $P(C)$, and the joint probability of two simultaneous events to occur is represented by $P(C, E)$. Finally, the conditional probability of one event in the occurrence of another is represented by $P(C|E)$. The conditional probability can be calculated by Bayes Theorem (see Eq 3-11), in which case the posterior probability is $P(C|E)$ and the prior probability is $P(C)$. $P(E|C)$ is referred to as the likelihood as $P(E)$, as evidence. Therefore, the posterior probability is calculated from the likelihood times the prior probability divided by the evidence. (Zhang, 2004). This theorem can be used for classification problems, to classify the probability of an observation E of belonging to class C .

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)} \quad (3-11)$$

There are other non-gaussian NB algorithms, i.e., Complement NB, Bernoulli NB, etc. However, these algorithms have not been explored for the work presented in this thesis.

Logistic regression (LR), despite its name, is only used for classification tasks, particularly for binary classification. It is a simple and efficient non-linear algorithm that uses a sigmoid function to classify observations.

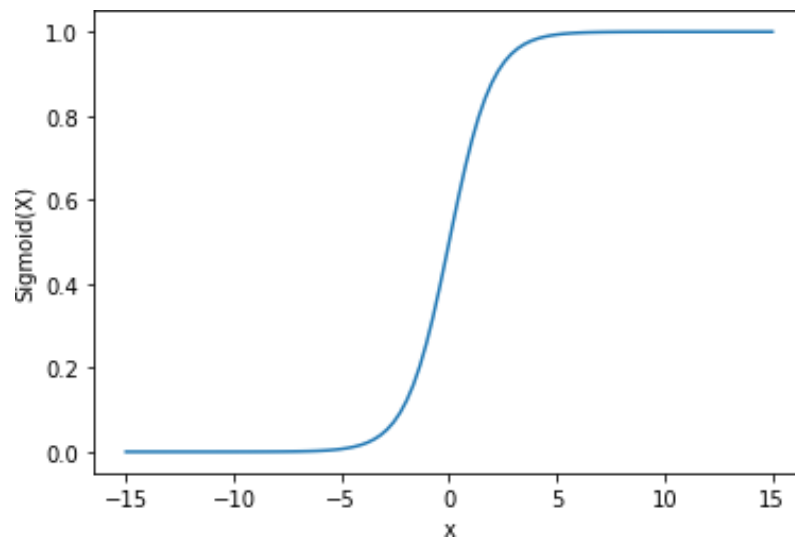


Figure 3-21: Sigmoid function used in LR to classify observations in class 0 or class 1, where “x” is an example of an independent variable.

Logistic regression (LR) is used in day-to-day examples such as identifying a poisonous mushroom or detecting if an email is a spam or not. To achieve the classification, the algorithm has a threshold to categorise the instances, and it could also be applied to multi-class classification following a “one vs all” approach (Dreiseitl & Ohno-Machado, 2002).

Boosting algorithms, i.e., AdaBoost (AB) and Gradient Boosting (GB), are based on the combination of weaker classifiers to build a strong algorithm. To achieve this, boosting algorithms assign weights to the training data. In AB, the majority vote of prediction made by the weak learners, weighted by their individual accuracy, is taken as the outcome of the model. In contrast, in GB, all the learners have the same weights.

The first boosting algorithm implemented was AB (Freund & Schapire, 1996). AB is built on decision stumps (incomplete decision trees that only have one node and two leaves). In the first decision stump, all the weights are the same, but as other stumps are built for all the variables, the weights assigned to the data vary: decision stumps that achieve higher accuracy get higher weights (Freund & Schapire, 1996; Kuhn & Johnson, 2013).

GB is also an ensemble algorithm that builds a stronger model by combining weak decision trees (decision trees that perform better than randomly guessing but worse than a stronger decision tree). The main difference between GB and RF is that in RF all the decision trees are created at once, whereas in GB, the decision trees are created subsequently. In GB, each new tree is built to better the previous one while minimising the gradient of the loss function.

All these classification algorithms will be explored in the thesis for different tasks. The performance of the trained algorithms will be assessed and compared to select the best-performing algorithm. The evaluation metrics to compare the performance of the classification algorithms are detailed in section 3.5.3.

3.5.2.2. Regression

Scientists have been using regression models before the rise of ML (Huang, Ko, Shu, & Hsu, 2020). Though, ML regression models are currently being used in different fields such as medical diagnosis or financial forecasting achieving good accuracy. Regression is a method for exploring the relationship between independent variables and a dependent variable. Hence, it can be used as a modelling technique in ML for a continuous response. For this work, Partial Least Square (PLS), RF, AB, and GB, have been used to build regression models. RF, AB, and GB have been described in the previous section, referring to classification. For regression, instead of taking the majority vote, the outcome of the model is calculated by averaging the results of the decision trees. PLS is a statistical method that reduces the number of variables into a smaller set of descriptors that will be used to predict the independent variable.

3.5.3. Model evaluation: metrics for classification and regression

The performance of the algorithms needs to be evaluated on test data. There are several methods to sample the data, being the most common methods random splitting or k-fold cross-validation. In random splitting, the dataset is split into training and testing randomly. The size of the train and test

sets is decided by the user, and common splits are 70:30, 80:20, and 90:20. In k-fold cross-validation, the dataset is split into “k” number of groups. $k - 1$ groups are used for training and the remaining fold is used for testing the model. This process is iterated until the model has been trained and tested on all folds. Once the model is evaluated, the actual values of the test set, and the values predicted by the model are reported in a confusion matrix (see Table 3-3). Usually, the predicted values are reported on the columns of the confusion matrix and the actual values of the test set are reported on the rows of the confusion matrix. Based on the layout of these values, four parameters are described:

- i. True Positives (TP): the instances that were *correctly* predicted as *positive*.
- ii. False Positives (FP): the instances that were *incorrectly* predicted as *positive*.
- iii. False Negatives (FN): the instances that were *incorrectly* predicted as *negative*.
- iv. True Negatives (TN): the instances that were *correctly* predicted as *negative*.

Table 3-3: Confusion matrix.

Confusion matrix

	Actually positive	Actually negative
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

These four parameters are used to calculate the classification evaluation metrics. The most common metrics are classification accuracy (CA), precision, recall, F-1 measure, and Area Under Curve – Receiver Operating Characteristics (AUC-ROC).

The CA of the prediction is calculated by the ratio of the corrected predicted observations (TP and TN) to the total of the observations (see Eq 3-12). The CA is one of the most used metrics to inform the performance of the model however, it can be misleading when there are data imbalance issues (one class is either underrepresented or overrepresented). This can be explained by the “1:100 class imbalance paradox”. Given the example of a dataset that contains 100 samples, of which only one sample belongs to class 0 and the rest of them belong to class 1. The CA of this model would be 99%; however, the model would not be accurate at predicting class 0 samples since this class is underrepresented in the training dataset (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2011).

$$\text{Classification accuracy (CA)} = \frac{TP+TN}{\text{Total of observations}} \tag{3-12}$$

To avoid the class imbalance paradox, there are other evaluation metrics that can provide more information about the performance of the model. Precision (see Eq 3-13) is calculated by dividing the TP by the FP and the TP, emphasising the importance of the samples that are correctly predicted as true (Sokolova & Lapalme, 2009).

$$\text{Precision} = \frac{TP}{FP+TP} \quad (3-13)$$

Moreover, recall, also known as sensitivity (shown in Eq 3-14), is calculated by the ratio of the TP to the TP and the FN, highlighting the number of positive samples that were correctly predicted among all the positive samples (Powers, 2020; Sokolova & Lapalme, 2009).

$$\text{Recall (sensitivity or true positive rate)} = \frac{TP}{TP+FN} \quad (3-14)$$

F-measure, also known as F1 Score, is the weighted average between precision and recall (Eq 3-15). It takes values from 0 to 1, and the greater the F1 Score value, the better is the performance of the model. Even though this metric might not be the most straight forward parameter, it is generally more useful than CA, particularly if there is a class imbalance (Powers, 2020).

$$F1 = 2 \times \frac{TP}{TP + \frac{1}{2}(FP+FN)} \quad (3-15)$$

AUC – ROC is a performance measurement for classification at different thresholds. ROC is the probability curve and AUC represents the ability to separate between classes. Hence, the greater the AUC – ROC value, the better the model is at discriminating between classes. A perfect model would have an AUC – ROC value of 1, whereas a poor model would have a value closer to 0, which means its ability to separate between classes would be very low. AUC – ROC is calculated by plotting the True Positive Rate (TPR) or recall (see Eq 3-14) against the False Positive Rate (FPR) (see Eq 3-16). This method has been widely applied in other fields such as in drug discovery for virtual screening to select or reject a molecule (Triballeau, Acher, Brabet, Pin, & Bertrand, 2005).

$$\text{False positive rate (FPR)} = 1 - \text{specificity} \quad (3-16)$$

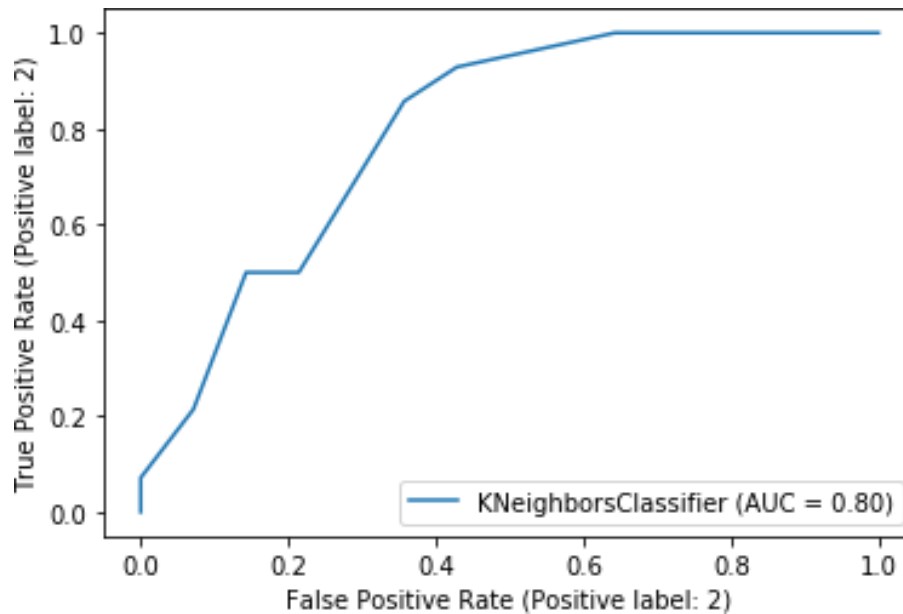


Figure 3-22: AUC – ROC curve, calculated from the TNR plotted against the FPR to study the performance of a supervised classification algorithm.

Figure 3-22 shows the AUC – ROC of a machine learning model. The AUC – ROC value is 0.8, which means that there is an 80% probability for the model to discriminate between classes. This metric is generally used for binary classification, but it can be used for multi-class problems. However, in a multiclass model, the AUC – ROC will indicate the probability of on class to be classified against the other classes.

A perhaps less common classification metric is the G-measure. The G-measure is calculated from the geometric mean of precision and recall (Powers, 2020). The geometric mean calculated from the sensitivity and the specificity (see Eq 3-17) is explained by Kubat *et al.* in the European conference on ML (Heidelberg, Germany) (Kubat, Holte, & Matwin, 1997). They proposed a new metric method to assess the performance of a learning model built on imbalanced training data.

$$\text{Geometric measure (G – measure)} = \sqrt{\text{sensitivity} \times \text{specificity}} \quad (3-17)$$

Cohen’s Kappa and Matthews Correlation Coefficient (MCC) are generally used for multi-class classification models, and usually they are correlated. However, when there is an imbalance issue, Cohen’s Kappa report less reliable results than MCC, i.e., a worse model gets better results in Cohen’s Kappa than MCC. Kappa was originally designed to measure the chance agreement between two judges, but this concept has evolved into the present and Kappa is now used in several fields such as

neuroscience, ML, or psychology. When this metric is used in classification to measure the agreement between the actual values and the predicted values, Cohen's kappa is described in Eq 3-18:

$$K = \frac{CA - P_e}{1 - P_e} \quad (3-18)$$

where P_e is the probability of chance of agreement between predicted and actual values, that is subtracted from the classification accuracy (Delgado & Tibau, 2019).

The main advantage of using Cohen's Kappa as the evaluation method is that it removes the possibility that the model is predicting by guessing randomly. It takes values from -1 to 1; however, the values are not similarly reachable: it is easier to get higher values when the class distribution is balanced. However, Cohen's Kappa will not give reliable information about the accuracy of the prediction of a single value.

The MCC was introduced by Matthews to evaluate the performance of binary classification models (Matthews, 1975) as an evaluation metric that accounts for class imbalance. It was then re-proposed and extended for multi-class classification in 2000 by Baldi *et al.* (Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000). MCC is calculated following Eq 3-19, and it takes values from -1 to 1. Since MCC measures the agreement between actual and predicted values, when MCC is 1, it shows that the predicted and actual values are in perfect agreement, whereas when MCC is -1, the predicted and actual values are in disagreement.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3-19)$$

The presented evaluation methods only apply for classification models, and therefore different metrics are needed for regression. The most widely used metrics to assess the performance of regression models are: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (R^2). The training is sampled following the same sampling methods as described for classification, and hence, the evaluation metrics stem from the comparison (or the error) between the actual values and the predicted values (Botchkarev, 2019).

The MSE measures is calculated by the squared difference between the predicted and the actual values following Eq 3-20. The difference between the predicted values and the actual values is squared to ensure a positive number and to punish large errors.

$$MSE = \frac{1}{\text{number of observations}} \sum_{i=1}^n (\text{actual} - \text{predicted})^2 \quad (3-20)$$

where n is the number of observations.

The RMSE is an extension of the MSE calculated from the square root of the MSE (see Eq 3-21). The main advantage of the RMSE over the MSE is that the RMSE is reported in the same units as the target variable. The RMSE still punishes large errors.

$$RMSE = \sqrt{\frac{1}{\text{number of observations}} \sum_{i=1}^n (\text{actual} - \text{predicted})^2} \quad (3-21)$$

where n is the number of observations.

The MAE measures the absolute error between the predictions and the actual values (see Eq 3-22), and it is also reported in the same units as the target variable. The main difference between the MAE and the RMSE is that the changes in MAE are linear and therefore larger errors have the same importance than smaller errors.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\text{actual} - \text{predicted}| \quad (3-22)$$

where n is the number of observations.

Finally, R^2 studies the proportion of variance in the dependent variable that is predicted from the independent variable. However, it has been advised that R^2 might not be the appropriate measure to evaluate the accuracy of predictive models (Li, 2017).

3.5.4. Model interpretability: Shapley Additive exPlanations (SHAP) values

Data-driven models have proven their potential applicability in several fields. However, there is a lack of trust that leads to not using data-driven models as much as they could be used. This lack of trust stems from the lack of explainability of the predictions, giving name to “black-box” models. Understanding how a ML model makes the prediction plays a crucial part in the implementation of these models. Therefore, several methods have been proposed to understand complex models (Ribeiro, Singh, & Guestrin, 2016; Shrikumar, Greenside, Shcherbina, & Kundaje, 2016). SHAP (Shapley Additive exPlanations) methods have been used in several chapters of this thesis.

The SHAP methodology assigns an importance score, also known as SHAP value, to each feature of the training dataset. These SHAP values are used to rank the variables in order of importance. Additionally, the impact of the variables (positive or negative) on the model output can be analysed (S. M. Lundberg & Lee, 2017).

There several benefits of using SHAP values over using other interpretability methods are:

- i. Global interpretability: the SHAP values can be used to analyse the feature importance of all variables and the impact (positive or negative) of the variables on the model output. For instance, this analysis allows to understand if a high value of a variable will increase or decrease the predicted value.
- ii. Local interpretability: the SHAP values can be calculated for each individual prediction. The results are reported using force plots. The force plots indicate which variables contributed towards the prediction and how much these variables contributed. For a given observation, the results of the local interpretability analysis might differ from the results of the global interpretability analysis, which had been performed on the total dataset.
- iii. Compatibility with different algorithms: the SHAP values are compatible with a wide range of algorithms (decision-trees-based algorithms, kernel-based algorithms, neural network), for classification and regression (S. Lundberg, 2018).

3.6. Deep Learning (DL) in pharmaceutical manufacturing

Deep Learning (DL), as a subset on ML, has been successfully applied to a wide range of different fields, and it refers to learning algorithms that have several hidden layers or artificial neurons that are interconnected. The first layer of these interconnected neurons is the input data. The information travels from the input layer to the hidden layers. The hidden layers contain activation functions, also known as transfer functions, and the selection of the appropriate one has an important impact on the performance of the neural network model. There are three main types of activation functions: rectified linear activation (ReLU), logistic (sigmoid), and hyperbolic tangent (Tanh). The last layer, i.e., the outcome of the model, is the output layer (Goodfellow, Bengio, & Courville, 2016).


There are five different types of neural networks: feedforward neural network, recurrent neural network, radial basis function neural network, Kohonen self-organized neural network, and modular neural network.

- i. Feed forward network: the information goes from the input layer to the output layer, in only one direction, without loopbacks. This was the first and more simplistic DL algorithm (Schmidhuber, 2015).
- ii. Recurrent neural network: the connections between the neurons can create circles, and the where the output layer becomes the input of the next layer, forming a feedback loop. This feedback loop allows the network to have memory of the previous steps and influence the coming steps (Tsoi, 1997).
- iii. Radial basis function neural network: radial basis functions are used as the activation functions of this network (T. Xie, Yu, & Wilamowski, 2011).
- iv. Kohonen self-organising network is self-organised by using unsupervised learning methods. It is formed by two connected layers (one input later and one output layer).
- v. Modular neural network breaks down large network into smaller independent network modules.

These different types of neural networks can be implemented in different architectures. In this thesis, feedforward neural networks haven been implemented using Convolutional Neural Network (CNN). The CNN neural network has an input later, hidden layers, and an output layer. In the hidden layers, the inputs and outputs of each layer are masked by the activation function and the convolution. A convolution is a mathematical operation that produces a function from two other function. This produced function shows how one function modifies the other (Shrestha & Mahmood, 2019).

3.7. References

- Alyami, H., Dahmash, E., Bowen, J., & Mohammed, A. R. (2017). An investigation into the effects of excipient particle size, blending techniques and processing parameters on the homogeneity and content uniformity of a blend containing low-dose model drug. *PLoS one*, *12*(6), e0178772.
- Augsburger, L. L., & Hoag, S. W. (2016). *Pharmaceutical dosage forms-tablets*: CRC press.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, *16*(5), 412-424.
- Ben-Israel, A., & Iyigun, C. (2008). Probabilistic d-clustering. *Journal of Classification*, *25*(1), 5-26.
- Berrar, D. (2018). Bayes' theorem and naive Bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 403.
- Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*, 45.
- Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
- Breiman, L., & Cutler, A. (2016). Random Forests for scientific Discovery. *linea*]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/berkeley_files/frame.htm.
- Chattoraj, S., Daugherty, P., McDermott, T., Olsofsky, A., Roth, W. J., & Toby, M. (2018). Sticking and picking in pharmaceutical tablet compression: an IQ consortium review. *Journal of pharmaceutical sciences*, *107*(9), 2267-2282.
- Clarke, J., Gamble, J. F., Jones, J. W., Toby, M., Greenwood, R., & Ingram, A. (2019). Alternative approach for defining the particle population requirements for static image analysis based particle characterization methods. *Advanced Powder Technology*, *30*(5), 920-929.
- Davidson, I. (2002). Understanding K-means non-hierarchical clustering. *Computer Science Department of State University of New York (SUNY), Albany*.
- Delgado, R., & Tibau, X.-A. (2019). Why Cohen's Kappa should be avoided as performance measure in classification. *PLoS one*, *14*(9), e0222916.
- Dorris, G. M., & Gray, D. G. (1980). Adsorption of n-alkanes at zero surface coverage on cellulose paper and wood fibers. *Journal of Colloid and Interface Science*, *77*(2), 353-362.
- Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), 352-359.
- Fichtner, F., Mahlin, D., Welch, K., Gaisford, S., & Alderborn, G. (2008). Effect of surface energy on powder compactibility. *Pharmaceutical research*, *25*(12), 2750-2759.
- Fitzpatrick, J. (2013). Powder properties in food production systems. In *Handbook of food powders* (pp. 285-308): Elsevier.
- Freund, Y., & Schapire, R. E. (1996). *Experiments with a new boosting algorithm*. Paper presented at the icml.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463-484.
- Gamble, J. F., Toby, M., & Hamey, R. (2015). Application of image-based particle size and shape characterization systems in the development of small molecule pharmaceuticals. *Journal of pharmaceutical sciences*, *104*(5), 1563-1574.
- Gardner, M. W., & Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.
- Ghahramani, Z. (2003). *Unsupervised learning*. Paper presented at the Summer school on machine learning.

- Godoy, J. L., Vega, J. R., & Marchetti, J. L. (2014). Relationships between PCA and PLS-regression. *Chemometrics and Intelligent Laboratory Systems*, *130*, 182-191.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.
- Grimsey, I. M., Feeley, J. C., & York, P. (2002). Analysis of the surface energy of pharmaceutical powders by inverse gas chromatography. *Journal of pharmaceutical sciences*, *91*(2), 571-583.
- Girra, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content*, *1*, 9-16.
- Hofmann, M. (2006). Support vector machines-kernels and the kernel trick. *Notes*, *26*(3), 1-16.
- Holland, S. M. (2008). Principal components analysis (PCA). *Department of Geology, University of Georgia, Athens, GA*, 30602-32501.
- Höskuldsson, A. (1995). A combined theory for PCA and PLS. *Journal of Chemometrics*, *9*(2), 91-123.
- Hu, L.-Y., Huang, M.-W., Ke, S.-W., & Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, *5*(1), 1-9.
- Huang, J.-C., Ko, K.-M., Shu, M.-H., & Hsu, B.-M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. *Neural Computing and Applications*, *32*(10), 5461-5469.
- ISO, B. (2009). Particle size analysis—Laser diffraction methods. In.
- Jenike, A. W. (1964). Storage and flow of solids. *Bulletin No. 123, Utah State University*.
- Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An improved overlapping k-means clustering method for medical applications. *Expert Systems with Applications*, *67*, 12-18.
- Kozbial, A., Li, Z., Conaway, C., McGinley, R., Dhingra, S., Vahdat, V., . . . Li, L. (2014). Study on the surface energy of graphene by contact angle measurements. *Langmuir*, *30*(28), 8598-8606.
- Kubat, M., Holte, R., & Matwin, S. (1997). *Learning when negative examples abound*. Paper presented at the European conference on machine learning.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26): Springer.
- Kwok, D., Gietzelt, T., Grundke, K., Jacobasch, H.-J., & Neumann, A. W. (1997). Contact angle measurements and contact angle interpretation. 1. Contact angle measurements by axisymmetric drop shape analysis and a goniometer sessile drop technique. *Langmuir*, *13*(10), 2880-2894.
- Legras, A., Kondor, A., Heitzmann, M., & Truss, R. (2015). Inverse gas chromatography for natural fibre characterisation: Identification of the critical parameters to determine the Brunauer–Emmett–Teller specific surface area. *Journal of Chromatography A*, *1425*, 273-279.
- Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r^2 , why not? Then what? *PloS one*, *12*(8), e0183250.
- López Córdoba, A. F., & Goyanes, S. N. (2017). Food Powder Properties.
- Lundberg, S. (2018). Welcome to the SHAP documentation .
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
- Massart, P. (2000). About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, *28*(2), 863-884.
- Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442-451.
- Max, J. (1960). Quantizing for minimum distortion. *IRE Transactions on Information Theory*, *6*(1), 7-12.
- Mohammadi-Jam, S., & Waters, K. (2014). Inverse gas chromatography applications: A review. *Advances in colloid and interface science*, *212*, 21-44.
- Nielsen, F. (2016). Hierarchical clustering. In *Introduction to HPC with MPI for Data Science* (pp. 195-211): Springer.

- Norouzi, M., Fleet, D. J., & Salakhutdinov, R. R. (2012). Hamming distance metric learning. *Advances in neural information processing systems*, 25.
- Peck, G. E., Baley, G. J., McCurdy, V. E., & Banker, G. S. (1989). Tablet formulation and design. *Pharmaceutical dosage forms*. Marcel Dekker, New York, 75-130.
- Pharmacopoeia, J. (2002). European pharmacopoeia. *Strasbourg: Council of Europe*.
- Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" *Explaining the predictions of any classifier*. Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61, 85-117.
- Schultz, J. a., Lavielle, L., & Martin, C. (1987). The role of the interface in carbon fibre-epoxy composites. *The Journal of Adhesion*, 23(1), 45-60.
- Shah, U. V., Olusanmi, D., Narang, A. S., Hussain, M. A., Gamble, J. F., Tobyn, M. J., & Heng, J. Y. (2014). Effect of crystal habits on the surface energy and cohesion of crystalline powders. *International journal of pharmaceutics*, 472(1-2), 140-147.
- Shenoy, P., Viau, M., Tammel, K., Innings, F., Fitzpatrick, J., & Ahrné, L. (2015). Effect of powder densities, particle size and shape on mixture quality of binary food powder mixtures. *Powder Technology*, 272, 165-172.
- Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *Ieee Access*, 7, 53040-53065.
- Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4), 427-437.
- Standard, I. (2004). Particle Size Analysis–Image Analysis Methods–Part 1: Static Image Analysis Methods. *BS ISO*, 13322-13321.
- Suwanda, R., Syahputra, Z., & Zamzami, E. (2020). *Analysis of euclidean distance and manhattan distance in the K-means algorithm for variations number of centroid K*. Paper presented at the Journal of Physics: Conference Series.
- Triballeau, N., Acher, F., Brabet, I., Pin, J.-P., & Bertrand, H.-O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of medicinal chemistry*, 48(7), 2534-2547.
- Tsoi, A. C. (1997). Recurrent neural network architectures: an overview. *International School on Neural Networks, Initiated by IIASS and EMFCSC*, 1-26.
- Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, 31(1), 76-77.
- Voelkel, A. (2021). Physicochemical measurements (inverse gas chromatography). *In Gas chromatography* (pp. 561-579). Elsevier.
- Wriedt, T. (2012). Mie theory: a review. *The Mie Theory*, 53-71.
- Xie, T., Yu, H., & Wilamowski, B. (2011). *Comparison between traditional neural networks and radial basis function networks*. Paper presented at the 2011 IEEE international symposium on industrial electronics.
- Xie, X. (2018). A k-nearest neighbor technique for brain tumor segmentation using minkowski distance. *Journal of Medical Imaging and Health Informatics*, 8(2), 180-185.
- Zhang, H. (2004). The optimality of naive Bayes. *Aa*, 1(2), 3.

4. Prediction of powder flow of pharmaceutical powders using machine learning

4.1. Introduction

In recent years, the pharmaceutical industry has increasingly explored Industry 4.0 technologies with the goal of using digital design to improve the prediction of bulk materials properties and minimise the amount of time and material required in early-stage development (Maier, 2017). Machine learning (ML) models can help inform and minimise extensive early-stage development experimental, water and power consumption.

Understanding powder flow of pharmaceutical materials is necessary when developing robust manufacturing processes (Abe, Yasui, Kuwata, & Takeuchi, 2009). Powder flow, typically characterised by the flow function coefficient (FFc), impacts the manufacturability of drug compounds, and optimising powder flow improves the likelihood that streamlined manufacturing processes can be developed successfully and operated consistently. For example, powder flow has a significant impact on steps involving tablet formation. Tablets can be manufactured using several techniques with direct compression (DC), wet granulation (WG), and roller compaction (RC) being the most widely used in industrial operations (Shangraw, 1989). Using DC for tablet manufacture requires that material properties, such as blend uniformity, compactability, and lubrication are tightly controlled (Schaller et al., 2019). By contrast, WG and RC improve powder flow and compactability prior to tablet compression while preventing segregation with the use of binding agents and secondary wetting, drying, and milling steps. As a result, WG and RC are more expensive and time-consuming, and they are not suitable for materials that are sensitive to heat and/or moisture. Thus, DC offers a streamlined process with fewer steps than WG for example. However, to use DC, powders must flow well.

The ability to predict flow properties of powders or powder blends using straightforward routine measurements is therefore of increasing importance (Trementozzi et al., 2017). A variety of particle and bulk properties are known to affect flowability, powder behaviour and process performance in DC. For example, particle size distribution (PSD) has a significant impact on powder behaviour (Goh, Heng, & Liew, 2018) and hence, PSD has traditionally been a key property considered when predicting powder behaviour (Yu, Muteki, Zhang, & Kim, 2011). However other physical properties can also affect powder flow behaviour and process performance, including shape, surface texture, surface area, density, cohesivity, adhesivity, elasticity, plasticity, porosity, charge potential, hardness, and hygroscopicity (Shah, Karde, Ghoroi, & Heng, 2017). These physical properties can have complex

effects on powder behaviour, which have been described in many publications (Crouter & Briens, 2014; Guo, Beddow, & Vetter, 1985; Kunnath, Chen, Zheng, & Davé, 2021).

PSD has a significant impact on powder flowability, however the relationship between these properties is not directly predictable (Bellamy, Nordon, & Littlejohn, 2008; Kaerger, Edge, & Price, 2004; Sun & Grant, 2001; Yu et al., 2011). The effect of PSD in manufacturing processes such as compression has been demonstrated previously, and therefore, the effects of PSD should be carefully studied to ensure good manufacturing properties to achieve the desired dosage form (Hlinak, Kuriyan, Morris, Reklaitis, & Basu, 2006; Masuda, Higashitani, & Yoshida, 2006; Shekunov, Chattopadhyay, Tong, & Chow, 2007). The guidelines proposed by Leane *et al.* indicated that powders with a PSD D90 greater than 1000 μm are ideal for DC, but no other PSD targets were established for other manufacturing techniques, such as wet granulation or roller compaction (Leane, Pitt, Reynolds, & Group, 2015).

We present herein an assessment of ML modelling for predicting FFC as a reliable, generally applicable method for a wide range of pharmaceutical powders. The proposed models aim to predict the FFC of new materials from the material's physical properties. Usually, materials that have a value of FFC greater than 10 are considered free-flowing (Jenike, 1964), and therefore, easy to manufacture. Here, by combining ML models with experimental measurements, the amount of material and time required to estimate powder flow was significantly decreased from 30 g and 2 hours to 2 g and 5 minutes. Implementing such models in the early stages of drug development could help target particle engineering or improve decision-making for formulation and processing technology selection while reducing the time and material required.

The aim of this chapter can be summarised in the following points:

- (i) Predict powder flow of pharmaceutical materials from simple, routinely characterised physical particle properties, allowing to accelerate the decision making of the appropriate manufacturing route, with special focus on DC.
- (ii) Implement SHAP values to assist model interpretability to ensure an understanding of how the model makes the predictions and draw recommendations for designing direct-compressible APIs.

4.2. Materials and methods

4.2.1. Materials

The materials displayed in Table 3-1 (Chapter 3) were used to build the dataset. These materials were included as single components and some of them were mixed and included as blends, resulting in a total of 112 observations. The reason for including blends of the pre-existing materials was two-fold: to broaden the variety of the observations included for training, and to gain knowledge of the changes in flowability when materials with different FFC were mixed. These formulations were treated as “single component” materials, meaning that the particle and bulk properties of the blends were used as the input data for training de ML models.

Blends were made for Ibuprofen 50, Paracetamol Powder, Paracetamol Granular Special, Mefenamic acid, and Ibuprofen sodium salt at different drug loadings (5%, 20%, 40%) for binary mixtures with FastFlo 316 and multicomponent mixtures including FastFlo 316, Croscarmellose sodium, Avicel PH-102, and Magnesium Stearate. The blends were prepared using a 1L bin blender (Pharmatech AB-105, UK). The composition of the blends is described in Tables 4-1 and 4-2.

Table 4-1: The composition of binary blends. All binary blends included FastFlo 316 and one of the following APIs: Ibuprofen 50, Paracetamol Granular Special, Paracetamol Powder, Mefenamic Acid, Calcium Carbonate.

Binary mixture	Drug Loading	Fast Flo 316
Low drug dosage	5%	95%
Medium drug dosage	20%	80%
High drug dosage	40%	60%

Table 4-2: The composition of multi-component blends. All multicomponent blends included FastFlo 316, one of the following APIs: Ibuprofen 50, Paracetamol Granular Special, Paracetamol Powder, Mefenamic Acid, Calcium Carbonate, and the remaining 25% of a combination of 20% Avicel PH-102, 3.5% Croscarmellose Sodium, and 1.5% Magnesium Stearate.

Binary mixture	Drug Loading	Fast Flo 316	Other excipients
Low drug dosage	5%	70%	25%
Medium drug dosage	20%	55%	25%
High drug dosage	40%	35%	25%

4.2.2. Experimental methods

The experimental methods described below were used to generate the input variables used for the prediction of powder flow of pharmaceutical materials.

4.2.2.1. Particle size and shape – QICPIC®, Sympatec: image analysis characterisation

Particle size and shape were analysed using QICPIC®, Sympatec. QICPIC® captures the physical properties of the particles by using a high-speed camera that performs dynamic image analysis. The measurements were done in triplicate. The PSD is represented with the values PSD D10, PSD D50, PSD D90, and Sauter Mean Diameter (SMD), volume-based. Among the particle shape descriptors that can be analysed using this equipment, aspect ratio and sphericity were selected as they were proven to have the biggest impact on powder flow.

4.2.2.2. Surface area and energy measurements – Surface Energy Analyser

The surface area and energy of 35 materials were measured using inverse gas chromatography (iGC – Surface Energy Analyser, Surface Measurement Systems Ltd.) For the surface energy measurements, the method selected was Dorris-Gray (Dorris & Gray, 1980).

4.2.2.3. Powder flow and bulk density – Powder Rheometer FT4

An FT4 Powder Rheometer (Freeman Technology, Malvern, UK) was used to carry out the shear cell test to measure the FFC and the bulk density of the different materials. The consolidation stress was set at 9 kPa, and the normal stress for shearing was set at 7, 6, 5, 4, and 3 kPa. The sample was sheared to obtain five yield points at different normal and shear stress. The 25 mm x 10 ml split-vessel was selected to carry out triplicates of each sample using the Freeman Technology user manual (Ltd).

Powders with a value of FFC below 4 have poor flow; between 4 and 10, they are fairly flowable; and above 10, free-flowing (Jenike, 1964). The FFC has been correlated with the manufacturing process by the Manufacturing Classification System (Leane et al., 2015; Zegzulka, Gelnar, Jezerska, Prokes, & Rozbroj, 2020), assigning to each API FFC and drug loading a suitable manufacturing process.

The consolidated bulk density was also measured using the FT4 Powder Rheometer (Freeman Technology Ltd.). The results of bulk density calculated using this method are generally more accurate and reproducible than the conventional measurements, such as the measurement in a graduated cylinder, or in a volumeter (Ltd; Organization, 2012). The test was repeated at least 2 more times, until the results were consistent, and the average value was calculated and taken as the result.

4.2.3. Data analytics: correlation between features and correlation filtering.

The Pearson correlation coefficient (PCC) reports the correlation between two variables. It is calculated from the ratio of the covariance of the two variables of study to their standard deviations. The PCC takes values from -1 to 1, being 1 perfect correlation, 0 (“zero”), non-correlation, and -1, perfect inverse correlation. The correlation between variables is visualised in a heatmap plot. Another use of this statistical analysis is filtering high-correlated variables. The idea of filtering using correlation values was first introduced to biological research in the 1950s, and it is still widely used. By removing highly correlated variables, we remove variables that are not adding new information, and their presence in the dataset can be detrimental for the performance of ML models (Benesty, Chen, & Huang, 2008; Sheugh & Alizadeh, 2015). For pairs of variables that have a PCC greater than 0.9, one of the variables was randomly removed.

4.2.4. Machine learning methods

Unsupervised and supervised ML models were built to investigate FFC prediction from PSD and particle shape.

4.2.4.1. Unsupervised learning approaches

For unsupervised learning, Principal Component Analysis (PCA), hierarchical clustering and Louvain clustering were applied to the data. PCA was performed using Anaconda Spyder (Scientific Python Development Environment), matplotlib (Matplotlib, 2012-2022), and sci-kit learn (Pedregosa et al., 2011). Louvain clustering was performed using Orange Data Mining (Demšar et al., 2013). For Louvain clustering, the data were normalised and a 3-component PCA was applied as pre-processing. To plot the graph, the distance metric used was Euclidean and 30 neighbours.

4.2.4.2. Classification models

Support Vector Machines (SVM), Random Forests (RF), neural networks, Naïve Bayes (NB), k-Nearest Neighbours (kNN), Logistic Regression (LR), and Adaboost (AB) were all investigated for classification capabilities of powder flow into three categories: cohesive, easy-flowing, free-flowing (as defined in section 4.2.2.3). Python 3.7 (Van Rossum & Drake, 2009) was used to write the code for the algorithms, using libraries including pandas, NumPy (Harris et al., 2020), matplotlib, and sci-kit learn.

The performance of each algorithm was evaluated using the area under the curve receiver operating characteristics (AUC – ROC)(Lever, 2016). This metric was calculated from the corresponding model's confusion matrix. The total of 112 pharmaceutical powders were included in these models, sampled using 10-fold cross-validation to test the performance.

4.2.4.3. Regression models

Linear regression (LR) (Montgomery, Peck, & Vining, 2021), Gradient Boosting (GB) (Sigrist, 2018) (Zhang & Haghani, 2015), Random Forest (RF) (Jaiswal & Samikannu, 2017; Liu, Wang, Huang, & Yin, 2020), and AdaBoost (AB) (Schapire, 2013) were used for FFC value prediction. Python, sci-kit learn, and Orange Data Mining software were used to implement these models.

4.2.4.4. Model interpretability

Shapley Additive Explanation (SHAP) (S. Lundberg, 2018; S. M. Lundberg & Lee, 2017) values were used to increase the interpretability of the models here and move away from the lack of understanding behind model decision making. This method has been used in this chapter for global interpretability, i.e., to identify the most important variables during training; for dependence interpretability, i.e., to understand how the effect of a single variable has on the predictions made by the model; and for local interpretability, i.e., to understand how the model made the prediction for a selected test powder.

4.2.4.5. External Validation

External validation was used to assess the performance of both the classification and regression models as standard practice to demonstrate the applicability of the model in unseen data. Prior to the test/train split, 8 pharmaceutical powders were removed from the dataset. These 8 “unseen” pharmaceutical powders were used for external validation of the highest performing classification and regression models.

4.3. Results and discussion

4.3.1. Experimental results

4.3.1.1. Particle size distribution (PSD) – QICPIC®, Sympatec

The 112 materials were analysed using the QICPIC® instrument including 30 active pharmaceutical ingredients, 43 excipients and 40 blends. The results of the experimental measurements of PSD D10, PSD D50, PSD D90, Sauter Mean Diameter (SMD), aspect ratio D50, and sphericity D50 are presented in this section (see Figs 4-1 to 4-8). The range of volume-based PSD descriptors is presented in Table 4-3.

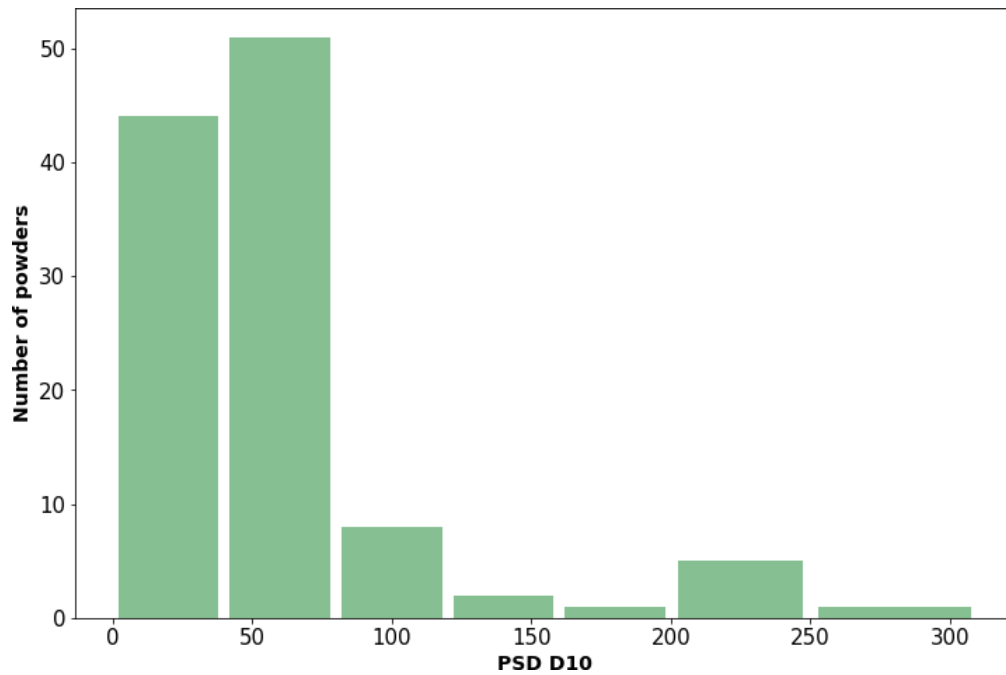


Figure 4-1: The distribution of the PSD D10. The values of PSD D10 ranged between 0 and 300 μm; and most of the powders had a PSD D10 smaller than 100 μm.

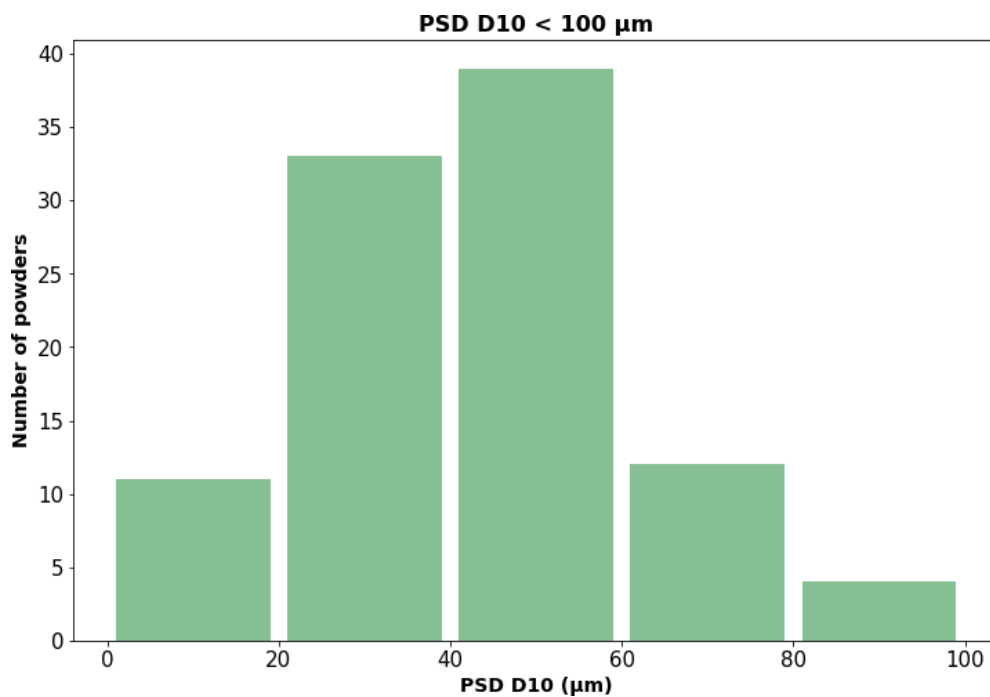


Figure 4-2: The distribution of the PSD D10 values smaller than 100 μm. The histogram shows that most of the powders had a PSD D10 between 20 and 60 μm (60% of the total of the pharmaceutical powders included for this study).

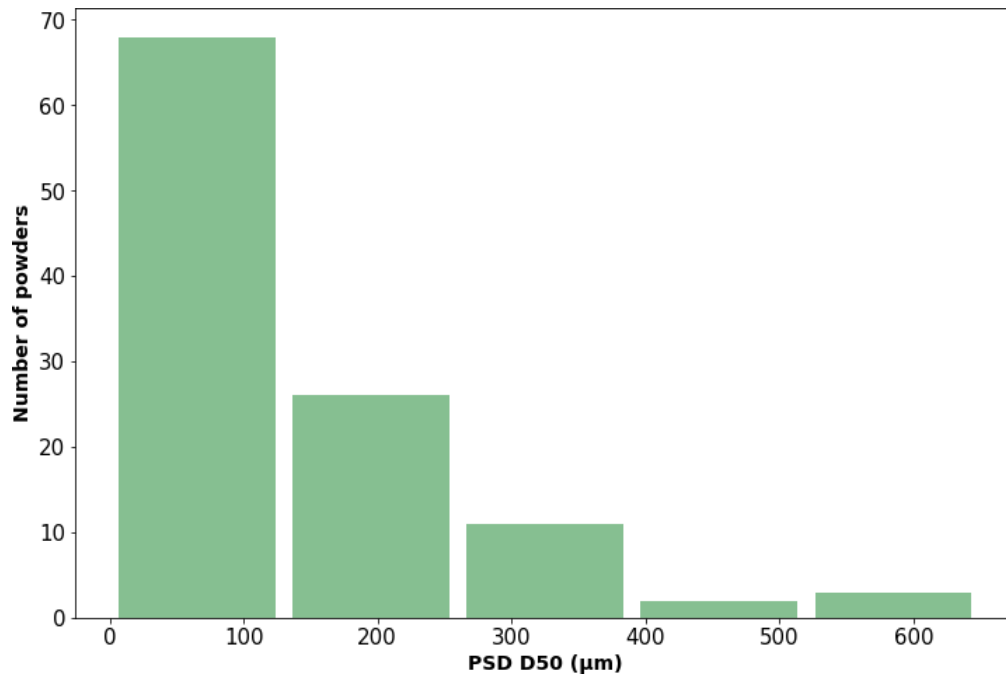


Figure 4-3: The distribution of the PSD D50. The PSD D50 values ranged from 0 to 600 µm, and most of the powders analysed had a PSD D50 smaller than 300 µm.

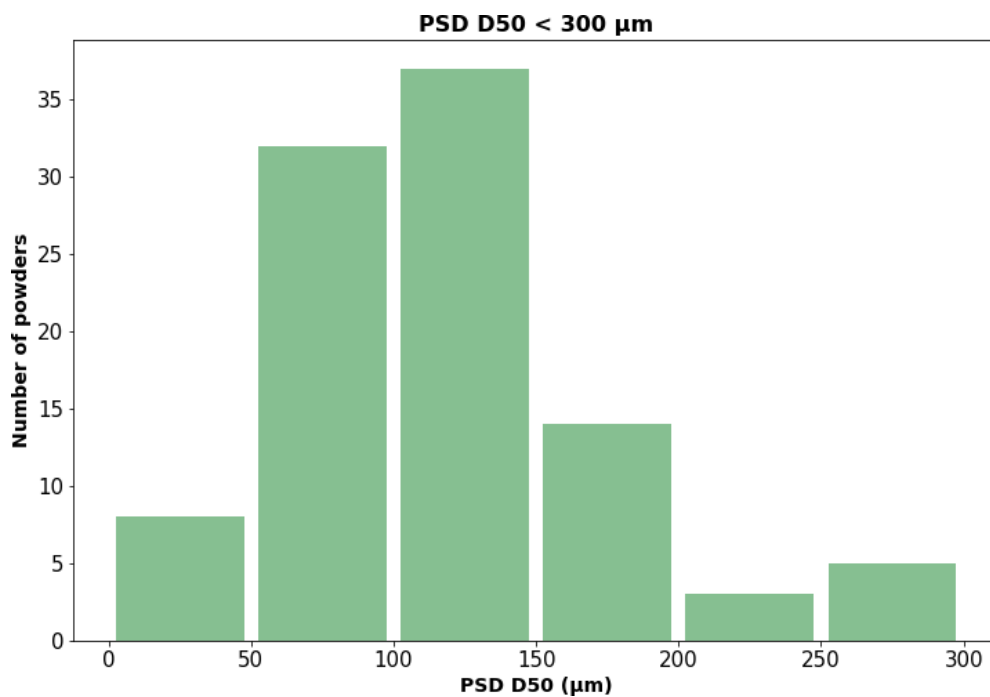


Figure 4-4: The distribution of the PSD D50 of the powders that exhibited a PSD D50 smaller than 300 µm. Of these powders, most of them had a PSD D50 between 50 and 150 µm.

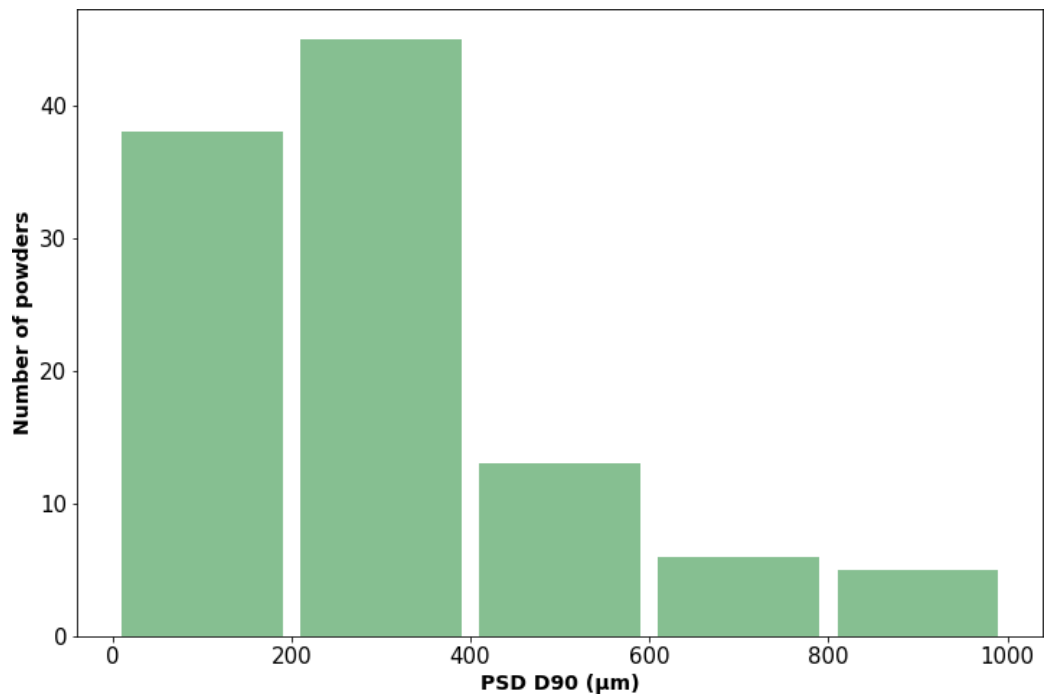


Figure 4-5: The distribution of the PSD D90 ranged between 0 and 1000 µm, and most of the materials had a PSD D90 value smaller than 400 µm.

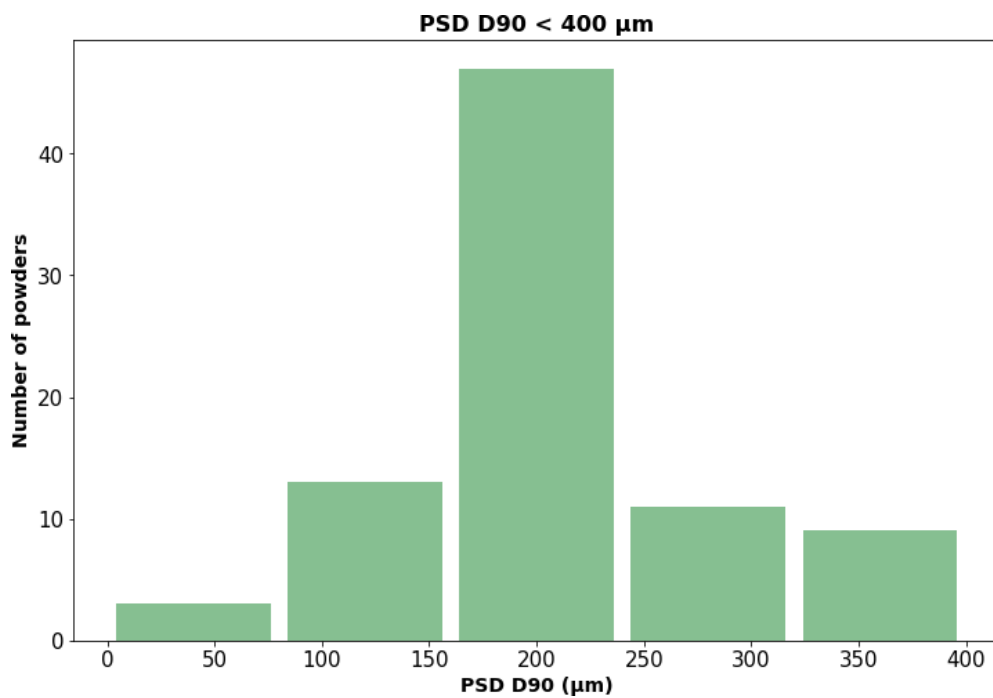


Figure 4-6: The distribution of the PSD D90 values smaller than 400 µm and most of the materials had a PSD D90 between 200 and 250 µm.

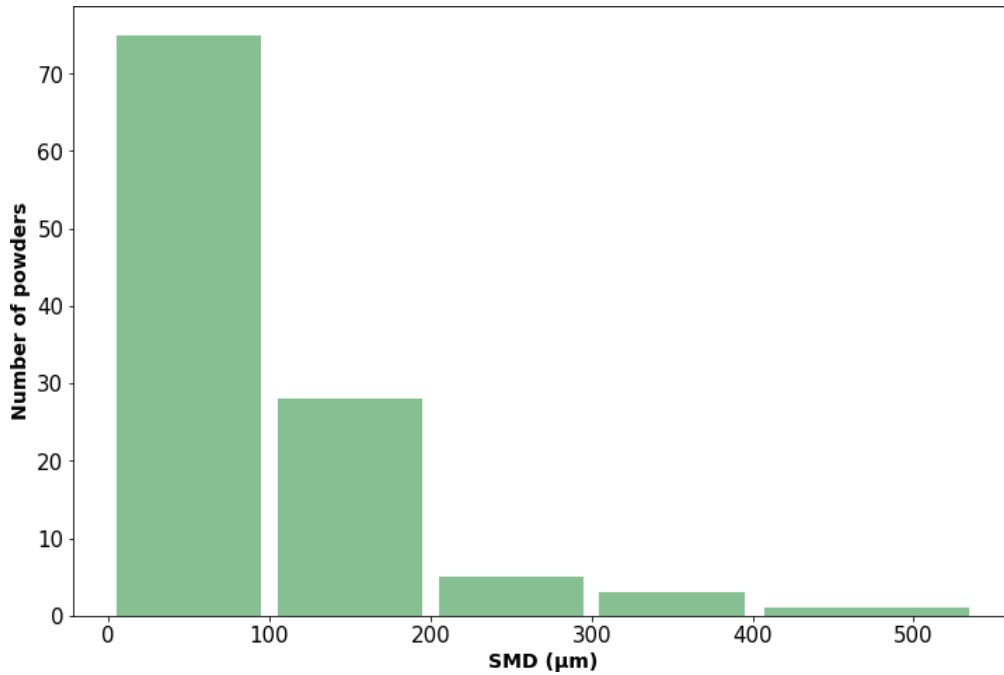


Figure 4-7: The distribution of the SMD values ranged from 0 to 500 μm. Most of the powders had a SMD smaller than 200 μm.

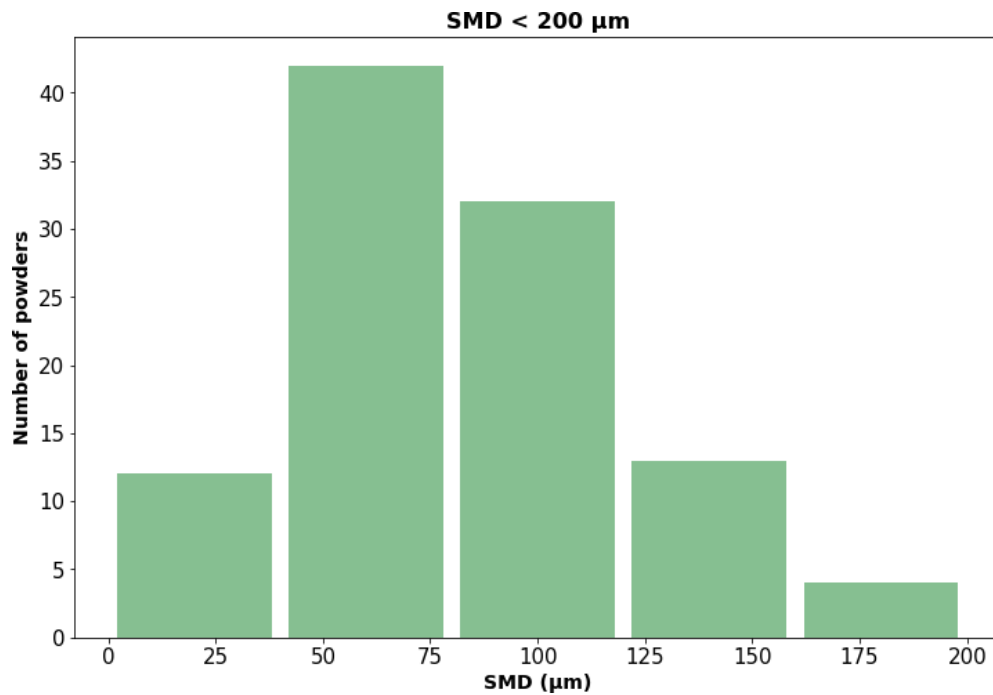


Figure 4-8: Sauter Mean Diameter (SMD) values smaller than 200 μm. The histogram shows that most of the powders had a SMD between 40 and 120 μm.

Table 4-3: PSD results, including the range of values and the median value for each parameter.

Parameter	Range of values (μm)	Median (μm)
PSD D10	9 – 225	54.84
PSD D50	25 – 644	149.19
PSD D90	53 – 1892	328.87
Sauter Mean Diameter (SMD)	19 – 541	94.63

The aspect ratio distribution and sphericity distribution were included as these parameters were found to have high importance in the ML models. Figs 4-9 to 4-11 show distribution of these parameters. Most of the materials considered for this study had an aspect ratio D50 value between 0.6 and 0.7 (see Fig 4-9). We observed a similar trend for sphericity D50 (see Fig 4-10): almost of 50% the powders had a sphericity D50 value between 0.7 and 0.8 (see Fig 4-11). According to literature, these sphericity and aspect ratios suggest that the majority of particles studied here were tetrahedral or hexahedral (Pinto, Lima, & Leal Filho, 2009).

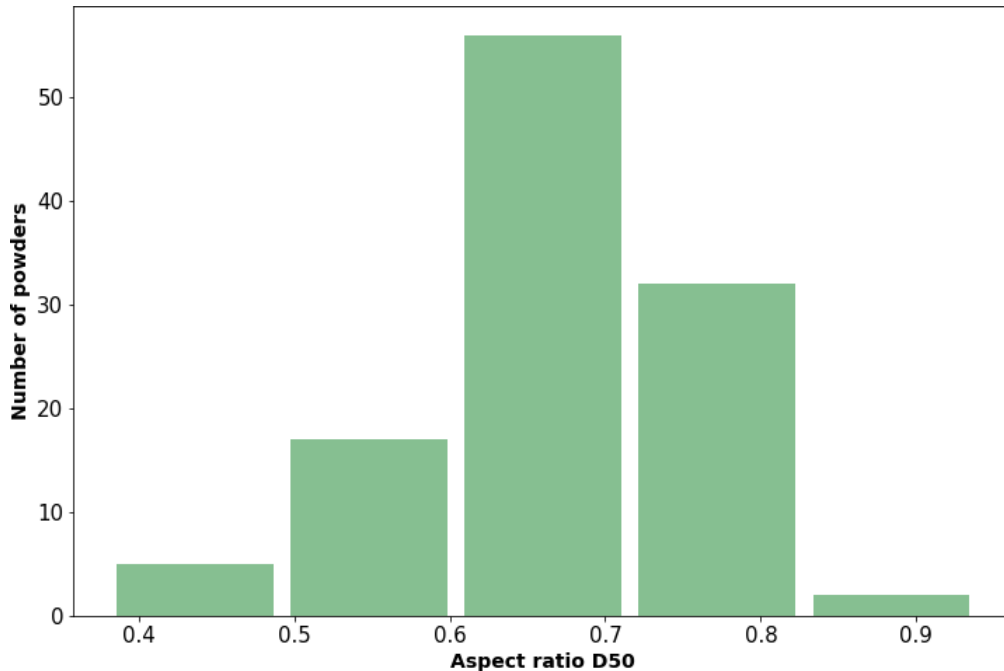


Figure 4-9: The distribution of the aspect ratio D50 values across the pharmaceutical powder included in the training dataset.

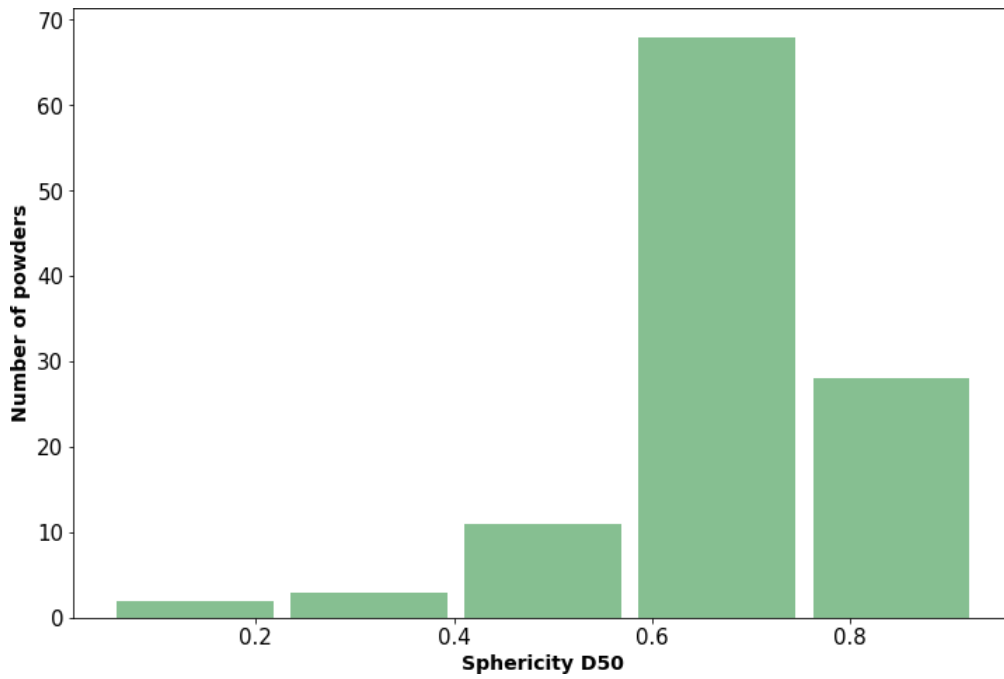


Figure 4-10: The distribution of the sphericity D50 values of the materials included in the training dataset.

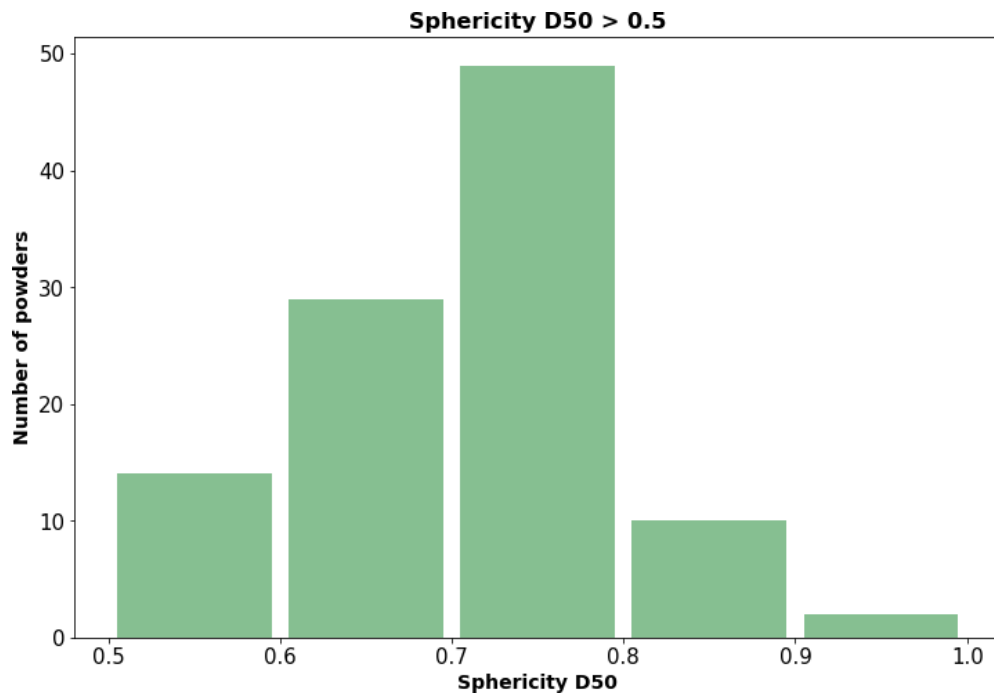


Figure 4-11: The distribution of the sphericity D50 values greater than 0.5. Most of the materials had a sphericity D50 between 0.7 and 0.8.

4.3.1.2. Surface area and energy measurements – Surface Energy Analyzer

A representative sample of 35 powders were selected for the measurement of surface area and surface energy. The surface energy parameters measured were specific surface energy (mJ/m^2), surface energy (mJ/m^2), and dispersive surface energy (mJ/m^2) at 0%, 3%, 5%, and 10% of coverage. The surface area, specific surface energy, and surface energy (com) are reported in Table 4-4.

Table 4-4: The surface area and surface energy measurements for the 35 powders analysed.

Parameter	Range of values	Mean
Surface area	0.17 to 2.76 m^2/g	0.64 m^2/g
Specific surface energy	2.94 to 16.81 mJ/m^2	7.07 \pm 0.48 mJ/m^2
Surface energy (com)	0.06 to 140.73 mJ/m^2	41.62 \pm 0.66 mJ/m^2

4.3.1.3. Powder flow and bulk density – Powder Rheometer FT4

The bulk density and powder flow of the 112 pharmaceutical powders were analysed using Powder Rheometer FT4 – Freeman Technology Ltd. Both types of measurements were done in triplicate. Fig 4-12 shows the distribution of bulk density across the powders included in the training dataset. Following the MCS, the recommendation for manufacturing efficiency is bulk density greater than 0.3 g/ml, and to achieve direct-compressible materials, 0.5 g/ml.

The Jenike’s classification (Jenike, 1964) of powder flow was adapted to label the materials included in the training dataset (see Table 4-5). The MCS recommends a FF_c greater than 4 to achieve a roller-compactable material, but it is not specified for any other manufacturing techniques. Following Jenike’s classification, a free-flowing material should be direct-compressible. These thresholds can be further adapted depending on the equipment that is intended to use but for this work, the classification found in literature was used.

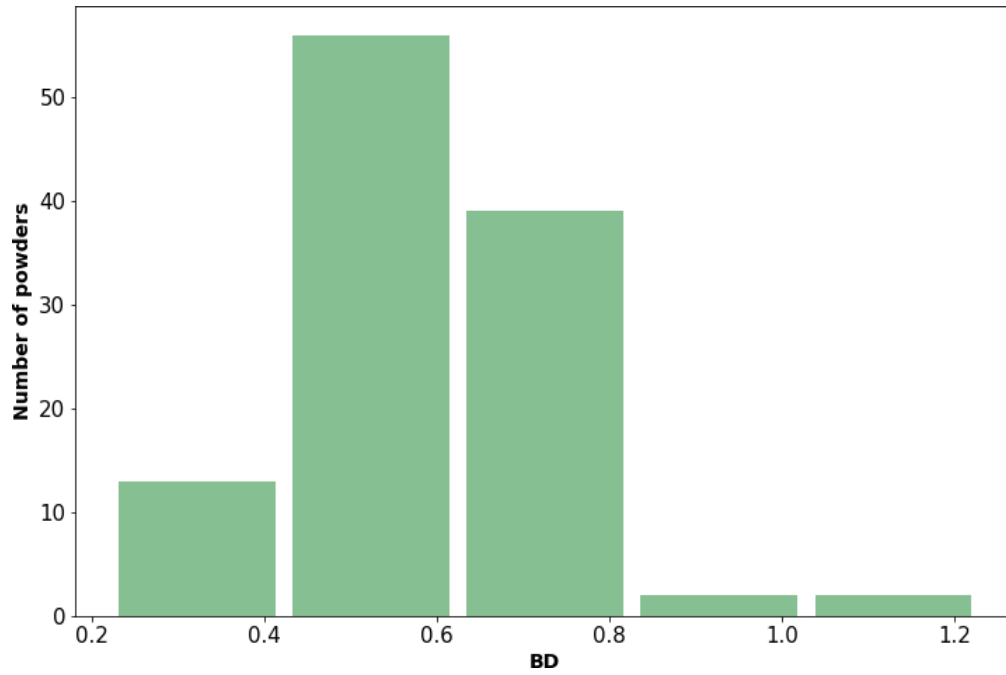


Figure 4-12: The distribution of bulk density of the pharmaceutical powders of study.

Table 4-5: The number of pharmaceutical powders in each range of interest of FFc.

Flow function coefficient	Powder behaviour	Number of observations
$FFc \leq 4$	Cohesive	29
$4 < FFc < 10$	Easy-flowing	32
$FFc \geq 10$	Free-flowing	51

4.3.2. Machine learning results

4.3.2.1. Data-driven model workflow

In this section, a workflow to guide for the development of data-driven models is presented (see Fig 4-13). The workflow includes four stages focusing in the four main areas: data, model training, model evaluation, and model interpretability.

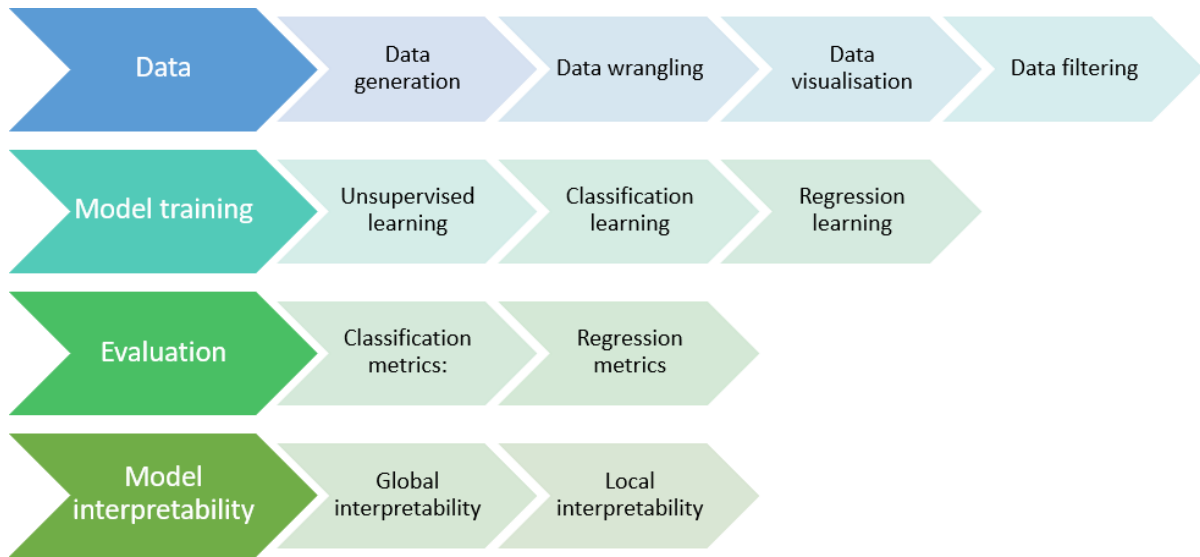


Figure 4-13: 4-stage data-driven model workflow: data, model training, model evaluation, and model interpretability.

The first step includes the generation, wrangling, visualisation, and filtering of the input data. Data wrangling involves mapping the data and gather it in the same usable file, enabling the next sub-steps of the workflow. Data visualisation is a graphical representation of the data from which information can be extracted using charts or graphs. This graphical representation makes the data more understandable, contributes to the identification of the task, and elucidates whether a data-driven model is the best approach to tackle the problem of study. Data filtering involves selecting a subset of the full dataset based on specific conditions, i.e., removing highly correlated variables and non-variance variables from the dataset.

The second step refers to model training. There are two types of ML models: unsupervised learning or supervised learning. In this thesis, unsupervised and supervised learning algorithms have been trained as described in Chapter 3.

The third step of the workflow is the evaluation of the model. To evaluate the performance of a model, suitable metrics must be chosen according to the type of ML model that was trained (classification or regression metrics). Moreover, the method used to sample the data (random split, k-fold cross-validation, bootstrap sampling) must be carefully chosen according to the data requirements.

The fourth and last step of the workflow focuses on the interpretability of the models. ML models are still perceived as a “black box” difficult to trust. Model interpretability explains how the model makes the prediction, what variables are important and how these variables affect the model’s prediction. In this chapter, SHAP values are used for model interpretability.

4.3.2.2. Data filtering using PCC

The identification of highly correlated variables is key in the data pre-processing step before training machine learning models since having correlated variables in the training data will not add information, and for some algorithms, highly correlated variables can mask the importance that other variables have on the dependent variable. To identify the highly correlated variables included in the training dataset, the PCC was calculated for each pair of variables (see Fig 4-14). The variables included in the dataset were PSD (D10, D50, D90), sphericity distribution (D10, D50, D90), aspect ratio distribution (D10, D50, D90), bulk density, the concentration of the API when the observation is a blend, and the response (FFc). High values of the PCC are presented in light orange, whereas low values of the PCC are presented in dark red. The concentration of the excipients included in the blends was not included in this analysis.

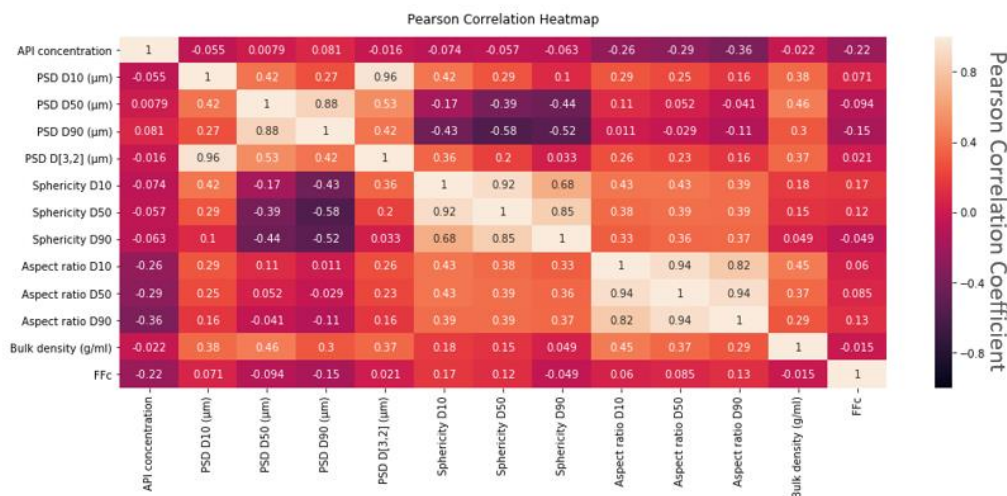


Figure 4-14: The PCC heatmap. High values of the Pearson correlation coefficient are presented in light orange, whereas low values of the PCC are presented in dark red.

The PCC was used as a filter to remove highly correlated variables (PCC > 0.9) from the training dataset. For each pair of high-correlated variables, one of the variables was randomly removed. The remaining variables were the API concentration, PSD D10, PSD D50, PSD D90, sphericity D10, sphericity D90, aspect ratio D10 and bulk density. These remaining variables were used as input variables for the machine learning models. The PCC heatmap of these final variables was plotted to confirm that no highly correlated variables remained in the dataset (see Fig 4-15). These remaining variables were used to train the ML models.

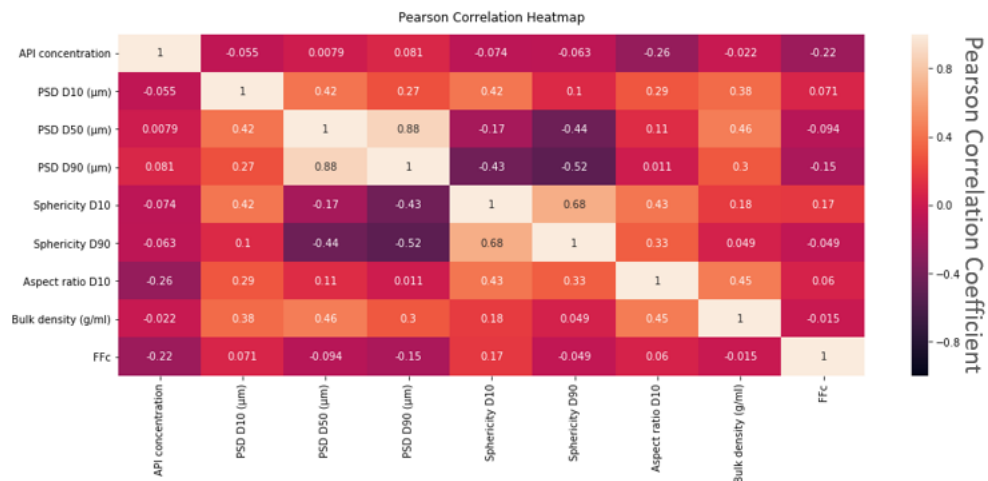


Figure 4-15: The PCC heatmap after filtering out highly correlated variables (PPC > 0.9).

4.3.2.3. Unsupervised learning

PCA, Louvain and hierarchical clustering analysis were performed using Orange Data Mining software to determine if the data could be clustered based on powder flow behaviour. In PCA, two principal components only accounted for 45% of the variance. The number of components was increased incrementally to 6 where 88% of the variance was accounted for. Louvain clustering showed the data clustered into 4 groups (see Fig 4-16), and hierarchical clustering resulted in 3 groups of data. For all methods, groups did not correlate with powder flow behaviour.

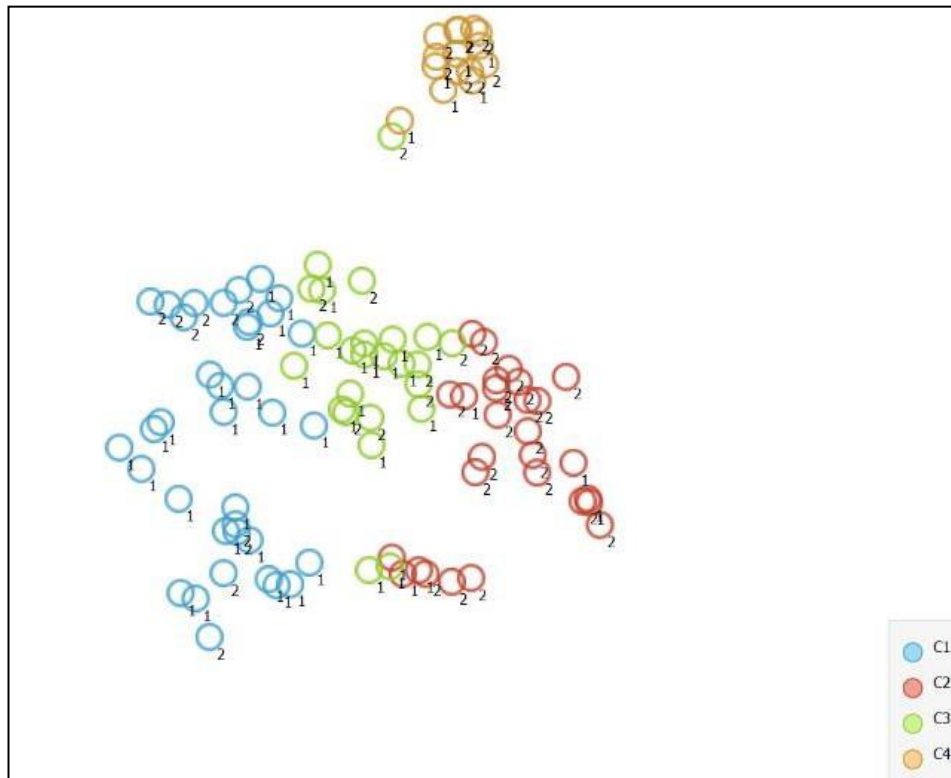


Figure 4-16: Louvain clustering analysis

4.3.2.4. Powder flow classification by a variety of supervised learning methods

Two types of models, namely a single-step and two-step classification, were investigated using supervised learning algorithms described in the ML methods section (section 3.5.2. of Chapter 3). The first model developed a single-step classification in which materials were classified into one of the three FFC classes described in Table 4-5 and illustrated in Fig 4-17. The input variables to the model are the remaining variables after the PCC filtering and the concentration of the APIs and excipients when the observation was a blend.

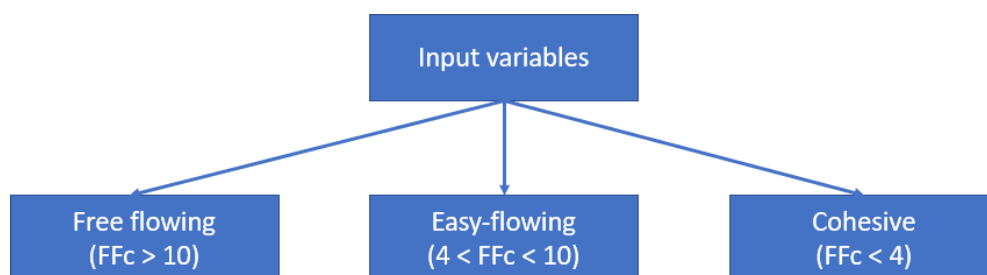


Figure 4-17: The diagram of the single-step classification model.

The performance of this single-step classification model was assessed by calculating AUC – ROC (see Table 4-6 and Fig 4-18). The Multilayer Perceptron (MLP) model achieved the highest performance (0.823). The confusion matrix of the MLP model was calculated since MLP was the best-performing algorithm (see Fig 4-19).

Table 4-6: The performance of the algorithm involved in model training. MLP achieved the highest performance (AUC-ROC = 0.823).

Model	ROC_AUC Performance	Standard deviation
kNN	0.761	0.06
RF	0.786	0.08
MLP	0.823	0.06
NB	0.726	0.09
AB	0.729	0.07
GB	0.762	0.08

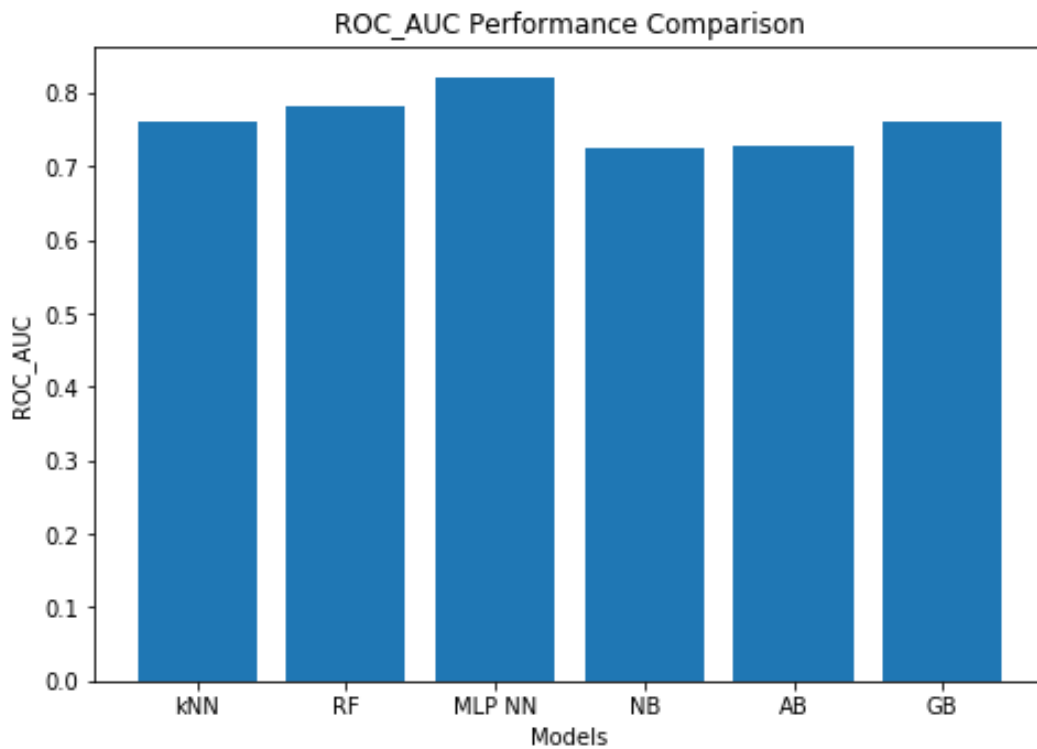


Figure 4-18: The comparison of the performance of the algorithms selected for the single-step classification model, from left to right: kNN, RF, MLP NN, NB, AB, and GB. Even though almost all of them exhibited a similar performance, MLP was selected for further analysis.

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	13	8	8	29
	Easy-flowing	5	14	13	32
	Free-flowing	3	8	40	51
Σ		21	30	61	112

Figure 4-19: MLP confusion matrix.

For classes “cohesive” and “free-flowing”, over 60% of the instances were correctly classified; however, for class “easy-flowing”, less than 45% of the materials were correctly classified by the model. The model therefore appeared to be better at predicting the FFC classes of cohesive and free-flowing materials but struggled to classify the easy-flowing (Zegzulka et al., 2020) materials across the transition from cohesive to free flowing.

As the MLP neural network confusion matrix indicated that easy-flowing materials were difficult to distinguish from free-flowing materials, a two-step classification model was developed as following Jenike's classification of powder flow (Jenike, 1964).

4.3.2.5. Two-step classification: classification model, evaluation, and model interpretability.

The input variables to the two-step classification model were the remaining variables after the PCC filtering (see Fig 4-15) and the concentration of the APIs and excipients when the observation was a blend. This two-step classification model was developed following Jenike's classification of powder flow (Jenike, 1964): Step 1 classified materials into free-flowing ($FFc \geq 10$) or non-free-flowing ($FFc < 10$), cohesive and easy-flowing powders were included in the latter category. Step 2 classified the material into cohesive materials ($FFc \leq 4$) and non-cohesive materials ($FF > 4$), easy-flowing and cohesive powders were included in the latter category (see Fig 4-20). According to the literature, easy and free-flowing powders are suitable for manufacturing with free-flowing powders being most suitable for DC (Leane et al., 2015). The performance of the algorithms included in Step 1 and Step 2 was again assessed using AUC – ROC (see Figs 4-21 and 4-25). The results showed that by separating the classification decisions, the two-step model was able to perform better than the single-step classification model. This improvement in the performance of the two-step model could be explained by considering that the imbalanced training dataset used for the single-step model affected the performance of the model, and when the model was split into steps, the detrimental impact of the imbalanced data was minimised.

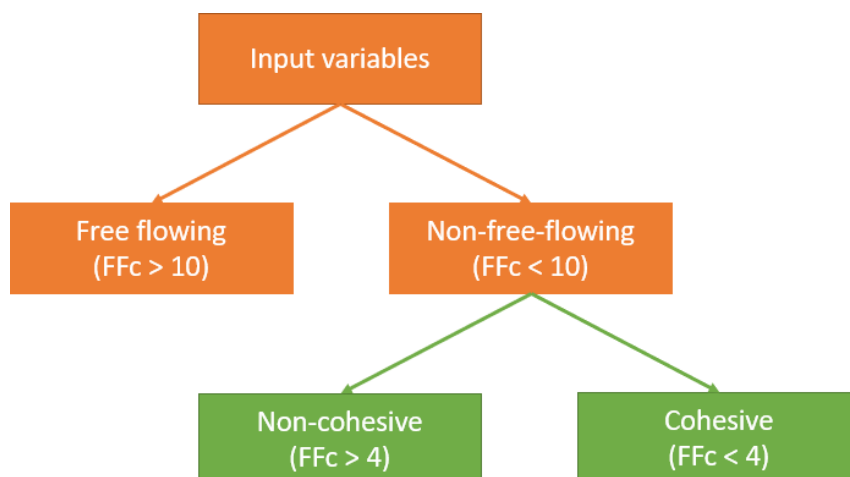


Figure 4-20: The diagram of the two-step classification model.

Since more algorithms are available for binary classification than for multiclass classification, two algorithms more were trained for the two-step model, resulting in a total of eight algorithms (compare to the six algorithms trained for the single-step model). The performance of the models was evaluated using 10-fold cross-validation and the results were reported using AUC-ROC. Table 4-7 shows the performance achieved by the algorithms trained on Step 1 (free-flowing vs. non-free-flowing materials). Overall, the performance achieved by Step 1 was better than the performance achieved by the single-step model. Fig 4-21 shows that MLP neural network model was the best performing algorithm (0.831), followed by RF (0.815). The confusion matrix of the MLP neural network model calculated using 10-fold cross-validation is presented in Fig 4-22.

Table 4-7: The performance of the classification algorithms included in Step 1 of the two-step model evaluated reported using the mean and the standard deviation of the ROC AUC calculated from the 10-fold cross-validation.

Model	ROC AUC Mean	ROC AUC Standard deviation
kNN	0.796	0.08
SVM	0.743	0.09
RF	0.815	0.09
MLP	0.831	0.08
NB	0.775	0.1
LR	0.807	0.05
AB	0.804	0.07
GB	0.805	0.08

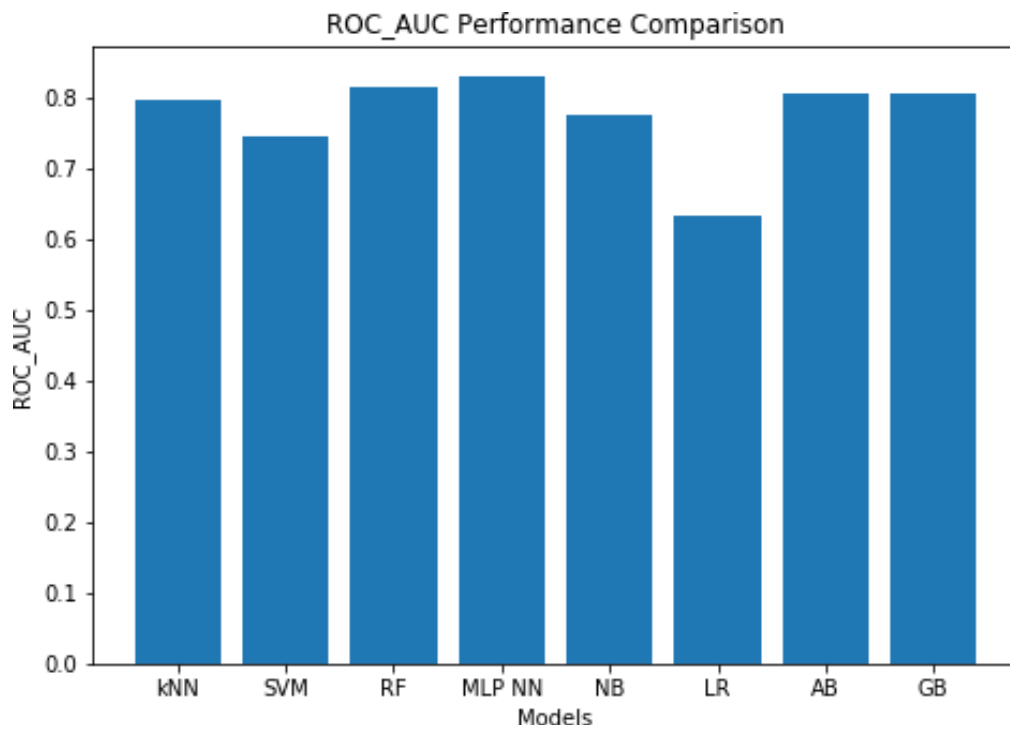


Figure 4-21: The mean ROC AUC performance of the classification algorithms included in Step 1 of the two-step model evaluated using 10-fold cross-validation.

		Predicted		Σ
		Non-free-flowing	Free-flowing	
Actual	Non-free-flowing	45	16	61
	Free-flowing	17	34	51
Σ		62	50	112

Figure 4-22: The MLP neural network confusion matrix of Step 1 of the two-step classification model evaluated using 10-fold cross-validation.

The MLP neural network model was initially used for external validation (see Fig 4-23). However, as the external validation classification accuracy was significantly worse (62.5%) than the classification accuracy for the test set, we hypothesize that the MLP neural network algorithm was overfitting the data. As the model with the next highest performance, the external validation was performed and explored for the RF model. Fig 4-24 shows that the RF model did not overfit the training data, and therefore RF was chosen as the best-performing algorithm for Step 1.

		Predicted		Σ
		Non-free-flowing	Free-flowing	
Actual	Non-free-flowing	3	2	5
	Free-flowing	1	2	3
Σ		3	5	8

Figure 4-23: The confusion matrix reporting the results of the external validation of Step 1 of the two-step classification model of the MLP neural network model. Only 62.5% of the materials were correctly classified.

		Predicted		Σ
		Non-free-flowing	Free-flowing	
Actual	Non-free-flowing	5	0	5
	Free-flowing	1	2	3
Σ		6	2	8

Figure 4-24: The confusion matrix reporting the results of the external validation of Step 1 of the two-step classification model of the RF model, where 87.5% of the observations were correctly classified.

Subsequently, the second step of the model (Step 2). The same classification algorithms were trained, using the same hyperparameters but the boundary was shifted to $FFc = 4$ to classify between cohesive and non-cohesive materials. The performance of these algorithms was assessed using 10-fold cross-validation and reported by AUC-ROC. The best-performing algorithm was RF, achieving an AUC-ROC of 0.876 (see Table 4-8 and Fig 4-25).

Table 4-8: The performance of the classification algorithms included in Step 2 of the two-step model reported by the mean and standard deviation of the ROC AUC metric using 10-fold cross-validation.

Model	ROC AUC Mean	ROC AUC Standard deviation
kNN	0.839	0.06
SVM	0.802	0.07
RF	0.876	0.09
MLP	0.861	0.08
NB	0.766	0.11
LR	0.622	0.13
AB	0.804	0.09
GB	0.839	0.07

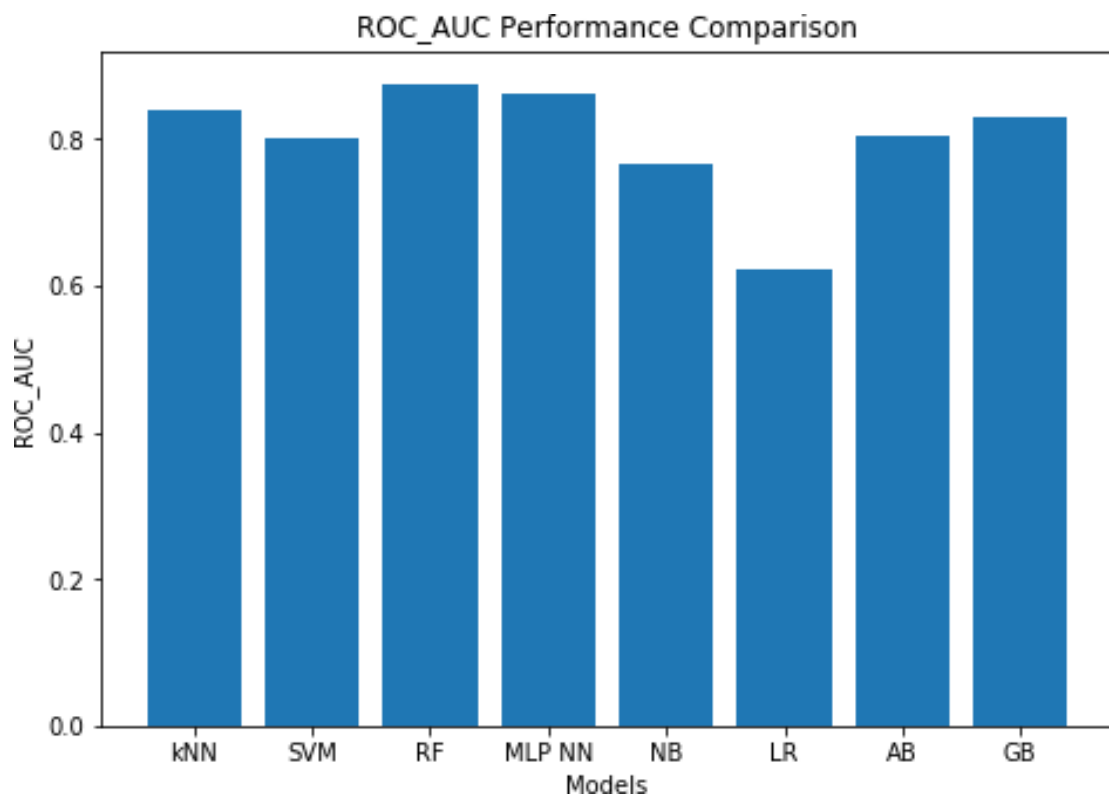


Figure 4-25: The comparison of the performance reported using the ROC AUC metric of the algorithms selected to train Step 2. RF was slightly higher than the other algorithms.

The RF models confusion matrices for Step 1 and Step 2 have been combined to have a better overview of the performance of the two-step model. The results of the 10-fold cross-validation are presented in Fig 4-26 and the results of the external testing are presented in Fig 4-27.

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	16	0	13	29
	Easy-flowing	0	29	3	32
	Free-flowing	6	13	32	51
Σ		22	42	68	112

Figure 4-26: The combined confusion matrices for Step 1 and Step 2 for the RF models, evaluated by 10-fold cross-validation.

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	0	2	0	2
	Easy-flowing	0	3	0	3
	Free-flowing	0	1	2	3
Σ		2	3	3	8

Figure 4-27: The combined confusion matrices for Step 1 and Step 2 for the RF models, evaluated by the external dataset.

SHAP values were calculated to help understand the predictions from the external validation (S. Lundberg, 2018). This method improves the interpretability of the predictive models by showing feature importance analysis. The feature that had the biggest impact on model performance in Step 1 was PSD D10 (i.e., 10% of the particles have a particle size smaller than this value; Fig 4-28(a)). Therefore, this analysis indicated that the model’s prediction between free-flowing and non-free-flowing powders was impacted significantly by the presence of fines in the material as captured in the elevated impact of the PSD D10 value. Similar results were obtained for Step 2 (Fig 4-28(b)). Hence, the presence of fines impacted both the prediction between free-flowing and non-free flowing materials and the prediction between cohesive and non-cohesive materials.

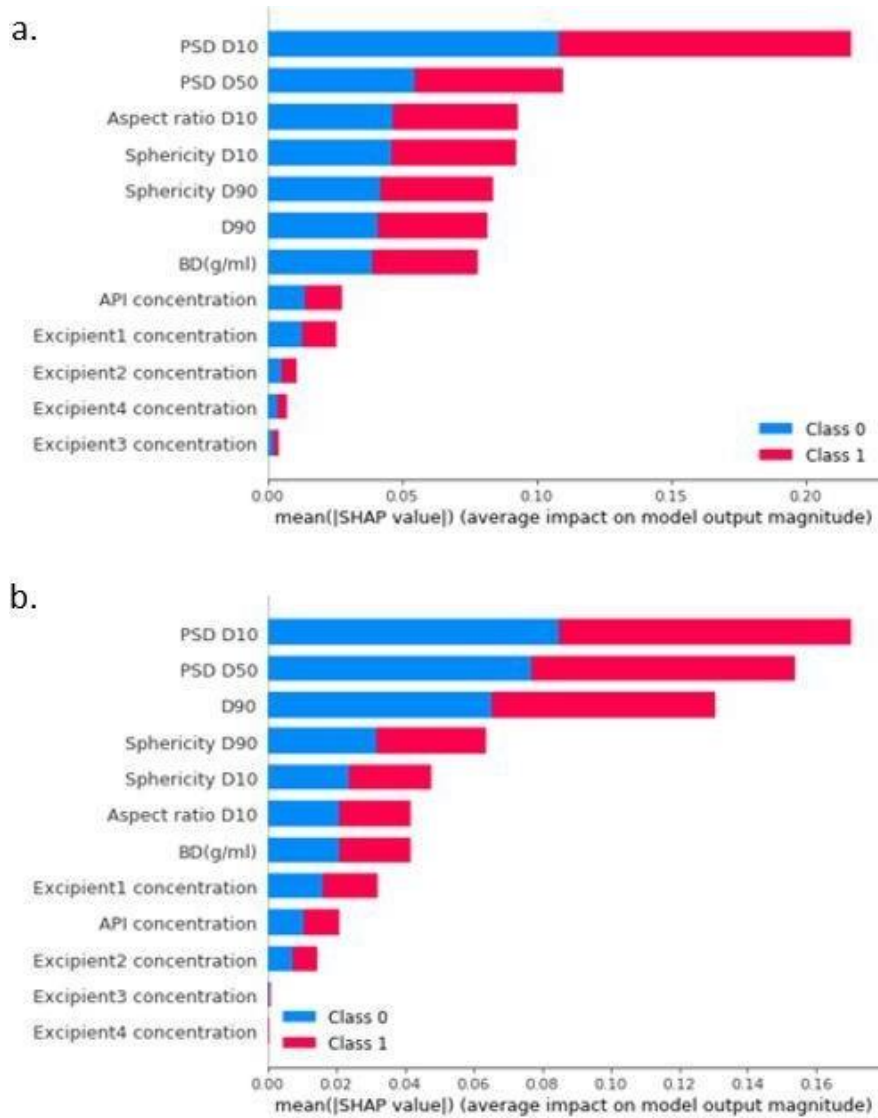


Figure 4-28: Feature importance analysis for the RF model in a) Step 1 and b) Step 2. The features are ranked based on their absolute mean SHAP score. The impact of each feature on each class is represented using colours and hence, the impact of the prediction on the output non-free-flowing is represented in blue and the impact of the output free-flowing is represented in red for Step 1. The impact of the prediction on the output cohesive is represented in blue and the impact of the output non-cohesive is represented in red for Step 2.

4.3.2.6. Adding surface area and surface energy data as input variables

Surface area and surface energy data were also added to the training set of the single-step and the two-step models because it has been previously shown that surface parameters have a significant influence on powder behaviour (Fichtner, Mahlin, Welch, Gaisford, & Alderborn, 2008; Jange &

Ambrose, 2019). The addition of these parameters resulted in a decrease in model performance for all algorithms, except kNN in Step 1, and SVM and LR in Step 2 (see Fig 4-29). While unexpected, this result can be explained by other publications that demonstrate the small influence of surface area and surface energy on the prediction of powder flow (Barjat et al., 2021). The decrease in performance due to the addition of more data can be a result of the small correlation between surface area and surface energy with powder flow, because the information we introduce to the model is effectively noise. This result suggests that powder flowability is more strongly dependent on size and shape than it on surface area and surface energy. As the addition of these parameters did not improve performance, they were not included in later training datasets.

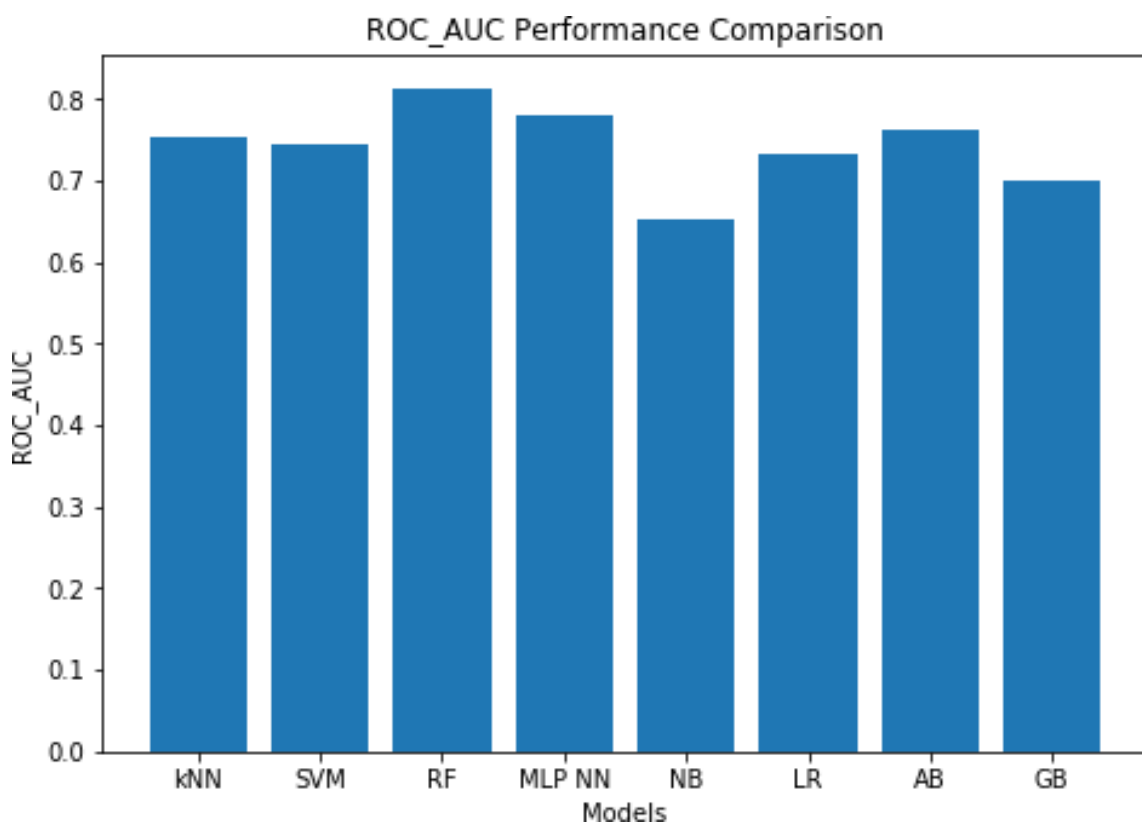


Figure 4-29: Comparison of the algorithms trained on particle size, particle shape, bulk density, surface area, specific surface energy, surface energy (com), and dispersive surface energy at 0, 3, 5 and 10% of coverage for Step 1 of the RF two-step classification model.

To confirm the hypothesis that powder flow is not correlated with surface area and surface energy, the feature performance analysis was performed using SHAP values (see Fig 4-30). Indeed, none of the surface area or surface energy descriptors were more important than size and shape descriptors for the classification of powder flow.

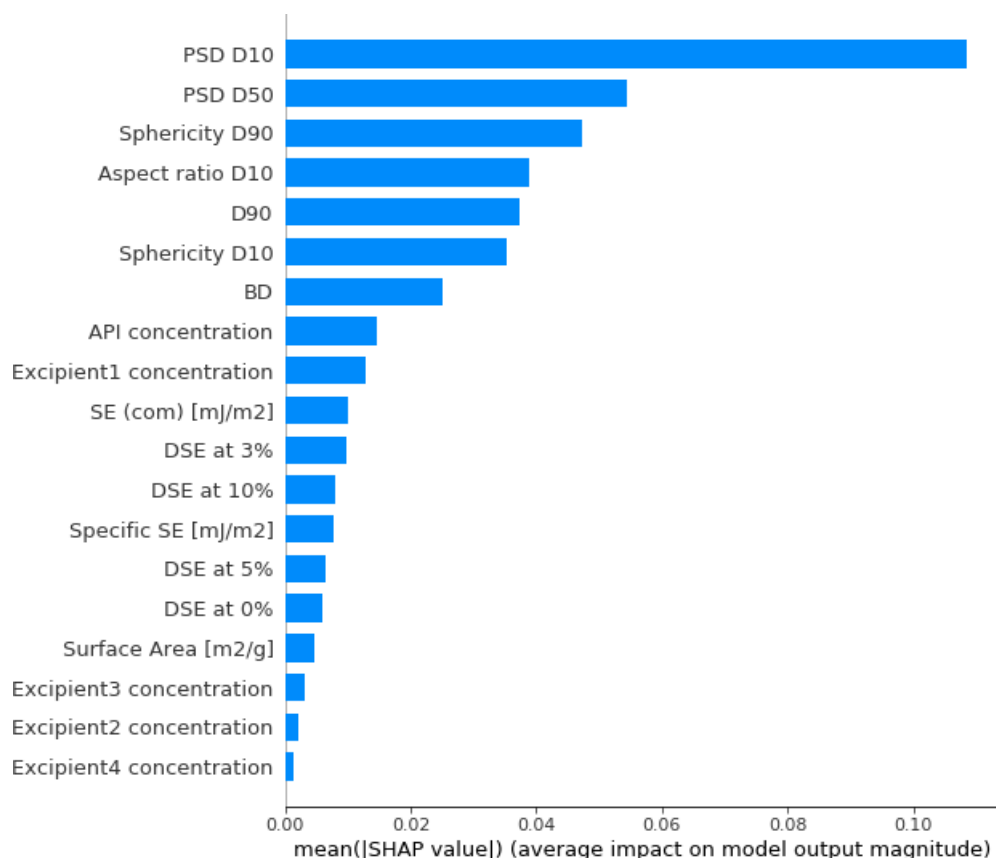


Figure 4-30: Feature importance analysis for the Step 1 of the two-step RF classification model, including surface area and surface energy as inputs of the model.

4.3.2.7. Regression models: model training, evaluation, and model interpretability.

For regression models the dependent variable (or predicted response) was the FFC value or the reciprocal of the FFC. All regression models performed poorly for FFC prediction (see Table 4-9). The reciprocal of the FFC was then calculated and used as the dependent variable. This transformation was carried out to emphasise small differences in low FFC values. For example, the difference between the behaviour of two materials with FFC 3 and 6 is more significant than the difference in the behaviour between two materials with an FFC of 13 and 16 (Zegzulka et al., 2020). Moreover, this data transformation helps normalise the distribution of the target variable and therefore potentially optimising the performance of the regression models.

Table 4-9: Regression metrics to evaluate the performance of the algorithms used to build the regression model, using FFC as the independent variable.

Model	R²	MSE	RMSE	MAE
RF	0.445	77.97	8.83	6.85
GB	0.232	107.82	10.38	7.76
AB	0.058	132.20	11.50	9.12

Table 4-10 shows the results of the regression models that have the reciprocal of the FFC as the dependent variable with performances evaluated by 10-fold cross-validation. From these results, we see that using the reciprocal of the FFC decreased prediction error compared to the models that predict FFC directly. For these models, RF exhibited the best performance, with an R² value of 0.553, and an RMSE of 0.13.

Table 4-10: Regression metrics to evaluate the performance of the algorithms used to build the regression model, using 1/FFC as the independent variable.

Model	R²	MSE	RMSE	MAE
RF	0.553	0.02	0.13	0.09
GB	0.385	0.02	0.15	0.10
AB	0.479	0.02	0.14	0.10

Although the regression models did not perform as well as the classification models, a further exploration of the regression models was carried out to better understand how the performance of these models could be improved. The same new, external dataset that was used in the previous section was used in this section to validate the regression models. Here, the FFC and the reciprocal of the FFC values for the 8 materials were predicted using the RF regression model, as it was the regression model with the highest performance (see Table 4-11).

Table 4-11: Results of the external validation performed with the regression model, setting the target variable first as “FFc”, and then as “1/FFc”.

Actual FFc	Predicted FFc	Actual 1/FFc	Predicted 1/FFc
1.90	15.94	0.526	0.127
2.28	15.85	0.438	0.203
7.42	8.24	0.135	0.148
7.46	9.73	0.134	0.053
8.17	15.83	0.122	0.088
32.14	13.78	0.031	0.079
38.21	27.84	0.026	0.075
23.00	17.27	0.043	0.040

The RF model is further analysed to gain a deeper understanding of how the model makes predictions using SHAP values to rank the most important features (see Fig 4-31). Once again, the model’s prediction in the regression models was impacted significantly by the presence of fines in the material as captured in the elevated impact of the PSD D10 value.

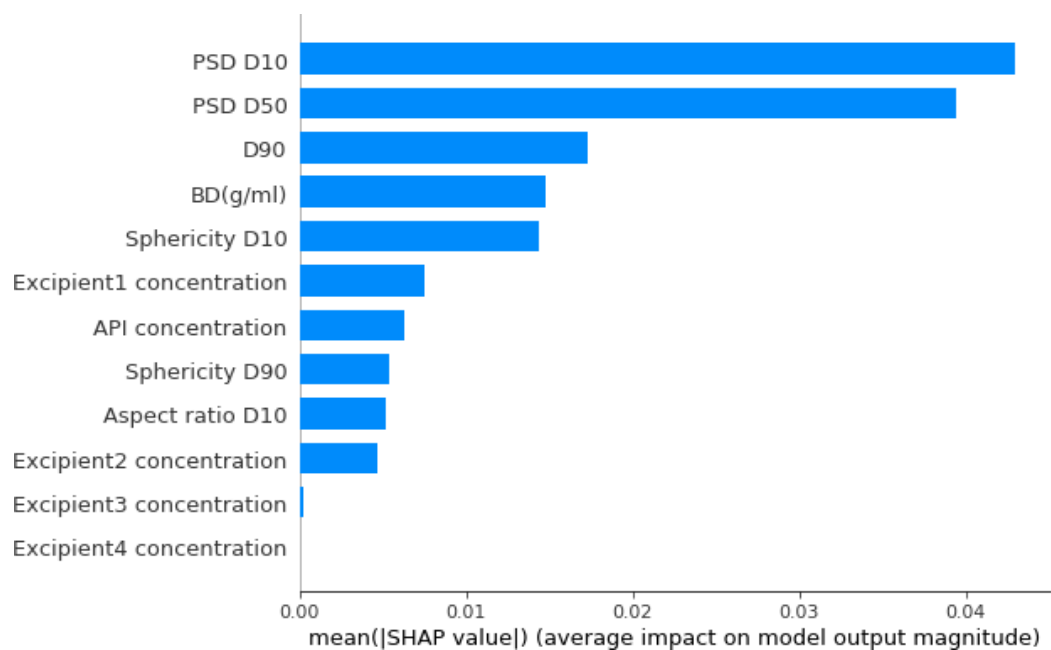


Figure 4-31: Feature importance analysis for the RF regression model. The features are ranked based on their absolute mean SHAP score.

The third most influential variable was the bulk density [BD (g/ml)]. This model was the first model that showed the bulk density as one of the most important features, and hence, this bulk property was selected for a deeper analysis using the SHAP dependence plot (see Fig 4-32). In this plot, the values of bulk density are plotted on the x-axis and the SHAP values for the bulk density, which represents how much the feature's value changes the output of the model, are plotted on the y-axis. The colour corresponds with the PSD D10 to explore its interaction with bulk density. Here, we selected PSD D10 as it was the most important variable for the regression model. High values of PSD D10 are presented in pink and low values of PSD D10 are presented in blue. We observe that values of bulk density between 0.4 and 0.8 g/ml had a negative SHAP value, which indicates a negative impact on the prediction of the reciprocal of FFC, meaning that the predicted value of the reciprocal of the FFC is likely to be low when the bulk density ranges between 0.4 and 0.8 g/ml. The observations that had a bulk density between 0.4 and 0.8 g/ml also had a low value of PSD D10. The relationship between particle attributes and bulk density is further explored in Chapter 7.

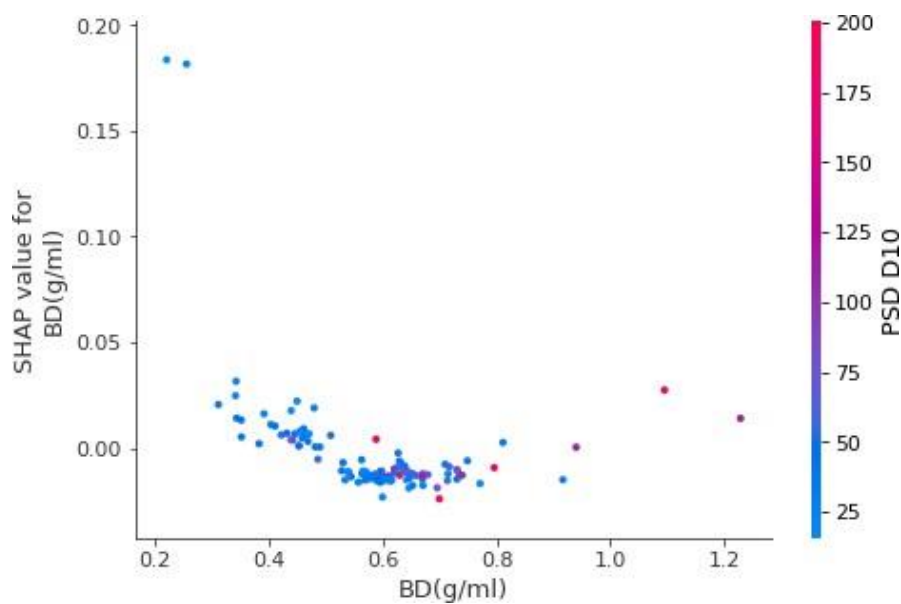


Figure 4-32: SHAP dependence plot of the RF regression model for the prediction of the reciprocal of the FFC with interaction visualisation between the bulk density and the PSD D10.

4.3.3. Exploration of the sensitivity of machine learning models to the instrument used to generate the training dataset: understanding the impact of changes in input variables

Considering the potential of modelling particle size and shape to determine manufacturability, the use of data generated by different instruments was explored to investigate the dependency of the two-step RF classification model described in section 4.3.2.5 on the particle characterisation instrument used to train the model (QICPIC®). Common particle characterization instruments, i.e., Morphologi® G3 (static image analysis) and Mastersizer® 3000 (laser diffraction) were used to study the performance of the model trained on QICPIC® data. The methodological differences of the instruments have been detailed in Chapter 3. A total of 12 powders were analysed with the Morphologi® G3 and the Mastersizer® 3000 and labelled based on their FFc (see Table 4-12) to perform explore the sensitivity of the classification model.

Table 4-12: The powders selected for the sensitivity analysis (4 cohesive, 4 easy-flowing and 4 free-flowing powders).

Powders for sensitivity analysis		
Cohesive (FFc ≤ 4)	Easy-flowing (4 < FFc < 10)	Free-flowing (FFc ≥ 10)
Span 60	Affinisol	Ac-Di-Sol
Cellulose	Soluplus	D-sorbitol
Caffeine	Ibuprofen 70	Pearlitol 300DC
Calcium Carbonate	4-amino benzoic acid	PVP

4.3.3.1. Classification models using measurements collected on three different instruments

As described in Table 4-12, the dataset for this study included four cohesive powders, four easy-flowing powders and four free-flowing powders. These powders were analysed with the Morphologi® G3 and with the Mastersizer® 3000 to evaluate the accuracy of the performance of the ML model trained on QICPIC® data.

The two-step classification model for predicting FFc class that was trained on QICPIC® data was applied to the Morphologi® G3 data and the Mastersizer® 3000 data for FFc class prediction. The results of the classification of the data are presented in the combined confusion matrices for Step 1 and Step 2 of

the classification models. The first challenge that we encountered was that the Mastersizer® 3000 only provides size data. Therefore, the required shape descriptors to run the RF classification model (sphericity and aspect ratio distribution) were taken from the Morphologi G3.

In Fig 4-33, the results of the classification of the powder flow of the 12 materials measured with the Morphologi® G3 is reported. This classification was obtained by running the two-step RF classification model trained on QICPIC®, using the Morphologi® G3 data as a test dataset. We observe that only 50% of the materials were correctly classified. Most of the misclassifications happened for the easy-flowing materials, since only one out of the four easy-flowing powders was correctly classified. The accuracy obtained here was lower than the accuracy obtained in the classification of materials measured with the QICPIC® (75%, see Fig 4-27).

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	2	2	0	4
	Easy-flowing	0	1	3	4
	Free-flowing	1	0	3	4
		3	3	6	12

Figure 4-33: Combined results of the RF model Step 1 and Step 2 for the sensitivity analysis carried out with Morphologi® G3 data.

The same methodology was applied to analyse the sensitivity of the RF classification model to the Mastersizer® 3000 data. The two-step RF classification model trained on QICPIC® data was run on Mastersizer® 3000 particle size data and Morphologi G3 shape data, since the Mastersizer® 3000 does not provide shape data. Here, the classification accuracy obtained was 41.6% (see Fig 4-34), which was a lower performance than the analysis performed on Morphologi G3 size and shape data.

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	4	0	0	4
	Easy-flowing	2	1	1	4
	Free-flowing	1	3	0	4
		7	4	1	12

Figure 4-34: Combined results of the Step 1 and Step 2 RF models for the sensitivity analysis carried out with Mastersizer® 3000 PSD data combined with the particle shape Morphologi G3 data.

The results obtained in this section indicate that the RF two-step classification model is sensitive to which instrument the input data was collected on. Therefore, to achieve the best performance of the model in classifying the flowability of a new pharmaceutical powder, we recommend that the analysis of the particle size and shape is carried out using the QICPIC® and that these data are used as input values for FFC class prediction. These results give rise to a new question: why does the model perform poorly on data collected on different instruments, and moreover, why is the class prediction accuracy better for Morphologi G3 input data than on Mastersizer® 3000 input data? This question will be explored in the next section by analysing the correlations between the different results obtained by the particle characterisation instruments.

4.3.3.2. Particle attribute measurement correlations across different instruments (QICPIC®, Morphologi® G3, and Mastersizer® 3000) to investigate the difference in performance of the classification model

A statistical analysis of the particle size and shape descriptors measured with the three instruments used as input for the RF two-step classification model is reported in this section.

Analysis of particle size distribution measurements as measured by QICPIC®, Morphologi® G3 and Mastersizer® 3000

The correlation of the PSD D10 measured with the QICPIC® (instrument used to generate the training dataset of the classification model) and the Morphologi® G3 and Mastersizer® 3000 (used to analyse how sensitive the model was to input data) is shown in Fig 4-35. In these results, we observe a high level of correlation of PSD D10 (R^2 greater than 0.9) for the three possible combinations of instruments. The highest correlation was achieved between the PSD D10 measured by the QICPIC® and the PSD D10 by the Mastersizer® 3000 ($R^2 = 0.975$). We hypothesized that this high correlation between the QICPIC® and the Mastersizer® 3000 could be due to the amount of sample used in each technique: both the Mastersizer® 3000 and QICPIC® require a few grams of sample that is streamed in front of a speed camera (QICPIC®) or analysed by laser diffraction (Mastersizer® 3000). In contrast, the Morphologi® G3 requires only approximately 500 mg of sample which is scattered onto a plate for analysis (static image analysis). Nonetheless, these results were surprising since laser diffraction techniques (Mastersizer® 3000) assume spherical particles which impacts the PSD; therefore, we would have expected to see a better correlation between image analysis based instruments (QICPIC® and Morphologi® G3).

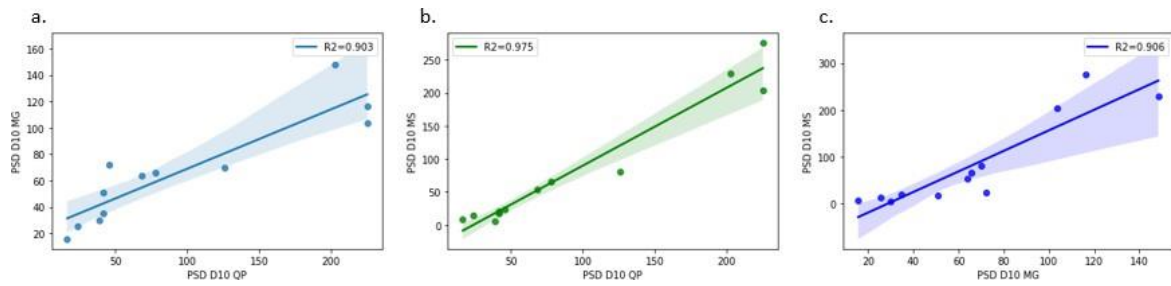


Figure 4-35: Scatter plot of PSD D10 measured using a) QICPIC® (x-axis) and Morphologi® G3 (y-axis), achieving a R^2 of 0.903; b) QICPIC® (x-axis) and Mastersizer® 3000 (y-axis), achieving a R^2 of 0.975; c) Morphologi® G3 represented (x-axis), and Mastersizer® 3000 (y-axis), achieving a R^2 of 0.906.

The results for each of the 12 materials included in the sensitivity analysis, the PSD D10 measured by the QICPIC® (shown in light blue), Morphologi® G3 (shown in cobalt blue), and Mastersizer® 3000 (shown in indigo blue) were plotted in Fig 4-36. Even though we see a good correlation overall between the PSD D10 values, for some powders, the PSD10 results of individual powders differ significantly. The PSD D10 value of Span 60 measured by the Morphologi® G3 was significantly lower than the PSD D10 measured by the QICPIC® or by the Mastersizer®. The Span 60 results may suggest that the difference in particle size observed between instruments could be explained by the presence of a mixture of primary particles and aggregates, i.e., the dispersion in the Mastersizer® 3000 may have contained aggregates whereas for the Morphologi® G3, the dispersion may have only contained primary particles.

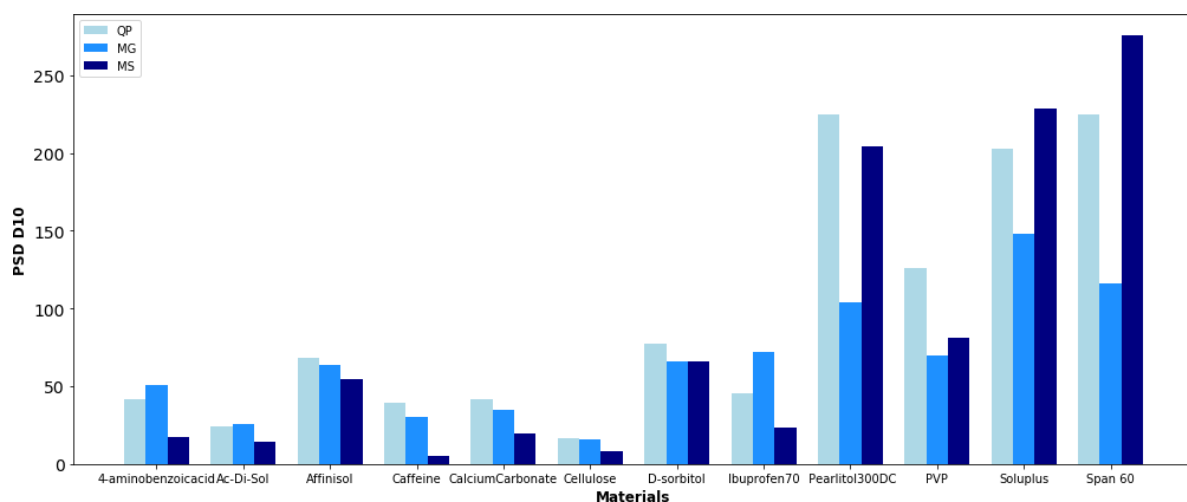


Figure 4-36: The PSD D10 measured by the QICPIC® (light blue), Morphologi® G3 (cobalt blue), and Mastersizer® (indigo blue) of the 12 powders included in this analysis.

The same analysis was performed for PSD D50 and PSD D90. The correlation of the PSD D50 measured by the three instruments was a little lower than the correlation obtained for the PSD D10 (see Fig 4-37). In this case, the highest correlation was achieved by the comparison between QICPIC® and Morphologi® G3 (see Fig 4-37(a)), with an R^2 of 0.904. Furthermore, the correlations were considerably lower for PSD D90 values. The lower correlation observed could be explained by the fact that larger measurement values have larger uncertainties attached to them. This lack of correlation was particularly evident for the comparison between QICPIC® and Mastersizer® 3000 (see Fig 4-38 (b)).

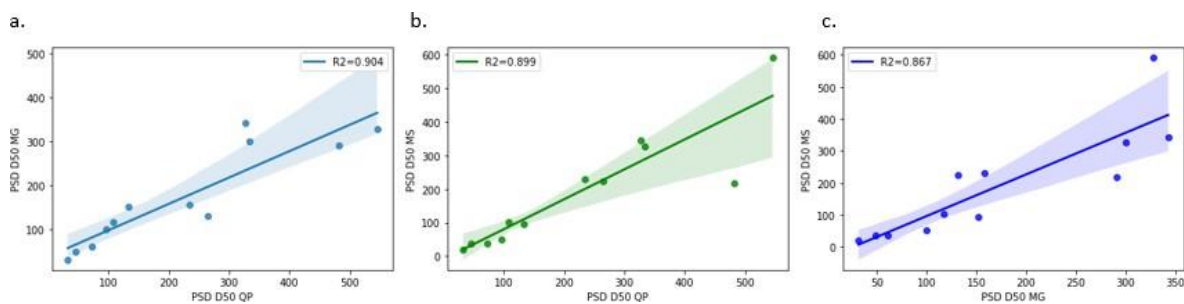


Figure 4-37: Scatter plot of PSD D50 measured using a) QICPIC® (x-axis) and Morphologi® G3 (y-axis), achieving a R^2 of 0.905; b) QICPIC® (x-axis) and Mastersizer® 3000 (y-axis), achieving a R^2 of 0.899; c) Morphologi® G3 represented (x-axis), and Mastersizer® 3000 (y-axis), achieving a R^2 of 0.867.

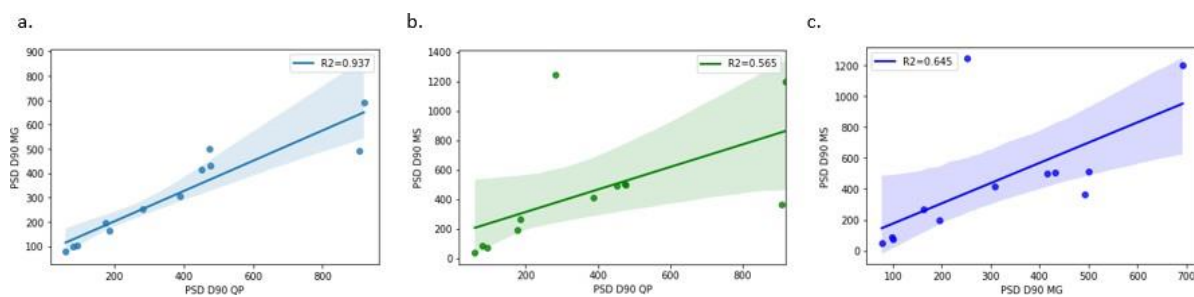


Figure 4-38: a) Scatter plot of PSD D90 measured using a) QICPIC® (x-axis) and Morphologi® G3 (y-axis), achieving a R^2 of 0.937; b) QICPIC® (x-axis) and Mastersizer® 3000 (y-axis), achieving a R^2 of 0.565; c) Morphologi® G3 represented (x-axis), and Mastersizer® 3000 (y-axis), achieving a R^2 of 0.645.

The individual results of PSD D50 and PSD D90 were compared for the 12 materials of the study (see Fig 4-39 and Fig 4-40, respectively). Once again, even though a good correlation was found for PSD D50 values across different instruments, some powders yielded very different results. For caffeine, the QICPIC® results were higher than the results of the other two methods (Morphologi® G3 and

Mastersizer® 3000). The correlation for PSD D90, as expected, was lower than for the other PSD descriptors. This difference in results was also observed by the individual PSD D90 measurements reported in Fig 4-40. Particularly, Ibuprofen 70 showed the biggest difference in results: the PSD D90 value obtained from the Mastersizer® 3000 was significantly higher than the results obtained from either of the other two instruments.

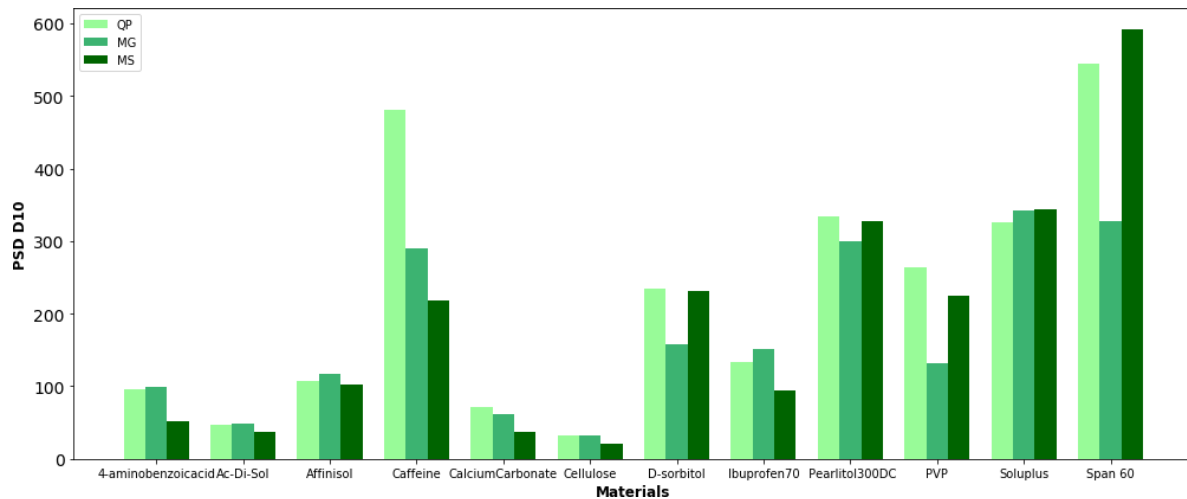


Figure 4-39: The PSD D50 measured by the QICPIC® (light green), Morphologi® G3 (medium green), and Mastersizer® (dark green) of the 12 powders included in this analysis.

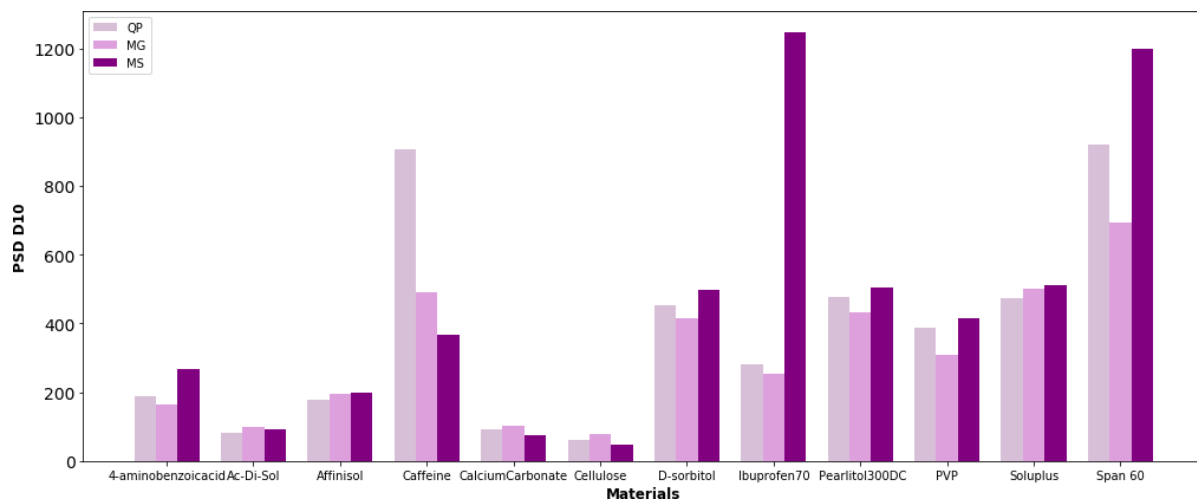


Figure 4-40: The PSD D90 measured by the QICPIC® (light pink), Morphologi® G3 (dark pink), and Mastersizer® (purple) of the 12 powders included in this analysis.

Analysis of sphericity measurements as measured by Morphologi® G3 and QICPIC®

Since the Mastersizer® 3000 does not provide shape data, the correlation between the sphericity values was calculated for the results obtained with the QICPIC® and with the Morphologi® G3 (see Fig 4-41). We observed that there was no correlation between the results obtained with these two instruments. In this section we refer to sphericity because that is the metric reported by the QICPIC®, which was used to build the ML models. The Morphologi® G3 reports this same measurement, but it is named circularity.

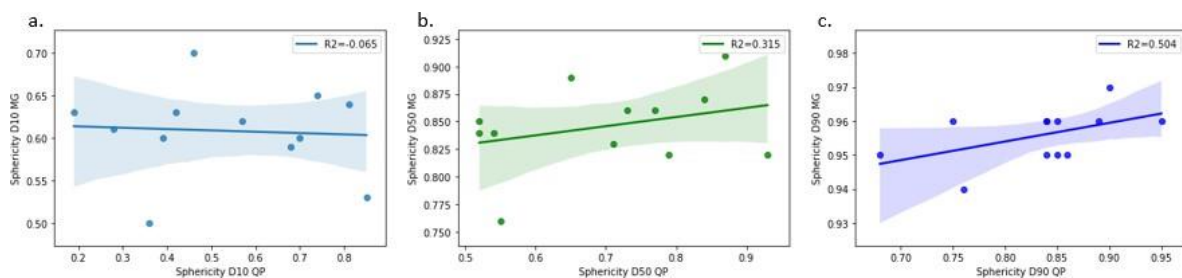


Figure 4-41: Scatter plot of sphericity results measured using QICPIC® (x-axis) and Morphologi® G3 (y-axis) of a) sphericity D10, achieving a R^2 of -0.065; b) sphericity D50, achieving a R^2 of -0.315; c) sphericity D90, achieving a R^2 of -0.504.

Analysis of aspect ratio measurements as measured by Morphologi® G3 and QICPIC®

Since the Mastersizer® 3000 does not report shape data, the correlation between the aspect ratio values obtained with the QICPIC® and with the Morphologi® G3 was calculated (see Fig 4-42). We observed that there was only a slightly better correlation between aspect ratio results than between sphericity results.

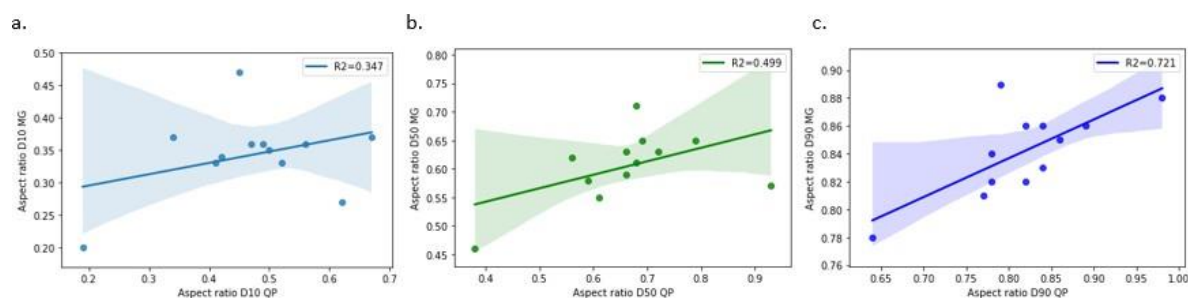


Figure 4-42: Scatter plot of aspect ratio results measured using QICPIC® (x-axis) and Morphologi® G3 (y-axis) of a) aspect ratio D10, achieving a R^2 of -0.347; b) aspect ratio D50, achieving a R^2 of -0.499; c) aspect ratio D90, achieving a R^2 of -0.721.

The results presented in this section explained why the performance of the two-step classification model trained on QICPIC® data performed better on Morphologi® G3 data than on Mastersizer® 3000 data. Even though the correlation of PSD D10 was better between the QICPIC® and Mastersizer® 3000 data, the correlation between PSD D50 and PSD D90 was better between QICPIC® and Morphologi® G3. Additionally, Mastersizer® 3000 does not provide shape data, so the Morphologi® G3 shape data was used instead. Additionally, larger deviations were found for larger particle size results, and Wadams *et al.* showed that for values greater than 40 μm laser diffraction did not perform effectively (Wadams *et al.*, 2022). Finally, these results suggest that percentile descriptors might not be appropriate to compare the particle size measurements between different instruments, and the whole distribution could be explored instead. The exploration of the robustness of these descriptors to report particle size distribution will be later presented in Chapter 6.

4.3.4. Reverse engineering: guidelines to design a direct-compressible material.

The main objective of pharmaceutical development is to manufacture high-quality drug products with the desired performance in downstream processes. Here, we propose a systematic approach to guide particle engineering scientist to obtain APIs with suitable particle properties for the direct compression. The guidelines for particle property values have been determined based on the results achieved by the Step 1 of the two-step RF classification model described in section 4.3.2.5. The desired values for each property were determined by assessing the SHAP dependence plots for each of the important variables (see Fig 4- 28(a)) on the classification between free-flowing and non-free-flowing materials. For each SHAP dependence plot, the impact of the individual value of the given powder

property on the flowability classification was considered, and the powder property values that contributed most towards a ‘free-flowing’ classification are presented here as the recommended values for this powder property (as reported in Table 4-13). Since the materials should be free-flowing to be suitable for DC, positive SHAP values are more desirable as they indicate that the model’s prediction is more likely to be “free-flowing”. Importantly, our recommended powder property values as determined by our analysis of the SHAP dependence plots (reported in Table 4-13) agree with proposed in the MCS publication (Leane et al., 2015). However, we acknowledge that achieving these properties is highly unlikely given that achieving the desired particle size and shape is not always possible even when particle engineering is used to design the API. These recommendations are meant to serve as guidelines for particle engineering scientists.

Table 4-13: Properties of a direct-compressible material.

Properties	Ranges
PSD D10	> 50 μm
PSD D50	> 100 μm
PSD D90	< 700 μm
Sphericity D10	0.5 – 0.8
Sphericity D90	0.7 – 0.8
Aspect ratio D10	0.5 – 0.7
Bulk density	0.5 – 0.7 g/mL

The analysis of the SHAP dependence value plots results indicated that to achieve a direct-compressible powder, the PSD D10 should be greater than 50 μm , the PSD D50 should be greater than 100 μm , and the PSD D90 should be smaller than 700 μm . Regarding particle shape, sphericity D10 should be between 0.5 and 0.8, and sphericity D90 should be between 0.7 and 0.8. Finally, to achieve a direct-compressible powder, its bulk density should be between 0.5 and 0.7 g/ml.

The RF SHAP dependence values were calculated for each powder property variable to determine the optimal values of this powder property for achieving a direct-compressible powder. For each property of interest, the powder property values are plotted on the x-axis, and the impact of the value of this variable on the prediction of the model (i.e., the corresponding SHAP value) is plotted on the y-axis.

Thus, Fig 4-43 shows the impact of the PSD D10 values on the model outcome (in this case, whether a powder is free flowing). In the RF SHAP dependence plot of the PSD D10, we observed that as the PSD D10 values increase, the impact of the PSD D10 value on the model outcome of being free-flowing increases (as indicated by the increase of the associated SHAP values). Therefore, the higher the PSD D10, the more likely the model’s prediction is “free-flowing”. For values of PSD D10 smaller than 50 μm , the impact of the given PSD D10 value on the model outcome was negative, meaning that when the powder had a PSD D10 value smaller than 50 μm , the model’s prediction was more likely to be “non-free-flowing”. Following this procedure, the recommendation of PSD D10 value greater than 50 μm is proposed in Table 4-13.

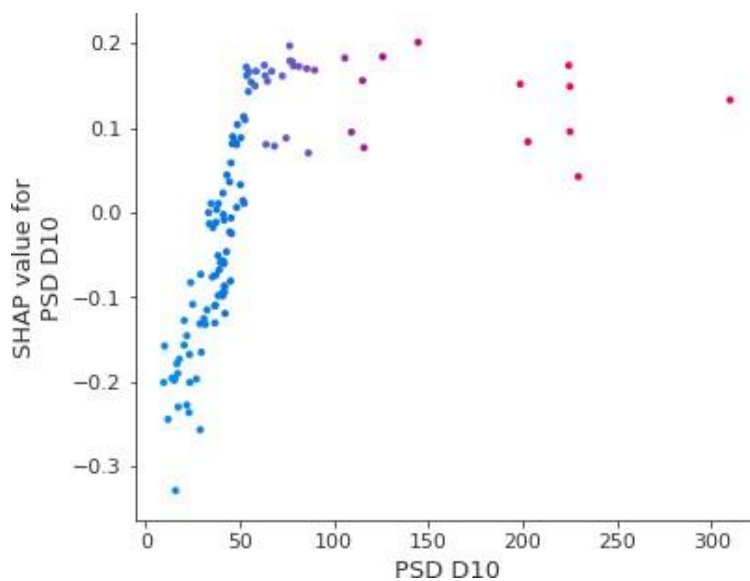


Figure 4-43: RF Step 1 classification model SHAP dependence plot of PSD D10. The colour of the data points corresponds with the x-axis values in that data points are presented in blue for low values of PSD D10, and pink for high values of PSD D10.

The same reasoning was applied for the remaining important variables for the first step of the two-step RF classification model presented in section 4.3.2.5 (again, see Figure 4-28 for the list of these variables).

We present here how the recommendation of the optimum values of sphericity D90 to achieve a direct-compressible powder was established as, in this case, the analysis yielded unexpected results. Again, the RF SHAP dependence plot was used (see Fig 4-44). The values of sphericity D90 were plotted on the x-axis and the SHAP values was plotted on the y-axis. Low values of sphericity D90 are presented in blue and high values of sphericity D90 are presented in pink. Unexpectedly, for values of sphericity

D90 greater than 0.7, the impact that the given sphericity D90 value had on the model outcome decreases. In particular, we see that values of sphericity D90 greater than 0.8 have a negative impact on the model outcome, meaning that if the value of sphericity of D90 was greater than 0.8, the model's prediction was more likely to be "non-free-flowing". This is unexpected as we would assume that as sphericity increases, the powder would likely to be more free flowing. Values between 0.7 and 0.8 had a positive impact on the model outcome, meaning that the model's prediction was more likely to be "free-flowing". Therefore, this range of sphericity D90 between 0.7 and 0.8 was taken as the recommendation to achieve a direct-compressible powder. For values smaller than 0.75, it was difficult to draw conclusions because only eight materials that had this value of sphericity D90 were included in the training dataset.

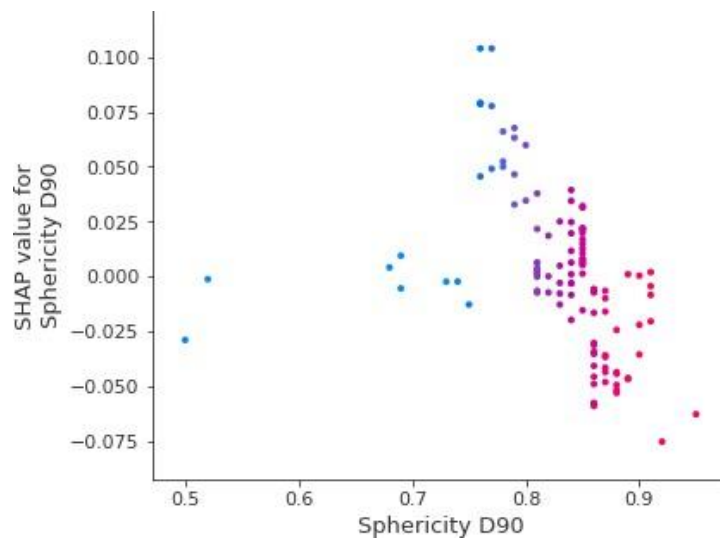


Figure 4-44: RF Step 1 model SHAP dependence plot of Sphericity D90. The colour of the data points corresponds with the x-axis values in that data points are presented in blue for low values of Sphericity D90, and pink for high values of Sphericity D90.

The rest of the value ranges of the properties were determined following the same methodology. The plots of these powder properties are not presented here since the result obtained were not unexpected.

To test if these collective recommendations result in the expected model predictions, three fictional materials were tested. Three materials hereinafter referred as "Material 1", "Material 2", and "Material 3" were created with different particle properties that were either within the recommended ranges of values or outside the recommended values:

- i. All the particle properties of *Material 1* were within the direct-compressible targets presented in Table 4-13.
- ii. The PSD of *Material 2* was within the range of recommended values, but the particle shape distribution (aspect ratio and sphericity) was outside the recommended values.
- iii. All the particle properties of *Material 3* were outside the recommended values to achieve a direct-compressible material.

The results of the predictions were analysed using SHAP force plots (see Fig 45). In these plots, the variables that have a positive impact on the predicted class (increase the probability of being direct-compressible) are represented in pink, whereas the variables that have a negative impact on the predicted class (decrease the probability of being direct-compressible) are represented in blue. The plots also show the actual value of the variables below the bars and the probability of being classified as direct compressible above the bars. Therefore, SHAP force plots help understand the individual contribution of each input variable towards the probability of the material of being direct-compressible.

Material 1 was classified as free-flowing using the first step of the two-step RF classification model presented in section 4.3.2.5., with a probability of 61% (Fig 4-45(a)). For *Material 1*, the main driver that contributed towards the prediction was PSD D10 (longest pink bar), which had a positive impact on the predicted class increasing the likelihood of the material of being direct-compressible. The value of PSD D10 of *Material 1* was 60 μm , which is within the recommendations presented in Table 4-13 (PSD D10 should be greater than 50 μm to achieve a direct-compressible material).

Fig 4-45(b) shows that for *Material 2*, the probability of the prediction of being free-flowing was 50%, and the model classified it as non-free-flowing. In this case, the PSD D10 was within the recommendations, but the particle shape descriptors were not. Once again, the PSD D10 had the biggest impact on the model's prediction (the longest pink bar), but the shape descriptors that were outside the recommended values for DC (sphericity and aspect ratio) drove the prediction towards non-free-flowing (represented in blue).

Finally, Fig 4-45(c) shows that *Material 3* was classified as non-free-flowing, with a probability of being free-flowing of 29%. The main drivers that contributed to the prediction were the PSD D90, the PSD D50 and the PSD D10. These PSD values were smaller than the recommended values. These PSD values drove the prediction towards non-free-flowing.



Figure 4-45: Flowability predictions for *Materials 1 (a), 2 (b), and 3 (c)*. Variables that increased the probability of being free-flowing were presented in pink, and variables that decreased the probability of being free-flowing were presented in blue. The length of the bars of the variables represented how much the variables impact the model outcome.

The results presented in this section showed that when all the powder properties of a test material were within the recommendations proposed in Table 4-13, the model's prediction was that the material would be free flowing. When only the particle size was within the recommendations of Table 4-13, the model predicted a probability of being free flowing of 50%, showing uncertainty. Finally, when all the values were outside of the recommendations made in Table 4-13, the model predicted that the material would be non-free-flowing.

4.4. Conclusions

Implementing ML models in the early stages of drug development can help determine suitable manufacturing strategies for a given material and provide rapid digital screening tools for advanced pharmaceutical development. In this work, FFC classes of pharmaceutical materials were predicted from routine, widely available, material-sparing analytical measurements. The 112 materials analysed exhibited a wide range of PSDs, particle shape distributions, and bulk densities and covered 3 classes of FFC that reflect what is captured in the literature (Jenike, 1964).

This work suggests that particle size and shape distribution measured with dynamic image analysis are sufficient to enable the prediction of flow properties. The best performing model presented in this work was achieved by the combination of RF models for Step 1 and Step 2, with over 80% probability of distinguishing between classes for each step. Further improvements to model performance could be made with more data from cohesive materials as this would help address class imbalance in the training dataset. Additionally, including training data with different combinations of particle size and shape with differing bulk behaviour could also reduce misclassifications in future models. The FFC boundaries of the classes of powder flow could also be adapted to specific industry needs; for example, optimal FFC values will vary depending on the different pieces of equipment that might be available. In this work, propagation of analytical measurement error has not been included in the model training, and this research angle could be interesting to explore in further work. Moreover, the model could be extended to inform formulation optimization or even to provide a performance target for particle engineering efforts to develop materials for direct compression.

Additionally, an analysis of how sensitive the RF classification model was to the input data was performed. This analysis revealed that the model was sensitive to the data used to make predictions. The model achieved a better prediction with Morphologi[®] G3 data than with Mastersizer[®] 3000. However, these predictions were worse than the results obtained when QICPIC[®]) data was used. Lastly, guidelines to ensure the development of direct-compressible APIs were proposed in the final section of this chapter. These recommendations agreed and expanded the proposed guidelines in the MCS.

The ML model's implementation enables the prediction of the material flow properties (FFC) from size and shape allowing early decision-making regarding manufacturing route selection. The implementation of the models presented here in industry applications, particularly in early-stage development, could help reduce the amount of materials needed and potentially eliminating measurements, which are costly, by serving as a screening of the most optimal powders for the development of a new medicine. The work presented in this chapter illustrates the benefits of implementing digital design workflows for the prediction of material properties in the pharmaceutical industry where the availability of data is often limited. This work highlighted multiple potential applications that could result from increasing the available FAIR data in this industry and how it can help to digitalise pharmaceutical manufacturing.

4.5. References

- Abe, H., Yasui, S., Kuwata, A., & Takeuchi, H. (2009). Improving powder flow properties of a direct compression formulation using a two-step glidant mixing process. *Chemical and Pharmaceutical Bulletin*, 57(7), 647-652.
- Barjat, H., Checkley, S., Chitu, T., Dawson, N., Farshchi, A., Ferreira, A., . . . Tobyn, M. (2021). Demonstration of the Feasibility of Predicting the Flow of Pharmaceutically Relevant Powders from Particle and Bulk Physical Properties. *Journal of pharmaceutical innovation*, 16(1), 181-196. doi:10.1007/s12247-020-09433-5
- Bellamy, L. J., Nordon, A., & Littlejohn, D. (2008). Effects of particle size and cohesive properties on mixing studied by non-contact NIR. *International journal of pharmaceutics*, 361(1-2), 87-91.
- Benesty, J., Chen, J., & Huang, Y. (2008). On the importance of the Pearson correlation coefficient in noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4), 757-765.
- Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2), 1-50.
- Crouter, A., & Briens, L. (2014). The effect of moisture on the flowability of pharmaceutical excipients. *Aaps Pharmscitech*, 15(1), 65-74.
- Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., . . . Starič, A. (2013). Orange: data mining toolbox in Python. *the Journal of machine Learning research*, 14(1), 2349-2353.
- Dorris, G. M., & Gray, D. G. (1980). Adsorption of n-alkanes at zero surface coverage on cellulose paper and wood fibers. *Journal of Colloid and Interface Science*, 77(2), 353-362.
- Fichtner, F., Mahlin, D., Welch, K., Gaisford, S., & Alderborn, G. (2008). Effect of surface energy on powder compactibility. *Pharmaceutical research*, 25(12), 2750-2759.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4), 463-484.
- Goh, H. P., Heng, P. W. S., & Liew, C. V. (2018). Comparative evaluation of powder flow parameters with reference to particle size and shape. *International journal of pharmaceutics*, 547(1-2), 133-141.
- Guo, A., Beddow, J., & Vetter, A. (1985). A simple relationship between particle shape effects and density, flow rate and Hausner ratio. *Powder Technology*, 43(3), 279-284.
- Harris, C., Millman, K., van der Walt, S., Gommers, R., Virtanen, P., Cournapeau, D., . . . Berg, S. (2020). Smith 474 nj. *Kern R, Picus M, Hoyer S, van Kerkwijk MH, Brett M, Haldane A, del R'io JF, Wiebe M, Peterson P, G'erard-475 Marchant P, et al. Array programming with NumPy. Nature*, 585(7825), 357-362.
- Hlinak, A. J., Kuriyan, K., Morris, K. R., Reklaitis, G. V., & Basu, P. K. (2006). Understanding critical material properties for solid dosage form design. *Journal of pharmaceutical innovation*, 1(1), 12-17.
- Jaiswal, J. K., & Samikannu, R. (2017). *Application of random forest algorithm on feature subset selection and classification and regression*. Paper presented at the 2017 world congress on computing and communication technologies (WCCCT).
- Jange, C. G., & Ambrose, R. K. (2019). Effect of surface compositional difference on powder flow properties. *Powder Technology*, 344, 363-372.
- Jenike, A. W. (1964). Storage and flow of solids. *Bulletin No. 123, Utah State University*.
- Kaerger, J. S., Edge, S., & Price, R. (2004). Influence of particle size and shape on flowability and compactibility of binary mixtures of paracetamol and microcrystalline cellulose. *European Journal of Pharmaceutical Sciences*, 22(2-3), 173-179.

- Kunnath, K., Chen, L., Zheng, K., & Davé, R. N. (2021). Assessing predictability of packing porosity and bulk density enhancements after dry coating of pharmaceutical powders. *Powder Technology*, 377, 709-722.
- Leane, M., Pitt, K., Reynolds, G., & Group, M. C. S. W. (2015). A proposal for a drug product Manufacturing Classification System (MCS) for oral solid dosage forms. *Pharmaceutical development and technology*, 20(1), 12-21.
- Lever, J. (2016). Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. *Nature methods*, 13(8), 603-605.
- Liu, Q., Wang, X., Huang, X., & Yin, X. (2020). Prediction model of rock mass class using classification and regression tree integrated AdaBoost algorithm based on TBM driving data. *Tunnelling and Underground Space Technology*, 106, 103595.
- Ltd, F. T. Shear Testing. Retrieved from <https://www.freemantech.co.uk/powder-testing/ft4-powder-rheometer-powder-flow-tester/shear-testing><https://www.freemantech.co.uk/powder-testing/ft4-powder-rheometer-powder-flow-tester/shear-testing>
- Lundberg, S. (2018). Welcome to the SHAP documentation🌐.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Maier, J. (2017). Made smarter review. *UK Industrial Digitalisation Review*.
- Masuda, H., Higashitani, K., & Yoshida, H. (2006). *Powder technology handbook*: CRC press.
- Matplotlib. (2012-2022). Matplotlib. Retrieved from <https://matplotlib.org/>
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*: John Wiley & Sons.
- Organization, W. H. (2012). S. 3.6. Bulk Density and Tapped Density of Powders. *The International Pharmacopoeia*, 6.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
- Pinto, T. S., Lima, O., & Leal Filho, L. (2009). Sphericity of apatite particles determined by gas permeability through packed beds. *Mining, Metallurgy & Exploration*, 26(2), 105-108.
- Schaller, B. E., Moroney, K. M., Castro-Dominguez, B., Cronin, P., Belen-Girona, J., Ruane, P., . . . Walker, G. M. (2019). Systematic development of a high dosage formulation to enable direct compression of a poorly flowing API: A case study. *INT J PHARMAcEUT*, 566, 615-630. doi:10.1016/j.ijpharm.2019.05.073
- Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37-52): Springer.
- Shah, U. V., Karde, V., Ghoroi, C., & Heng, J. Y. (2017). Influence of particle properties on powder bulk behaviour and processability. *International journal of pharmaceuticals*, 518(1-2), 138-154.
- Shangraw, R. F. (1989). Compressed tablets by direct compression. *Pharmaceutical dosage forms: Tablets*, 1, 195-246.
- Shekunov, B. Y., Chattopadhyay, P., Tong, H. H., & Chow, A. H. (2007). Particle size analysis in pharmaceuticals: principles, methods and applications. *Pharmaceutical research*, 24(2), 203-227.
- Sheugh, L., & Alizadeh, S. H. (2015). *A note on pearson correlation coefficient as a metric of similarity in recommender system*. Paper presented at the 2015 AI & Robotics (IRANOPEN).
- Sigrist, F. (2018). Gradient and newton boosting for classification and regression. *arXiv preprint arXiv:1808.03064*.
- Sun, C., & Grant, D. J. (2001). Effects of initial particle size on the tableting properties of L-lysine monohydrochloride dihydrate powder. *International journal of pharmaceuticals*, 215(1-2), 221-228.
- Trementozzi, A. N., Leung, C.-Y., Osei-Yeboah, F., Irdam, E., Lin, Y., MacPhee, J. M., . . . Zawaneh, P. N. (2017). Engineered particles demonstrate improved flow properties at elevated drug loadings

- for direct compression manufacturing. *Int J Pharm*, 523(1), 133-141. doi:10.1016/j.ijpharm.2017.03.011
- Van Rossum, G., & Drake, F. L. (2009). *Python/C Api Manual-Python 3*: CreateSpace.
- Wadams, R. C., Akseli, I., Albrecht, J., Ferreira, A. P., Gamble, J. F., Leane, M., . . . Tobyn, M. (2022). Particle Property Characterization and Data Curation for Effective Powder Property Modeling in the Pharmaceutical Industry. *Aaps Pharmscitech*, 23(8), 286.
- Yu, W., Muteki, K., Zhang, L., & Kim, G. (2011). Prediction of Bulk Powder Flow Performance Using Comprehensive Particle Size and Particle Shape Distributions. *J. Pharm. Sci*, 100(1), 284-293. doi:10.1002/jps.22254
- Zegzulka, J., Gelnar, D., Jezerska, L., Prokes, R., & Rozbroj, J. (2020). Characterization and flowability methods for metal powders. *Scientific Reports*, 10(1), 1-19.
- Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324.

5. Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

***Declaration:** This chapter is the result of a collaboration between Laura Pereira Diaz (writing, investigation, data curation, methodology, validation), Stephanie Marchal (supervision, editing), Paul Kroll (supervision, validation, editing), Albert Hofstetter (supervision, editing), Moritz Lang (supervision, validation, editing), and Patrick M. Piccione (supervision, editing).*

5.1. Introduction

Pharmaceutical formulation is the process in which the chemical substances that form a final medicinal product are combined, including the active pharmaceutical ingredient (API) and the other components, referred to as excipients (Afrin & Gupta, 2020). Artificial intelligence (AI) and Machine Learning (ML) have emerged as potential tools to optimize the transition from formulation development to manufacturing and thus, the use of digital design and data-driven models provides the prospect to accelerate these important development steps (Yang et al., 2019). Changes in formulation from variations in excipients, their composition, or variations in drug loading can impact bulk properties such as powder flowability and therefore risk impacting subsequent manufacturing processes (Leane, Pitt, Reynolds, & Group, 2015). More subtle changes, for instance in the physical properties of API, such as particle size and shape, can also influence the manufacturability of the drug product (Leane et al., 2018; Ticehurst & Marziano, 2015). This chapter focuses on three different ML models that used in combination can facilitate formulation development.

Formulations have a substantial impact on the quality of life of patients including the outcome of the disease and the adherence to the treatment. Formulation and process development activities are necessary to make a drug substance into a safe, effective, and quality drug product yet can be a resource-consuming, iterative process. A structured, systematic approach during development allows researchers to develop products by describing the phenomena of interest to the level of understanding currently available (ten Kate, Piccione, Westbye, & Becker, 2022). Pharmaceutical formulation is a process that involves several steps where the API is blended with excipients. To achieve a successful formulation, the properties and the stability of the API and the excipients and any interaction between the components and manufacturability of the blend are key aspects to consider (Afrin & Gupta, 2020). The development of formulations is still heavily reliant on time consuming trial-

and-error methods. Consequently, to relieve pressure on the pharmaceutical industry to accelerate the introduction of new medicines and reduce the cost of healthcare, there is a pressing need to optimise the time-consuming process of formulation development (Stewart et al., 2016; Yang et al., 2019; Zhang et al., 2017).

As mentioned in previous chapters, powder flow is a complex property that is key to ensure the manufacturability of a formulation. Several studies have been published regarding the flowability characterisation of powder blends. Nalluri et al., have focused on an avalanche testing instrument to characterise the flowability of pharmaceutical binary blends (Nalluri & Kuentz, 2010). Morin *et al.* also used avalanche measurements to evaluate the effect of lubricants on powder formulations (Morin & Briens, 2013). Vasilenko *et al.* characterised a gravitational displacement rheometer flow index and compared it to shear cell measurements of fourteen blends (Vasilenko, Glasser, & Muzzio, 2011).

Recently, the granular Bond number has been used to predict the powder flow of binary and ternary mixtures from their single component's particle and bulk properties. Whilst effective, this mechanistic model has some limitations for pharmaceutical systems as it assumes even contact between particles regardless of their size and shape (Giraud et al., 2021). Furthermore, Discrete Element Modelling (DEM) has been used to predict bulk powder flow of pharmaceutical materials in rotary tablet presses (Hildebrandt, Gopireddy, Scherließ, & Urbanetz, 2019), and to predict the powder flow of binary and ternary mixtures to foresee the mechanical work of a rotating impeller penetrating a packed bed, based on the mean Bond number of the individual components (Hildebrandt et al., 2019; Pasha, Hekiem, Jia, & Ghadiri, 2020). The increasing number of publications on this topic highlights the importance on the characterisation and prediction of powder flow for mixtures, yet there is still a need for a generally applicable method for the prediction of powder flow for pharmaceutical materials and processes.

In the work presented in this chapter, our efforts have been focused in two aims:

- (i) The assessment of the viability of pharmaceutical formulations for continuous direct compression (cDC) based on the prediction of their value of flow function coefficient (FFc), and the wall friction angle (WFA).
- (ii) The consideration of the analytical error of the measurement instruments for the prediction of the viability of the formulations for cDC by calculating the confidence intervals of the predictions.

The first aim of this chapter was achieved by predicting and studying the FFc and the WFA, measured using the Ring Shear Tester from D. Schulze, to ensure viability for an industrial mini-batch cDC

manufacturing line. In DC, the components of a blend (API and excipients) are mixed without any intermediate step and, to be successfully applied, blends must be free flowing. The implementation of cDC reduces the amount of material required to make the tablet, but the process becomes challenging when the blends exhibit poor flowability (Andrews, 2022). Thus, formulations that have an FFC greater than 5, and a WFA less than 20° were considered viable for cDC. These values were selected based on previous industrial experience and considering the requirements of the cDC manufacturing line. The unit operations of the cDC process are feeding, blending and compression, with automated transfers between the unit operations. The main disadvantage of cDC is the time and waste during the start-up and ramp down of the process. To minimise the cDC disadvantages while achieving the benefits of continuous compression, a semi-continuous Mini-Batch cDC is being implemented by Roche (Bautista et al., 2022).

For this study, both classification and regression ML models have been developed. The variables included in the dataset to train the models were selected considering the availability of particle size and shape data and the feasibility of the prediction of powder flow from size and shape descriptors demonstrated in previous publications (Barjat et al., 2021; Pereira Diaz, Brown, & Florence, 2021). The implementation of ML models would assist formulators to elucidate the optimum composition of excipients and their concentration for a given API in order to achieve a viable formulation for cDC.

As the adoption of ML methods has increased substantially in recent years, interpretability methods to better understand how these models make predictions have become key to building trust and confidence in the application of ML models (Kaur et al., 2020). Understanding how ML models make predictions of bulk behaviour has the potential of accelerating placing these new medicines in the market. A traditional approach of extracting information from ML models is estimating the probability that an observation will fall into a given class. A common method to calculating such probabilities is training logistic regression models; however, logistic regression is usually not suitable for the type of high-dimensional, non-linear data used in this study (Kruppa et al., 2014). Therefore, for this study, the probabilities of an observation belonging to a given class were calculated using the Random Forest (RF) models, coupled with Monte Carlo (MC) simulations of test formulations that capture the error of the analytical measurements.

The second aim focuses on the calculation of the confidence intervals of the predictions accounting for the error of the analytical instruments used to measure the API properties. For each test formulation, MC methods were used to simulate one thousand formulations considering the analytical error of the instruments. The RF model was used to classify these simulated formulations into viable or non-viable for cDC, and the classification probability was calculated for each simulated formulation.

The range between the minimum and maximum probability of being viable for cDC was reported as the range of probabilities in which the model was 90% confident that the actual probability will be found. The estimation of probabilities was coupled with SHAP analysis (S. Lundberg, 2018; S. M. Lundberg & Lee, 2017), to increase the confidence on the outcomes of the models.

5.2. Materials and methods

A total of 83 Roche proprietary formulations have been used to build the dataset to train and test the ML models, formed by the combination of 24 different APIs and 35 excipients. To extend this dataset, 20 CMAC formulations were added comprising 5 APIs and 4 excipients, achieving a total of 103 experimental formulations, of which 99 formulations were used for training the ML models and 4 formulations were used for external testing. The training datasets contained the particle size and shape measurements of the APIs included in the formulations, the concentration of the APIs, the FFC of the APIs, the concentration of each excipient included in the formulations and the volume of the measurement cell of the instrument used to measure the FFC of the formulation. The target or dependent variable of the dataset was the FFC of the formulations,

5.2.1. Experimental methods

Particle size and shape of the APIs were analysed using a Morphologi[®] G3, Malvern. Morphologi[®] G3 is a static image analysis technique that captures the properties of the particles. The primary particle size distribution (PSD) is represented with the values of CE Diameter $D[n, x]$ (μm), and the particle shape is represented with the following descriptors: Length $D[n, x]$ (μm), Width $D[n, x]$ (μm), Aspect Ratio $D[n, x]$, Circularity $D[n, x]$, Convexity $D[n, x]$, Elongation $D[n, x]$, HS Circularity $D[n, x]$, Area $D[n, x]$ (μm^2), and Perimeter $D[n, x]$ (μm). The PSD was also analysed using a Mastersizer[®] 3000, Malvern. Mastersizer[®] 3000 uses laser diffraction, in this case, in wet dispersion to report the PSD of a sample in 101 particle size bins. The particle size and shape of the APIs included in the formulations (see Table 5-1) in combination with the concentration of the excipients in each formulation, were used to train the ML models, for the prediction of the viability of pharmaceutical formulations for cDC.

Table 5-1: The independent variables used to train the ML model measured with the Morphologi[®] G3 and the Mastersizer[®] 3000.

Instrument	Independent variables
Morphologi[®] G3, Malvern	Length D[n, x] (μm), Width D[n, x] (μm), Aspect Ratio D[n, x], CE Diameter D[n, x] (μm), Circularity D[n, x], Convexity D[n, x], Elongation D[n, x], HS Circularity D[n, x], Area D[n, x] (μm ²), Perimeter D[n, x] (μm)
Mastersizer[®] 3000, Malvern	101 particle size bins that include the whole distribution.

The FFc and the WFA were measured with the ring shear tester created by Dietmar Schulze at normal consolidation stress of 1000 Pa, since this was the relevant pressure of the cDC manufacturing line under consideration at Roche. The FFc of the formulations was the dependent variable of the ML models developed to predict the viability of the formulations.

5.2.2. Machine learning methods: data curation, unsupervised (PCA) and supervised (classification and regression) learning algorithms.

Data curation involved filtering and preparing the final dataset that will be used in the training step, i.e., highly correlated variables (PCC > 0.95), and non-variance variables were removed to avoid noise in the dataset. Initially, all the properties reported in Table 5-1, the FFc of the API, the FFc of each excipient, and the concentration of the API and the excipients, were included in the training dataset. After filtering, these variables decreased in number and the final independent variables considered are reported in Table 5-2. 101 particle size variables were initially considered from the Mastersizer[®] 3000, but only 6 particle size bins remained after filtering (0.099 μm, 0.214 μm, 0.243 μm, 40.146 μm, 163.490 μm, 756.449 μm). These variables represent particle size bins, and their values indicate the upper limit size of the particles included in the bins. The concentration of the excipients included in the blends were included as independent variables of the model.

Table 5-2: The independent variables included in the training dataset after filtering, including the particle size and shape variables of the API included in the from the Morphologi® G3 and the Mastersizer® 3000 particle size bins (of the API), the FFC of the APIs included in the formulations, and the concentration of each of the 35 different excipients included in the formulations.

Instrument	Independent variables
Morphologi® G3, Malvern	Length D[n, x] (um) D10, Aspect Ratio D[n, x] D10, Aspect Ratio D[n, x] D50, CE Diameter D[n, x] (um) D10, Circularity D[n, x] D10, HS Circularity D[n, x] D10, HS Circularity D[n, x] D50, HS Circularity D[n, x] D90, Area D[n, x] (um ²) D90
Mastersizer® 3000, Malvern	0.099 µm, 0.214 µm, 0.243 µm, 40.146 µm, 163.490 µm, 756.449 µm.
RST-Xs.® Dietmar Schulze	API FFC
Concentrations	Drug loading and concentration of each excipient (total of 35 different excipients)

Subsequently, Principal Component Analysis (PCA) was performed to identify smaller groups of variables and the hidden patterns within the data (Abdi & Williams, 2010; Education, 2020). PCA was applied using the python package Sci-kit learn (version 1.0.2) decomposition and visualized using the python (version 3.7) library matplotlib (version 3.1.1.) (Matplotlib, 2012-2022). Hotelling's *T*-squared test was applied to define the space of the model's domain. This method works by computing the chi-square test across the *n*-components, in this case, 18 (variance explained = 96.26%). The detection of outliers was determined using Fisher's method, and the alpha parameter for the detection of outliers was the default 0.05.

Successively, supervised learning algorithms were used to build the FFC models (classification and regression models), and the WFA regression model. In the FFC classification model, the formulations were classified into viable and non-viable classes, with a threshold defined at FFC = 5. The first step was the selection of the best-performing algorithm, using precision as the metric to compare algorithms' performance. Here, precision is defined as the ratio of the true positives to the total instances predicted as positives. High precision ensures that the formulations that are predicted as viable are indeed viable and therefore, suitable for cDC. The main reason to choose this metric was to avoid false positives (non-viable formulations predicted as viable), which would lead to a waste of time and resources and jeopardize user confidence in the model. The models were trained using RF and

they were robustly validated using bootstrap sampling and 10-fold cross-validation. After training, the models were further validated with an external dataset to reduce the risk of overfitting.

Two regression models for the prediction of the FFC were built: the first model aimed to predict the unmodified FFC value of the formulations, whereas the second model aimed to predict the reciprocal of the FFC of the formulations. This data transformation is appropriate (Barjat et al., 2021) for this dependent variable, since it emphasizes changes that occur at lower FFC values and hence, matches the reality, i.e., changes at low FFC values have more impact on the behaviour of a formulation than changes at high FFC values, and thus, two formulations that have an FFC of 4 and 6, have a bigger difference in their behaviour than two formulations that have an FFC of 14 and 16. For the applicability of the prediction, the global wall friction angle (PHIW), calculated following Eq 5-1, was used in the WFA regression model as the dependent variable.

$$PHIW = \arctan \left[\frac{\text{Wall friction coefficient}}{SIGMA_w (Pa)} + \tan (WFA) \right] \quad (5-1)$$

5.2.3. Monte Carlo Simulation considering the analytical error to estimate the probability of classification.

Computer simulations are implemented in our daily lives in airlines, or entertainment industries. Recently, these simulations are also used in the pharmaceutical industry for the discovery of new medicines, in molecular modelling or designing clinical trials (Bonate, 2001). In 1979, a framework for computer simulation practices in real-world problems was introduced by the Society of Computer Simulation. Computer simulations can be classified as deterministic (fixed parameters) or stochastic (random parameters), the latter being crucial to define real-world problems. MC simulations repeatedly simulate data from a pre-defined sampling distribution using stochastic parameters. The term “Monte Carlo” was coined by Ulam and von Neumann (Rubinstein, 1981) during the development of the atomic bomb for the Manhattan Project (Murphy, 2018), and since then, these methods have been extensively expanded to many fields, including engineering, biology and chemistry (Kroese, Brereton, Taimre, & Botev, 2014).

For this study, MC simulation has been used to generate 1000 formulations sampling from each of the four formulations included in the external test set, considering the value of each physical or bulk property that characterizes the sample as the mean of the distribution, and the analytical error of each instrument as the standard deviation of the distribution. The trained ML models were then run to classify these simulated 1000 formulations into viable or non-viable for cDC. The probability of classification of each simulated formulation was calculated as the mean predicted class probabilities of the trees in the forest (Olson & Wyner, 2018). The interval between the minimum and the maximum probability was taken as the confidence interval of the classification of the test formulations. Estimation of probabilities has been reported useful in several fields such as medicine (Gooley, Leisenring, Crowley, & Storer, 1999; Jiang, Osl, Kim, & Ohno-Machado, 2012; McGinn, Jervis, Wisnivesky, Keitz, & Wyer, 2008), ecology (Ellison, 2004; Waits, Luikart, & Taberlet, 2001), and sports forecasting (Garnica-Caparrós, Memmert, & Wunderlich, 2022; Lam, 2018; Štrumbelj, 2014). Moreover, the interval of probabilities can be categorised following a traffic-light system. Hence, formulations which the minimum value of their range of probabilities is greater than 50% are considered *green formulations*, formulations that have an interval of probabilities that includes 50% are considered *amber formulations*, and formulations that the maximum value of their interval of probabilities is smaller than 50% are considered *red formulations*. By reporting the results as probabilities coupled with the traffic-light system, we observed that the outcomes of the models were better received by the formulation scientists since more information was provided.

To simulate the test formulations, the standard deviation was calculated from the error of the analytical instruments. Particle size and shape analysis instruments performed by a Morphologi[®] G3 (Malvern) had an analytical error of 10% for number-based measurements and 20% for volume-based measurements. Particle size distribution was measured by laser diffraction (Mastersizer[®] 3000, Malvern), which had an analytical error of 10%. The FFC measured with the ring shear tester at normal consolidation stress of 1000 Pa (relevant pressure of the cDC manufacturing line under consideration) had an analytical error of 18% (see Table 5-3). The standard deviation calculated from the analytical error of each instrument was multiplied by 3.2 to ensure 90% confidence intervals (Higgins & Deeks, 2011).

Table 5-3: Standard deviation of the measurements performed by the three different instruments that were used to analyse the data required to build the model.

Instrument	Variables	Standard deviation
RST-Xs.® Dietmar Schulze	API FFc.	18%
Morphologi® G3, Malvern	Length D[n, x] (µm) D10, Length D[n, x] (µm) D90, Width D[n, x] (µm) D90, Aspect Ratio D[n, x] D10, Circularity D[n, x] D90, HS Circularity D[n, x] D10, HS Circularity D[n, x] D50, HS Circularity D[n, x] D90, Area D[n, x] (µm ²) D90, Perimeter D[n, x] D90.	10%
Mastersizer® 3000, Malvern	0.099 µm, 0.214 µm, 0.243 µm, 40.146 µm, 163.490 µm, 756.449 µm.	10%

5.3. Results and discussion

5.3.1. Powder flow: flow function coefficient (FFc) models

5.3.1.1. Data curation: removing highly correlated variables and non-variance variables.

Data curation is the process of cleaning and homogenising data to facilitate data analytics. One challenge encountered was that for this set of APIs, particle size, shape, and FFc data were not available for all the samples. Missing data have a negative impact on model performance causing unreliable results. Missing characterisation data accounted for 18% of non-random missing values in the total dataset, considering that for the pharmaceutical formulations that missed data, only missed 17% of their values. Missing values under a threshold of 20% were considered acceptable (Nelson, Taylor, & MacGregor, 1996; Van Snick et al., 2018). Interpolation was used to replace missing data with the mean value of the appropriate descriptor.

Subsequently, the PCC heatmap was calculated to check the correlations between variables. The training dataset is formed by 99 formulations and 203 variables, and hence, the visualisation of the full heatmap was challenging. For pairs of variables that have a PCC greater than 0.95, one of the variables was removed from the training dataset, and hence, the number of variables decreased from 203 to 97. Highly correlated variables do not add information to the dataset but instead they can create noise that can affect the performance of the model; hence, one variable of each pair of highly correlated variables was removed randomly, since the impact of removing either of them would be the same. Non-variance variables were also removed, further decreasing the number from 97 to 48. Thus, the final dataset used for unsupervised and supervised learning included 99 formulations and 48 variables (independent and dependent variables), as described in Table 5-2 in the methods section.

5.3.1.2. PCA: unsupervised learning for data visualisation

PCA was performed for data visualisation to study whether clusters of data were evident and to explore the feasibility of dimensionality reduction. Fig 5-1 shows that 12 principal components (PCs) were needed to explain 82% of the data variance. When reducing the number of variables from 48 to 12 components, the computational time did not decrease, and 18% of the information was lost. Hence, dimensionality reduction using PCA was discarded.

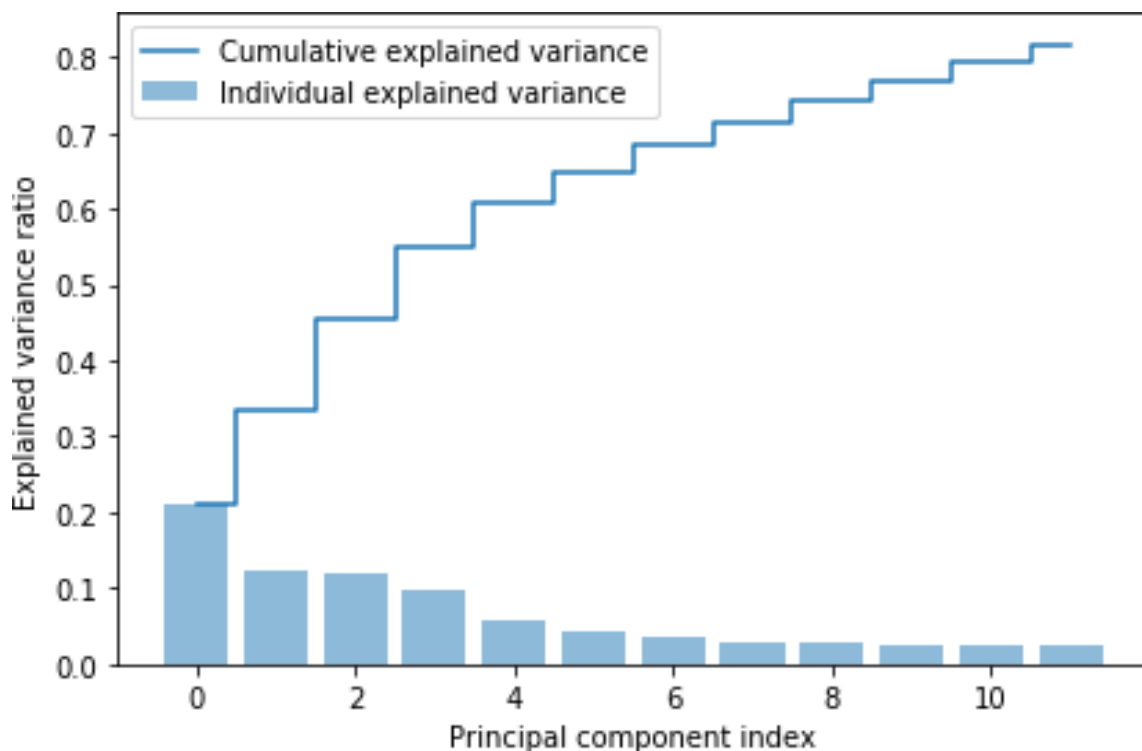


Figure 5-1: Number of PCs plotted against the explained variance ratio. The columns represent the individual explained variance by each component and the line represents the cumulative variance. The first PC explains most of the variance (approximately 20%), and when 12 PCs are used, 82% of the variance is explained.

To visualise the data, PC1 was plotted against PC2 (see Fig 5-2). The Hotelling *T*-squared ellipse was also plotted to calculate the 95% confidence domain of the model, which showed that some of the formulations included for either training or testing were outliers. The group on the left side of the ellipse was formed by 10 formulations (red circle) that included the same API at different concentrations. The API included in these formulations was not included in any of the formulations that belonged to the main cluster. These 10 formulations appear in the plot in a tight group, showing that there were not many differences between them. When these 10 formulations were compared to the formulations of the main cluster, no other significant differences in physical and/or particle properties were found. On the right side outside of the confidence ellipse, four other formulations were also plotted (orange circle). Once more, these formulations had a different API that was not included in any of the formulations that belong to the main cluster. These results helped understand how the data was grouped but it was not useful to classify the flowability of the formulations.

Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

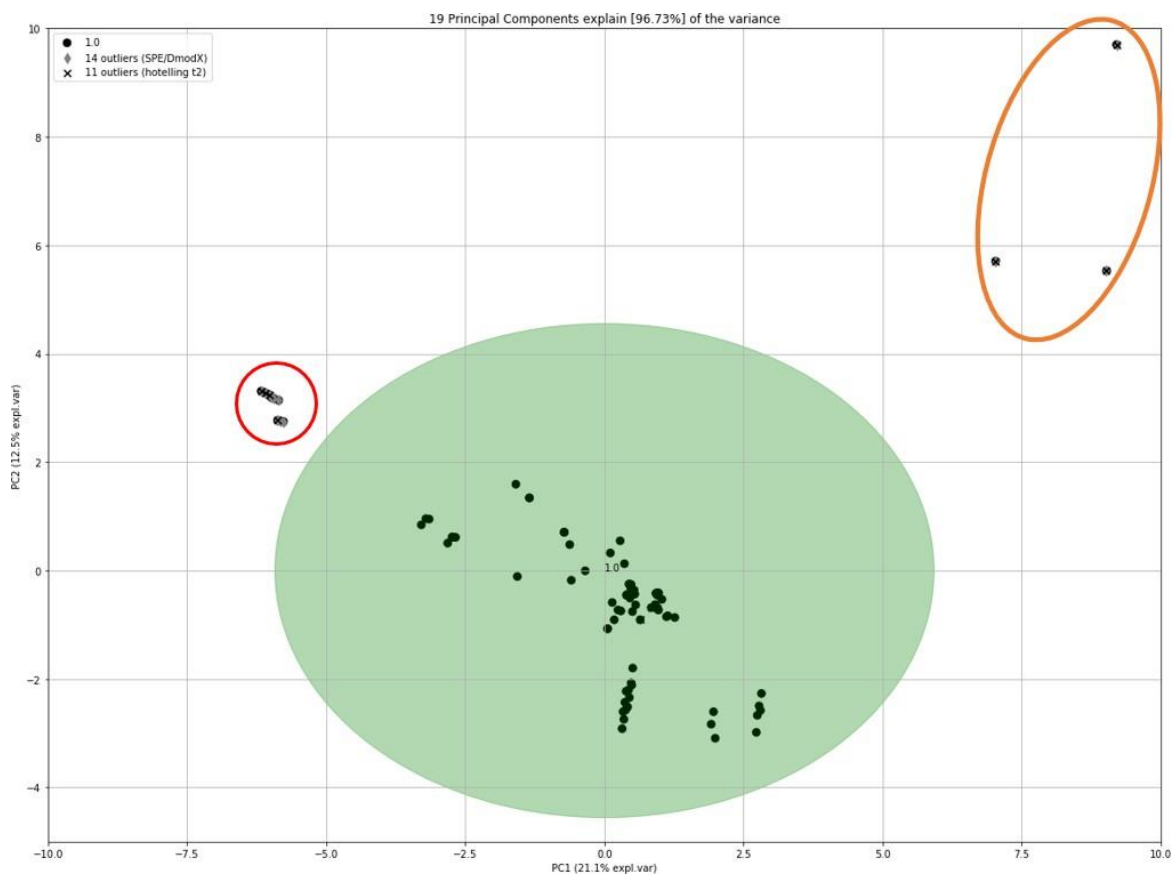


Figure 5-2: The PCA plot of PC1 vs PC2 for the whole dataset, without differentiation between classes.

The PCA loadings were computed to identify important features. The loadings describe how much information each independent variable provides to the PCs. The loading values and the squared loading values of the highest correlated variables with PC1 and PC2 are presented in Table 5-4. For PC1, the highest squared loading were HS Circularity D[n, x] D50, Circularity D[n, x] D10, and HS Circularity D[n, x] D90, so clearly the PC1 was mainly explained by the variables that describe the circularity of the particles of the API. For PC2, the highest squared loadings were the concentration of FlowLac 90, the concentration of Aerosil 200, and the CE Diameter D[n, x] (μm) D10 of the API. Hence, the PC2 was not only influenced by the shape of the API included in the formulations, but also by the concentration of excipients. Interestingly, FlowLac 90 was mainly present in viable formulations ($\text{FFc} > 5$), so it was not unexpected to see that the value of this descriptor would have an important impact. In contrast, of the 39 formulations that contained Aerosil 200 in the training dataset, 32 were non-viable ($\text{FFc} < 5$), likely due to the poor flowability of the other components of the formulation. This fact established a bias in the model and therefore highlights one of its limitations. To overcome this challenge, more data with different combinations of excipients resulting in better flowing formulations should be added to the training dataset to expand the applicability of this model.

Table 5-4: Loadings of the highest correlated variables with PC1 and PC2 and their squared loading scores for comparison. The impact of the independent variables, including the particle size and shape of the APIs included in the formulation and the concentration of the excipients included in the formulation are reported through the loadings. The highest correlated variable with PC1 was HS Circularity D[n, x] D50, and the highest correlated variable with PC2 was the concentration of FlowLac 90.

Variables	PC1 loadings	PC1 squared loadings	PC2 loadings	PC2 squared loadings
HS Circularity D[n, x] D50	-0.311	0.097	-0.085	0.007
Circularity D[n, x] D10	-0.275	0.076	0.219	0.048
HS Circularity D[n, x] D90	-0.271	0.073	-0.063	0.004
Aspect Ratio D[n, x] D50	-0.27	0.073	0.18	0.032
CE Diameter D[n, x] (µm) D10	0.236	0.056	0.239	0.057
Length D[n, x] (µm) D10	0.23	0.053	0.237	0.056
Concentration of Isomalt Galen IQ721	0.21	0.044	0.232	0.054
40.14 µm (PSD bin)	0.177	0.031	0.07	0.005
Concentration of Aerosil 200	0.135	0.018	0.294	0.086
Concentration of FlowLac 90	0.079	0.006	-0.3	0.090

5.3.1.3. Classification models

Classification models were built to categorise the formulations into viable ($FFc > 5$), and non-viable ($FFc < 5$) for cDC, using an RF Classifier. 67 formulations included in the training dataset belonged to the non-viable class, whereas only 30 belonged to the viable class. The distribution of the FFc of the formulations included in the training dataset is shown in Fig 5-3.

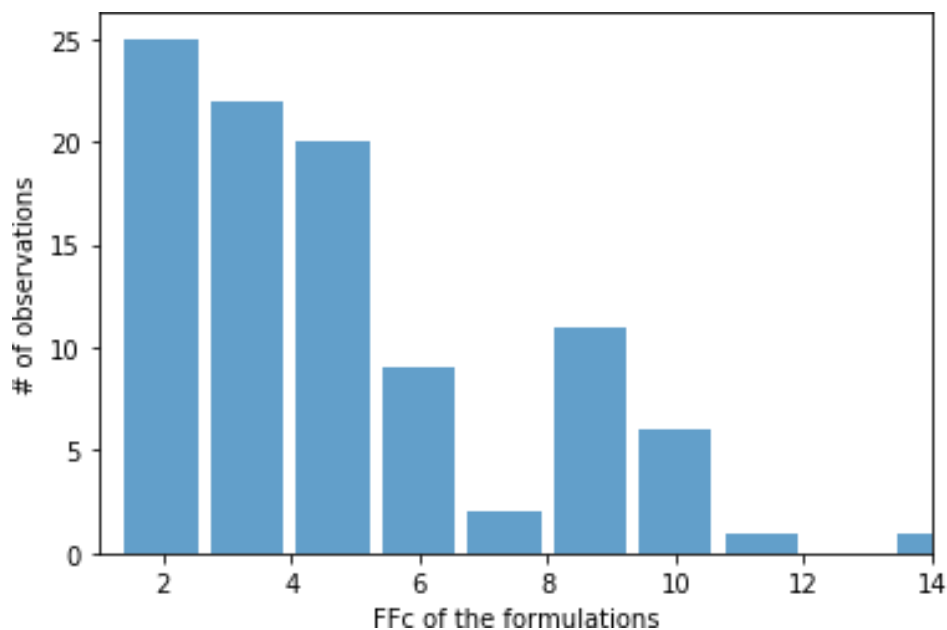


Figure 5-3: Distribution of the FFC of the formulations used for training, showing that the majority of the formulations had a low FFC (viable for cDC greater than 5).

The validation of the classification models involved four different methods:

- (i) Randomly splitting the dataset into training (75%) and testing (25%).
- (ii) Bootstrap sampling validation
- (iii) 10-fold cross-validation
- (iv) External validation

Classification metrics were calculated for each validation method, focusing on precision and recall scores, although more metrics were reported. Precision calculates the ratio of true positives (formulations that are correctly classified as viable) to the true positives and false positives (the total number of formulations classified as viable). Recall reports the ratio of the true positives to the true positives and false negatives (viable formulations that were misclassified as non-viable). These two metrics were of interest for our study because (i) precision identifies the false positives, which, in this case, are the non-viable formulations predicted as viable, and (ii) recall reports the ability to detect positive samples (viable formulations, $FFc > 5$). False positives are the most resource-consuming mistake that the model could make, since classifying one non-viable formulation ($FFc < 5$) as viable ($FFc > 5$) would result in a waste of time and material developing such formulation. Therefore, the higher the precision, the better the model is at detecting false positives. A higher recall score indicates more viable formulations detected. This metric relates to how many of the viable formulations are

successfully detected, i.e., to how useful the model is in finding “hits”. Reporting these two metrics provides an overview of the performance of the model. In the first validation method (randomly splitting between training and testing 75:25), a precision score of 0.89, and recall score of 0.80 were achieved (see Table 5-5). The confusion matrix reported that two non-viable formulations and one viable formulation were misclassified (see Fig 5-4).

Table 5-5: Classification metrics of the RF classification model using randomly splitting between training and testing (75:25).

Metric	Value
Precision	0.89
Accuracy	0.88
Recall	0.80
F1-measure	0.73
AUC-ROC	0.85

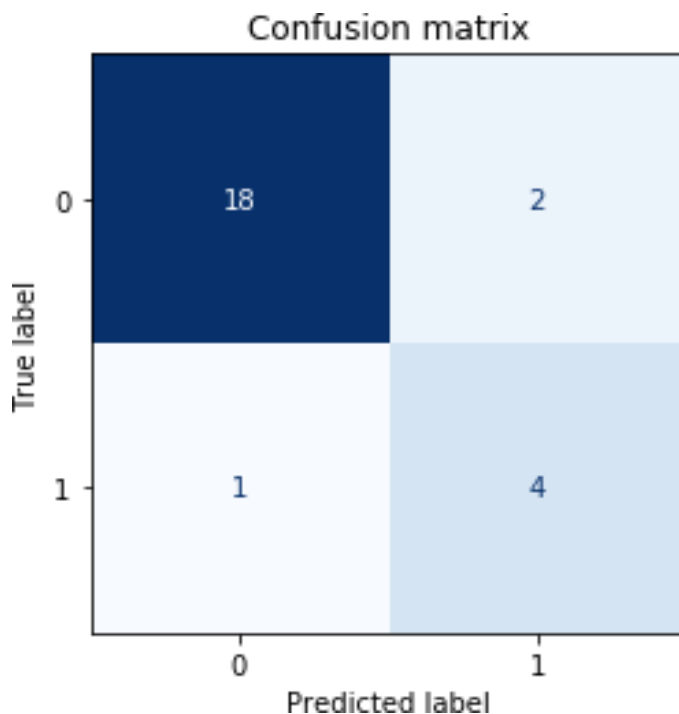


Figure 5-4: Confusion matrix of the RF classification model using randomly splitting between training and testing (75:25). In the testing set, 16 non-viable formulations ($FFc < 5$, non-viable) and 4 viable formulations ($FFc > 5$, viable) were included.

The next validation of the RF classification model was the bootstrap sampling. Fig 5-5 shows the variation of the precision of the RF classification model, plotted on the x-axis at different iterations of the sampling. The mean accuracy achieved was 0.85. Fig 5-6 shows the variation of the recall of the RF classification model, plotted on the x-axis, at different iterations. The mean value was 0.65.

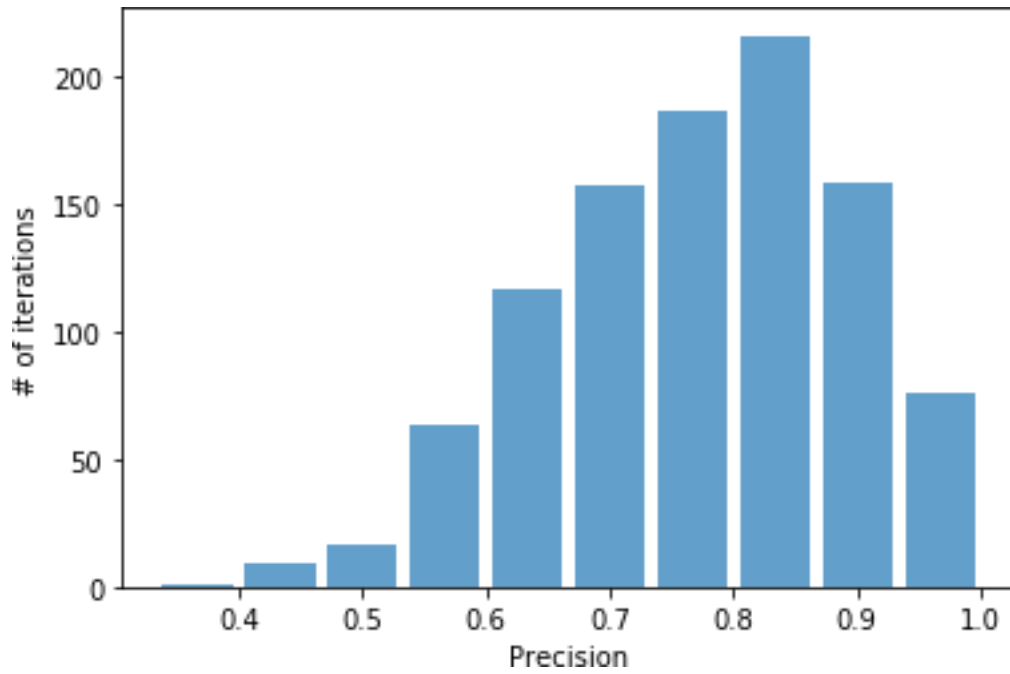


Figure 5-5: Precision distribution calculated using a bootstrapping method of the RF model.

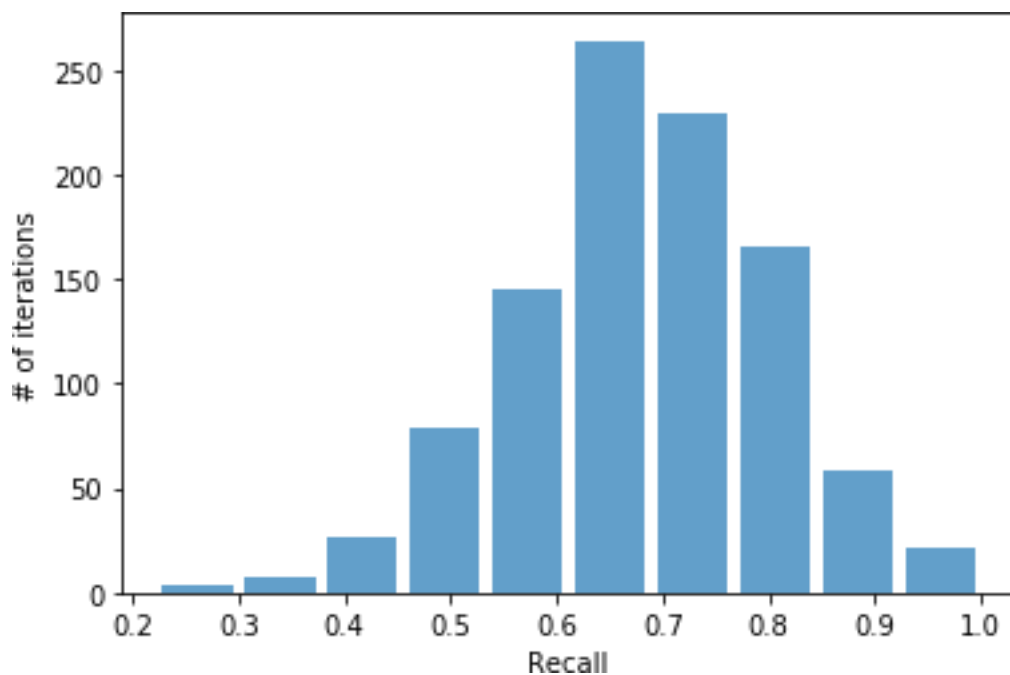


Figure 5-6: Recall distribution calculated using a bootstrapping method of the RF model.

As the last validation test method, a 10-fold cross-validation was performed. The precision score of the RF classification model achieved in this test was 0.9 with a standard deviation of 0.14, and the recall score was 0.83 with a standard deviation of 0.07 (see Table 5-6). The results of these three internal validation methods confirmed that the RF classification model performance was better than the results obtained in a “benchmark model”. The benchmark model was a majority model which calculated the probability of a formulation of being classified as non-viable ($FFc < 5$), accounting for the class imbalance present in the training dataset. The benchmark model achieved a 69% accuracy. The metrics obtained by the RF classification model were significantly better than the results of the benchmark model and therefore, the predictions made by the RF classification model were considered formally optimum.

Table 5-6: Classification metrics of the RF model using 10-fold cross-validation.

Metric	Score	Standard deviation
Accuracy	0.92	0.05
Precision	0.90	0.14
Recall	0.83	0.07

The last method used to assess the performance of the RF classification model was an external validation. The external set contained four randomly selected formulations that were held out from the training dataset. Three of these formulations were non-viable ($FFc < 5$), and one formulation was viable ($FFc > 5$). The results reported in Table 5-7 show that the model classified all the formulations as non-viable. The FFc of the misclassified viable formulation was 5.4, which was very close to the threshold and therefore, it was reasonable that the model misclassified this formulation. The classification probabilities were also reported to show how confident the model was on the predictions. No significant difference was observed between how confident the model was on the correct or incorrect classifications.

Table 5-7: Results of the classification of the four formulations included in the external validation using the RF classification model.

Formulation	FFc	Actual class	Predicted class	Probability
1	2.24	Non-viable	Non-viable	0.66
2	2.42	Non-viable	Non-viable	0.95
3	5.4	Viable	Non-viable	0.70
4	1.6	Non-viable	Non-viable	0.69

To aid the interpretability of the model, SHAP values were calculated (S. Lundberg, 2018). SHAP assigns a score (SHAP value) to all the variables involved in the training, as detailed in Chapter 3 (section 3.5.4.). Fig 5-7 shows that the Area $D[n,x]$ (μm^2) D90 of the API included in the formulation was the most important variable for the classification of the formulation flowability. Most of the important features referred to the particle size and shape of the APIs included in the formulation, i.e., HS circularity or length. The concentration of FlowLac 90 was the most important variable regarding to the concentration of the formulations. The importance of FlowLac 90 could be explained since most of the CMAC formulations that were included in the dataset were formulated with FlowLac 90, and almost all CMAC formulations had a better flowability than Roche formulations.

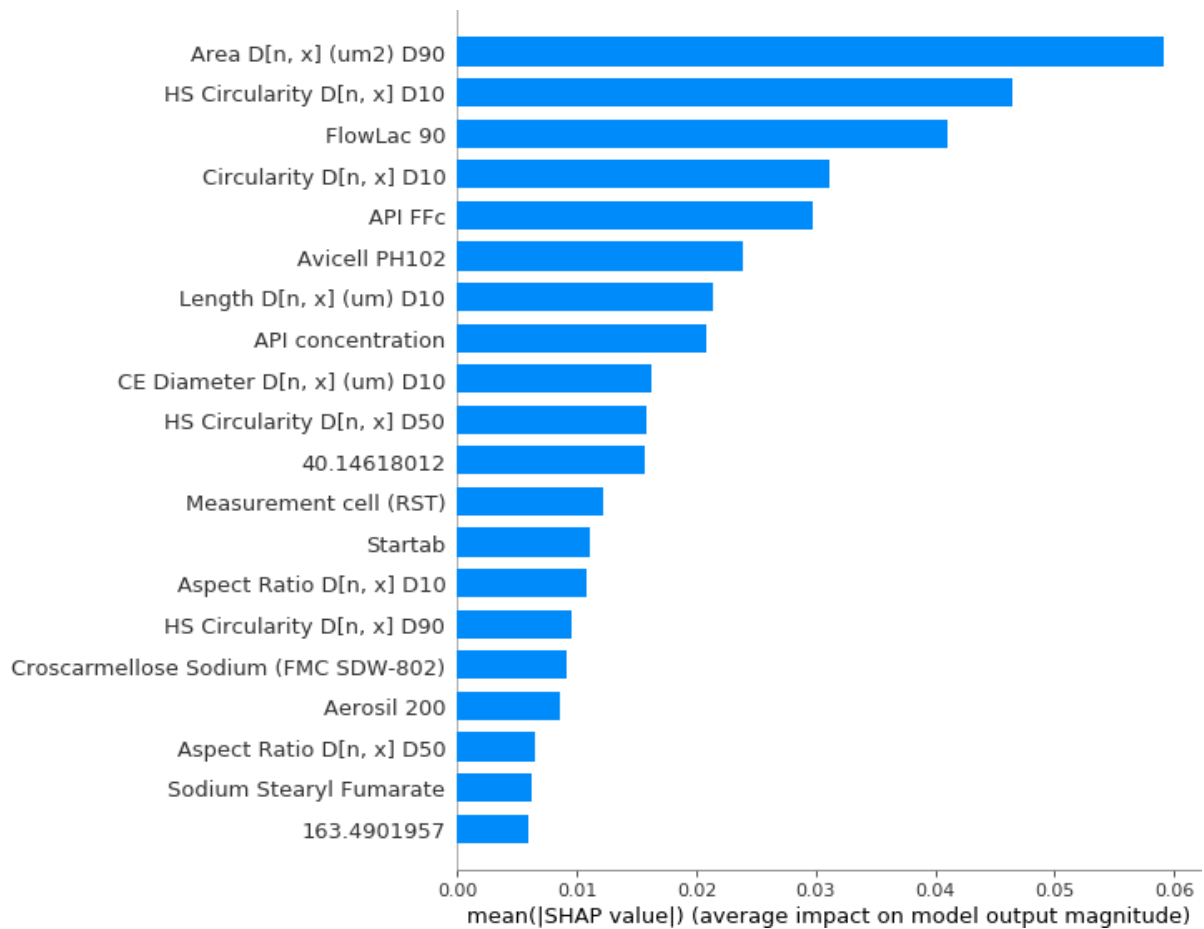


Figure 5-7: The SHAP values plot. The input variables (Table 5-2) are ranked based on their SHAP score: the larger the value, the more important the feature. The name of the excipients represents their concentrations in the formulations.

The importance of the variables for the prediction of the FFc class of the formulations can be better understood in Fig 5-8. High values of the Area D[n,x] (μm^2) D90 of the API included in the formulation led to viable formulations, as high values of this variable (plotted in pink) appear towards the right side of the graph. Similar trends were observed for the concentration of FlowLac 90 and for the FFc of the API included in the formulation.

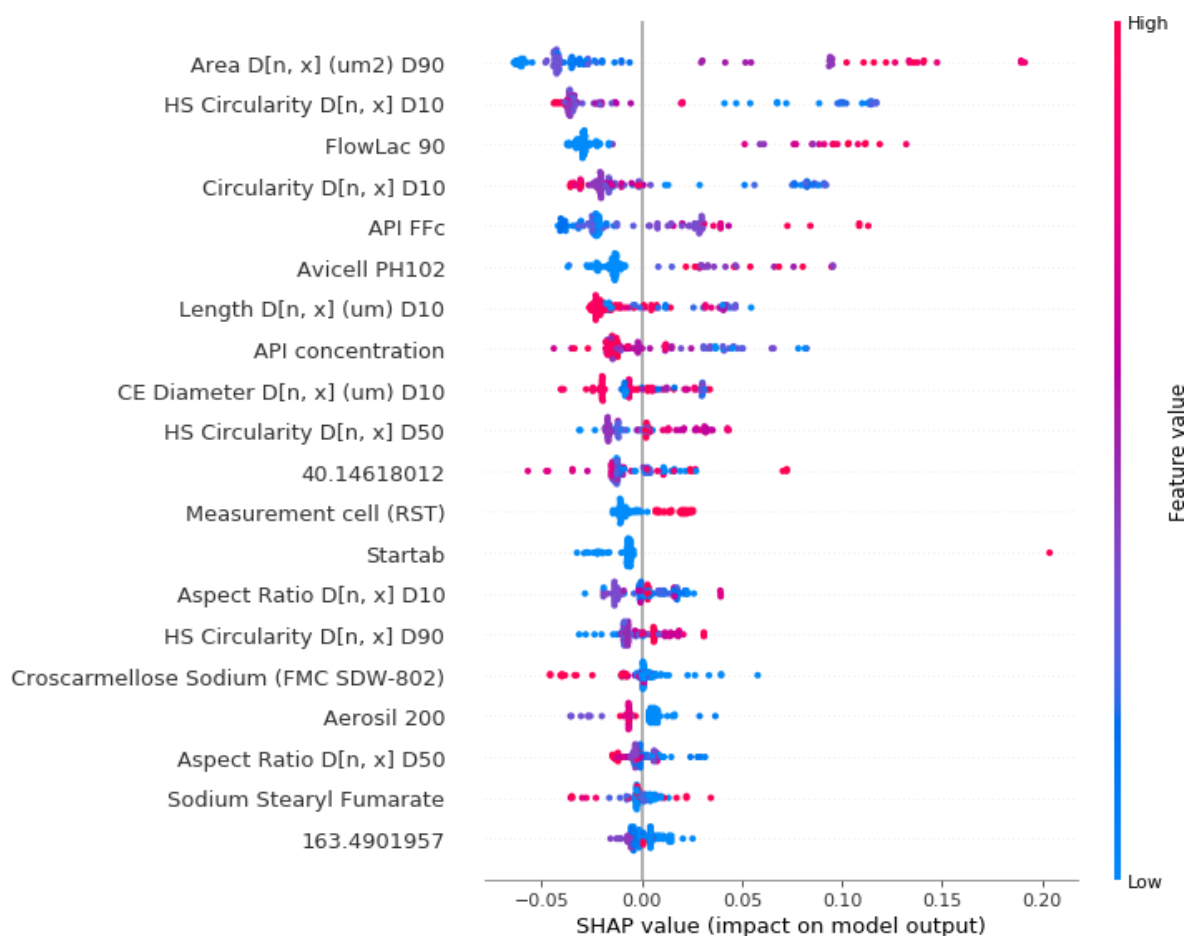


Figure 5-8: The SHAP bee swarm plot. The direction in which the input variables (Table 5-2) contribute to the prediction: the more scatter the data points, the more important the variable. High variables of the variables are represented in pink and low values are represented in blue.

The impact of the FFc of the API on the flowability of the formulation was analysed by calculating the SHAP dependence plot (see Fig 5-9). As expected, as the FFc of the API increases, the flowability of the formulation increases too. Likewise, the same trend is observed in the SHAP dependence plot of the Area D[n,x] (μm^2) D90 (see Fig 5-10): as the area of the API increases, the flowability of the formulation improves. Surprisingly, when the impact of the HS circularity D[n,x] D10 of the API on the flowability of the formulations was studied using the SHAP dependence plot, a negative correlation between these variables was observed (see Fig 5-11). This result was unexpected since spherical particles usually promote powder flowability (Brika, Letenneur, Dion, & Brailovski, 2020).

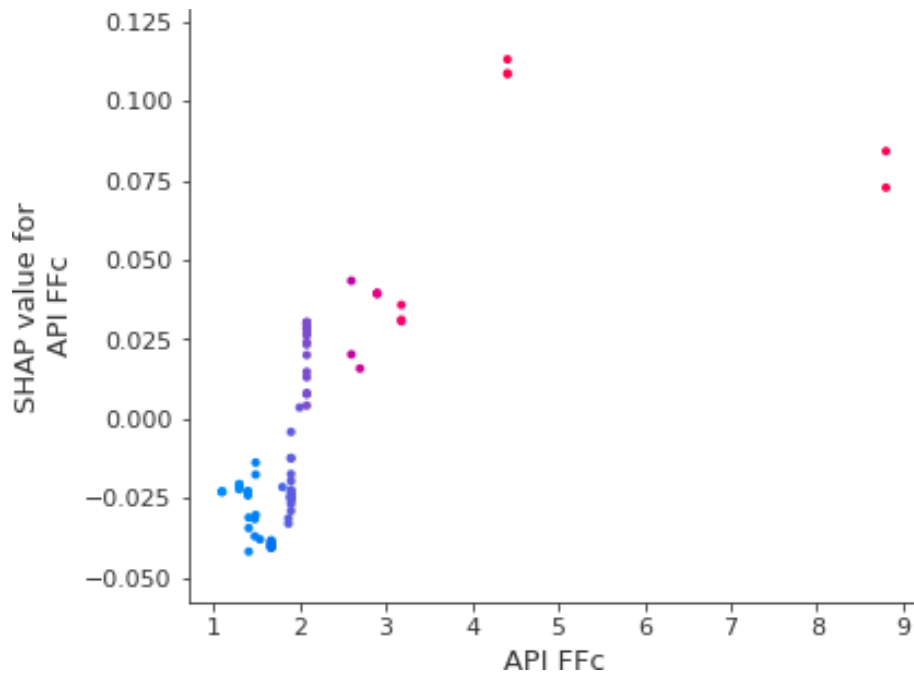


Figure 5-9: SHAP dependence plot to analyse the impact of the FFC of the API included in the formulation on the flowability of the blend.

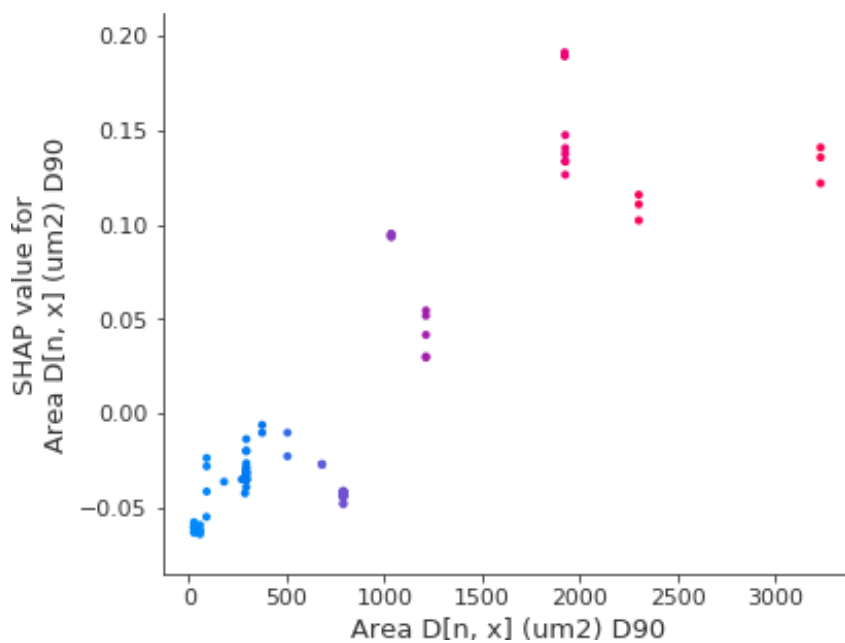


Figure 5-10: SHAP dependence plot of the impact of the Area D[n,x] (μm^2) D90 of the API on the flowability of the formulation.

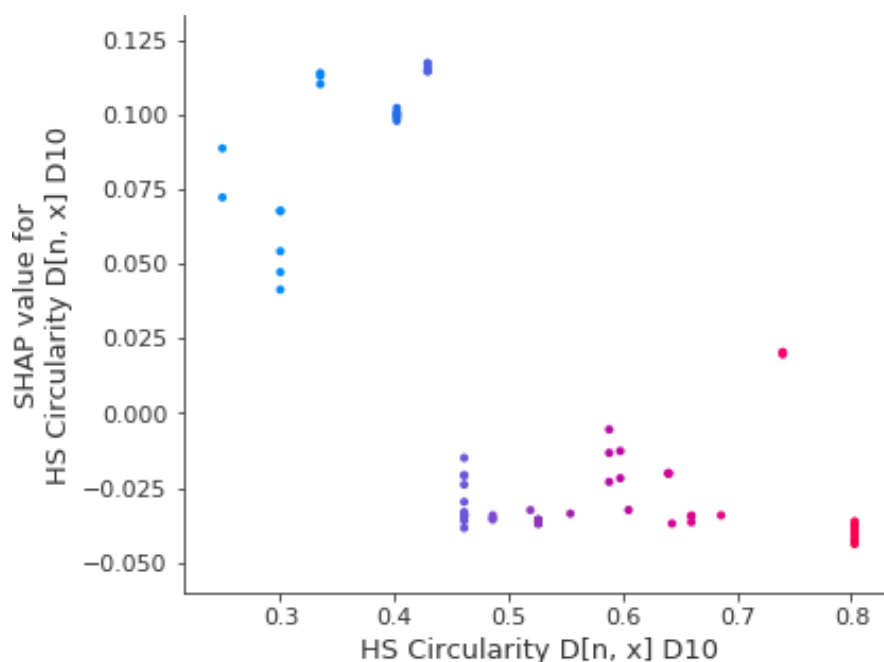


Figure 5-11: SHAP dependence plot of the impact of the HS Circularity D[n,x] D10 of the API on the flowability of the formulation.

5.3.1.4. Monte Carlo (MC) Simulations

One thousand simulated formulations were obtained per test formulation. Fig 5-14 shows the histogram distribution of the API FFC of formulation test 4. As mentioned in the methods section 5.2.1., the Dietmar Schulze RST had an analytical error of 18%. To simulate the formulations, the mean and the standard deviation of the distribution were required. For each test formulation of which the simulated formulations will be generated, the measured value of each variable was taken as the mean of the distribution for the simulation, and the standard deviation per each variable was calculated by multiplying the z-score for 90% interval (1.64), multiplied by two to broaden the distribution and be able to account for any potential measurement error, obtaining a final score of 3.2. The mean of each variable was multiplied by this final score to calculate the standard deviation of the distribution of the variable in the simulated formulations.

Hence, for the API FFC variable, the main value of the distribution is the measured value of the API included in the formulation test and the 18% analytical error value of the Dietmar Schulze RST was multiplied by 3.2, resulting in 0.55. The same procedure was carried out for all the independent variables shown in the methods section (Table 5-4), except the concentration of the formulation that was kept unmodified. Once the simulated values for all the variables were calculated, they were wrapped into

a new data frame, and the unmodified drug loading and concentration of the excipients were appended.

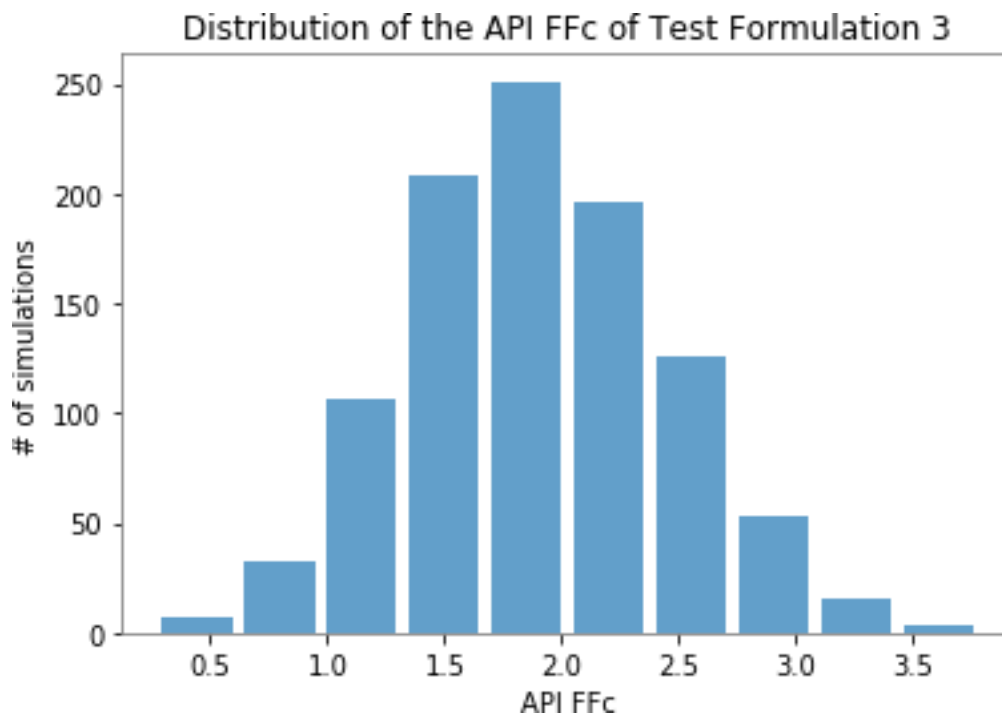


Figure 5-14: Histogram of the distribution of the FFc values of the API included in test formulation 3 across the 1000 simulated formulations (mean value = 1.87, standard deviation = 0.55).

The results of the predictions of the FFc class accounting for the error of the analytical instruments are reported for each formulation included in the external dataset (see Table 5-8). Four formulations were predicted as non-viable, however, only three of them were actually non-viable. The misclassified formulation (formulation test 3) had a FFc of 5.4, which is very close to the threshold used to divide the classes. The results of the classification were reported with the range of probabilities of the formulation of being viable for cDC. Following the traffic-light system described in section 5.2.3, the results showed that the three correctly classified non-viable formulations were classified based on their probabilities as *red formulations*, since their maximum probability was smaller than 0.5. The misclassified formulation was classified as *amber formulation*, as the range of probabilities included 0.5. This traffic-light system provides a simplified system to classify the viability of the formulations for cDC, accounting for the error of the analytical instruments.

Table 5-8: Classification using the RF model of the 4 formulations used for external validation after using MC methods to simulate formulations considering the analytical error of the measurements.

Formulation	FFc	Actual class	Predicted class (MC)	Probability of being viable
1	2.24	Non-viable	Non-viable	0.29-0.47
2	2.42	Non-viable	Non-viable	0.01-0.41
3	5.4	Viable	Non-viable	0.27-0.57
4	1.6	Non-viable	Non-viable	0.28-0.38

The SHAP plots were calculated for one simulated formulation per each test formulation included in the external dataset (see Figs 5-15 to 5-18). The SHAP individual plot shows the variables that had the biggest impact on the RF classification model's prediction. Variables that increased the probability of the classification of the formulation of being viable are represented in red, and variables that decreased the probability of the formulation of being viable are represented in blue. The length of the bar indicates the importance of the variable for the individual prediction. Thus, the variable that has the largest bar is the most important variable for the model for the classification of the formulation of study. The simulated formulations selected had a probability of being viable for cDC close to the mean probability of the test formulation of being viable for cDC. Hence, Fig 5-15 shows the SHAP individual plot for one simulated formulation of the test formulation 1, which indicated that the main driver that increased the probability of this formulation of being viable for cDC was the CE Diameter D10 of the API (14.86 μm), and the main driver that decreased the probability was the API FFc (0.75). The value of the CE Diameter D10 of the API included in test formulation 1 was considerably higher than the mean value of the CE Diameter D10 in the training dataset (4.89 μm), and the FFc of the API was lower than the mean value of the dataset (2.08). The comparison with the mean value explained the impact of these parameters in the prediction. The feature importance analysis (see Fig 5-10) showed that high values of API FFc lead to viable formulations, which explains why such a low value of API FFc decreased the probability of being viable. The feature importance analysis did not show clear results of the impact of the value of CE Diameter D10 on the viability of the formulation, but in this case, a significantly higher value than the mean of CE Diameter D10 increased the probability of being viable.

Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

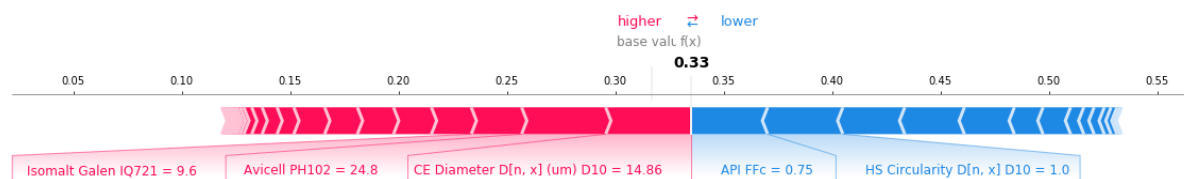


Figure 5-15: SHAP individual plot of one of the simulated formulations from the formulation test 1.

The same analysis was performed for the test formulation 2, where the main driver that increased the probability of the formulation of being viable was the concentration of Avicel PH102 (47.57%) (see Fig 5-16). Avicel PH102 is a microcrystalline cellulose excipient which is commonly used to improve tablet binding as a compression aid and flow aid for directly compressed tablets. Therefore, it was not surprising to see that the model's prediction was impacted by the high concentration of this excipient in test formulation 2. The feature importance analysis (see Fig 5-10) showed that a high concentration of Avicel PH-102 leads to viable formulations. On the other hand, the variable that decreased the probability of being viable was the Area D[n,x] (μm^2) D90 of the API (52.01 μm^2), which was significantly smaller than the mean Area D[n,x] (μm^2) D90 of the training set (790.80 μm^2). The feature importance analysis showed that the Area D[n,x] (μm^2) D90 of the API was the most important variable for the classification of the viability, and that large values of Area D[n,x] (μm^2) D90 lead to viable formulations, which explains why for test formulation 2, the low value of the Area D[n,x] (μm^2) D90 of the API decreased the probability of being viable.

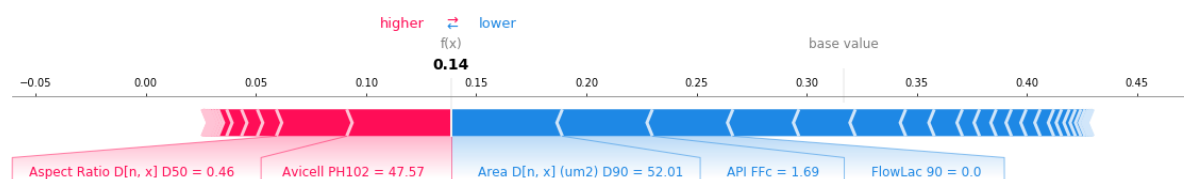


Figure 5-16: SHAP individual plot of one of the simulated formulations from the formulation test 2.

Like formulation 2, for test formulation 3, the main driver that decreased the viability for cDC was the Area D[n,x] (μm^2) D90 of the API (58.36 μm^2) (see Fig 5-17). Although for this formulation, the main

Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

impact on the classification was the HS circularity D[n,x] D10 of the API (0.42) and the concentration of Avicel PH102. Once more, the concentration of Avicel PH102 increased the probability of being viable. In the training dataset, the mean value of HS circularity D[n,x] D10 was 0.52 with a standard deviation of 0.13. The feature importance analysis (see Fig 5-10) showed that, counterintuitively, low values of HS circularity D[n,x] D10 of the API lead to viable formulations, which explains why a HS circularity D10 value of 0.42 increased the probability of being viable.

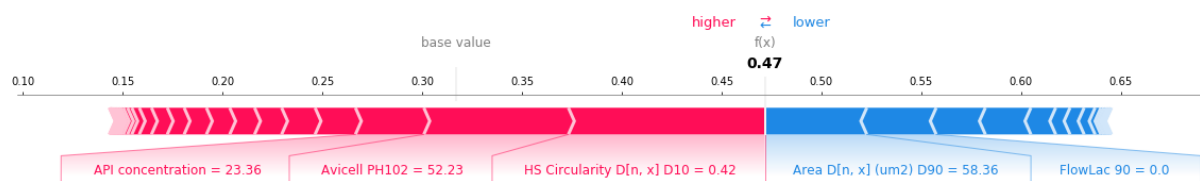


Figure 5-17: SHAP individual plot of one of the simulated formulations from the formulation test 3.

Finally, for test formulation 4, the main driver that decreased the prediction was the Area D[n,x] (μm^2) D90 of the API (see Fig 5-18). The main driver that increased the probability was the value of the particle size bin 40.14 μm of the API (9.99). In the training dataset, the mean value of this particle size bin was 3.05 with a standard deviation of 2.27. The feature importance analysis did not show a clear correlation between the values of this particle size bin and the viability of the formulation. For this formulation, the value of the particle size bin increased the probability of being viable and had more impact than any of the other variables.

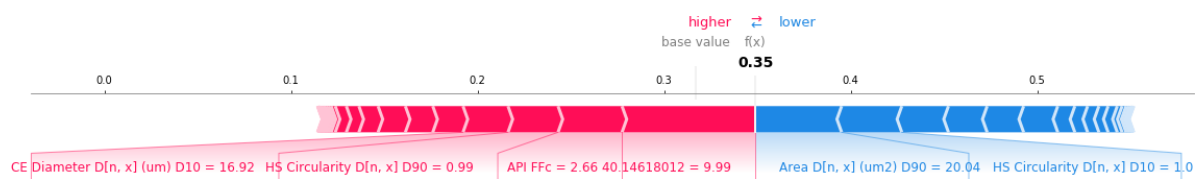


Figure 5-18: SHAP individual plot of one of the simulated formulations from the formulation test 4.

5.3.1.5. Regression models

RF Regressor was used to train a model that predicts the FFC of a pharmaceutical formulation based on selected variables (Section 5.2.2., Table 5-2). Table 5-9 shows that the regression model achieved an R^2 of 0.81, and the MAE was 0.17. The predicted values of FFC were plotted against the measured FFC values of the test formulations (see Fig 5-19).

Table 5-9: Regression metrics used to evaluate the performance of the model.

Metric	Value
R^2	0.81
MSE	1.67
RMSE	1.29
MAE	0.17

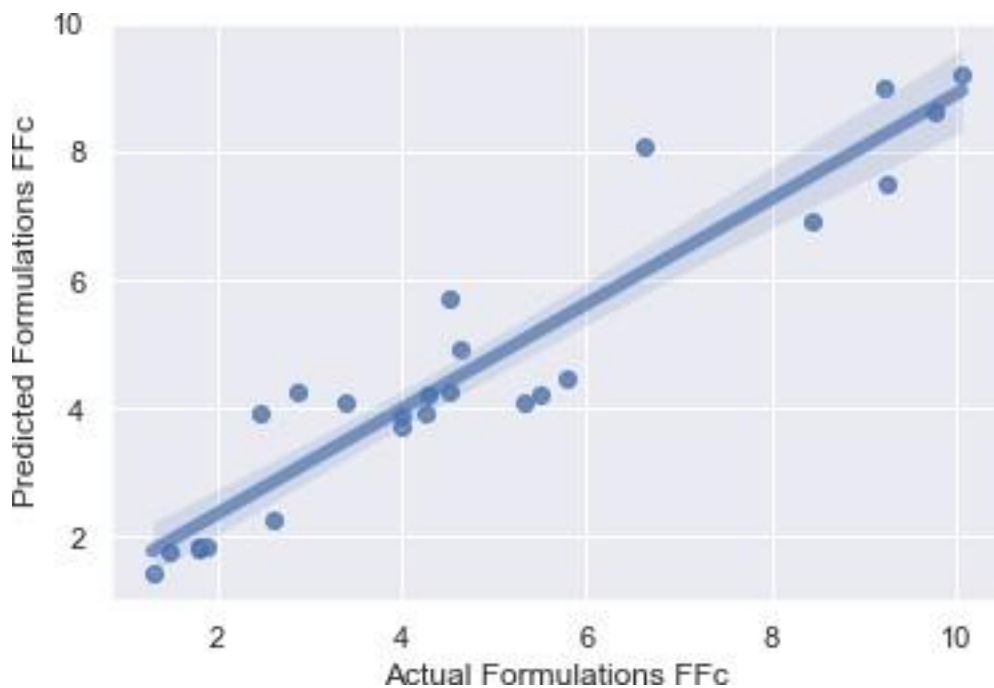


Figure 5-19: Scatter plot of the actual values of the FFC of the test formulations and the predicted value of the FFC of the formulations. The R^2 value achieved was 0.81.

The same materials and measurements used for external validation in the classification models were used to validate the regression model. Table 5-10 shows the result of the prediction of the FFC of the four formulations included in the external set. For the first three formulations (test formulations 1-3), the predicted values of FFC were close to the measured values FFC. For test formulation 4, the predicted FFC value was higher than the measured value, although this predicted FFC value would still classify test formulation 4 as non-viable for cDC. Fig 5-20 shows the measured values of the external formulations plotted against the predicted values by the RF regression model.

Table 5-10: Results of external validation of the FFC regression model.

Formulation	Measured FFC	Predicted FFC
1	2.2	2.9
2	2.4	2.8
3	5.4	5.1
4	1.6	3.6

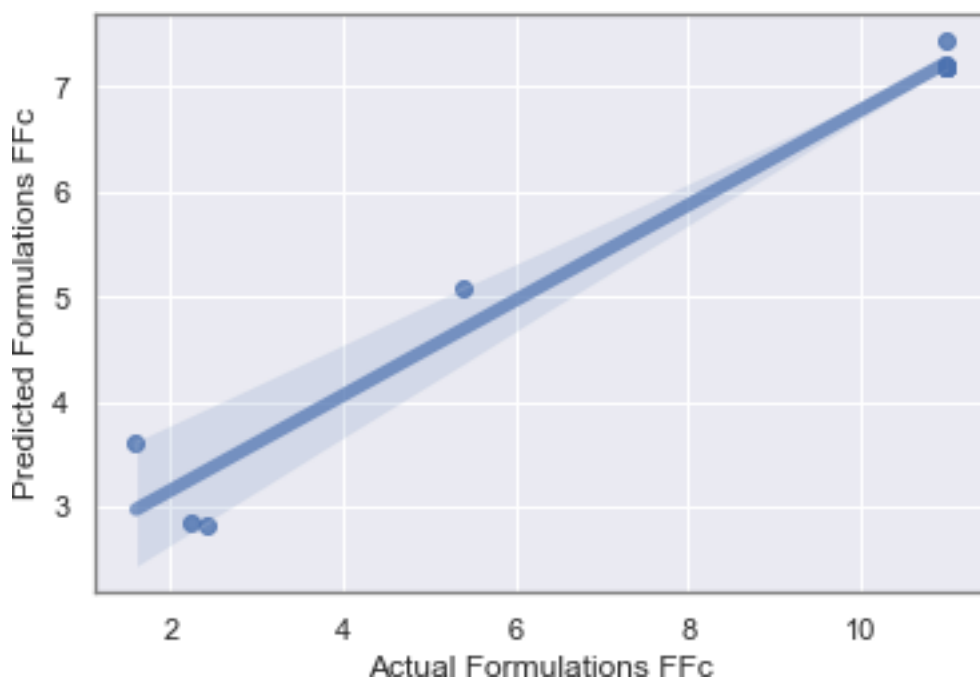


Figure 5-20: Measured values of the FFC of the formulations used for external validation plotted against the predicted values by the RF regressor.

The SHAP values for feature importance analysis were again calculated to check whether there were differences comparing the results to the classification model results (see Fig 5-21). Similar to the classification model, the Area D[n, x] (μm^2) D90 of the API included in the formulation was the most important variable for the prediction of the FFC of the formulation. In this case, the second most important variable was the FFC of the API included in the formulation. The impact of the API FFC was further investigated analysing the SHAP dependence plot (see Fig 5-22). The results of the SHAP dependence plot revealed that, as the FFC of the API increased, the impact on the FFC of the formulation increased too. Particularly, a value of API FFC below 2 had a negative impact on the FFC of the formulation, since the SHAP values were negative. From these results, an important conclusion can be drawn: if the formulation contains an API which FFC value is greater than 2, it will have a positive impact on the flowability of the formulation. And the opposite is true, if the formulation contains an API that has a FFC smaller than 2, this value will have a negative impact on the flowability of the formulation.

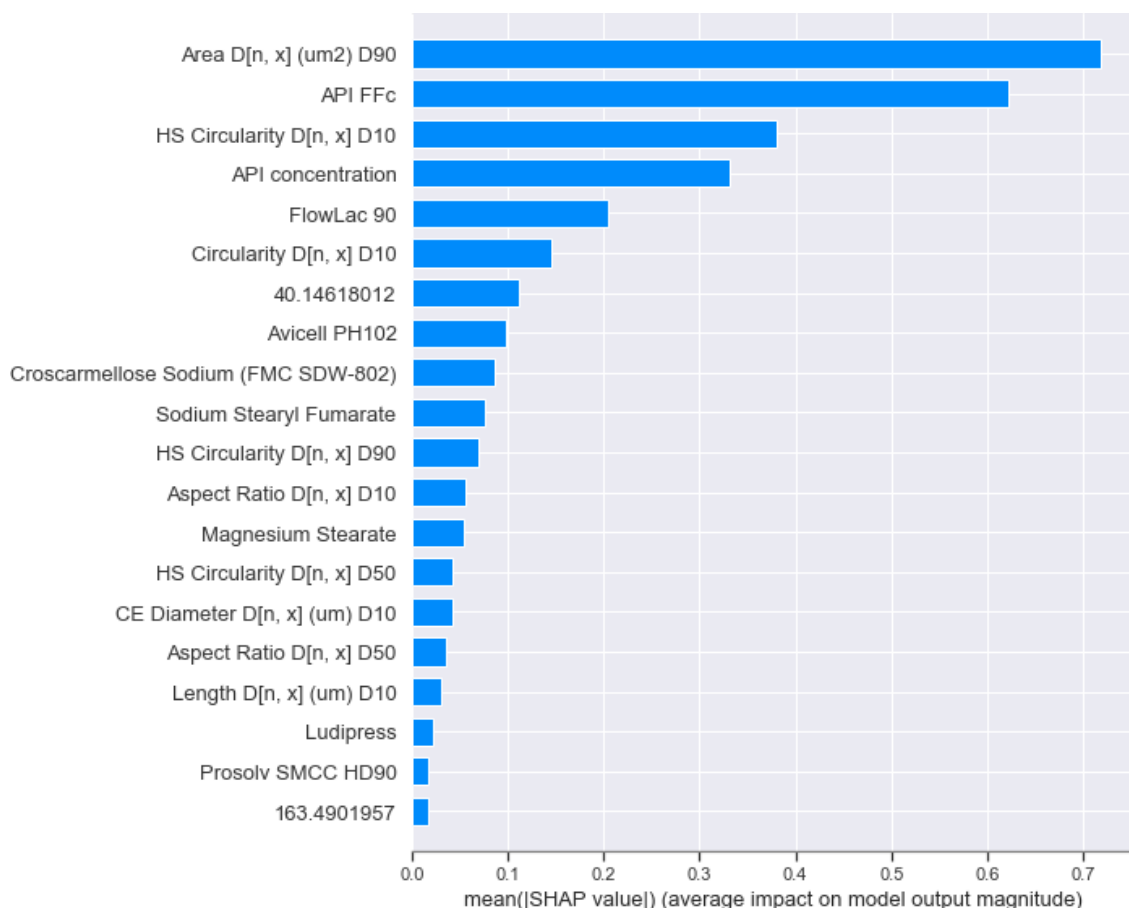


Figure 5-21: Ranking of the most important variables using SHAP values for the regression models.

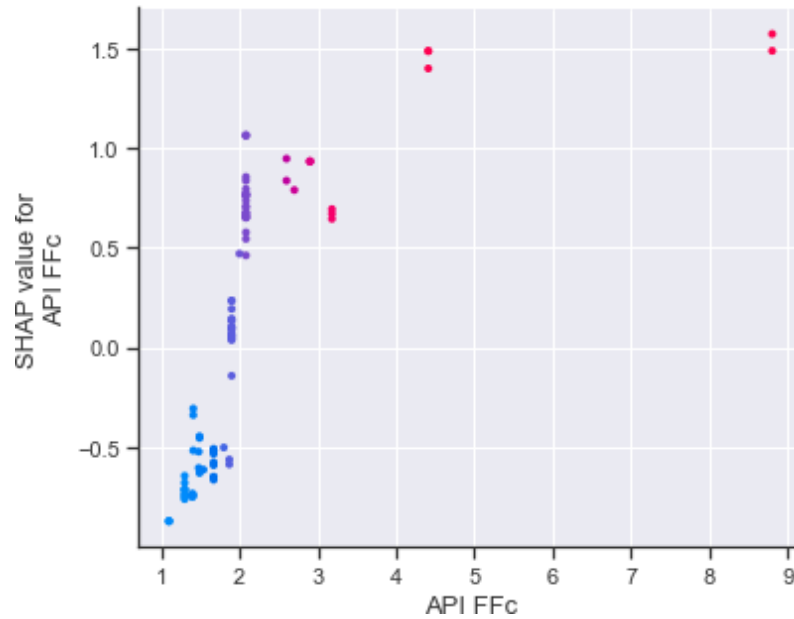


Figure 5-22: SHAP dependence plot of the impact of the value of the FFC of the API included in the formulation on the FFC of the formulation.

One of the test formulations was randomly selected for an individual analysis. Fig 5-23 shows the explanation of the prediction of the test formulation 1. The main drivers for this prediction were the HS Circularity D[n, x] D10 of the API (0.72) and the API FFC (1.5). As expected, the API FFC (1.5) decreased the predicted formulation FFC value, as it was lower than 2. Also, following the results obtained in the classification model, a high concentration of Avicel PH102 increased the predicted FFC, and a high value of HS Circularity D[n, x] D10 decreased the predicted FFC.

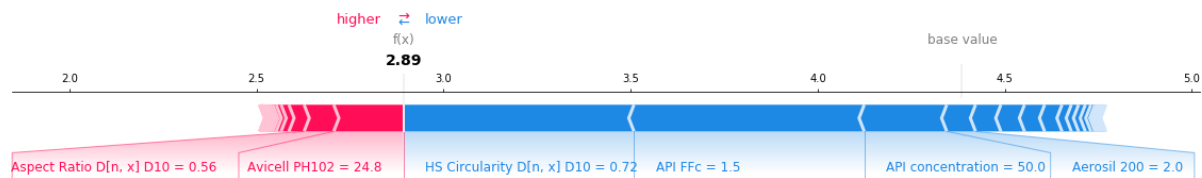


Figure 5-23: SHAP individual plot of the prediction of test formulation 1. The main drivers that increased the predicted value of the FFC of the formulation were the concentration of Avicel PH102 of the formulation and the value of Aspect ratio D[n, x] D10 of the API. On the other hand, the drivers that decreased the predicted value were the HS Circularity D[n, x] D10 of the API.

Aiming to improve the RF model's performance, the reciprocal of the FFC of the formulations was defined as the response of the regression model. The distribution of the reciprocal of the FFC is shown in Fig 5-24. Compared to the distribution of the FFC of the formulations presented in Fig 5-5, the reciprocal of the FFC of the formulations was better distributed. This data transformation not only improves the distribution of the dependent variable but also emphasises the differences at smaller values of FFC, where the changes have a bigger impact on powder flowability.

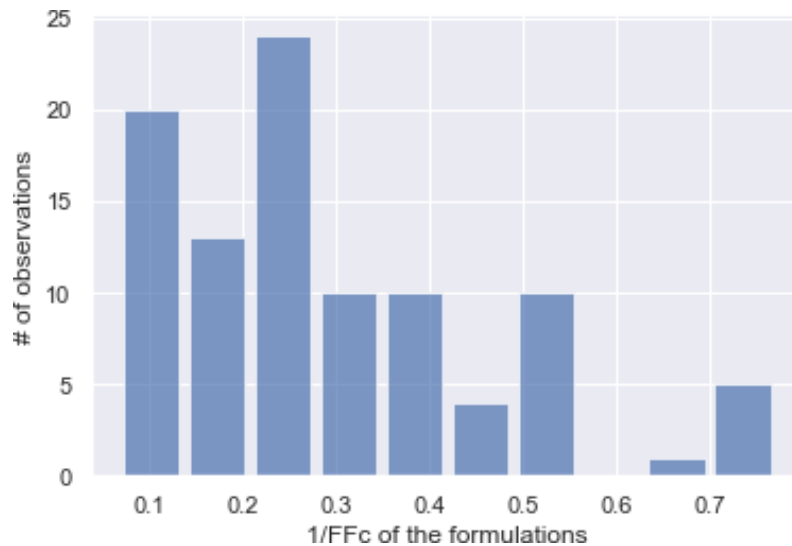


Figure 5-24: Distribution of the values of the reciprocal of the FFC of the formulations included in the training dataset.

The model was trained using RF Regressor, and it achieved a R^2 score of 0.856, a MSE of 0.004, a RMSE of 0.062, and a MAPE of 0.252 (see Table 5-11). The MSE and the RMSE are significantly smaller than the metrics achieved by the first FFC regression model described previously, but these errors report the results in the units of the dependent variable. The FFC values range from 1.3 to 14.69, whereas for the reciprocal of FFC, the values range from 0.06 to 0.76. Therefore, the only metric that was comparable across models was the mean absolute percentage error (MAPE). The MAPE is calculated by dividing the difference between the predicted and the measured value to the actual value of the measurement. For the previously described FFC regression model, the MAPE was 0.173, whereas for the reciprocal of FFC regression model, the MAPE was 0.252. Hence, we could conclude that the regression model that predicts the reciprocal of the FFC performed worse than the model that predicts the FFC.

Table 5-11: Results of the regression models for the prediction of the reciprocal value of the FFc of the formulations.

Metric	Value
R ²	0.856
MSE	0.004
RMSE	0.062
MAPE	0.252

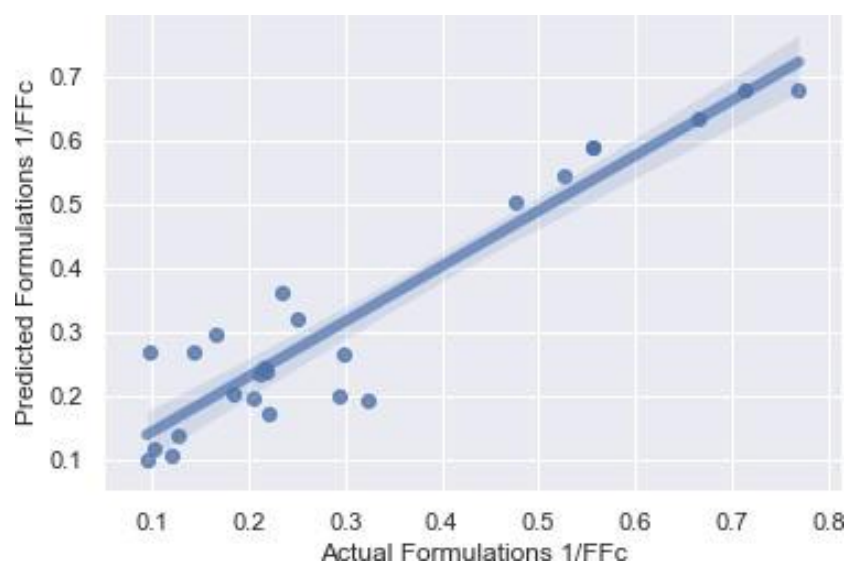


Figure 5-25: Scatter plot of the measured reciprocal of FFc against the predicted reciprocal value of FFc of the formulations of the internal validation set ($R^2=0.856$).

The SHAP values were once again used to rank the variables in order of importance. Interestingly, the most important variable for the prediction of the reciprocal of the FFc was the API FFc, followed by the concentration of the API in the formulation, and the concentration of Prosolv SMCC HD90 (see Fig 5-26). The plot shows that low values of API FFc and high concentration of API lead to higher reciprocal values of FFc (lower FFc), agreeing with the results obtained by the first FFc regression model.

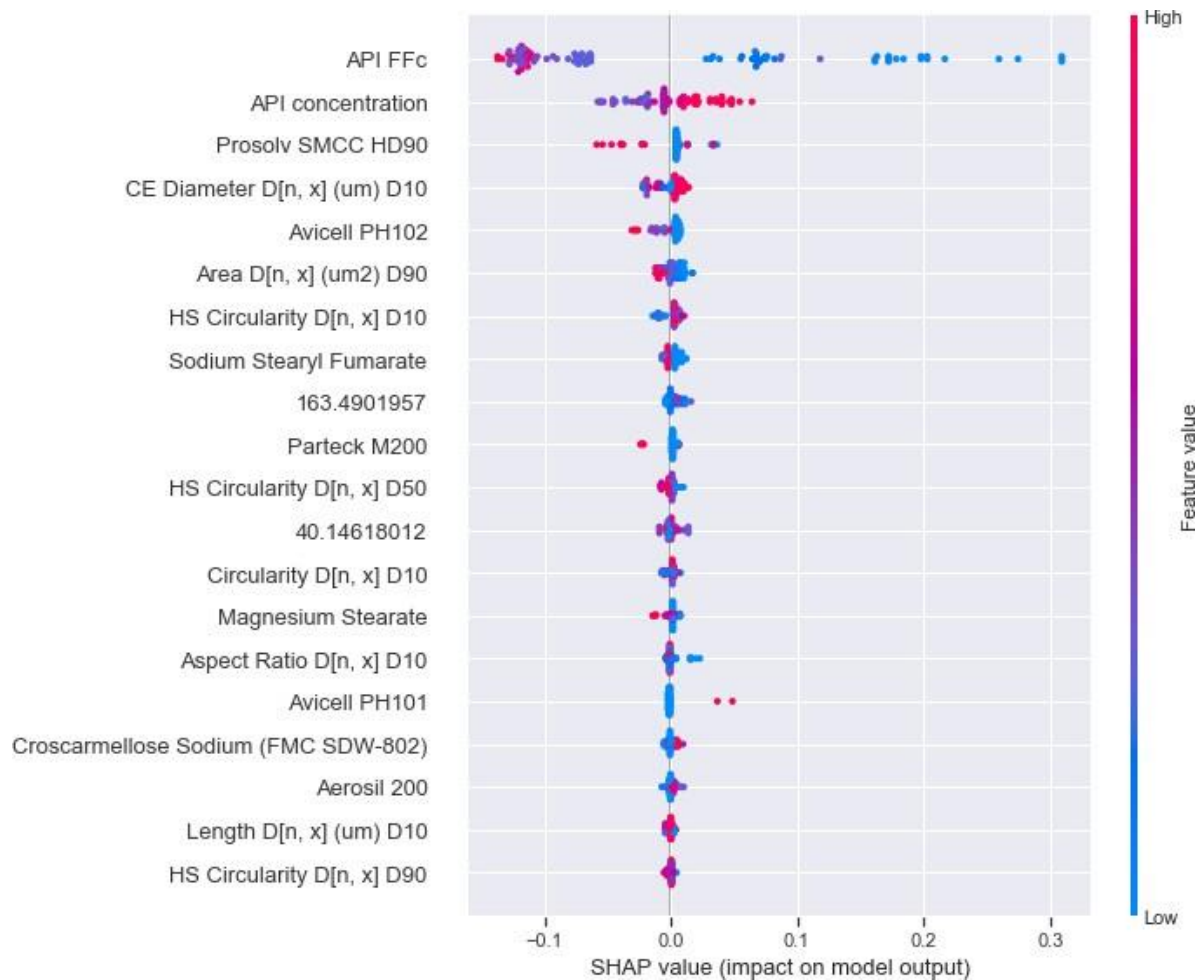


Figure 5-26: SHAP feature importance analysis of the reciprocal of FFC model.

An external validation was then performed to validate the results of the reciprocal of FFC model (see Table 5-12). All the predictions were reported after retransforming the data from the reciprocal of the FFC to the FFC of the formulation. The results of the predicted FFC were close to the measured values of FFC of these formulations, and unlike the results achieved by the first FFC regression model, the fourth test formulation predicted result was also close to the measured value of FFC. The first formulation was selected for a deeper analysis using individual SHAP values, to compare the results with the results obtained in the first FFC regression model. The main drivers for the prediction were the API FFC (1.5) and the concentration of the API (50%). These properties increased the predicted value of the reciprocal of FFC and therefore had a negative impact on the final FFC value (see Fig 5-27).

Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

Table 5-12: Results of the prediction of the FFc of the formulations included in the external set, after retransforming the data from the reciprocal of FFc to the FFc.

Formulation	FFc	Predicted FFc
1	2.2	2.1
2	2.4	2.7
3	5.4	5.1
4	1.6	1.9

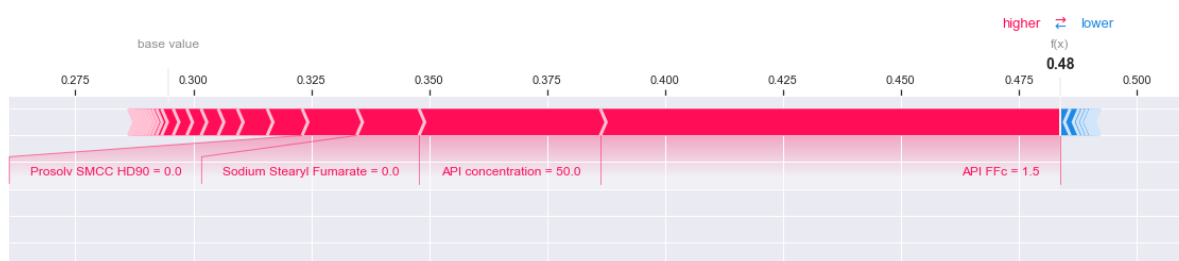


Figure 5-27: The individual SHAP analysis for the first formulation of the external dataset. The predicted value of the reciprocal of FFc was 0.48 (FFc = 2.07). The main drivers of this prediction were the API FFc (1.5), and the API concentration (50%), which increased the predicted value of reciprocal of FFc (decreased the FFc).

Lastly, a final data transformation was introduced to study its impact on the prediction of the reciprocal of the FFc of the formulations. The input variable “API FFc” was transformed to obtain the reciprocal of the FFc of the API and added to the dataset a training variable. The results obtained for this model were similar to the results obtained before this data transformation (see Table 5-13). Fig 5-28 shows that the FFc of the API, even when it was transformed, was still the most important variable for the prediction of the reciprocal of the FFc of the formulations.

Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

Table 5-13: The results of the performance of the RF regression model for the prediction of the reciprocal of the FFC of the formulation, considering the reciprocal of the FFC of the APIs as an input variable.

Metric	Value
R²	0.808
MSE	0.005
RMSE	0.067
MAPE	0.209

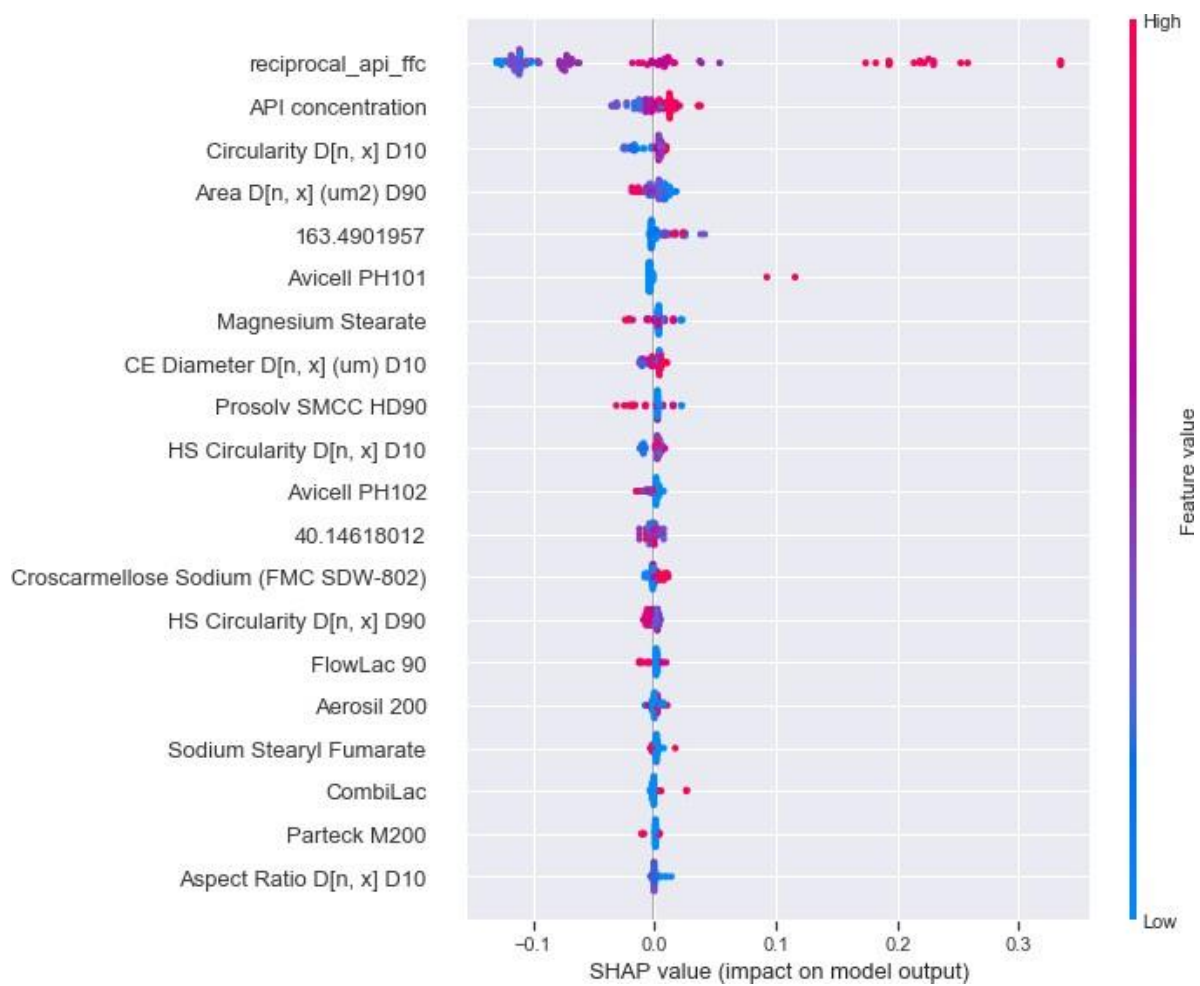


Figure 5-28: The SHAP feature importance analysis of the RF regression model when the reciprocal of the FFC of the API was an input variable.

Since the results of the model trained with the reciprocal of the FFC of the API did not achieve a better performance than the original model, this data transformation was not included in the final version of the model.

5.3.2. Powder flow: global wall friction angle (WFA) regression models

Low WFA values indicate a low adhesion of the powder to the equipment, and usually correlates with good flowability (Divya & Ganesh, 2019). Formulations that have a WFA smaller than 20° were considered viable for cDC and formulation with greater WFA than 20° were considered non-viable. As for the FFC, this threshold was defined by the formulation scientists at Roche. The results of this regression model are only applicable in an industrial setting when they are coupled with the prediction of FFC.

For WFA regression models, the limited data were available (34 formulations). Morphologi® G3 and Mastersizer® 3000 data were used as the inputs of the model (independent variables) as described in Table 5-1, and the WFA was set as the dependent variable.

To validate the model, the dataset was randomly split into 25% testing and 75% training. RF Regressor was used to train the model achieving a R² score value of 0.828 and a MAE of 1.83 (see Table 5-14). The WFA values range from 6.7° to 31.6°, and therefore an error of 1.83 was considered acceptable.

Table 5-14: Results of the evaluation of the WFA regression model on the testing set.

Metric	Value
R²	0.82
MSE	5.65
RMSE	2.38
MAE	1.83

Feature importance analysis was carried out using SHAP values. Fig 5-29 shows that the concentration of FastFlo 316 was the most important variable for the prediction of WFA. FastFlo 316 was the most common excipient in the training dataset, which explains its importance in the feature importance analysis. The particle properties of the API that had the biggest impact on the model were particle-size related (area and length), whereas the particle-shape related properties (circularity and aspect

ratio) were relegated to lower positions in the SHAP feature importance analysis. The SHAP feature importance analysis shows that as the length D10 of the API decreases, the WFA increases. This result agrees with previous publications, such as the study by Liu *et al.* that demonstrated that an increase in particle size led to a decrease of wall friction angle (Liu, Guo, Lu, & Gong, 2015). Nonetheless, the variables that had the biggest impact on the model output were the concentration of FastFlo 316, the concentration of sodium stearyl fumarate and concentration of Aerosil 200 in the formulation. While more data are needed to improve the model performance, this approach explores the feasibility of the classification of pharmaceutical formulations based on their WFA.

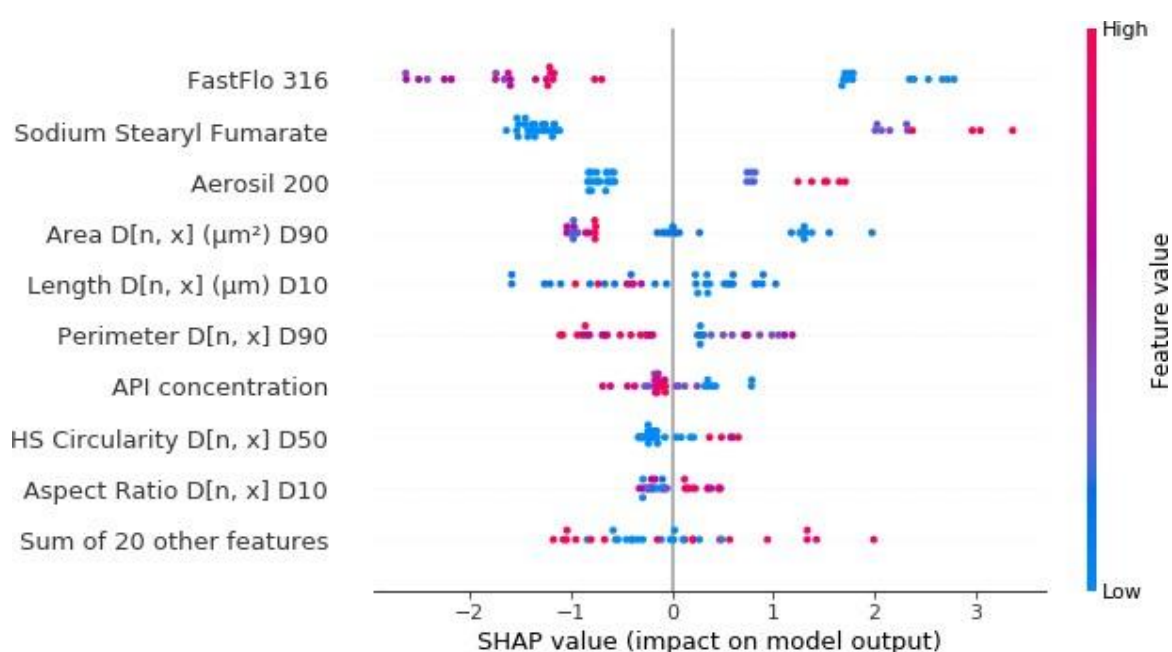


Figure 5-29: SHAP bee swarm plot for the PHIW regression model.

The predicted values of WFA were plotted against the measured values of WFA (see Fig 5-30). These results can be better interpreted by using SHAP individual plots. The formulation that had the biggest error in its prediction was selected for a further analysis (see Fig 5-31). The measured value of WFA of this formulation was 9.9°, and the predicted value by the RF model was 13.79°. The concentration of croscarmellose sodium, or better said, the absence of croscarmellose sodium increased the predicted value of WFA. On the other hand, the absence of sodium stearyl fumarate and the value of the Area D[n,x] (µm²) D90 of the API (2122 µm²) decreased the predicted value of FFC. Even though this prediction had a prediction error of 3.89°, the model would have classified this formulation as viable (WFA < 20°), and therefore, viable for cDC.

Predicting of the viability of pharmaceutical formulations for continuous direct compression (cDC) using machine learning approaches.

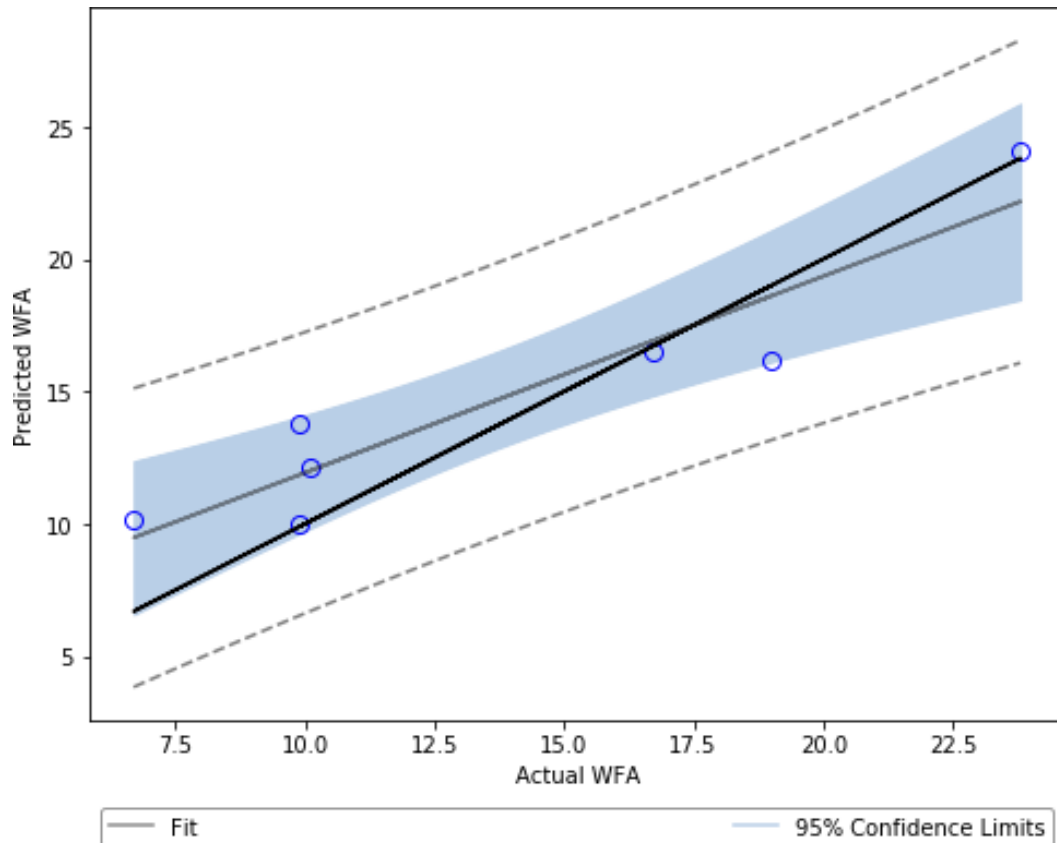


Figure 5-30: The scatter plot actual values of WFA plotted against the predicted values of WFA. The 95% confidence intervals and the best-fit line are represented.

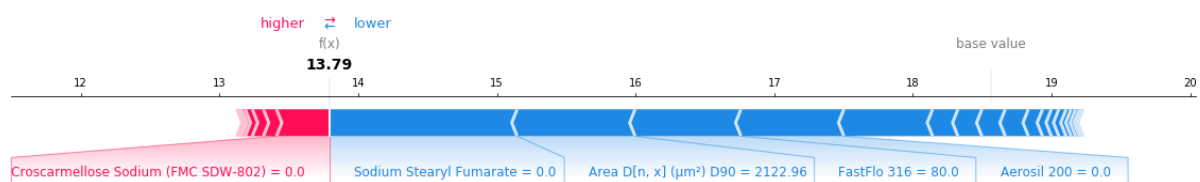


Figure 5-31: The SHAP individual plot of the testing formulation that was predicted with the largest error. The actual value of the WFA was 9.9° and the predicted value was 13.79°.

For comparison, the formulation that had the smallest error (MAE = 0.2°) in the prediction was further analysed by visualising the SHAP individual plot. The measured value of WFA for this formulation was 16.7°, and the predicted value by the RF model was 16.5°. The main variables that increased the predicted value were the Perimeter D[n, x] D90, and the Length D[n, x] D10 of the API included in the formulation (see Fig 5-32). The API included in this formulation had smaller perimeter and length

values than the main values of the training dataset: the Perimeter $D[n, x]$ D90 of the API of this formulation was $90.56 \mu\text{m}$, whereas the mean Perimeter $D[n, x]$ D90 of the APIs included in the training dataset was $120.96 \mu\text{m}$. Likewise, the Length $D[n, x]$ D10 of the API of this formulation was $5.9 \mu\text{m}$, whereas the mean Length $D[n, x]$ D10 of the APIs included in the training dataset was $9.8 \mu\text{m}$. It is known that smaller particles lead to stronger adhesion forces (Bowling, 1988; Katainen, Paajanen, Ahtola, Pore, & Lahtinen, 2006) and therefore, a higher WFA could be expected. The descriptors that decreased the predicted value of WFA, and therefore indicating a better flowability), were the concentration of sodium stearyl fumarate (0%), the concentration of FastFlo 316 (80%), and the concentration of Aerosil 200 (0%). Hence, for this test formulation, the absence of sodium stearyl fumarate and Aerosil 200, in combination with a high concentration of FastFlo 316 resulted in a lower WFA (better flowability).

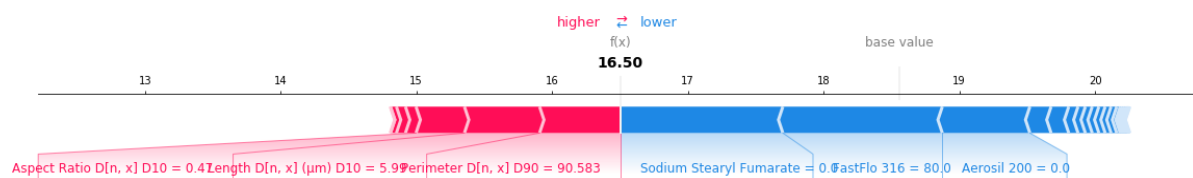


Figure 5-32: The SHAP individual plot the testing formulations that was predicted with the smallest error. The actual value of the WFA was 16.7° and the predicted value was 16.5° .

5.4. Conclusions

The implementation of data-driven models for the development of new pharmaceutical formulations can reduce the resources and waste involved in this process. The application of classification and regression models to rapidly assess the viability of a library of model pharmaceutical formulations for cDC from the physical properties of the API and the concentration of the formulation is presented. The dataset included a total of 103 formulations, of which 4 were held out to perform an external validation. The formulations were divided into viable (FFc greater than 5) or non-viable (FF smaller than 5) based on the experience of an industrial formulation scientist, according to the requirements of the cDC manufacturing line.

The results show that the particle size and shape of the API, measured with both static image analysis and laser diffraction analysis, coupled with the concentration of formulation components can enable rapid, predictive assessment of the viability of the formulations for cDC. The classification model achieved a precision of approximately 0.9 after several methods of validation. The results of the classification models were reported by the interval of probabilities of being viable for cDC, providing

enough information to the formulation scientist to guide them in the development of a new formulation. The models were robustly validated by 10-fold cross-validation with randomly shuffled splits, by bootstrap sampling and by external validation on four formulations that were held out of the dataset, concluding that the model showed a strong performance across the multiple validation techniques, and therefore guaranteeing that the model was not overfitting. Regarding the regression models, the best FFC regression model had an MAPE of 0.173, which was better than the reciprocal FFC regression model (MAPE = 0.252), and the regression model that included the reciprocal of the FFC of the APIs of the formulations as input (MAPE = 0.209). Finally, the WFA regression model had a MAE of 1.83. The implementation of these models in early-stage development of new formulations will save resources in the assessment of the viability of the formulations for cDC by moving away from the current trial-and-error approaches, and thereby reducing the requirements of amount and materials and measurements.

One of the limitations of the current models is the lack of viable formulations (FFC > 5) included in the training dataset, which results in a class imbalance issue. These models could be improved by adding more viable formulations (FFC > 5) to the dataset to address the aforementioned bias in the training set available. The threshold of the FFC classification model could be adjusted if needed, based on requirements of different equipment or for modification of the current guidelines.

The classification model has been deployed into a user-friendly app and implemented on-site at Roche, allowing the early assessment of the viability of pharmaceutical formulations without needing to make and analyse the blend. This advancement is particularly beneficial in the early-stage development of a new API, and the results can be fed back into particle design to modify the properties of the API and thus ensure the viability of cDC. The work presented herein can serve as a showcase of the potential applicability of the implementation of digital workflows for the prediction of the behaviour of powders in downstream processes.

5.5. References

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
- Afrin, S., & Gupta, V. (2020). Pharmaceutical Formulation.
- Andrews, G. (2022). Continuous Pharmaceutical Processing and Process Analytical Technology. In *Drugs and Pharmaceutical Sciences*: CRC Press.
- Barjat, H., Checkley, S., Chitu, T., Dawson, N., Farshchi, A., Ferreira, A., . . . Toby, M. (2021). Demonstration of the Feasibility of Predicting the Flow of Pharmaceutically Relevant Powders from Particle and Bulk Physical Properties. *Journal of pharmaceutical innovation*, 16(1), 181-196. doi:10.1007/s12247-020-09433-5
- Bonate, P. L. (2001). A brief introduction to Monte Carlo simulation. *Clinical pharmacokinetics*, 40(1), 15-22.
- Bowling, R. A. (1988). A theoretical review of particle adhesion. *Particles on surfaces* 1, 129-142.
- Brika, S. E., Letenneur, M., Dion, C. A., & Brailovski, V. (2020). Influence of particle morphology and size distribution on the powder flowability and laser powder bed fusion manufacturability of Ti-6Al-4V alloy. *Additive Manufacturing*, 31, 100929.
- Divya, S., & Ganesh, G. (2019). Characterization of Powder Flowability Using FT4–Powder Rheometer. *Journal of Pharmaceutical Sciences and Research*, 11(1), 25-29.
- Education, I. C. (2020, 21/09/2020). Unsupervised Learning. Retrieved from <https://www.ibm.com/cloud/learn/unsupervised-learning#:~:text=Unsupervised%20learning%2C%20also%20known%20as,the%20need%20for%20human%20intervention.>
- Ellison, A. M. (2004). Bayesian inference in ecology. *Ecology letters*, 7(6), 509-520.
- Garnica-Caparrós, M., Memmert, D., & Wunderlich, F. (2022). Artificial data in sports forecasting: a simulation framework for analysing predictive models in sports. *Information Systems and e-Business Management*, 1-30.
- Giraud, M., Vaudez, S., Gatumel, C., Nos, J., Gervais, T., Bernard-Granger, G., & Berthiaux, H. (2021). Predicting the flowability of powder mixtures from their single components properties through the multi-component population-dependent granular bond number; extension to ground powder mixtures. *Powder Technology*, 379, 26-37.
- Gooley, T. A., Leisenring, W., Crowley, J., & Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in medicine*, 18(6), 695-706.
- Higgins, J., & Deeks, J. (2011). Obtaining standard deviations from standard errors and confidence intervals for group means. *Cochrane handbook for systematic reviews of interventions*, 5.
- Hildebrandt, C., Gopireddy, S. R., Scherließ, R., & Urbanetz, N. A. (2019). Investigation of powder flow within a pharmaceutical tablet press force feeder—A DEM approach. *Powder Technology*, 345, 616-632.
- Jiang, X., Osl, M., Kim, J., & Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2), 263-274.
- Katainen, J., Paajanen, M., Ahtola, E., Pore, V., & Lahtinen, J. (2006). Adhesion as an interplay between particle size and surface roughness. *Journal of Colloid and Interface Science*, 304(2), 524-529.
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). *Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning*. Paper presented at the Proceedings of the 2020 CHI conference on human factors in computing systems.
- Kroese, D. P., Brereton, T., Taimre, T., & Botev, Z. I. (2014). Why the Monte Carlo method is so important today. *Wiley interdisciplinary reviews: computational statistics*, 6(6), 386-392.

- Kruppa, J., Liu, Y., Biau, G., Kohler, M., Koenig, I. R., Malley, J. D., & Ziegler, A. (2014). Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. *Biometrical Journal*, *56*(4), 534-563.
- Lam, M. W. (2018). One-match-ahead forecasting in two-team sports with stacked Bayesian regressions. *Journal of Artificial Intelligence and Soft Computing Research*, *8*.
- Leane, M., Pitt, K., Reynolds, G., & Group, M. C. S. W. (2015). A proposal for a drug product Manufacturing Classification System (MCS) for oral solid dosage forms. *Pharmaceutical development and technology*, *20*(1), 12-21.
- Leane, M., Pitt, K., Reynolds, G. K., Dawson, N., Ziegler, I., Szepes, A., . . . Group, M. C. S. W. (2018). Manufacturing classification system in the real world: factors influencing manufacturing process choices for filed commercial oral solid dosage formulations, case studies from industry and considerations for continuous processing. *Pharmaceutical development and technology*, *23*(10), 964-977.
- Liu, Y., Guo, X., Lu, H., & Gong, X. (2015). An investigation of the effect of particle size on the flow behavior of pulverized coal. *Procedia engineering*, *102*, 698-713.
- Lundberg, S. (2018). Welcome to the SHAP documentation[Ⓢ].
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
- Matplotlib. (2012-2022). Matplotlib. Retrieved from <https://matplotlib.org/>
- McGinn, T., Jervis, R., Wisnivesky, J., Keitz, S., & Wyer, P. C. (2008). Tips for teachers of evidence-based medicine: clinical prediction rules (CPRs) and estimating pretest probability. *Journal of general internal medicine*, *23*(8), 1261-1268.
- Mehos, G., Eggleston, M., Grenier, S., Malanga, C., Shrestha, G., & Trautman, T. (2018). Designing hoppers, bins, and silos for reliable flow. *The Best of Equipment Series*, *33*.
- Morin, G., & Briens, L. (2013). The effect of lubricants on powder flowability for pharmaceutical application. *Aaps Pharmscitech*, *14*(3), 1158-1168.
- Murphy, K. P. (2018). Machine learning: A probabilistic perspective (adaptive computation and machine learning series). In: The MIT Press: London, UK.
- Nalluri, V. R., & Kuentz, M. (2010). Flowability characterisation of drug–excipient blends using a novel powder avalanching method. *European Journal of Pharmaceutics and Biopharmaceutics*, *74*(2), 388-396.
- Nelson, P. R., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*(1), 45-65.
- Olson, M., & Wyner, A. J. (2018). Making Sense of Random Forest Probabilities: A Kernel Perspective. *stat*, *1050*, 14.
- Pasha, M., Hekiem, N. L., Jia, X., & Ghadiri, M. (2020). Prediction of flowability of cohesive powder mixtures at high strain rate conditions by discrete element method. *Powder Technology*, *372*, 59-67.
- Pereira Diaz, L., Brown, C., & Florence, A. (2021). *Prediction of powder flow of pharmaceutical materials from physical particle properties using machine learning*. Paper presented at the SPHERe Proceedings: 4rd International Symposium on Pharmaceutical Engineering Research.
- Rubinstein, R. (1981). Simulation and Monte Carlo Method. New York: John & Wiley & Sons. In: Inc.
- Stewart, K. D., Johnston, J. A., Matza, L. S., Curtis, S. E., Havel, H. A., Sweetana, S. A., & Gelhorn, H. L. (2016). Preference for pharmaceutical formulation and treatment process attributes. *Patient preference and adherence*, *10*, 1385.
- Štrumbelj, E. (2014). On determining probability forecasts from betting odds. *International journal of forecasting*, *30*(4), 934-943.
- ten Kate, A. J. B., Piccione, P. M., Westbye, P., & Becker, A. F. A. (2022). An industrial and chemical engineering perspective on the formulation of active ingredients in pharmaceuticals and agrochemicals. *Current Opinion in Chemical Engineering*, *36*, 100747.

- Ticehurst, M. D., & Marziano, I. (2015). Integration of active pharmaceutical ingredient solid form selection and particle engineering into drug product design. *Journal of Pharmacy and Pharmacology*, *67*(6), 782-802.
- Van Snick, B., Dhondt, J., Pandelaere, K., Bertels, J., Mertens, R., Klingeleers, D., . . . De Beer, T. (2018). A multivariate raw material property database to facilitate drug product development and enable in-silico design of pharmaceutical dry powder processes. *International journal of pharmaceutics*, *549*(1-2), 415-435.
- Vasilenko, A., Glasser, B. J., & Muzzio, F. J. (2011). Shear and flow behavior of pharmaceutical blends—Method comparison study. *Powder Technology*, *208*(3), 628-636.
- Waits, L. P., Luikart, G., & Taberlet, P. (2001). Estimating the probability of identity among genotypes in natural populations: cautions and guidelines. *Molecular ecology*, *10*(1), 249-256.
- Yang, Y., Ye, Z., Su, Y., Zhao, Q., Li, X., & Ouyang, D. (2019). Deep learning for in vitro prediction of pharmaceutical formulations. *Acta pharmaceutica sinica B*, *9*(1), 177-185.
- Zhang, W., Zhao, Q., Deng, J., Hu, Y., Wang, Y., & Ouyang, D. (2017). Big data analysis of global advances in pharmaceutics and drug delivery 1980-2014.

6. Deep learning (DL) approaches for the investigation of particle size and shape descriptors

6.1. Introduction

Oral dosage forms, particularly tablets, are one of the most widely used pharmaceutical products (Aulton & Taylor, 2013) and can be produced via different techniques, such as wet granulation, roller compaction or direct compression. Direct compression (DC) requires few unit operations and therefore is increasingly preferred by pharmaceutical companies (Shangraw, 1989). The main disadvantage of direct compression is that this manufacturing technique requires better powder flow and highly uniform blend content throughout all stages than other techniques (Abe, Yasui, Kuwata, & Takeuchi, 2009; Clayton, 2019; Gad, 2008; Shangraw, 1989).

The physical properties of powders including particle size and particle shape are key material attributes in the manufacturing of tablets due to their impact on the physicochemical and biopharmaceutical properties of drug substances and on the manufacturability of the blend or product. It has been widely shown that particle size and particle shape impact physical properties (Goh, Heng, & Liew, 2018; Kurek, Wyrwiz, Piwińska, & Wierzbicka, 2016; Swaminathan & Kildsig, 2002), quality attributes (i.e., bioavailability, dissolution rates, drug release profile) (Hintz & Johnson, 1989; Kesisoglou & Wu, 2008; Peng et al., 2016; Sandri, Bonferoni, Ferrari, Rossi, & Caramella, 2014), and manufacturing processability (Bassini et al., 2022; Fonteyne et al., 2014; Hejduk, Czajka, & Lulek, 2021) of pharmaceutical products and therefore, the efficacy and safety of drugs. The results of Chapter 4 showed that materials with a PSD D10 greater than 50 μm , a PSD D50 between 100 and 300 μm and a PSD D90 between 300 and 700 μm were likely to be suitable for DC due to their desirable flow properties enabling effective feeding of materials through the DC process stages. Suitability for DC is also dependent on particle shape (Guo, Beddow, & Vetter, 1985; Kaerger, Edge, & Price, 2004); indeed, the results in Chapter 4 also showed that powders comprising particles with sphericity D10 between 0.5 and 0.8, and aspect ratio D10 between 0.5 and 0.7 are more likely to be suitable for DC than powder whose shape parameters lie out with these ranges.

One of the main challenges in particle size and shape characterization is the lack of method standardisation and dependence of the specific measured values on the particular particle size characterization instrument used (Shekunov, Chattopadhyay, Tong, & Chow, 2007). This is particularly relevant in the context of producing robust data to better understanding the relationship between drug substance properties, particle attributes, bulk properties and manufacturing process and product performance through data-driven models given the reliability and accuracy of these models

depends intrinsically on the quality of the data that they are built upon (Wadams et al., 2022). Leane *et al.* identified particle size as the most important quality attribute for the manufacturability of drug substances followed by particle shape (Leane, Pitt, Reynolds, & Group, 2015; Leane et al., 2018). Hence, providing robust particle size and morphology data is key to achieving the desired properties of the end product (Burgess, Duffy, Etzler, & Hickey, 2004).

Particle size and shape data are traditionally reported using percentile descriptors, i.e., cumulative distribution (D10, D50, D90) number or volume weighted. Expecting that these descriptors represent the actual particle size distribution might be unrealistic for a material (Ferreira et al., 2018). Morphometric features, i.e., aspect ratio, and elongation, are also reported using percentile descriptors. These features might not capture the real complexity of the shape of the particles. Additionally, materials that have equivalent values of D10 and D90 can have different particle habits (Gamble, Tobyn, & Hamey, 2015), and therefore have a different impact on the bulk properties.

In light of these challenges another approach to describing particle size and shape, unrelated to the traditional particle size and shape descriptors, could be useful to enable the better application of computational models. This concept was investigated in this chapter by the comparison of two data-driven AI models: a random forest (RF) model built on particle size and shape traditional percentile descriptors, and a convolutional neural network (CNN) model built on analysis of powder images (see Fig 6-1). These models were trained to predict powder flow characteristics, typically reported as the flow function coefficient (FFc). This was previously modelled using machine learning (ML) models by (Barjat et al., 2021; Pereira Diaz, Brown, & Florence, 2021) and in Chapter 4 and 5. Using images for the prediction of FFc would be particularly useful in early-stage development when material quantity is scarce, and the definitive characterisation of particle size and shape is not a priority (and may not be possible as particle formation processes are still being developed).

Machine learning (ML) is one of the most exciting and widely applied research areas in recent years driven by the need to move beyond current limitations of *ab initio*, empirical, mechanistic, or often slow and expensive first principles simulations. ML can make data-driven predictions based on training data that can allow research scientists to make decisions in, or close to, real time. Importantly ML models cannot just give reliable, useful predictions but can be analysed to understand how the models make predictions. Deep learning (DL) is a type of ML used to build predictive models as it can extract and transform data by using multiple layers of neural networks. One of the main applications of DL is in image analysis where it has been applied in the field of pharmaceutical sciences to detect defective tablets (Quan, Huy, Hoan, & Duc, 2020), to develop calibration models based on online Raman

spectroscopy (Yan, Zhang, Fu, & Qu, 2020), or to classify the quality of coated tablets (Hirschberg, Edinger, Holmfred, Rantanen, & Boetker, 2020).

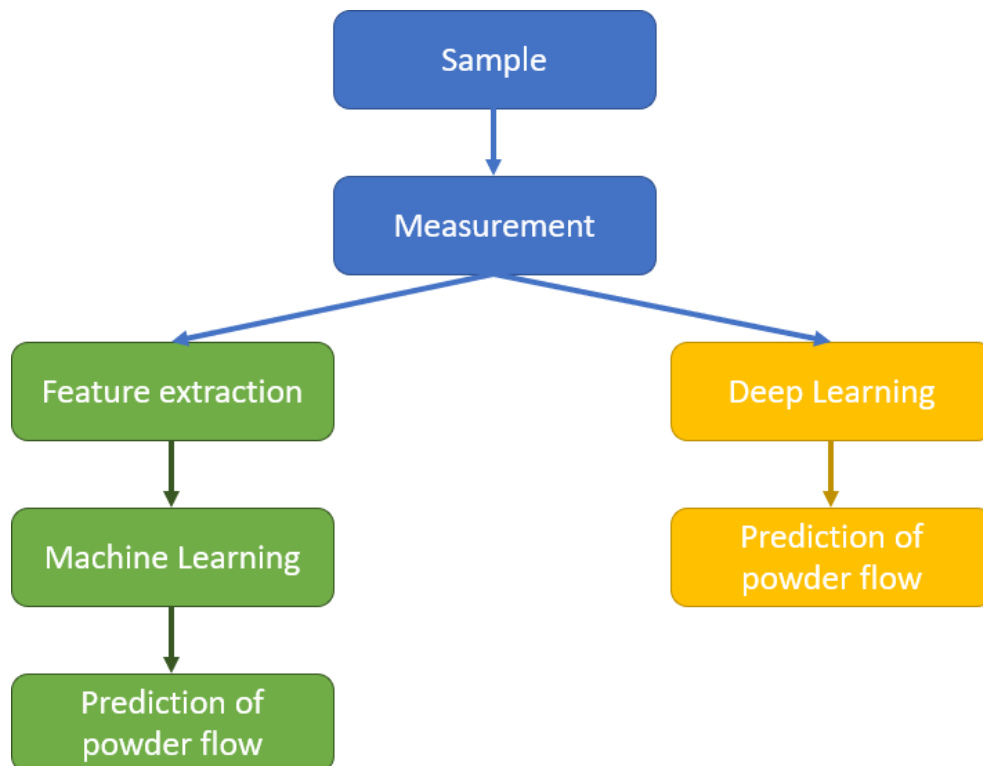


Figure 6-1: The difference in the methodology of ML and DL models. For ML models, the particle information is extracted from the images generated by the particle size characterisation instrument by the built-in software, whereas for the DL models, this first feature extraction step is already performed by the DL model and therefore, the images can be fed directly to the model. Therefore, the inputs of these models would be different, i.e., the inputs of the ML models are particle size and shape descriptors, and the inputs for the DL model are pixel information.

DL has already been used in many areas of technology and science, achieving the best results in robotics, autonomous driving, image, text, and voice recognition. Traditionally, machine learning (ML) models have performed well for classification and regression tasks. The main difference between DL and non-deep ML models is the feature extraction step (Shinde & Shah, 2018). DL models, as opposed to non-deep learning models, do not need a manual feature extraction step, which allows this method to learn features to understand the complexities of the data rather than trying to learn from user-defined features. DL has been successfully used in biomedicine and genomics to predict the 3D structure of proteins and has the potential to become an invaluable tool for the discovery and development of new drugs (Aliper et al., 2016).

Researchers have utilized AI to successfully predict the performance of various pharmaceutical products, such as stability, dissolution, and porosity in advancing product development (Paul et al., 2021). In pharmaceutical manufacturing, DL has also been used, although less extensively, for the optimization and digitalization of process analytical technology (PAT) tools to monitor in real time manufacturing processes. PAT tools are used to monitor the efficiency of tableting, which is influenced by particle size and powder behaviour. DL has been used to monitor certain undesirable processes that might happen during tableting, such as capping (premature detachment of the top layer of the tablet) (Belič, Škrjanc, Božič, Karba, & Vrečer, 2009), or to identify the parameters that have the biggest impact on the porosity of the tablet and the tensile strength (Zawbaa et al., 2018).

Despite providing very accurate models, DL has some disadvantages that must not be overlooked. The algorithms usually struggle at generalizing: even though they are good at interpolation, they do not perform well in extrapolation beyond the bounds included in the training data. If the training data are too similar, there is a risk of underfitting (undergeneralizing); on the other hand, if the data is too different, there is a risk of decreasing the performance while overfitting (Tsimenidis, 2020). Additionally, the learning approach only creates predictions, not recommendations, so the output of the model should not be understood as the answer to a question, but as a prediction that must be interpreted. Another critical point are the loops of feedback: when the model interacts with the environment, by adding the predictions to the training dataset and hence, the model becomes biased. Therefore, the model becomes affected by its own predictions. An example of these loops of feedback could be the use of a DL model for the prediction of crime area. The DL model predicts that crime will occur in a specific area. If this prediction is added to the training dataset in a feedforward loop, the model can become bias towards that area.

The work in this chapter is focused on the comparison of the performance between an RF model, built on particle size and shape descriptors extracted from a particle size characterisation instrument, and a CNN model, trained on the images obtained from the same instrument but with no physical descriptors provided to the training model. Both models aim to classify pharmaceutical powders into cohesive, easy flowing or free flowing, based on their FFC: smaller than 4, between 4 and 10, or greater than 10, respectively. The performance of the models was compared by analysing the classification of 9 pharmaceutical included in an external dataset.

6.2. Materials and methods

6.2.1. Materials

A total of 67 powders were analysed to build the dataset for the RF and the CNN model. These powders were either individual materials (see Table 6-1) or formulations. The formulations contained ibuprofen, Paracetamol Granular Special, mefenamic acid or calcium carbonate, at different drug loadings (5%, 20%, 40%), and with different combination of excipients, i.e., binary mixtures contained lactose, and multicomponent mixtures contained lactose and a defined mixture of Avicel PH-102, croscarmellose sodium and magnesium stearate (20%) (see Table 4-1, and Table 4-2). 23 of these formulations were included in this training dataset. The materials were measured using different volume samples, adding up to 98 measurements.

Table 6-1: Powders included in the training dataset for the RF and the CNN model.

Individual materials included in the dataset		
1-octadecanol	Cellulose	Lubritose mannitol
4-aminobenzoic acid	Cholic acid	LUBRITOSE MCC
Ac-Di-Sol SD	D-glucose	LUBRITOSE PB
Affinisol HPMC	D-sorbitol	Magnesium stearate
Avicel PH-101	FastFlo 316	Methocel DC2
Azelaic acid	Granulac 230	Microcel MC-102
Benece K100M	HPMC	Microcel MC-200
Caffeine	Ibuprofen 70	Mowiol 18-88
Calcium carbonate	Lidocaine	Paracetamol Granular Special
Calcium phosphate dibasic		

6.2.2. Particle size and shape analysis: Morphologi® G3, Malvern.

Static image analysis was used to generate the dataset for the RF and the CNN models. This method involves the characterization of particles that have been dispersed onto a plate. This method typically allows better quality images can be obtained from the particles compared to dynamic image analysis techniques collected from snapshots of moving particles. Additionally, static image analysis is the most appropriate technique to support modeling of particle morphology (Wadams et al., 2022). Images that are taken at different focal planes can be compiled to create a composite image that includes all the particles of the sample. However, static image analysis has some challenges, such as dealing with large particles, and the limitation of a long time required (typically 1 hour).

To train the RF and DL models, the powders presented in Table 6-1 were characterised using Malvern Morphologi® G3 particle characterisation system (Malvern Panalytical, Malvern, UK). Once the measurement was done, the tabular data (particle size and shape variables) were used as the training dataset for the RF classifier, and the area composite images were used as the training dataset for the DL models.

To perform the measurements a 500 mg sample was prepared and placed on the dry dispersive unit with a spatula of 5, 7, 9 or 13 mm³. The sample was then dispersed with an injection pressure of 0.8 bar, for 20 milliseconds, followed by a settling time of 60 seconds, onto a glass plate (180 x 110 mm). The compensation for plate tilt was enabled and the illumination was diastopic. The optic selection was 5x (6.5 µm – 420 µm) with an overlap of 40 % and the threshold intensity was 105. The scan area was selected to capture an area of 2894.348 mm². The software disregarded any measurement that contained less than 10 pixels by setting the minimum trash size to 10 pixels. The segmentation method was disabled, and hole filling was enabled, so the software filled the background in the case that any particle contained areas where the background shows through. No filters or classification settings were selected. After the measurement, the area composite was extracted, obtaining the 5359 x 3491 pixel size images that then were used to train the deep learning models. A total of 97 images were used.

6.2.3. Powder flow analysis: FT4 Powder Rheometer, Freeman Technology Ltd.

Since the main goal of the present paper was to predict the powder behaviour of pharmaceutical materials and blends, the training data incorporating the measured powder attributes and/or images needed to be labelled with the response being predicted. The FT4 Powder Rheometer – Freeman

Technology Ltd. was used to measure the powder flow behaviour using the shear cell test. The details of the measurement were described in Chapter 3.

6.2.4. Machine learning methods: RF classification models

Random Forest (RF) classifiers were built to establish a benchmark for the CNN classification models built on images. RF classification models were built for the prediction of powder flow on the physical properties obtained with the Morphologi G3[®]. The number of features to consider when looking at the best split was set as “sqrt” (squared root), and thus the number of features was calculated from the squared root of the total number of features. The number of estimators (number of trees) was set at 100.

For practical purposes, Jenike’s classification (Jenike, 1964) was adapted to create three categories that describe powder behaviour as follows: materials that had a FFc smaller than 4 were labelled as cohesive, materials that had a FFc between 4 and 10 were labelled as easy flowing, and materials that had a FFc greater than 10 were labelled as free flowing. Initially, a RF model that distinguishes between the three categories described above was built (single-step model, see Fig 6-2 (a)). A two-step classification model was also used to train the model. The first step of the two-step model classified between free-flowing materials ($FFc > 10$), or non-free-flowing materials ($FFc < 10$), and the second step that classified into cohesive materials ($FFc < 4$), and non-cohesive materials ($FFc > 4$) (see Fig 6-2(b)) (Pereira Diaz et al., 2021). For both models the input variables to the model included particle size distribution D10, D50, D90, and Sauter Mean Diameter, particle shape distribution (aspect ratio and circularity) D10, D50, and D90 (see Table 6-2). All the models were internally validated using a 10-fold cross-validation method, and externally dataset test set of 9 materials that were held out of the training set.

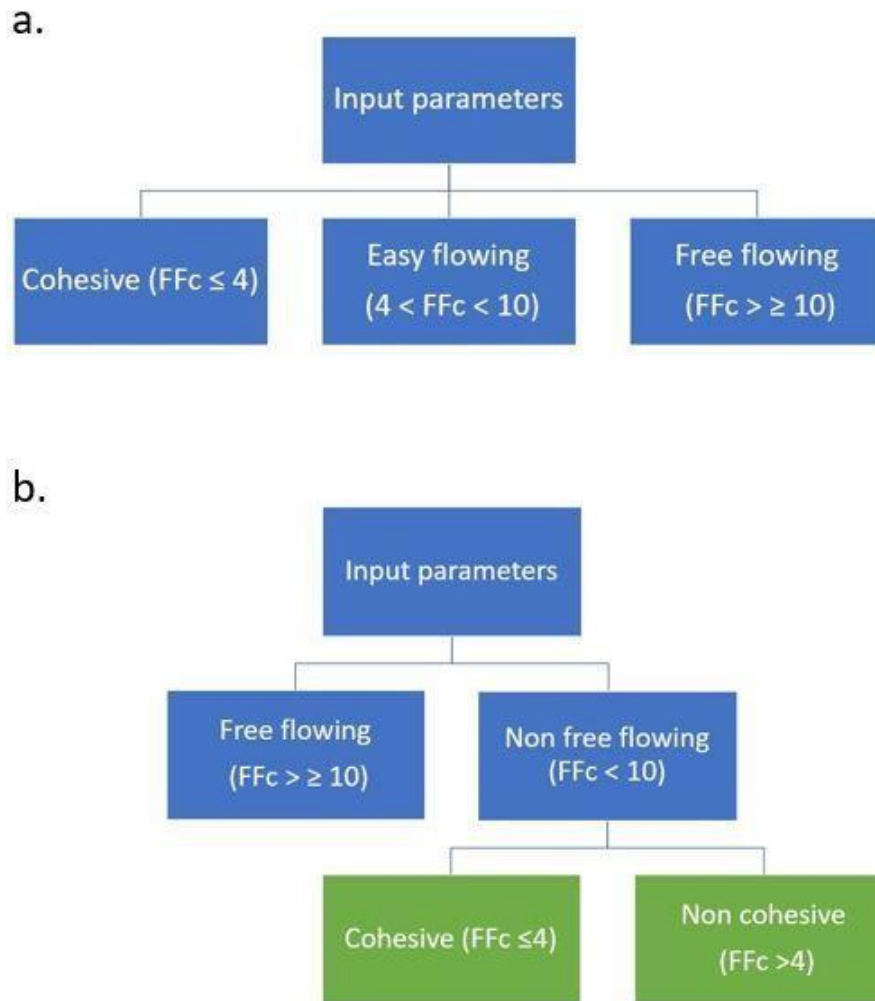


Figure 6-2: The diagrams of the RF models: a) multi-step classification model: the images are classified into cohesive, easy flowing or free flowing, and b) two-step classification model: the images are initially classified into free flowing or non-free flowing, and if they are classified as non-free-flowing, a second step is applied to classify them into cohesive or non-cohesive.

Table 6-2: The Morphologi® G3 descriptors used to train the RF classifier model.

Independent variables	<ul style="list-style-type: none"> - Particle size distribution (D10, D50, D90, Sauter Mean Diameter) - Aspect ratio distribution (D10, D50, D90) - Circularity distribution (D10, D50, D90)
Dependent variable	<ul style="list-style-type: none"> - FFc measured at 9 kPa.

6.2.5. Machine learning methods: CNN classification model.

As shown in Figure 1, the main goal of the DL models is the prediction of the flowability of a pharmaceutical material without the need to define instrument-specific descriptors. To achieve this, the models were trained using the raw microscopy images, all obtained using the same collection procedure using a Morphologi® G3 instrument as input and the FFc classes as defined in section 6.2.3 as the output. These images present a challenge for DL as their size ($\approx 4k \times 6k$ pixels) incur a very large computational overhead, particularly when computing neural network gradients, which measure the change of the weights of the network based on the change in the error. To circumvent this issue, each large image was segmented into 24 1024×1024 segments which were treated as individual samples for training. This image segmentation also overcame the issue of the large images not having dimensions that were divisible by 32, which is required by the DL architecture used. To test the model, the large images in the external test set were segmented and labelled using the three FFc classes. A majority vote was then taken across all segments to determine the overall class of the full image. The results for the external test set presented in Section 6.3.2 are majority votes for each material.

For comparison with the RF models, three Convolutional Neural Networks (CNNs) were trained following the same structure shown in Fig 6-2: one for the multiclass classification model and two to perform the two-step classification model. The CNN architecture used for all models was the VGG-11 model (Simonyan & Zisserman, 2014) with batch normalisation (Ioffe & Szegedy, 2015), implemented via PyTorch’s torchvision package (Paszke et al., 2019). Two main alterations were then made to the network:

- (i) The first feature extraction layer was replaced with a convolutional layer with the same number of output channels, kernel size and stride but with the number of input channels changed from 3 to 1 to reflect our images being greyscale.

- (ii) The last fully-connected layer of the network was replaced with one that maps to either 2 or 3 classes depending on whether the network is for the two-step or multiclass model (respectively) rather than the 1000 classes that the implementation assumes by default.

All the other layers were initialised by PyTorch's pretrained ImageNet (Deng et al., 2009) model, except the last convolutional layer that was initialised using He initialisation (He, Zhang, Ren, & Sun, 2015). Each model was trained using stochastic gradient descent (SGD) with momentum (Qian, 1999). The hyperparameters used in the optimiser were taken from (Simonyan & Zisserman, 2014), namely a starting learning rate of 0.01, weight decay of 5×10^{-4} and momentum coefficient of 0.9. The learning rate was reduced by a factor of 10 when the validation loss stagnated for two epochs. Each model was trained for 25 epochs with a batch size of 24 on 2 NVIDIA RTX A6000s. The optimal model was taken to be the first instance where the gradient of the validation loss was minimal. This corresponded to 13 for the multiclass model, 11 for the first step of the two-class model (free-flowing vs. non-free-flowing) and 7 for the second step of the two-step model (cohesive vs non-cohesive).

6.2.6. External Validation

External validation was used to assess the performance of both the RF and the CNN models as standard practice to demonstrate the applicability of the model on unseen data. Three powders of each powder flow class (cohesive, easy-flowing and free-flowing) were held out from the main dataset to perform the validation of the algorithms (see Table 6-3). Statistical metrics (Area Under the Receiver Operating Characteristic Curve (AUC – ROC), Accuracy, Precision and Recall) were calculated to compare the performance of the RF and the CNN performance. AUC – ROC is useful to know how much the results of the model can be trusted. This metric reported the probability of the model of distinguishing between classes. Furthermore, AUC – ROC has been recommended as the preference to report overall accuracy for evaluation of machine learning algorithms (Bradley, 1997). Accuracy reported the proportion of the correct predictions to the total number of predictions made by the model. This metric is perhaps one of the most used, however it does not account for class imbalance, which might result in misleading results. Precision is the proportion of True Positives divided by the True Positives and False Negatives. This metric quantified the proportion of correct predictions made by the model. Finally, recall is the ratio of true positives to all the positive predictions. Recall reported the ability of the model to find positive observations.

Table 6-3: The powders included in the external validation set, classified into cohesive, easy-flowing or free-flowing based on their FFc value.

Cohesive (FFc ≤ 4)	Easy flowing (4 < FFc < 10)	Free flowing (FFc ≥ 10)
Alcohol cetyl	Avicel PH-101	Mefenamic acid
Lidocaine	Ibuprofen 50	Pearlitol 200SD
Calcium Carbonate (40%) - multicomponent	Soluplus	Dimethyl fumarate

6.3. Results

6.3.1. Experimental methods

The 97 measurements done with the Morphologi[®] G3 were classified based on their FFc value. As mentioned, the dataset was split into training and validation datasets. The training dataset contained 24 cohesive powders, 31 easy-flowing powders, and 33 free-flowing powders (see Table 6-4). The validation dataset contained 3 cohesive, 3 easy-flowing and 3 free-flowing powders (see Tables 6-5 and 6-6).

Table 6-4: Number of powders per class used to train the RF model.

Classes	FFc values	Number of powders
Cohesive	FFc < 4	24
Easy-flowing	4 < FF < 10	31
Free-flowing	FFc > 10	33

Table 6-5: Number of powders per class used to validate the RF model.

Classes	FFc values	Number of powders
Cohesive	FFc < 4	3
Easy-flowing	4 < FF < 10	3
Free-flowing	FFc > 10	3

Table 6-6: Materials included in the external test set.

Material	FFc	Class
Alcohol cetyl	1.86	Cohesive
Lidocaine	2.33	Cohesive
Calcium Carbonate (40%) - multicomponent	3.16	Cohesive
Ibuprofen 50	7.42	Easy flowing
Avicel PH-101	7.46	Easy flowing
Soluplus	8.47	Easy flowing
Mefenamic acid	12.07	Free flowing
Dimethyl fumarate	13.02	Free flowing
Pearlitol 200SD	20.92	Free flowing

Table 6-7 summarize the values for each variable included in the training dataset. The range values of the descriptors and the mean and median values are reported in the table. The results of the measurements show that the materials included in the training dataset cover a wide range of particle sizes and shapes. The results also showed that most of the properties were positively skewed. Non-normal distributed could have an impact for the development of machine learning models, but RF can deal with non-normal distributed data. Fig 6-3 shows the distribution of PSD D50, which was positive skewed, and Fig 6-4 shows the distribution of aspect ratio D50, which was normal distributed.

Table 6-7: Descriptors used to train the RF classifier model. The range of values and the mean value are presented in this table.

Descriptor	Min	Max	Mean	Median
Particle size distribution D10 (μm)	13.24	247.00	61.17	54.57
Particle size distribution D50 (μm)	25.36	604.20	154.20	128.90
Particle size distribution D90 (μm)	57.79	745.80	304.87	252.20
Sauter Mean Diameter (μm)	22.98	359.20	109.18	94.26
Aspect ratio D10	0.21	0.66	0.40	0.38
Aspect ratio D50	0.47	0.85	0.65	0.65
Aspect ratio D90	0.79	0.96	0.87	0.87
Circularity D10	0.45	0.88	0.68	0.67
Circularity D50	0.73	0.96	0.88	0.88
Circularity D90	0.81	0.99	0.96	0.96
FFc	1.16	50.91	12.15	7.00

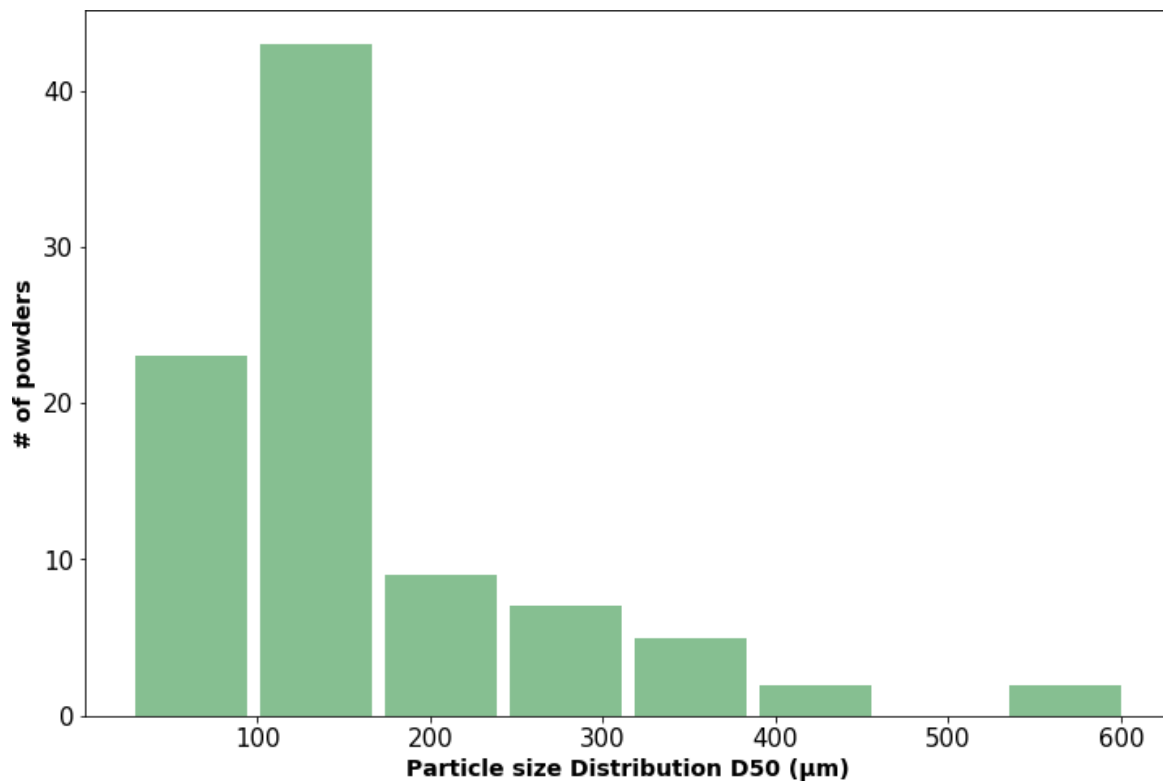


Figure 6-3: Particle size distribution (D50 values) across the powders included in the training dataset. Approximately 45% of the materials had a D50 value greater than 100 µm and smaller than 200 µm.

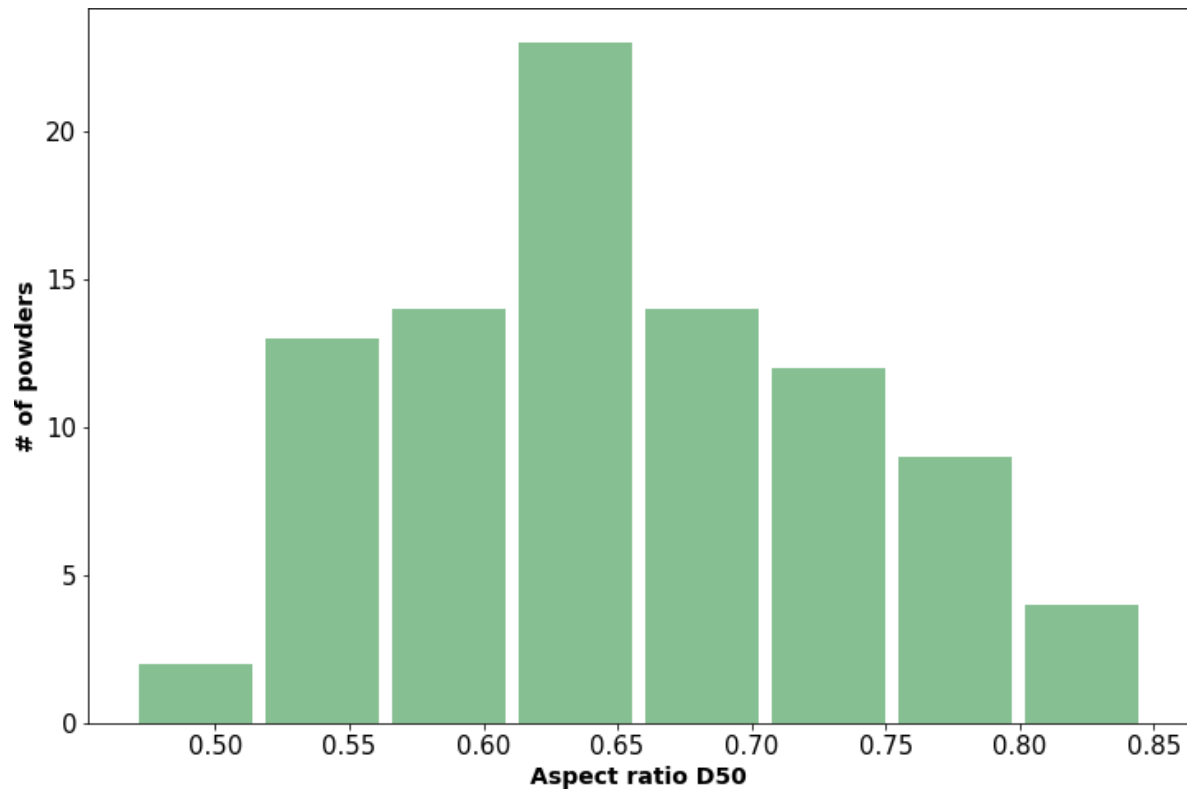


Figure 6-4: Distribution of the aspect ratio D50 values across the materials included in the training dataset. Approximately 25% of the materials had an aspect ratio D50 value between 0.6 and 0.65.

Finally, Fig 6-5 shows how the area composite extracted from the Morphologi[®] G3 of four different materials (one cohesive, two easy-flowing and one free-flowing) look like. These images are very difficult to differentiate without any additional analysis or information, which increases the complexity of the task for the DL models.

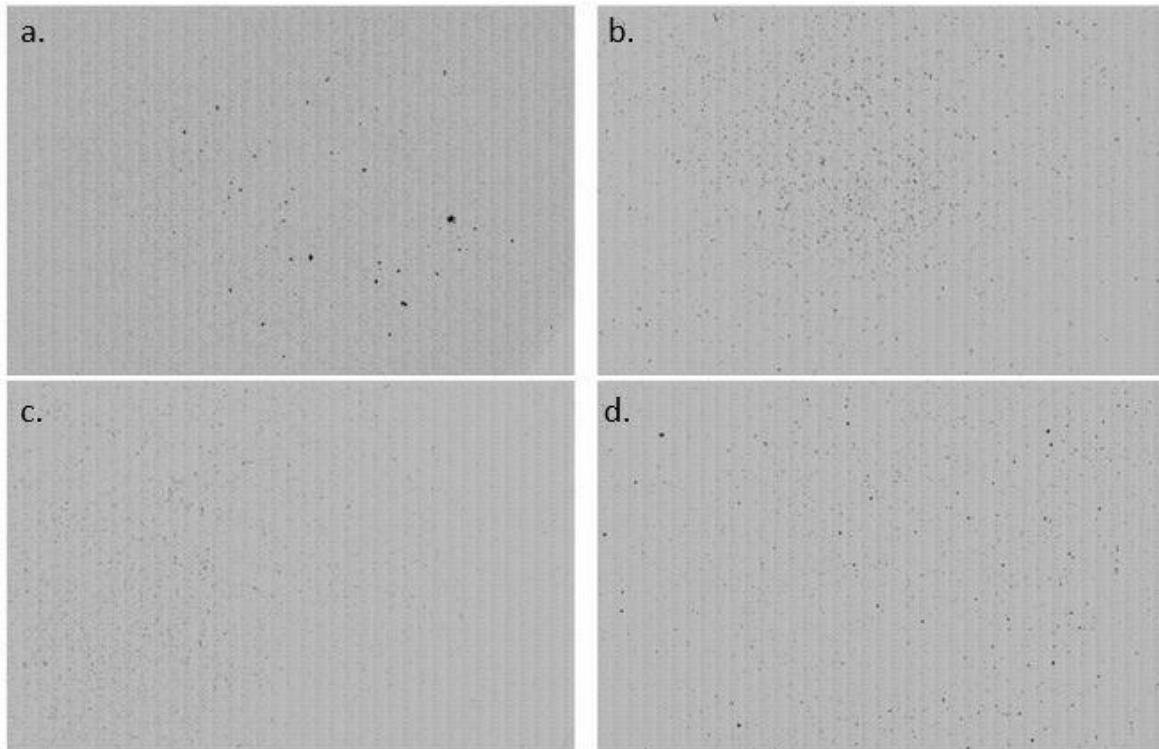


Figure 6-5: The area composite images taken using the Morphologi[®] G3 of a) cohesive material, b) easy-flowing material, c) easy-flowing material, and d) free-flowing material.

6.3.2. Machine learning results: comparison between the RF and CNN models for the prediction of powder flow of pharmaceutical materials

To develop the ML models, two approaches were taken. The first approach was the multi-class classification model, where both RF and CNN tried to classify a given set of descriptors or images, respectively, into cohesive, easy-flowing or free-flowing (see Fig 6-6). We observe that the results obtained by RF and CNN were very similar, with the same values of accuracy, precision, and recall (0.556), and with a slightly higher value of AUC – ROC of RF (0.694) than CNN (0.602). Fig 6-6 also shows that for both algorithms, 5 out of the 9 powders included in the validation dataset were correctly classified, and both algorithms made the same classification mistakes: 1 free flowing powder was classified as cohesive, 2 easy-flowing powders were classified as free-flowing and one cohesive powder was classified as easy-flowing. the RF model was more reliable than the CNN for the classification of powder flow in multi-class classification as the RF model achieved a higher score of AUC-ROC than CNN.

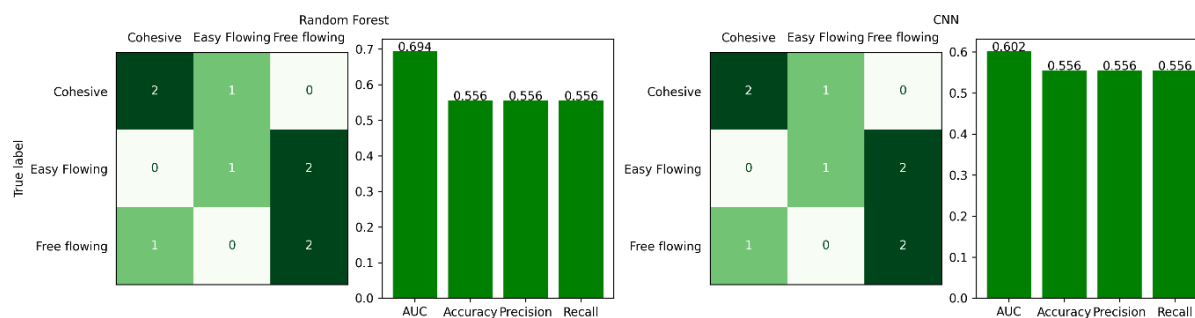


Figure 6-6: The results of the external validation of a) the RF multiclass classification model and b) the CNN multiclass classification model.

Since the performance of neither of the multi-class models was satisfactory, the second approach was tested using a two-step model (see Fig 6-3). The first step of this model, hereinafter Step 1, classifies powders into free-flowing ($FFc \geq 10$) or non-free-flowing ($FFc < 10$). The second step of the model, hereinafter Step 2, classifies powders into cohesive ($FFc \leq 4$) or non-cohesive ($FFc > 4$).

The results of these models are presented at the top (Step 1) and bottom (Step 2) of Fig 6-7. There is an improvement in the performance of these models compared to the performance of the multi-class classification models for both algorithms. In RF Step 1 (Fig 6-7(a)), 7 out of the 9 powders included in the external dataset were correctly classified, whereas for CNN (Fig 6-7(b)), only 6 out of 9 were correctly classified. The difference between the results was that, for RF, one more non-free-flowing powder was correctly classified. Regarding the metrics calculated from the confusion matrices, the CNN achieved a higher value of AUC - ROC denoting that this model was more reliable for the distinction between classes. The analysis of the metrics showed that, even though the results achieved in the external validation were worse for CNN, this algorithm was more reliable than RF for the classification of powders into free-flowing and non-free-flowing. In Step 2, CNN (Fig 6-7(d)) achieved better results than RF (Fig 6-7(c)). In this case, all the CNN metrics were higher than the RF metrics.

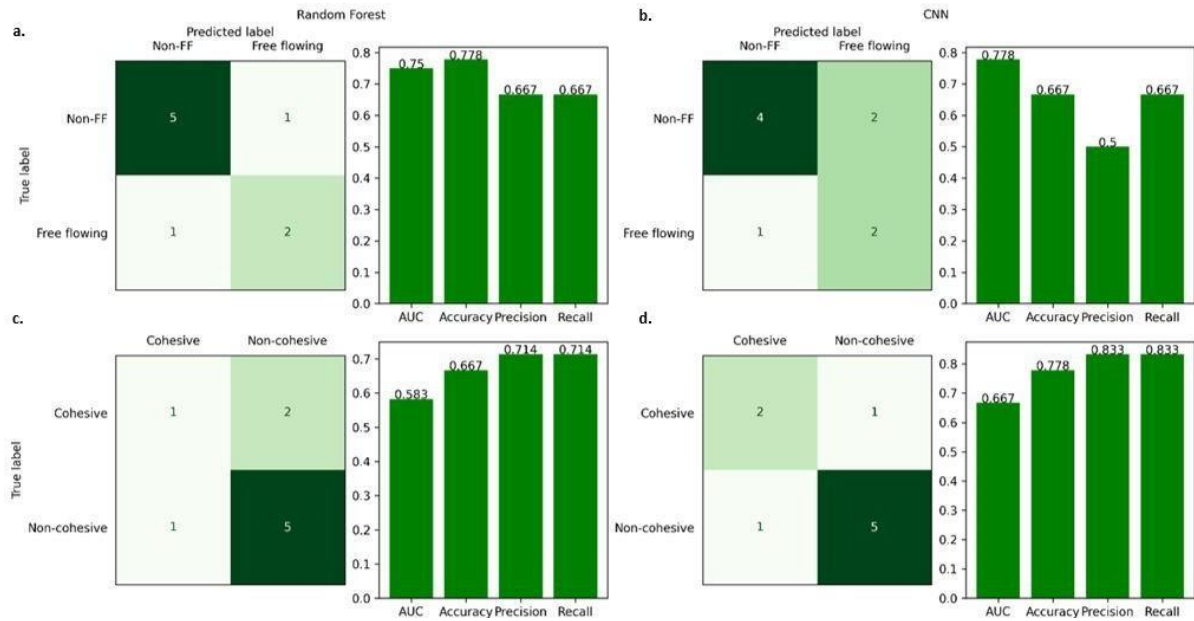


Figure 6-7: The performance of a) Step 1 of the two-step RF classification model, b) Step 1 of the two-step CNN classification model, c) Step 2 of the two-step RF classification model, and d) Step 2 of the two-step CNN classification model.

Finally, the results of the Step 1 and Step 2 were combined and reported in Fig 6-8. In this case, the number of correctly classified powders by RF and CNN were the same. However, the CNN metrics were higher than the RF metrics. Therefore, we hypothesized that the CNN model had a better performance than the RF model for the classification of powder flow. This hypothesis should be further investigated by the addition of more data and by increasing the complexity of the CNN models.

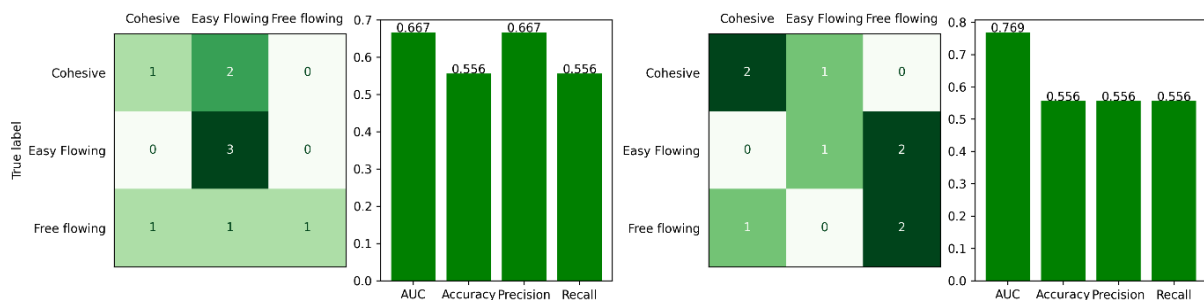


Figure 6-8: The results of the combination of the performance of Step 1 and Step 2 for a) the two-step RF classification model and b) the two-step CNN classification model.

6.4. Conclusions

In the work described herein, FFC classes of pharmaceutical materials were predicted from both particle size and shape descriptors and images of the powder sample. A total of 97 powders were analysed. These materials exhibited a wide range of particle size distributions, from 13.2 μm to 759.2 μm , and particle shape distributions, with circularity values between 0.42 and 0.99, and aspect ratio values between 0.21 and 0.98, and covered the 3 classes of FFC adapted from Jenike's classification (Jenike, 1964).

The results suggest that the prediction of powder flow can be achieved from both particle properties descriptors and images of powder, measured, and taken with static image analysis. Based on the AUC - ROC score, CNN model performed better when combining Step 1 and Step 2, with a 77% probability of distinguishing between the three classes (cohesive, easy flowing and free flowing). However, this result should be further investigated by the addition of more data into the dataset or the development of more complex DL architectures. Additionally, the quality of the images could be further analysed and improved to provide more information for the CNN model, and hence increasing the performance of the model. The performance of the CNN models of this chapter were below the performance of other DL models published in literature (Yang et al., 2019), and therefore if better quality images of particles were obtained, the performance of the DL models would be expected to increase.

Acknowledging that these models could be further improved, we believe that the application of the RF and CNN models in combination for the prediction of the FFC of pharmaceutical powders would accelerate the measurement of powder flow, and therefore reducing time and material required in the development of a new API. The use in combination of the two models could be done by the application of the two models at the same time or by using first the CNN model for feature extraction, coupled with the RF model for the classification of powder flow. The work presented in this paper illustrates the potential benefits of the implementation of data-driven models in early-stage development of pharmaceutical products.

6.5. References

- Abe, H., Yasui, S., Kuwata, A., & Takeuchi, H. (2009). Improving powder flow properties of a direct compression formulation using a two-step glidant mixing process. *Chemical and Pharmaceutical Bulletin*, 57(7), 647-652.
- Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular pharmaceutics*, 13(7), 2524-2530.
- Aulton, M. E., & Taylor, K. (2013). *Aulton's pharmaceuticals: the design and manufacture of medicines*: Elsevier Health Sciences.
- Barjat, H., Checkley, S., Chitu, T., Dawson, N., Farshchi, A., Ferreira, A., . . . Toby, M. (2021). Demonstration of the Feasibility of Predicting the Flow of Pharmaceutically Relevant Powders from Particle and Bulk Physical Properties. *Journal of pharmaceutical innovation*, 16(1), 181-196. doi:10.1007/s12247-020-09433-5
- Bassini, E., Galech, U., Soria, T., Aristizabal, M., Iturriza, I., Biamino, S., & Ugues, D. (2022). Effect of the particle size distribution on physical properties, composition, and quality of gas atomized Astroloy powders for HIP application. *Journal of Alloys and Compounds*, 890, 161631.
- Belič, A., Škrjanc, I., Božič, D. Z., Karba, R., & Vrečer, F. (2009). Minimisation of the capping tendency by tableting process optimisation with the application of artificial neural networks and fuzzy models. *European Journal of Pharmaceutics and Biopharmaceutics*, 73(1), 172-178.
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145-1159.
- Burgess, D. J., Duffy, E., Etzler, F., & Hickey, A. J. (2004). Particle size analysis: AAPS workshop report, cosponsored by the Food and Drug Administration and the United States Pharmacopeia. *The AAPS journal*, 6(3), 23-34.
- Clayton, J. (2019). An introduction to powder characterization. In *Handbook of pharmaceutical wet granulation* (pp. 569-613): Elsevier.
- Deng, J., Dong, W., Socher, R., Li, L. J., Kai, L., & Li, F.-F. (2009, 20-25 June 2009). *ImageNet: A large-scale hierarchical image database*. Paper presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition.
- Ferreira, A. P., Gamble, J. F., Leane, M. M., Park, H., Olusanmi, D., & Toby, M. (2018). Enhanced understanding of pharmaceutical materials through advanced characterisation and analysis. *Aaps Pharmscitech*, 19(8), 3462-3480.
- Fonteyne, M., Wickström, H., Peeters, E., Vercruyse, J., Ehlers, H., Peters, B.-H., . . . Sandler, N. (2014). Influence of raw material properties upon critical quality attributes of continuously produced granules and tablets. *European Journal of Pharmaceutics and Biopharmaceutics*, 87(2), 252-263.
- Gad, S. C. (2008). *Pharmaceutical manufacturing handbook: production and processes* (Vol. 5): John Wiley & Sons.
- Gamble, J. F., Toby, M., & Hamey, R. (2015). Application of image-based particle size and shape characterization systems in the development of small molecule pharmaceuticals. *Journal of pharmaceutical sciences*, 104(5), 1563-1574.
- Goh, H. P., Heng, P. W. S., & Liew, C. V. (2018). Comparative evaluation of powder flow parameters with reference to particle size and shape. *International journal of pharmaceutics*, 547(1-2), 133-141.
- Guo, A., Beddow, J., & Vetter, A. (1985). A simple relationship between particle shape effects and density, flow rate and Hausner ratio. *Powder Technology*, 43(3), 279-284.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv:1502.01852 [cs]*. Retrieved from <http://arxiv.org/abs/1502.01852>

- Hejduk, A., Czajka, S., & Lulek, J. (2021). Impact of co-processed excipient particles solidity and circularity on critical quality attributes of orodispersible minitables. *Powder Technology*, 387, 494-508.
- Hintz, R. J., & Johnson, K. C. (1989). The effect of particle size distribution on dissolution rate and oral absorption. *International journal of pharmaceuticals*, 51(1), 9-17.
- Hirschberg, C., Edinger, M., Holmfred, E., Rantanen, J., & Boetker, J. (2020). Image-Based Artificial Intelligence Methods for Product Control of Tablet Coating Quality. *Pharmaceutics*, 12(9), 877.
- Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*. Retrieved from <http://arxiv.org/abs/1502.03167>
- Jenike, A. W. (1964). Storage and flow of solids. *Bulletin No. 123, Utah State University*.
- Kaerger, J. S., Edge, S., & Price, R. (2004). Influence of particle size and shape on flowability and compactibility of binary mixtures of paracetamol and microcrystalline cellulose. *European Journal of Pharmaceutical Sciences*, 22(2-3), 173-179.
- Kesisoglou, F., & Wu, Y. (2008). Understanding the effect of API properties on bioavailability through absorption modeling. *The AAPS journal*, 10(4), 516-525.
- Kurek, M., Wyrwiz, J., Piwińska, M., & Wierzbicka, A. (2016). The effect of oat fibre powder particle size on the physical properties of wheat bread rolls. *Food Technology and Biotechnology*, 54(1), 45-51.
- Leane, M., Pitt, K., Reynolds, G., & Group, M. C. S. W. (2015). A proposal for a drug product Manufacturing Classification System (MCS) for oral solid dosage forms. *Pharmaceutical development and technology*, 20(1), 12-21.
- Leane, M., Pitt, K., Reynolds, G. K., Dawson, N., Ziegler, I., Szepes, A., . . . Group, M. C. S. W. (2018). Manufacturing classification system in the real world: factors influencing manufacturing process choices for filed commercial oral solid dosage formulations, case studies from industry and considerations for continuous processing. *Pharmaceutical development and technology*, 23(10), 964-977.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26(1), 80.
- Peng, T., Lin, S., Niu, B., Wang, X., Huang, Y., Zhang, X., . . . Wu, C. (2016). Influence of physical properties of carrier on the performance of dry powder inhalers. *Acta pharmaceutica sinica B*, 6(4), 308-318.
- Pereira Diaz, L., Brown, C., & Florence, A. (2021). *Prediction of powder flow of pharmaceutical materials from physical particle properties using machine learning*. Paper presented at the SPHERe Proceedings: 4rd International Symposium on Pharmaceutical Engineering Research.
- Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1), 145-151.
- Quan, H. T., Huy, D. D., Hoan, N. T., & Duc, N. T. (2020). *Deep Learning-Based Automatic Detection of Defective Tablets in Pharmaceutical Manufacturing*. Paper presented at the International Conference on the Development of Biomedical Engineering in Vietnam.
- Sandri, G., Bonferoni, M. C., Ferrari, F., Rossi, S., & Caramella, C. M. (2014). The role of particle size in drug release and absorption. In *Particulate Products* (pp. 323-341): Springer.
- Shangraw, R. F. (1989). Compressed tablets by direct compression. *Pharmaceutical dosage forms: Tablets*, 1, 195-246.
- Shekunov, B. Y., Chattopadhyay, P., Tong, H. H., & Chow, A. H. (2007). Particle size analysis in pharmaceuticals: principles, methods and applications. *Pharmaceutical research*, 24(2), 203-227.

- Shinde, P. P., & Shah, S. (2018). *A review of machine learning and deep learning applications*. Paper presented at the 2018 Fourth international conference on computing communication control and automation (ICCUBEA).
- Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*. Retrieved from <http://arxiv.org/abs/1409.1556>
- Swaminathan, V., & Kildsig, D. O. (2002). Polydisperse powder mixtures: effect of particle size and shape on mixture stability. *Drug Development and Industrial Pharmacy*, 28(1), 41-48.
- Tsimenidis, S. (2020). Limitations of Deep Neural Networks: a discussion of G. Marcus' critical appraisal of deep learning. *arXiv preprint arXiv:2012.15754*.
- Wadams, R. C., Akseli, I., Albrecht, J., Ferreira, A. P., Gamble, J. F., Leane, M., . . . Tobyn, M. (2022). Particle Property Characterization and Data Curation for Effective Powder Property Modeling in the Pharmaceutical Industry. *Aaps Pharmscitech*, 23(8), 286.
- Yan, X., Zhang, S., Fu, H., & Qu, H. (2020). Combining convolutional neural networks and on-line Raman spectroscopy for monitoring the Cornu Caprae Hircus hydrolysis process. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 226, 117589.
- Yang, Y., Ye, Z., Su, Y., Zhao, Q., Li, X., & Ouyang, D. (2019). Deep learning for in vitro prediction of pharmaceutical formulations. *Acta pharmaceutica sinica B*, 9(1), 177-185.
- Zawbaa, H. M., Schiano, S., Perez-Gandarillas, L., Grosan, C., Michrafy, A., & Wu, C.-Y. (2018). Computational intelligence modelling of pharmaceutical tableting processes using bio-inspired optimization algorithms. *Advanced Powder Technology*, 29(12), 2966-2977.
- Zegzulka, J., Gelnar, D., Jezerska, L., Prokes, R., & Rozbroj, J. (2020). Characterization and flowability methods for metal powders. *Scientific Reports*, 10(1), 1-19.

6.6. Appendix

6.6.1. Machine learning results: benchmark of the prediction of powder flow classes from particle size and shape using data-driven models.

6.6.1.1. Single-step RF model: multi-class classification

A RF multi-class model was trained to classify pharmaceutical powders into cohesive ($FFc < 4$), easy-flowing ($4 < FFc < 10$), and free-flowing ($FFc > 10$) materials. Fig 6-9 shows that the results achieved by the model of accuracy, precision and recall were only just over 0.5, and over 0.7 for AUC-ROC. Fig 6-10 shows the confusion matrices of the 10-fold cross-validation internal test set (Fig 6-10(a)) and the external test set (Fig 6-10(b)). These confusion matrices show that the model struggles at classifying easy-flowing materials, both in the validation and in the external set. The external test set confusion matrix shows that only two easy-flowing materials were misclassified as free-flowing. These materials were Avicel PH-101 and Soluplus. The FFc of both materials (7.46 and 8.46, respectively) are very close to the threshold that divides the easy-flowing to the free-flowing materials ($FFc = 10$). The rest of the materials were correctly classified by the model. To overcome the misclassification of easy-flowing materials, the two-step model was trained.

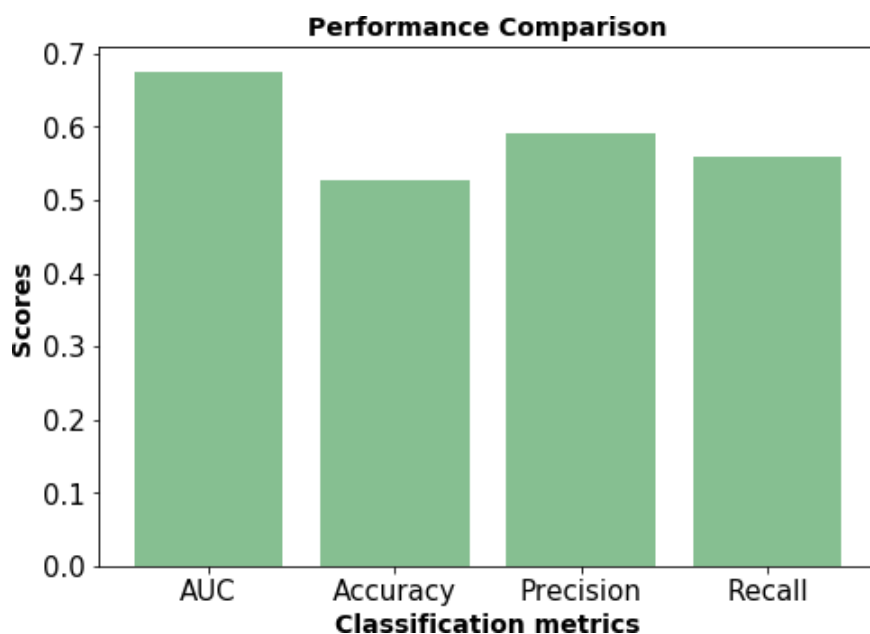


Figure 6-9: Classification metrics of the multi-step classification RF model, including accuracy (0.526 ± 0.11), AUC (0.676 ± 0.12), precision (0.591 ± 0.11), and recall (0.558 ± 0.08).

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	13	7	5	25
	Easy-flowing	5	15	12	32
	Free-flowing	6	9	19	34
Σ		24	31	36	91

		Predicted			Σ
		Cohesive	Easy-flowing	Free-flowing	
Actual	Cohesive	2	0	0	2
	Easy-flowing	0	1	2	3
	Free-flowing	0	0	2	2
Σ		2	1	4	7

Figure 6-10: RF confusion matrix for multi-class classification RF model (cohesive vs. easy-flowing vs. free-flowing materials) a) obtained from the validation and b) obtained from the external test.

6.6.1.2. Two-step RF models: binary classification

The results of the first step of the two-step classification model are shown in Fig 6-11. This step classifies between free-flowing and non-free-flowing materials. This figure shows that the metrics have improved compared to the single-step model metrics, achieving better results in accuracy (0.67), precision (0.71), and AUC (0.79) and recall (0.52). Fig 6-12(a) shows the 10-fold cross-validation confusion matrix. The model classified 80% of the non-free flowing materials correctly, but only 40% of the free-flowing materials. Fig 6-12(b) shows that all the materials included in the external test were correctly classified.

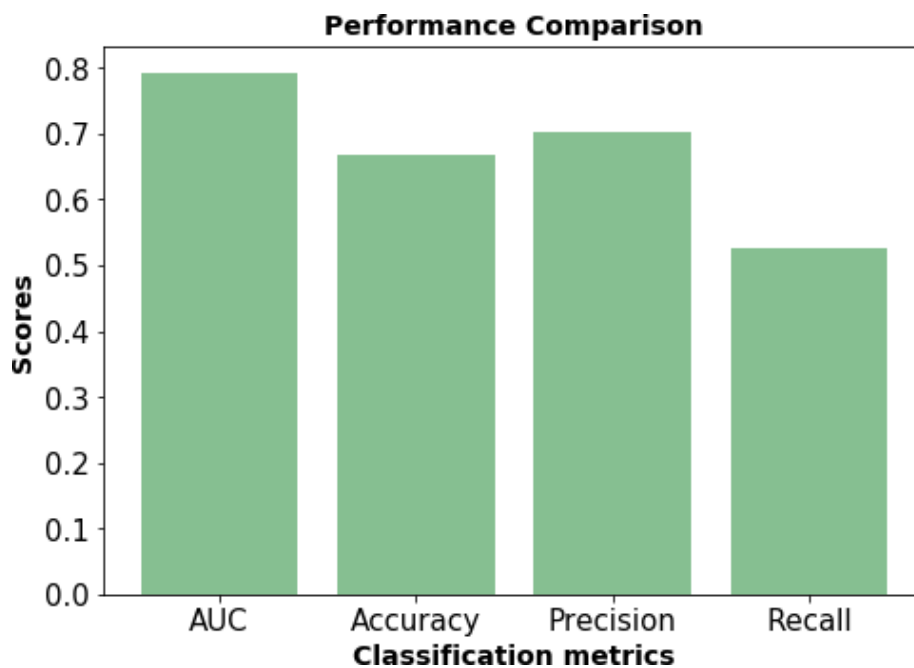


Figure 6-11: Classification metrics of the Step 1 of the RF model, including accuracy (0.668 ± 0.1), AUC-ROC (0.794 ± 0.09), Precision (0.703 ± 0.08) and Recall (0.525 ± 0.22).

		Predicted		Σ
		Non-free-flowing	Free-flowing	
Actual	Non-free-flowing	46	11	57
	Free-flowing	21	13	34
Σ		67	24	91

		Predicted		Σ
		Non-free-flowing	Free-flowing	
Actual	Non-free-flowing	5	0	5
	Free-flowing	0	2	2
Σ		5	2	7

Figure 6-12: RF confusion matrix for Step 1 (free-flowing vs. non-free-flowing materials) a) obtained from the validation and b) obtained from the external test.

The second step of the two-step model classified materials into cohesive and non-cohesive. As shown in Fig 6-13, the results of this step of the two-step model showed an improvement in comparison with the models described above, achieving better results in accuracy (0.75), precision (0.77), and AUC (0.85) and recall (0.89). Fig shows the result of the confusion matrices of the 10-fold cross validation (Fig 6-14(a)) and of the external test (6-14(b)). In this step, the best performance is achieved for the non-cohesive materials (over 80% of the materials were correctly classified in the 10-fold cross-validation), but the performance was poorer for the cohesive materials (40% correctly classified in the 10-fold cross-validation). Regarding the external test, only one cohesive material (cetyl alcohol) was misclassified as non-cohesive. Fig 6-15 shows the individual explanation of the cetyl alcohol prediction. The plot shows that the probability of being class 1 (non-cohesive) was 0.54, and therefore, the model was not confident in this prediction. The main drivers that increase the probability of being non-cohesive where the particle size distribution D10 (43.02 μm), and the Sauter Mean Diameter (92.94 μm). These values of PSD D10 were smaller than the mean of the PSD D10 values in the training dataset; however, regarding the mean values of PSD D10 per class in the second step of the two-step models, the difference between cohesive (65.18 μm), and non-cohesive (63.04 μm) was not very significant. Similar phenomenon happed regarding the values of Sauter mean diameter. The values of the cetyl alcohol are smaller than the mean of the training set, but the difference between the mean of the cohesive materials (114.78 μm) and non-cohesive (111.72 μm) was not significant either.

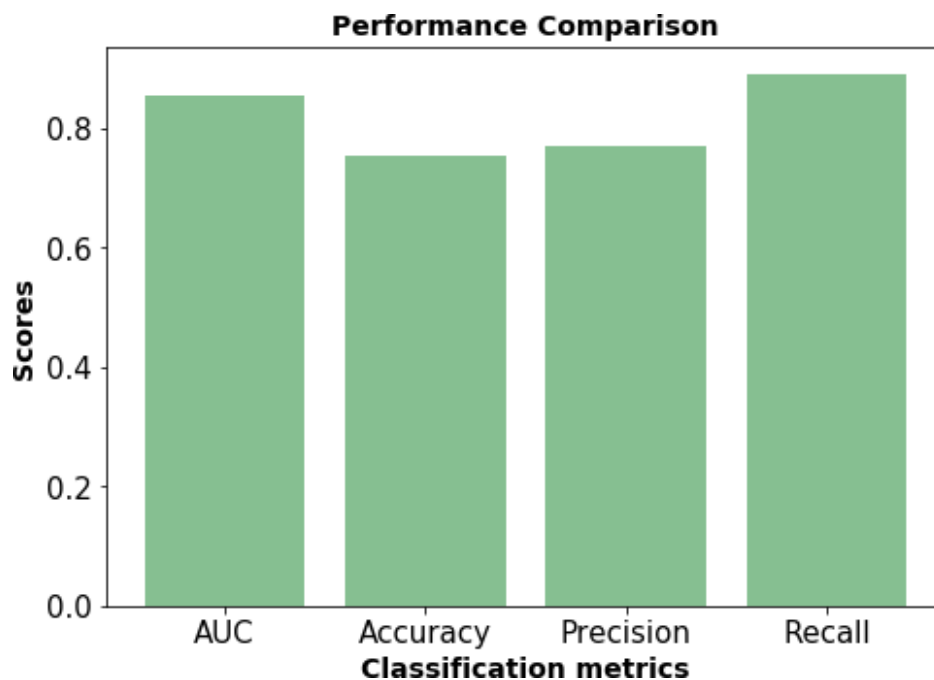


Figure 6-13: Classification metrics of the Step 2 of the RF model, including accuracy (0.753 \pm 0.09), AUC (0.855 \pm 0.08), precision (0.769 \pm 0.07), and recall (0.892 \pm 0.06).

		Predicted		Σ
		Cohesive	Non-cohesive	
Actual	Cohesive	10	15	25
	Non-cohesive	11	55	66
Σ		21	70	91

		Predicted		Σ
		Cohesive	Non-cohesive	
Actual	Cohesive	1	1	2
	Non-cohesive	0	5	5
Σ		1	6	7

Figure 6-14: RF confusion matrix for Step 2 (cohesive vs. non-cohesive materials) a) obtained from the validation and b) obtained from the external test.

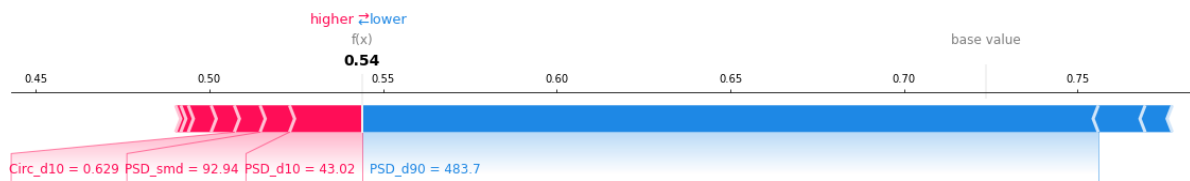


Figure 6-15: SHAP analysis for the misclassification of cetyl alcohol. The plot shows that the probability of being class 1 (non-cohesive) was 0.54, and therefore, the model was not confident in this prediction. The main drivers that increase the probability of being non-cohesive where the particle size distribution D10 (43.02 μm), and the Sauter Mean Diameter (92.94 μm).

7. Interrogation of particle and bulk property descriptors in the context of machine learning and prediction of pharmaceutical materials.

7.1. Introduction: the importance of the prediction of bulk density, surface area and surface energy on tablet manufacturability.

Digital design has become an important area of interest in pharmaceutical development and manufacturing to help develop medicines in a faster and smarter way. In the work described herein, the bulk density, the surface area and the surface energy of pharmaceutical materials are predicted to understand their influence on manufacturability using machine learning models.

This chapter has two main goals:

- (i) To explore the feasibility of the prediction of bulk density, surface area and surface energy from the physical particle properties of the pharmaceutical powder.
- (ii) To improve the understanding of the relationship between particle properties and bulk density.

The achievement of these aims would result in a reduction in the time of the bulk density, from 2 hours to 2 minutes, and surface properties measurements, from 24 hours to 2 hours, and the reduction of the measurements required. These measurements are essential to fully characterise pharmaceutical powders in order to predict their behaviour in downstream processes. Understanding bulk density is key to studying the flowability of a powder and it is used to calculate the Hausner ratio as an indicator of powder flow (Kalman, 2021), and therefore the prediction of bulk density from particle size and share would save time and resources, and thereby help to predict the manufacturability of new APIs or formulations.

Even though surface area and surface energy data was proven not to be useful for the prediction of powder flow since it did not improve the performance in comparison with the models trained on particle size and particle shape (Chapter 4), the prediction of surface area in initial stages of drug development would allow to estimate the drug bioavailability since surface area can be directly linked to the dissolution rate of the drug (Kumar, Chirravuri, & Shastri, 2014). Furthermore, it has been demonstrated that surface energy impacts compactability and hence, manufacturability (Fichtner, Mahlin, Welch, Gaisford, & Alderborn, 2008).

7.1.1. Bulk density

Achieving the desired bulk density is important when handling pharmaceutical powders. Consequently, the efforts of formulation departments of pharmaceutical companies have been focused on creating formulations with sufficiently large bulk density to enable manufacturability (Fitzpatrick, 2013; Kunnath, Chen, Zheng, & Davé, 2021). The ratio of tapped to bulk density is known as the Hausner ratio (Grey & Beddow, 1969); thus, an increase in bulk density would result in a decrease in the Hausner ratio. Table 7-1 gives the range in Hausner ratios observed for given powder flow behaviours. Here we see that as the Hausner ratio decreases, the flowability of the powder improves (Hausner, 1967). Bulk density is calculated by dividing the mass of the powder by the volume occupied by the powder, considering the voids. Tapped density is calculated by tapping the powder to consolidate the packed bed resulting in an increased bulk density (Pharmacopeia, 2012). Moreover, the MCS (Leane, Pitt, Reynolds, & Group, 2015) proposed guidelines for manufacturability, indicating that materials that have a bulk density greater than 0.3 g/ml will yield greater manufacturing efficiency, and that materials with a bulk density greater than 0.5 g/ml will be most suitable for direct compression.

Kunnath et al developed a first-order model to estimate powder bulk density improvements after dry coating (Kunnath et al., 2021). Yu et al. studied the empirical correlation between porosity and particle size (Yu, Zou, & Standish, 1996). Porosity is the proportion of the volume of the sample that does not contain solid, and thus, porosity is inversely correlated with bulk density. Yu et al. concluded that porosity decreases when sphericity increases (Yu et al., 1996), and therefore that bulk density should increase with sphericity. By contrast, German et al. concluded that the loose packing density of non-cohesive spherical powders does not depend on the particle shape as long as the particle size is large enough (German, 1989). The differences in conclusions among these experimental approaches highlighted the need for better understanding of the relationship between particle and bulk properties across the range of attributes of pharmaceutical relevance. The approach proposed in this chapter is the development of machine learning models coupled with interpretability methods (S. Lundberg, 2018) to predict and understand bulk density based on the particle size and shape of a given powder.

Table 7-1: Powder Flowability based on Hausner ratio from excellent flow (HR smaller than 1.11) to very poor flow (HR greater than 1.46) (Gorle & Chopade, 2020).

Flow behaviour	Hausner ratio
Excellent	1 – 1.11
Good	1.12 – 1.18
Fair	1.19 – 1.25
Passable	1.26 – 1.34
Poor	1.35 – 1.45
Very poor	> 1.46

7.1.2. Surface properties

The characterisation of surface area and surface energy for individual samples is relatively time consuming (close to 48 hours) and the instrument used is very sensitive to external conditions and sample preparation. Thus, even though these surface properties can be accurately measured, the implementation of a predictive model that gave reliable estimates of surface properties without the need for extensive measurement would be highly beneficial. Samiei et al. demonstrated the impact of electrostatic properties on tablettability, particularly on the punch sticking propensity (Samiei et al., 2017). For the direct compression process to be viable, the powder needs to have a low punch sticking propensity (Gohel & Jogani, 2002). Following the same reasoning, it could be expected that surface area and surface energy would also have an impact on powder flow. However, the results obtained by Barjat et al. (Barjat et al., 2021), and the results obtained in Chapter 4 of this thesis did not support this hypothesis: the addition of surface area and surface energy data to the training dataset did not improve the performance of the machine learning models for the prediction of powder flow of pharmaceutical materials. On the other hand, R Williams presented the importance of surface energy to understand the performance of particulated pharmaceutical products and processes that have not yet been clearly established (R Williams, 2015). Therefore, another hypothesis can be raised: surface area and surface energy are yet not well understood, and the exploration of these properties can be useful to better understand how surface properties interact with other particle and bulk properties.

Furthermore, studies have demonstrated a correlation between specific surface area (surface area per unit mass or per unit volume), compressibility characteristics (Sridharan, Rao, & Murthy, 1986)

and the angle of internal friction (Moore, 1991) and how these properties affect powder compaction. Powder compaction is achieved by the rearrangement of the particles followed by elastic and then plastic deformation (Felton, 2013). There are three processes that are used to explain the bonding mechanism in powder compaction: the creation of solid bridges between particles, interparticle forces, and mechanical interlocking. These three mechanisms depend on the effective amount of surface area available for bonding and interparticle contact (Nyström, Alderborn, Duberg, & Karehill, 1993). The surface area depends on particle properties such as size and texture, even though this relationship between the surface area and the compact strength is not yet well understood (Karehill, Glazer, & Nyström, 1990).

On a different note, the bioavailability of the drugs in the organism depends on the solubility and the permeability of the drug. Based on these two parameters, the Biopharmaceutical Classification System (BCS) was designed to select drugs that can potentially be successful in *in vivo* studies (Benet, 2013). It is estimated that 90% of drugs are poorly soluble (class II or class IV of the BCS), which can compromise their oral bioavailability (Dressman, Amidon, & Fleisher, 1985). The bioavailability of the drugs depends on their dissolution rate, which is calculated following the Noyes-Whitney equation (Noyes & Whitney, 1997), shown in Eq 7-1, where C and C_s are the concentration of the dissolved substance, D represents the diffusion coefficient of the substance, S_w represents the surface area, V represents the volume of the solution and h represents the thickness of the diffusion layer. To avoid these compromises in bioavailability, one of the most common used techniques is particle size reduction (Liversidge & Cundy, 1995): as the particle size decreases, the surface area increases, improving the dissolution rate and therefore oral bioavailability. For this reason, it becomes relevant to consider the optimum particle size for both manufacturability and bioavailability when reverse engineering the “ideal” properties of a new API.

$$\frac{dC}{dT} = \frac{DS_w}{Vh}(C_s - C) \quad (7-1)$$

The concept of “surface energy” was introduced by Gibbs and is defined as the difference in energy between a particle, and the same number of atoms of such particle extended in an infinite solid (Gibbs, 1875). The surface energy of powders is an important factor in the control of the pharmaceutical manufacturing performance and the final quality of the product (R Williams, 2015). Surface energy values have been used in granulation for binder (granulation agent) selection (Rowe, 1988; Wells & Walker, 1983; Zajic & Buckton, 1990). Moreover, they have been used for the prediction of dry powder inhaler formulation performance where the surface energies were inversely proportional to dispersion performance (Saleem, Smyth, & Telko, 2008).

Therefore, accurate prediction of the bulk density, the surface area and the surface energy could reduce the number of experiments required to characterise a pharmaceutical powder and is the motivation of this chapter. Using ML models as an interpretability tool, the work presented in this chapter paves the way for the prediction of these three properties, as well as for the increase in understanding of how these properties are correlated with other physical properties.

7.2. Materials and methods

7.2.1. Experimental methods for data generation

The USP recommends at least 100 grams of material for the measurement of bulk density, and if such quantity was not available, the measurement could be carried out in a 100 ml cylinder using 50 to 100 ml of material. However, these quantities can be challenging to obtain in early development of a new API, and therefore an alternative, more feasible, approach has been used in this thesis to measure bulk density. The FT4 was used to analyse the conditioned bulk density (CBD) as described in Chapter 3 (section 3.2.6).

CBD is the density of the powder's sample free of stress or excess air, obtained after the conditioning cycle of the FT4. A 10 ml vessel was used to measure the CBD, for which approximately 10 grams of sample were needed per measurement. The bulk density measured by the FT4 Powder Rheometer is not included in the Pharmacopoeia and the absolute results may differ from other more traditional methods but provide a consistent relative measure of the material property of measure. A total of 112 powders (same powders included in Chapter 4) were measured to build the bulk density models training dataset.

Surface area and surface energy have been analysed with the Surface Energy Analyzer (SEA) – iGC Surface Measurement Systems (section 3.2.4). A total of 31 powders were tested with the SEA instrument to build the surface area and surface energy models, using a range of dispersive probes (n-decane, n-nonane, n-octane, n-heptane, and n-hexane), injected at a range of 4 fractional surface coverages between 0 and 10%. The free dispersive surface energy was calculated using a best fit line through the data points, using Cirrus plus analysis software (version 1.2). The dispersive component was obtained using the Dorris-Gray approach (Shi, Wang, & Jia, 2011) and the polar components were calculated using the polarization approach, using the Cirrus plus analysis software (version 1.2) of the SEA.

7.2.2. Machine learning methods

Classification and regression models were built for the prediction of bulk density, surface area and surface energy. For bulk density classification models, two classes were divided at a threshold of 0.5 g/ml, creating a low-density class and high-density class. These classes were defined following the recommendations of the MCS for direct-compression materials (Leane et al., 2015), and the previous work obtained in Chapter 4 (section 3.4). For surface area classification models, two classes were divided at a threshold of 0.65 m²/g. Likewise, for surface energy, two classes were divided at a threshold of 6.97 mJ/m². The mean of the surface area and the mean of the surface energy of the training dataset were taken as the threshold values.

ML classification algorithms kNN, SVM, RF, MLP, NB, LR, AB, and GB (described in Chapter 3, section 3.5.2.) were explored for the classification task. The hyperparameters were tuned to improve algorithm performance unless no improvement was observed and therefore the default hyperparameters were used. In kNN, one of the most important parameters is the number of nearest neighbours, which is determined from the squared root of the total number of training observations. Considering the data used to build these models, the squared root of 112 is 10.6 (for bulk density models), and the squared root of 35 is 5.9 (for surface area and surface energy models). Since binary classification models were built, the odd number was selected (11 and 5, respectively). The rest of the hyperparameters remained as default. The kernel function used in SVM was linear, which is the most basic kernel in SVM. The rest of the parameters for SVM were set as default. The number of trees in RF was modified to 200, and the rest of the parameters were set as default. For the MLP classifier, the activation function of the hidden layer was the logistic sigmoid function, the solver selected was stochastic gradient descent; alpha was 4e-06, which refers to the strength of the L2 regularization term. Finally, the number of maximum iterations was set at 100000. For NB, the hyperparameters were set as default, since hyperparameter tuning did not achieve better results than default parameters. In LR, the solver was “liblinear”, which is a good choice for small datasets and binary classification, and a maximum of 1000 iterations was set. For GB, the default parameters were used. Lastly, for AB, the algorithm used was “SAMME.R”, which is a variant of the SAMME algorithm.

For model validation, the dataset was sampled using 10-fold cross-validation for bulk density models and 5-fold cross-validation for surface area and surface energy models. The difference in the size of the folds was due to data availability. The best-performing algorithm was chosen based on its AUC-ROC score, and it was further validated with an external dataset and analysed for model interpretability using SHAP values.

For regression models, PLS, RF, GB and AB were explored. For PLS, 2 components were calculated. For RF, GB and AB, the default hyperparameters were used. The dataset was randomly split into 25% testing and 75% training. The performance of the algorithms was compared based on their MAE.

Lastly, interpretability methods were used to understand how the classification and regression models were making predictions (see Chapter 3, section 3.5.4.). The interpretation of these models helped elucidate how physical and bulk properties interact with each other. SHAP values were used for global and local analyses. Global analysis refers to ranking the importance of the variables, to studying the direction (+/-) each variable contributed to the model output, to analyse the correlations between independent and dependent variables. Local analyses refer to understanding the prediction of an individual powder.

7.3. Results and discussion

7.3.1. Bulk density measurements and associated ML models

7.3.1.1. Bulk density experimental measurements

To establish the dataset of bulk density values for training ML models, the pharmaceutical powders described in Chapter 3 were measured with the powder rheometer to calculate their bulk density. The distribution of the bulk density of these powders is shown in Fig 7-1. As mentioned, the model was also validated with an external dataset. This external dataset contained 8 materials, and most of them had a bulk density higher than 0.5 g/ml (see Table 7-2). The materials included in the external dataset are within the boundaries of the domain of the bulk density model (0.2 – 1.2 g/ml). 16 variables were included in the training data set (see Table 7-3).

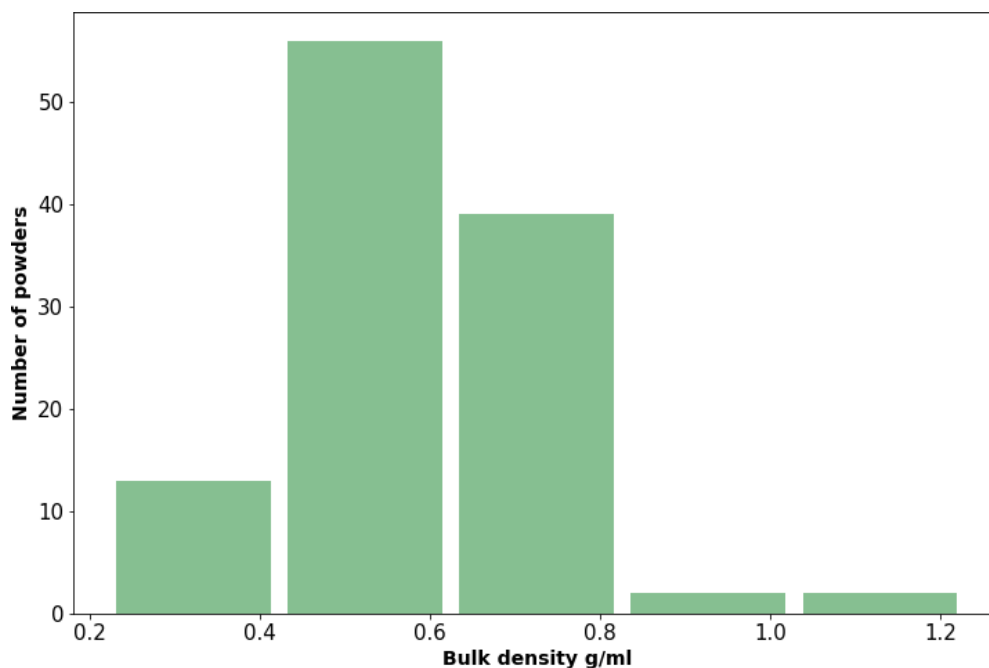


Figure 7-1: Bulk density (g/ml) experimental results.

Table 7-2: Bulk density experimental results of the external dataset used to validate the machine learning model.

Material	Bulk density (g/ml)
Span60	0.589
Calcium carbonate	0.563
Ibuprofen 50	0.531
Avicel PH-101	0.344
Ibuprofen 20 % + Povidone	0.528
Plasdone K29/32	0.453
Pearlitol 100 SD - Mannitol	0.585
Ibuprofen 50 (20%) - multicomponent	0.593

Table 7-3: Variables included in the bulk density model.

Variables included in the bulk density model	
Concentration of API	SMD
Concentration of excipient 1	Sphericity D10
Concentration of excipient 2	Sphericity D50
Concentration of excipient 3	Sphericity D90
Concentration of excipient 4	Aspect ratio D10
PSD D10	Aspect ratio D50
PSD D50	Aspect ratio D90
PSD D90	FFc

7.3.1.2. Data analysis

Statistical analysis was performed to understand the data, starting with Pearson correlation coefficient (PCC). Fig 7-2 shows the heatmap of the PCC for each pair of variables. Positively correlated variables were represented in light orange, and negatively correlated variables were represented in dark red. Only variables that belong to the same descriptor (sphericity or aspect ratio) were highly correlated with each other. One variable of each pair of highly correlated variables ($PCC > 0.9$) was randomly removed from the training dataset, because highly correlated variables do not improve the performance of the model. Indeed, including highly correlated variables to train ML models can even be detrimental for some machine learning algorithms. For example, RF is able to detect the interactions between variables, but if two variables that are highly correlated variables are included in the training set they can mask interactions between other data (Darst, Malecki, & Engelman, 2018). Alternatively, in the case of SHAP methods, importance scores are additive, so highly correlated variables will appear less important as the importance score will be split between the two highly correlated variables (S. M. Lundberg & Lee, 2017; Mase, Owen, & Seiler, 2019).

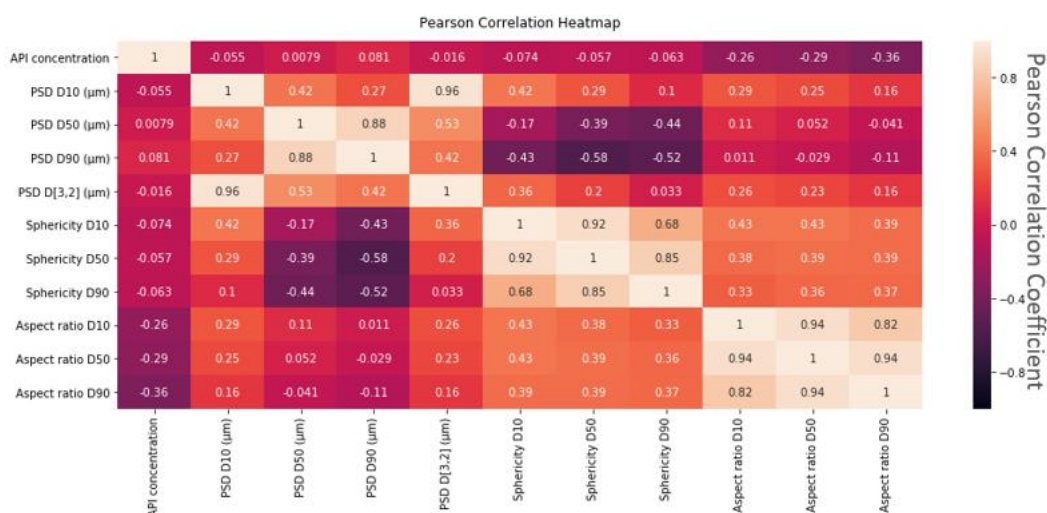


Figure 7-2: Pearson correlation coefficient heatmap of the features used for training bulk density models.

After 2 highly correlated variables were removed, the remaining 11 variables (see Table 7-4) were evaluated for any further correlations (see Fig 7-3). This plot includes the distribution plot for each variable (diagonal axis) and scatter plots of each pair of variables. No other highly correlated between variables were found in the data set.

Table 7-4: Variables included in the bulk density model after PCC

Variables included in the bulk density model after PCC	
Concentration of API	PSD D50
Concentration of excipient 1	PSD D90
Concentration of excipient 2	Sphericity D10
Concentration of excipient 3	Sphericity D90
Concentration of excipient 4	Aspect ratio D10
PSD D10	FFc

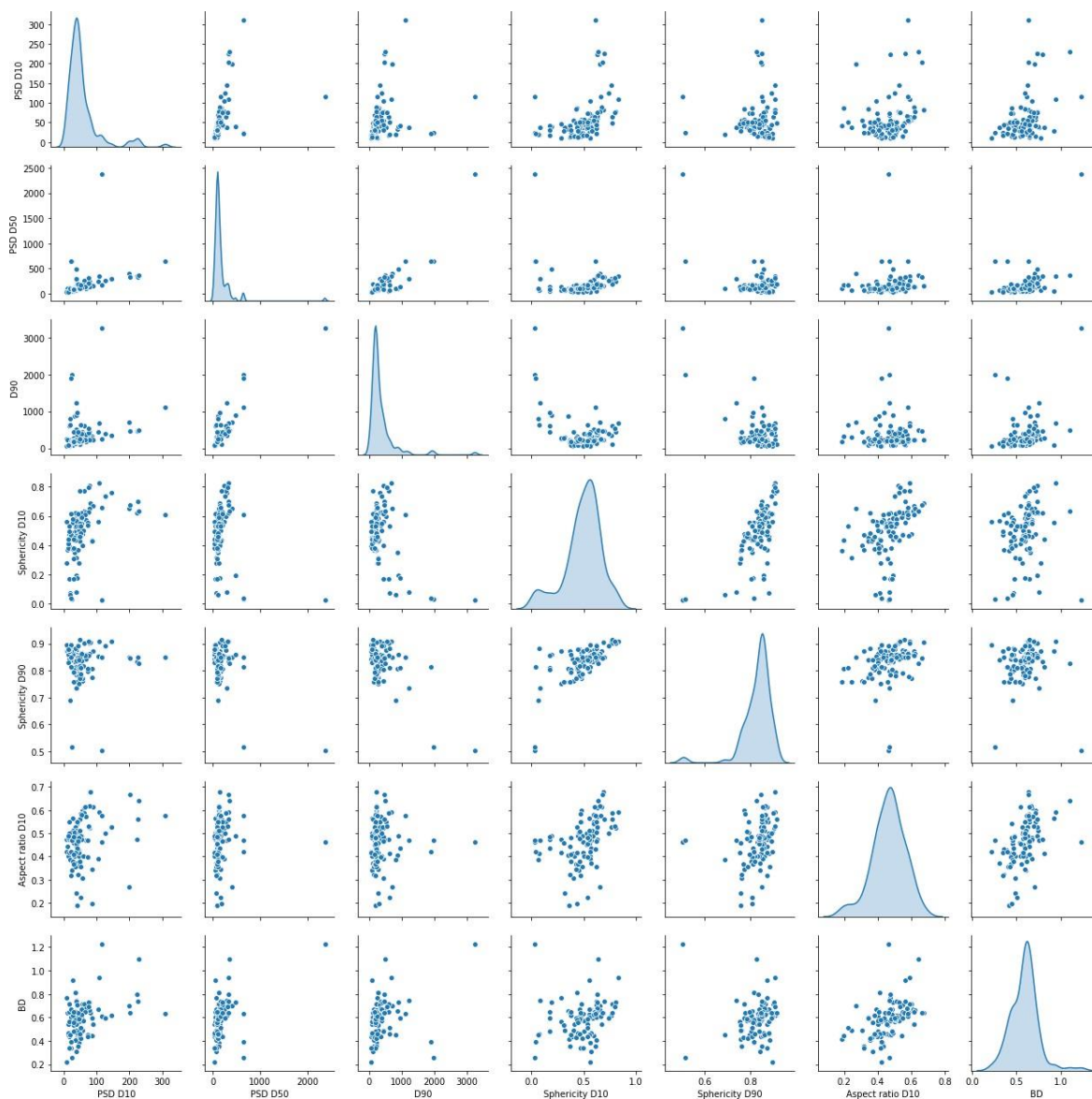


Figure 7-3: Pairplot of the particle size (D10, D50, D90), particle shape (sphericity D10 and D90, and aspect ratio D10) and bulk density. This figure shows the scatter plot of each pair of variables to explore correlations.

PCA analysis was then used as a clustering technique to identify potential groups of data points in the training data. The variables reported in Table 4 were used for this analysis. PC1 (x-axis) was plotted against PC2 (y-axis) (see Fig 7-4). The cumulative variance explained by the 2-component PCA was 48%, and the data points were plotted in green (bulk density greater than 0.5 g/ml) or red (bulk density lower than 0.5 g/ml) (see Table 7-5). The green and red data points overlapped and therefore, this analysis was not useful to cluster the data based on their bulk-density class. Considering that 12

variables were considered for the PCA, the small cumulative variance explained was not sufficient to justify the use of PCA as a dimensionality reduction technique.

Table 7-5: Bulk density categories. Powders with a bulk density less than 0.5 g/ml belong to the low-density class, and powders with bulk density greater than 0.5 g/ml belong to the high-density class.

Classes	Values	Number of powders	Suitable for Direct Compression?
Low bulk density	BD < 0.5 g/ml	29	No
High bulk density	BD > 0.5 g/ml	75	Yes

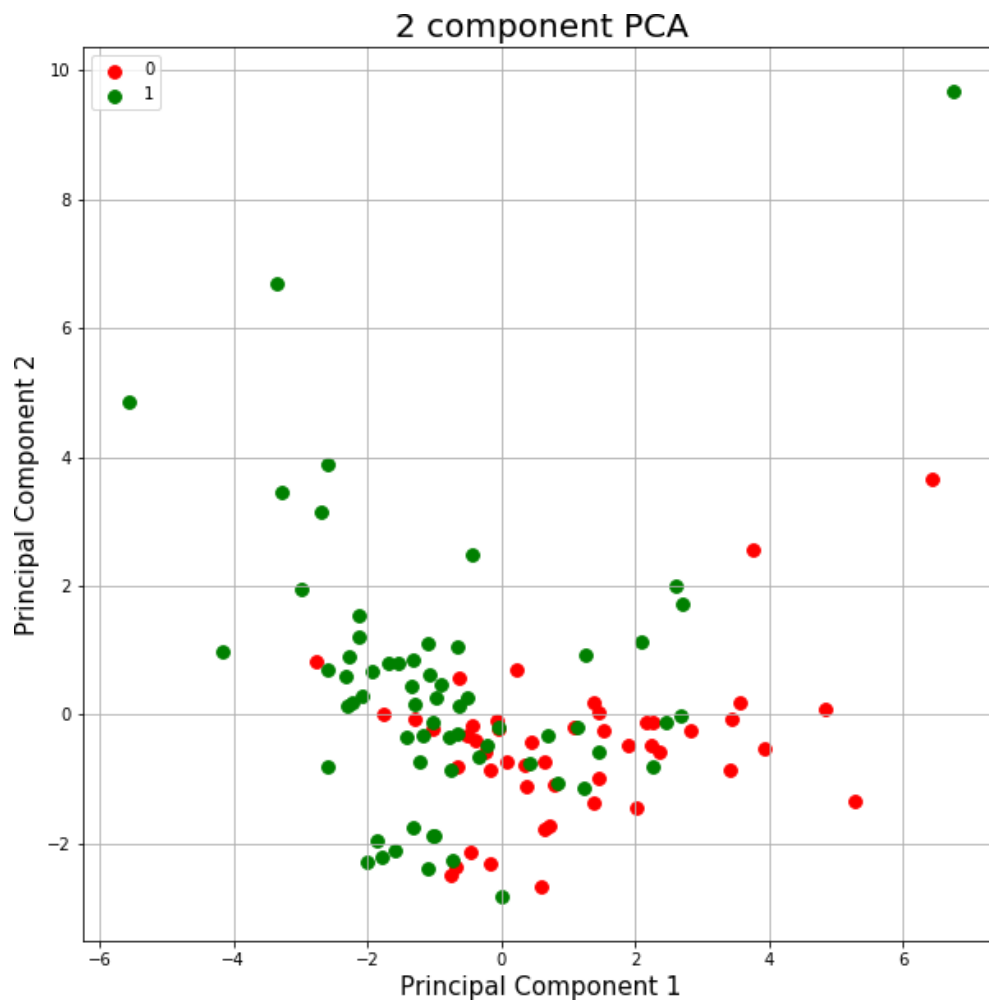


Figure 7-4: A plot of the top 2-components in the PCA analysis which can account for 48% of the variance of the dataset. Low-density powders (class 0) are represented in red, whereas high-density powders (class 1) are represented in green.

7.3.1.3. Machine learning models: classification models for the prediction of bulk density

Since PCA could not usefully classify the bulk density of the test pharmaceutical powders, the ability of ML models to predict bulk density was investigated. Firstly, classification models were built, for which two classes of bulk density based on the desired value for direct compression (> 0.5 g/ml) (Leane et al., 2015), as showed in Table 7-5.

Classification algorithms were trained to predict bulk density. The trained algorithms were compared to select the best-performing algorithm which was then used for validation and interpretation of the model. Fig 7-5 shows that both RF and GB achieved the highest performance as calculated by 10-fold cross-validation. RF achieved the slightly higher performance of 0.828 AUC – ROC, and hence, this was the algorithm used for external validation. The external validation was performed on the RF model to test for overfitting. Likewise, as the best-performing model, RF was also selected for model interpretability analysis.

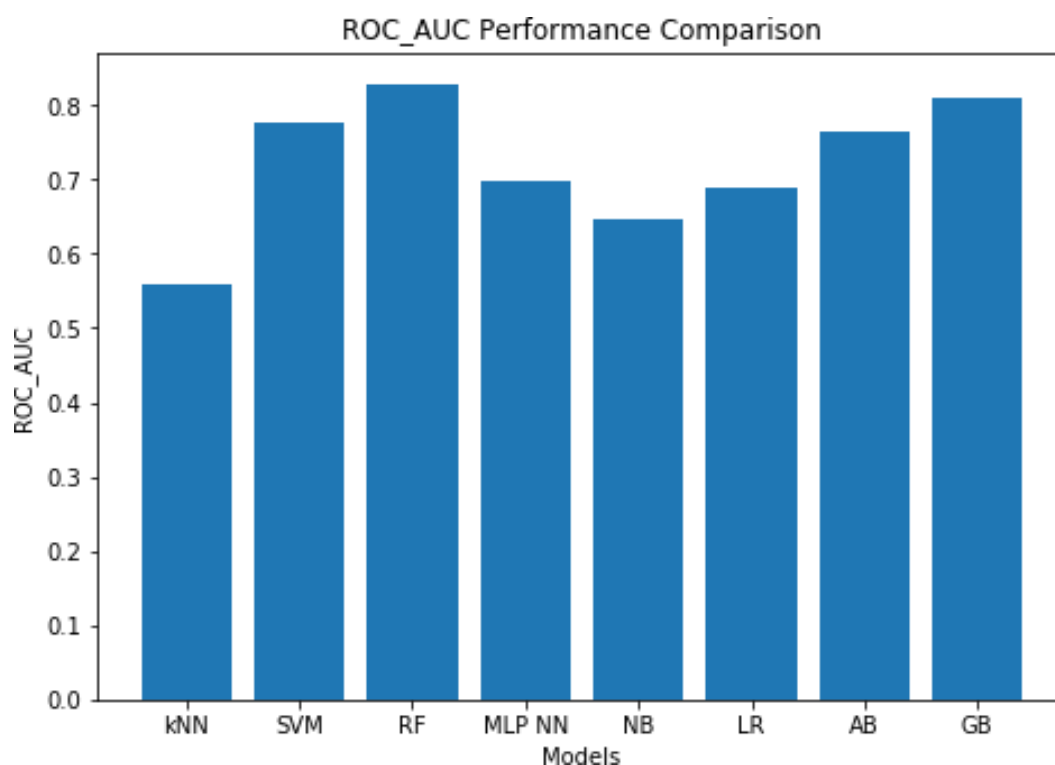


Figure 7-5: Comparison of the performance of the classification algorithms trained for the prediction of bulk density, calculated using 10-fold cross-validation. The metric analysed was AUC-ROC.

The same materials and measurements used in the external validation of Chapter 4 were used to validate the bulk density classification model. The confusion matrix reported in Fig 7-6 shows that 1 of the 2 low-density powders and 5 of the 6 high-density powders were correctly classified, resulting in a 75% accuracy (similar to the performance achieved in the testing set). Table 7-6 shows the classification scores for each prediction. The results show that there was not a significant difference between the classification scores of correctly and incorrectly classified powders.

		Predicted	
		Low bulk density	High bulk density
Actual	Low bulk density	1	1
	High bulk density	1	5

Figure 7-6: External validation results using the RF model. 5 of the powders included in the dataset had low density (BD > 0.5 g/ml), and 3 powders had high density (BD > 0.5 g/ml).

Table 7-6: The classification scores which predict whether or not a powder has a bulk density that is optimal for direct compression for each of the materials included in the external dataset. When the classification was correct, the classification score is reported in green, whereas when the classification was wrong, the classification score is reported in red. Classification scores above 0.5 indicate “high-density class” will be predicted class and classification scores below 0.5 indicate that “low-density class” will be the predicted class.

Material	Actual class	Predicted class	Classification Scores
Span60	High	High	0.780
Calcium carbonate	High	Low	0.675
Ibuprofen 50	High	High	0.745
Avicel PH-101	Low	Low	0.745
Ibuprofen 20 % + Povidone	High	High	0.515
Plasdone K29/32	Low	High	0.575
Pearlitol 100 SD - Mannitol	High	High	0.945
Ibuprofen 50 (20%) - multicomponent	High	High	0.835

One advantage of using ML models is the possibility of interpreting the way the models make the predictions. Here, SHAP methods were used to probe the RF models predicting bulk density. Fig 7-7 shows that the most important variable for the prediction of bulk density was the aspect ratio D10 and that high values of aspect ratio D10 led to high values of bulk density. These results are in agreement with previous studies; for example Yu et al. demonstrated the positive correlation between aspect ratio and bulk density (Yu et al., 1996). Moreover, these results can be linked back to the results obtained in the powder flow models in Chapter 4 where we observed that more spherical particles (high aspect ratio) are more likely to result in free-flowing powders. Here, we see that more rounded particles (higher aspect ratio) result in a smaller Hausner ratio and thus are also more free flowing, considering the correlation between bulk density and powder flow through the Hausner ratio equation (tapped density divided by bulk density).

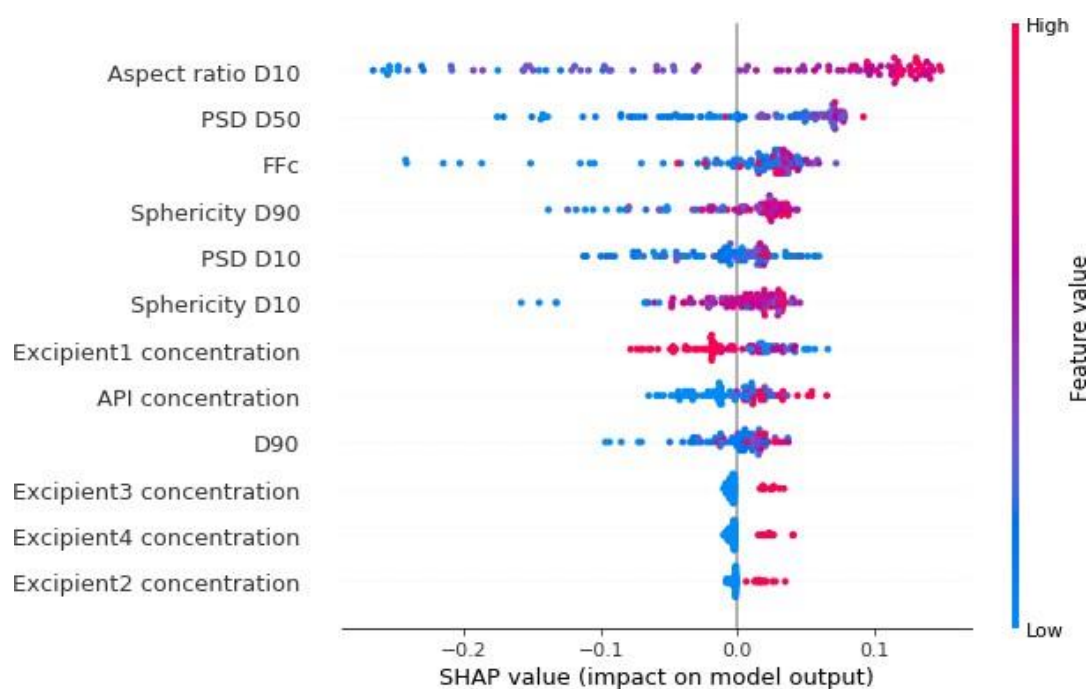


Figure 7-7: SHAP values plot (model output: class 1) for the RF model. Aspect ratio D10 and PSD D50 were the most important variables for the classification of bulk density. Moreover, high values of aspect ratio D10 and high values of PSD D50 led to higher values of bulk density.

Fig 7-8 shows the RF SHAP dependence plot of the aspect ratio D10. The SHAP dependence plot shows the impact of the aspect ratio D10 values (x-axis) on the bulk density (y-axis). The plot shows a positive trend between aspect ratio D10 and bulk density, indicating that a high the aspect ratio D10 leads to a high bulk density. The plot also shows that values of aspect ratio D10 smaller than ca. 0.45 have a

negative impact on the bulk density, making the prediction more likely to be smaller than 0.5 g/ml (low-bulk-density class). In contrast, values of aspect ratio D10 greater than 0.45 increase the probability of the classification of the powder into the high-bulk density class.

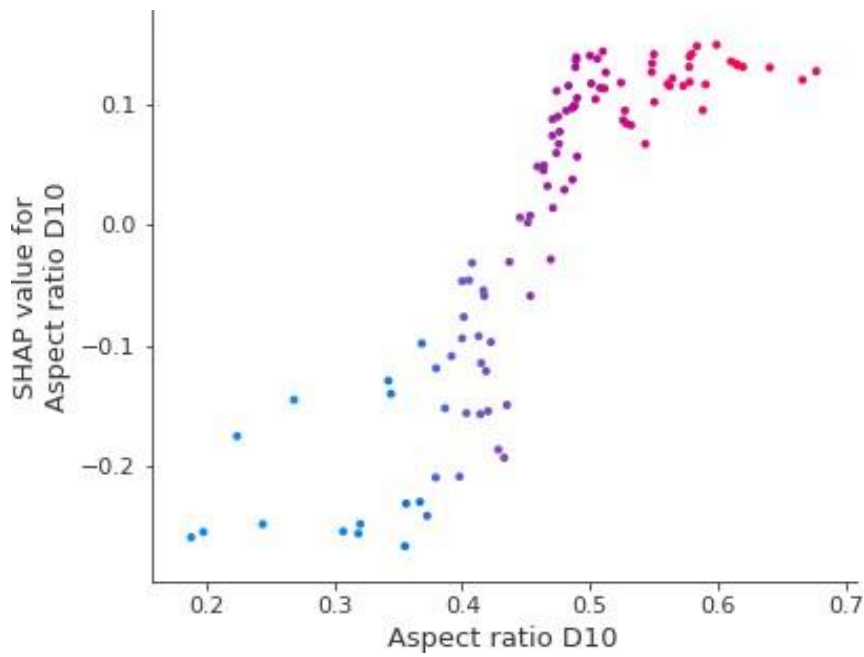


Figure 7-8: SHAP dependence plot for the RF model. The values of the aspect ratio D10 are represented on the x-axis, and the SHAP values of the aspect ratio D10 (“importance score”) are represented on the y-axis. The colour of the data points indicates the value of aspect ratio D10, with low values represented in blue and high values represented in pink.

The SHAP values plot indicated that FFc was the third most important variable for the classification of bulk density, but this plot did not show clear results of the impact of the FFc on the bulk density prediction. Thus, the RF SHAP dependence plot were calculated to investigate this impact (see Fig 7-9). The values of the FFc are plotted on the x-axis and the SHAP values (the impact of the FFc on the bulk density) are plotted on the y-axis. The plot shows that for values of FFc between 0 and 10, as the FFc increases, the bulk density increases too. Although these values of FFc had a negative impact on the classification of bulk density, increasing the likelihood of classification of the powder as low-density material (< 0.5 gm/ml). For values of FFc greater than 10, we observe that there is a lack of correlation and lack of impact of the FFc on the bulk density classification, since the data points seem to be scattered between the SHAP value 0 and 0.05.

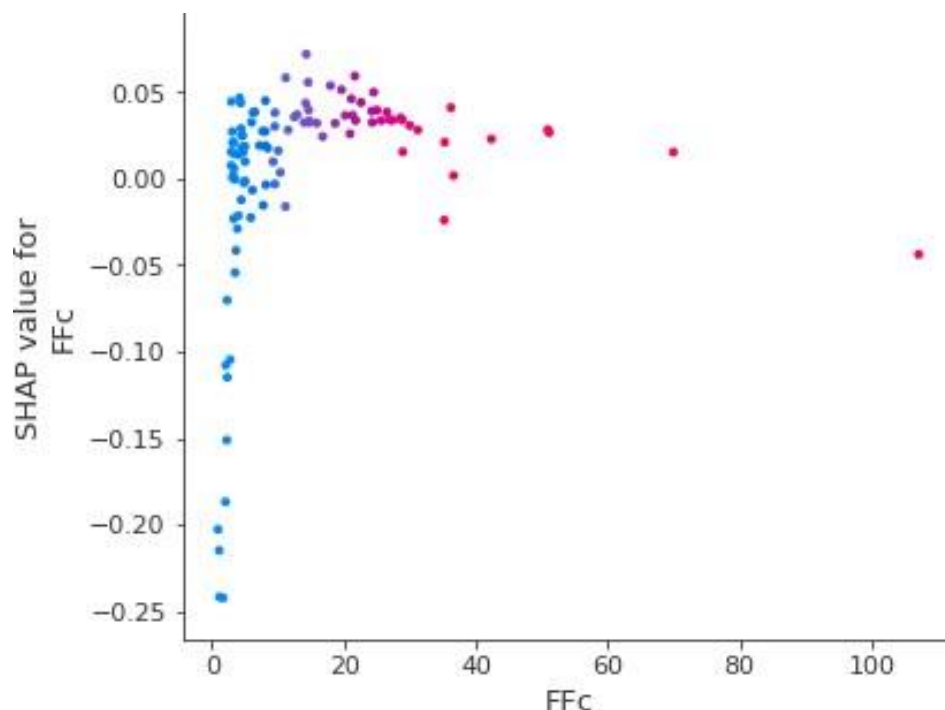


Figure 7-9: SHAP dependence plot for the RF model. The values of FFC are represented on the x-axis, and the SHAP values of the FFC (“importance score”) are represented on the y-axis. The colour of the data points depends on the value of FFC, with low values represented in blue and high values represented in pink.

Based on the results of the SHAP values feature importance analysis (Fig 7-7) and the analysis of the impact of the FFC on the bulk density prediction (Fig 7-9), we hypothesized that if the FFC was removed from the dataset, the performance of the RF model would not change noticeably. As mentioned in previous chapters, the measurement of the FFC requires more time and resources than the measurement of all the other properties included in the training dataset (particle size and shape). Therefore, not needing this property for the RF model will facilitate the classification of bulk density by reducing the data and experimental time required for model input. Hence, the models were retrained removing the FFC from the training dataset. As expected, the performance, reported in Fig 7-10, did not decrease significantly after the FFC was removed, and RF was again the best-performing algorithm. Thus, we see that the classification of bulk density by our RF model from only the particle size and shape of particles in pharmaceutical powders is possible, only requiring approximately 2 grams of materials and less than 5 minutes to predict bulk density, as opposed to at least 10 grams and 30 minutes per measurement.

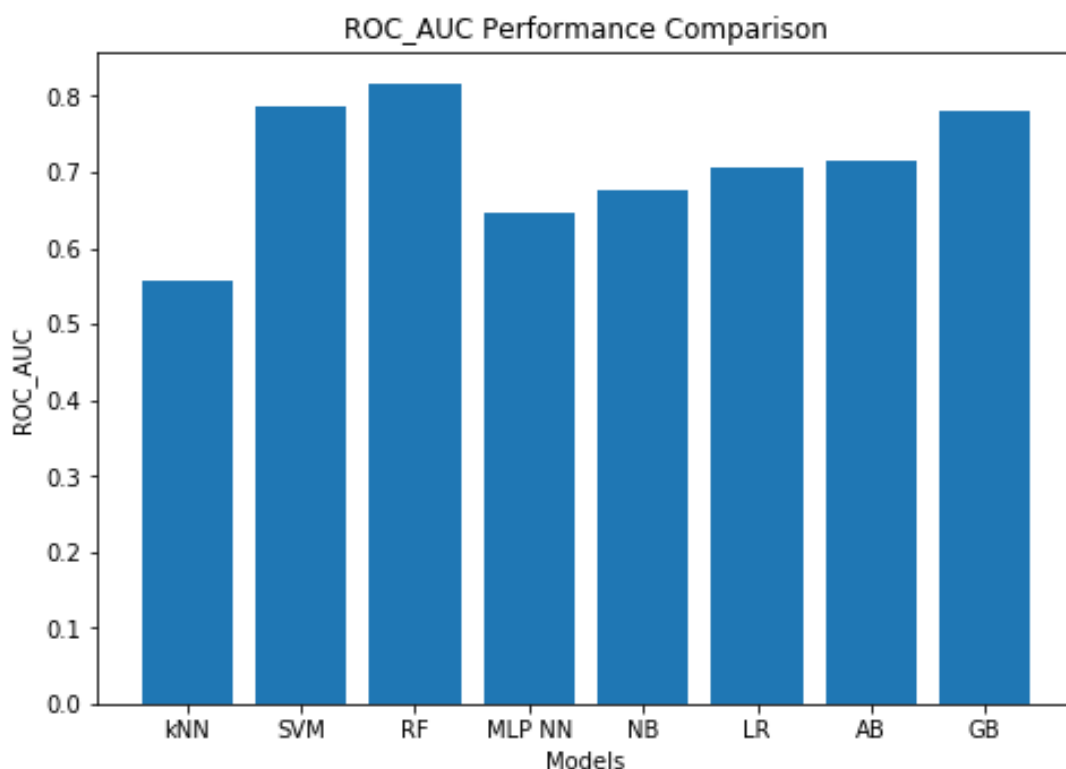


Figure 7-10: The comparison of the results of the classification models only including particle size, sphericity and aspect ratio, and the composition of the blends in the training dataset.

External validation of the RF model was carried out to ensure that the model was not overfitting. This external test achieved the same results to the external test obtained when the FFC was considered for training (see Fig 7-11). Therefore, these results confirmed that the FFC was not required for the prediction of bulk density. Table 7-7 shows the classification scores of the prediction for each of the materials included in the external validation. In this case, there was a difference between the misclassified and the correctly classified materials. For the two misclassified materials, the model was not as confident in the prediction as for the correctly classified materials.

		Predicted	
		Low bulk density	High bulk density
Actual	Low bulk density	1	1
	High bulk density	1	5

Figure 7-11: External validation results using the RF model. 2 of the powders included in the dataset had low density ($BD > 0.5 \text{ g/ml}$), and 6 powders had high density ($BD > 0.5 \text{ g/ml}$).

Table 7-7: Classification scores of predictions for each of the materials included in the external dataset.

Material	Actual class	Classification	Probability
Span60	1	1	0.98
Calcium carbonate	1	0	0.55
Ibuprofen 50	1	1	0.70
Avicel PH-101	0	0	0.83
Ibuprofen 20 % + Povidone	1	1	0.54
Plasdone K29/32	0	1	0.52
Pearlitol 100 SD - Mannitol	1	1	0.97
Ibuprofen 50 (20%) - multicomponent	1	1	0.81

7.3.1.4. Machine learning models: regression models for the prediction of bulk density

Regression models were also developed for the prediction of bulk density. Partial least squares (unsupervised learning), RF, GB, and AdaBoost (supervised learning) were trained and compared to find the best-performing algorithm. The performance of the algorithms was compared based on their MAE value (see Chapter 3, section 3.5.3.). The 2-component PLS had a MAE of 0.08 g/ml, RF had an error of 0.08 g/ml, GB had an error of 0.10 g/ml, and AB had an error of 0.09 g/ml (see Fig 7-12). The FFC was not included in the regression models, based on the results obtained by the classification models.

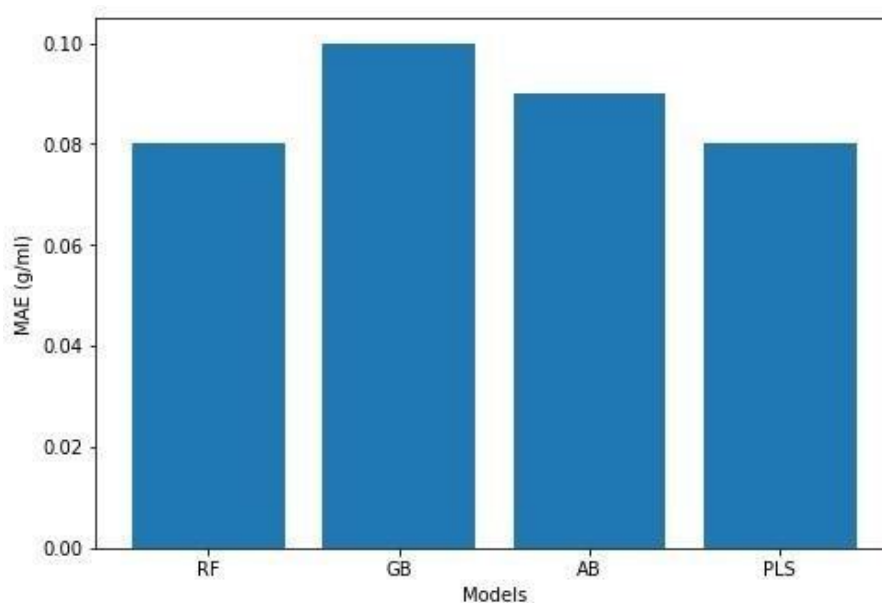


Figure 7-12: The comparison of the MAE achieved by supervised learning (RF, GB, AB), and unsupervised learning (PLS).

The RF SHAP values were used for feature importance analysis (see Fig 7-13). As for the RF classification model, aspect ratio D10 was the most important feature for the prediction of bulk density. The feature importance plot also revealed that high values of aspect ratio D10 led to higher values of bulk density. The second most important variable for the prediction of bulk density was the PSD D10. The correlation between PSD D10 and bulk density was further explored by calculating the dependence plots (see Fig 7-14). This plot showed that values of PSD D10 smaller than 100 μm had a negative impact (i.e., decreased the predicted value) on bulk density prediction. This dependence plot also showed that as the PSD D10 increased, the positive impact of this property on the bulk density increased too.

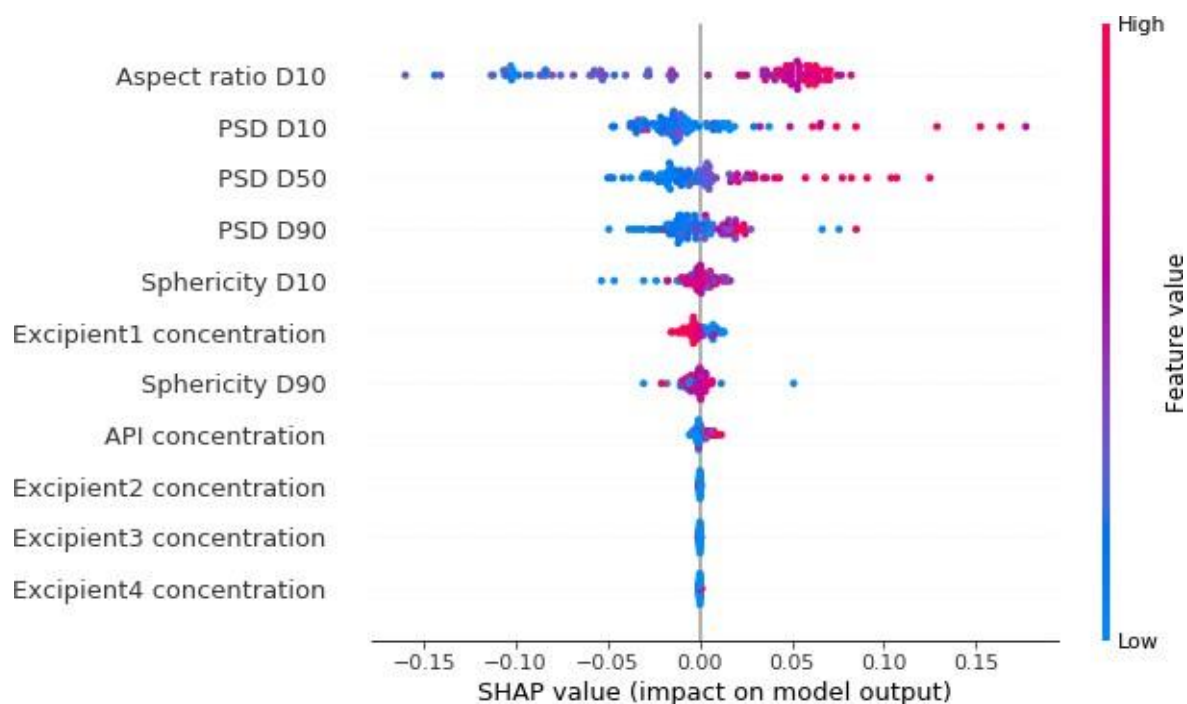


Figure 7-13: RF SHAP values plot. Aspect ratio D10 and PSD D10 were the most important variables for the prediction of bulk density. Moreover, high values of aspect ratio D10 and high values of PSD D10 led to higher values of bulk density.

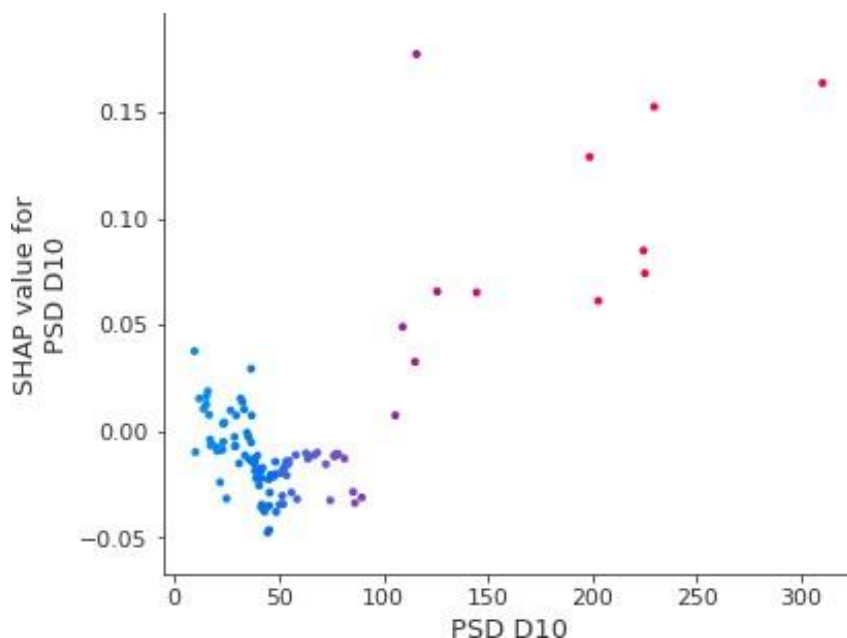


Figure 7-14: SHAP dependence plot for the RF model. The values of the PSD D10 are represented on the x-axis, and the SHAP values of the PSD (“importance score”) are represented on the y-axis. The

colour of the data points indicates the value of PSD D10, with low values represented in blue and high values represented in pink.

The same powders used to validate the classification models and to validate the powder flow models from Chapter 4 were used to validate the bulk density regression models. Table 7-8 shows that the error in the predictions for the validation set ranges from 0.03 to 0.15. This error is in the same range of the error obtained in the testing set and suggests that the model did not overfit the training set.

Table 7-8: External validation results of the regression model.

Material	Bulk density (g/ml)	Predicted bulk density (g/ml)	Absolute error
Span60	0.589	0.734	0.145
Calcium carbonate	0.563	0.455	0.108
Ibuprofen 50	0.531	0.567	0.037
Avicel PH-101	0.344	0.467	0.123
Ibuprofen 20 % + Povidone	0.528	0.540	0.011
Plasdone K29/32	0.453	0.552	0.098
Pearlitol 100 SD - Mannitol	0.585	0.635	0.051
Ibuprofen 50 (20%) - multicomponent	0.593	0.544	0.049

The predicted and actual values of the materials included in the external validation were plotted with a 95% confidence interval (see Fig 7-15). The predicted bulk density values for Span 60 and calcium carbonate were outside of this 95% confidence limit and were chosen for further investigation. The actual value of bulk density of Span 60 was 0.58 g/ml, and the predicted value was 0.73 g/ml (MAE = 0.415). The main drivers of this overprediction, as shown in Fig 7-16, were PSD D50, PSD D10, and PSD D90, since they increased the predicted value considerably. Span 60 had a larger particle size than most of the other powders considered for training, but it neither had the high bulk density nor the high flowability ($FFc = 1.9$, which would be considered cohesive following the adaptation of Jenike's classification described in Chapter 4) expected for powders with large particle sizes. Span 60 is a sorbitan mono-ester that is used as a non-ionic detergent. Possible reasons for the higher predicted value of bulk density for Span 60 include the presence of electrostatic interactions or possible misrepresentation of particle size in that the particle size recorded by the instruments may have been that of agglomerates not individual particles.

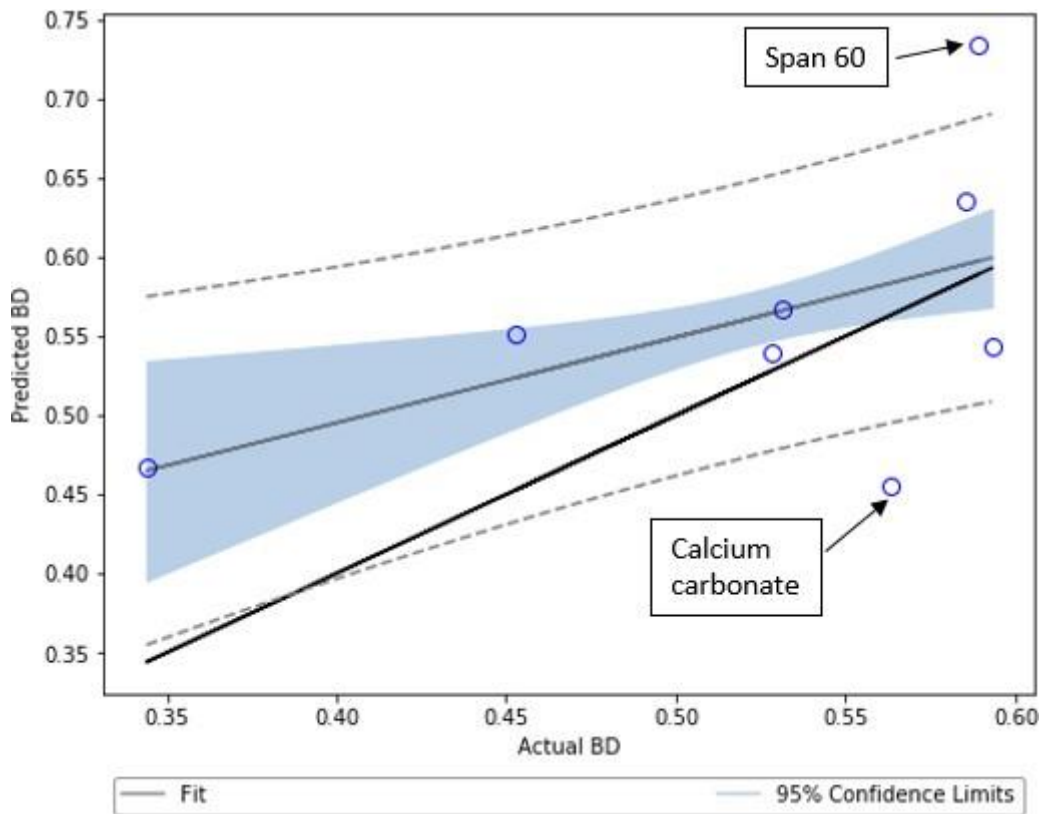


Figure 7-15: Scatter plot of the actual values of bulk density against the predicted values by the gradient boosting model.

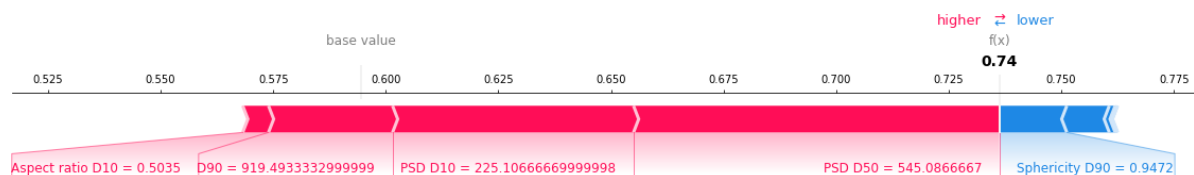


Figure 7-16: Force plot of the Span 60 bulk density prediction for the RF model. As seen in this plot, the main driver of the prediction is the PSD, which takes the prediction from the base value to the predicted value (0.74).

The RF model bulk density prediction for calcium carbonate was also outside the 95% confidence limit, but unlike the bulk density prediction for Span60, the bulk density prediction for calcium carbonate was underpredicted. The actual value of the bulk density of calcium carbonate was 0.56 g/ml while the predicted value was 0.41 g/ml. As can be seen by the SHAP analysis shown in Fig 7-17, the main

driver for the underprediction of the bulk density was the PSD D90. The value of PSD D90 of calcium carbonate was significantly smaller than the values of PSD D90 of the other powders included in the dataset. As shown in Fig 7-13, high values of PSD D90 lead to high bulk density, which explains why a low value of PSD D90 decreased the predicted bulk density value.

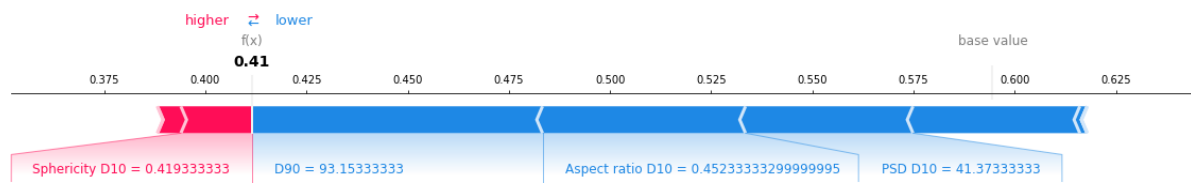


Figure 7-17: Force plot of Calcium carbonate. The main drivers for the underprediction were the particle size distribution (D90), the aspect ratio D10, and the particle size distribution D10.

Finally, for comparison to the previous examples, the prediction of the bulk density of the blend of ibuprofen and Plasdone povidone at 20% drug loading was also analysed (see Fig 7-18). This blend had a lower prediction error than the two powders analysed above. The actual bulk density of the blend was 0.52 g/ml, and the model predicted 0.5 g/ml.

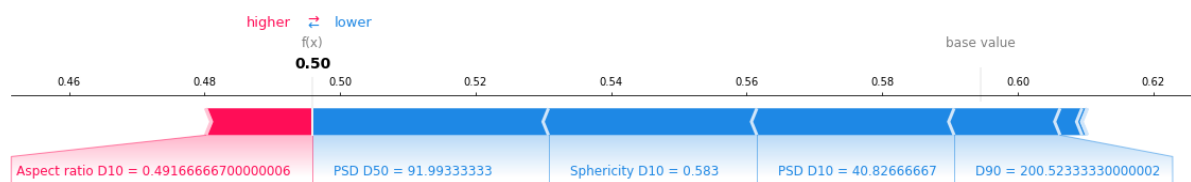


Figure 7-18: Force plot of ibuprofen (20%) and Plasdone povidone blend. The main drivers for the prediction were the particle size distribution D50, the sphericity D10, and the PSD D10.

The SHAP analysis of the RF regression models suggests that while the models perform well (as seen by external validation), the bulk density of two powders was not predicted within the 95% confidence interval due to their particle size and shape descriptors. The addition to the dataset of more powders that have a wider combination of particle and bulk properties could improve the performance of these bulk density models.

7.3.2. Surface area measurements and associated ML models

7.3.2.1. Surface area experimental measurements

To establish the dataset of surface area for training ML models, a total of 31 pharmaceutical powders were analysed with the SEA. The distribution of the measurements of the specific surface area (surface area per unit of mass or volume) for all the powders analysed is shown in Fig 7-19. Most of the materials exhibited a surface area smaller than 1 m²/g.

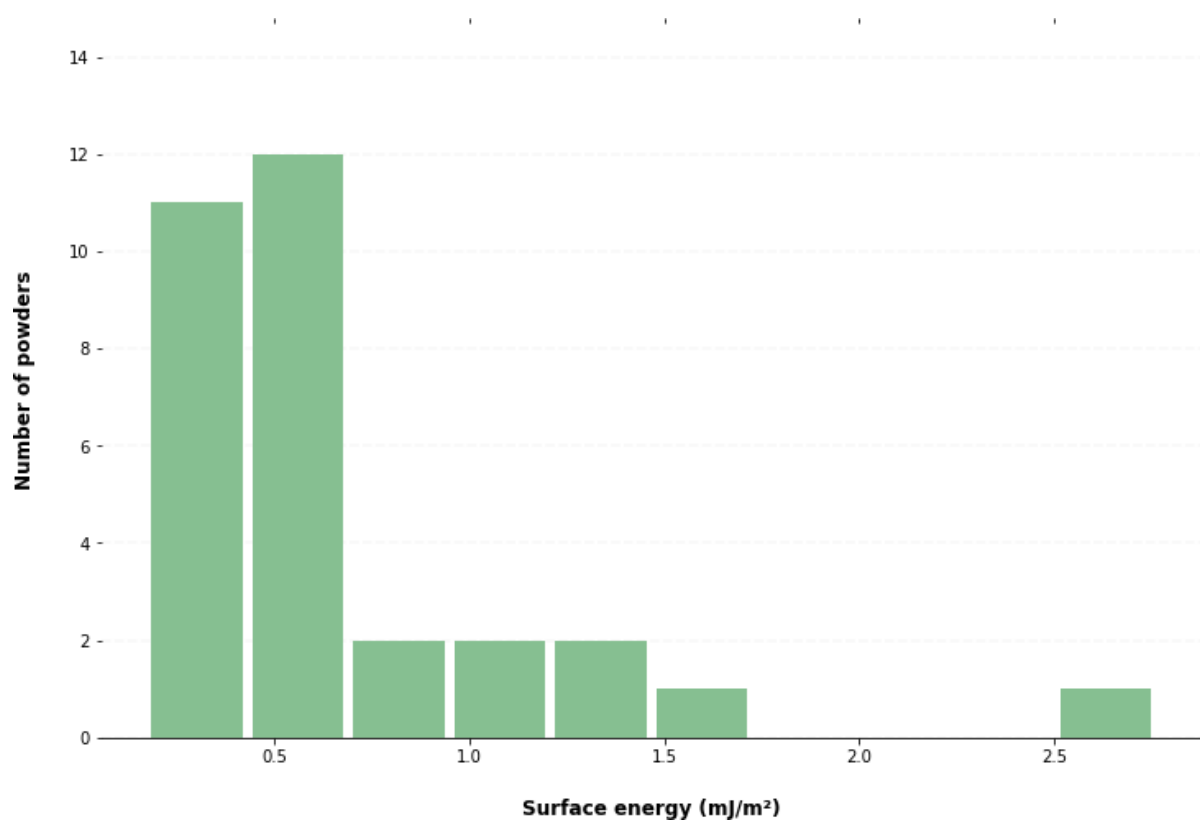


Figure 7-19: The distribution of the specific surface area (surface area per unit of mass or volume) of the powders included in the training dataset.

For the classification models, the pharmaceutical powders were divided into high-surface-area powders (a surface area greater than 0.65 m²/g), and low surface area powders (surface area lower than 0.65 m²/g). The mean of the surface area values of the training dataset was taken as the threshold to divide the classes. Based on this classification, 23 powders were classified as low-surface-area powders (class 0), and 8 were classified as high-surface-area powders (class 1) (see Table 7-9). The

criteria followed to divide the classes resulted in class imbalance that can jeopardize the performance of the model. The variables included for this study are presented in Table 7-10.

Table 7-9: Number of powders in each class of surface area.

Classes	Values	Number of powders
Class 0	Surface area < 0.65 m ² /g	23
Class 1	Surface area > 0.65 m ² /g	8

Table 7-10: The variables included in the surface area models.

Variables included in the surface area models		
Concentration of API	SMD	FFc
Concentration of excipient 1	Sphericity D10	Surface Area
Concentration of excipient 2	Sphericity D50	Specific SE
Concentration of excipient 3	Sphericity D90	SE com
Concentration of excipient 4	Aspect ratio D10	DSEat0%
PSD D10	Aspect ratio D50	DSEat3%
PSD D50	Aspect ratio D90	DSEat5%
PSD D90	BD (g/ml)	DSEat10%

7.3.2.2. Data analysis

The first statistical analysis performed was PCC. The correlation between variables is presented in a heatmap (see Fig 7-20). Variables that have a high positive correlation are represented in light orange, whereas variables that have a high negative correlation are represented in dark red. Variables that belong to the same descriptor (sphericity or aspect ratio) were highly correlated, but none of these were correlated with surface area. The heatmap also shows surface energy descriptors (specific SE, SE(com), dispersive surface energy (DSE) at 0%, 3%, 5%, and 10% coverage). The surface energy descriptors were not included in the training dataset for the prediction of surface area. As was done

for the bulk density model, the highly correlated variables were removed from the dataset before training the model (Darst et al., 2018). One variable of each pair of highly correlated variables was removed randomly before training the surface area models. The remaining variables are reported in Table 7-11.

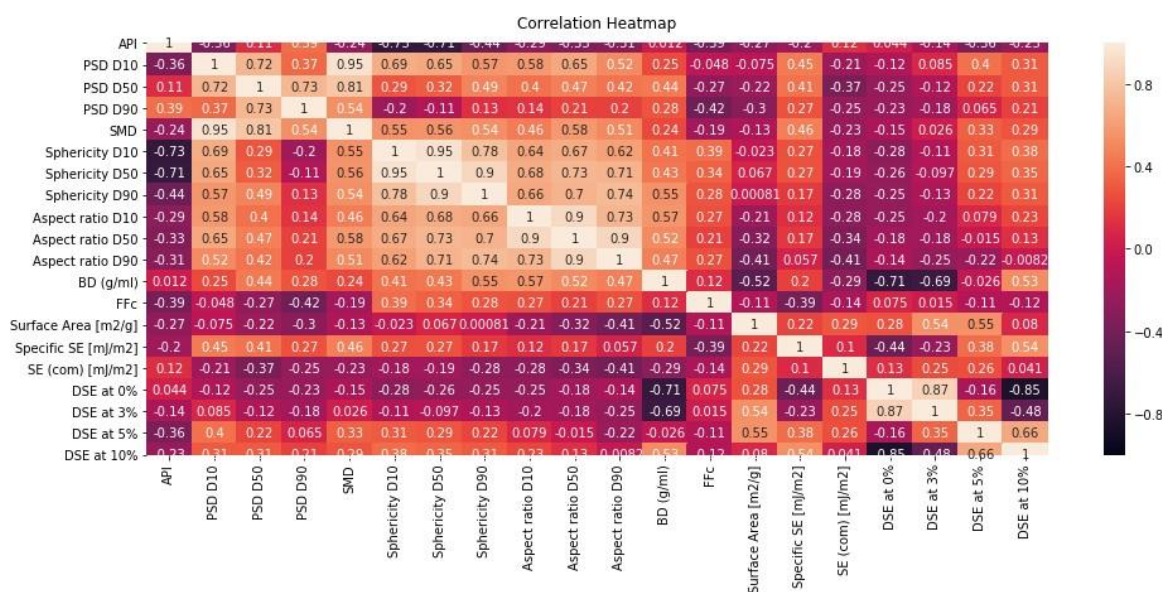


Figure 7-20: The heatmap of the PCC of the variables used to train surface area models.

Table 7-11: The variables considered to train the surface area model after removing highly correlated variables.

Variables included in the surface area models	
API	PSD D90
Excipient1	Sphericity D10
Excipient2	Sphericity D90
Excipient3	Aspect ratio D10
Excipient4	Aspect ratio D90
PSD D10	Ffc
PSD D50	BD (g/ml)

A 2-component PCA was then performed to understand different clusters within the data. This analysis explained the 60% of the cumulative variance. The red data points represent the low-surface area powders (surface area smaller $0.65 \text{ m}^2/\text{g}$) and the green data points represent the high-surface area powders (surface area greater than $0.65 \text{ m}^2/\text{g}$), as shown in Fig 7-21. The red and green data points show significant overlap and therefore PCA was not useful for clustering the powders based on their surface area.

Interestingly, a group of 5 low-surface area powders was separated from the other powders. These powders were multicomponent blends, instead of single component powders. The blends have a variable drug loading (between 5 and 40%) and fixed concentration of a mixture of Avicel PH-102, Croscarmellose sodium, and magnesium stearate, and a variable FastFlo 316 concentration. These results showed that while PCA was not useful to classify the powders based on their surface area, this analysis was able to differentiate the individual materials from the blends.

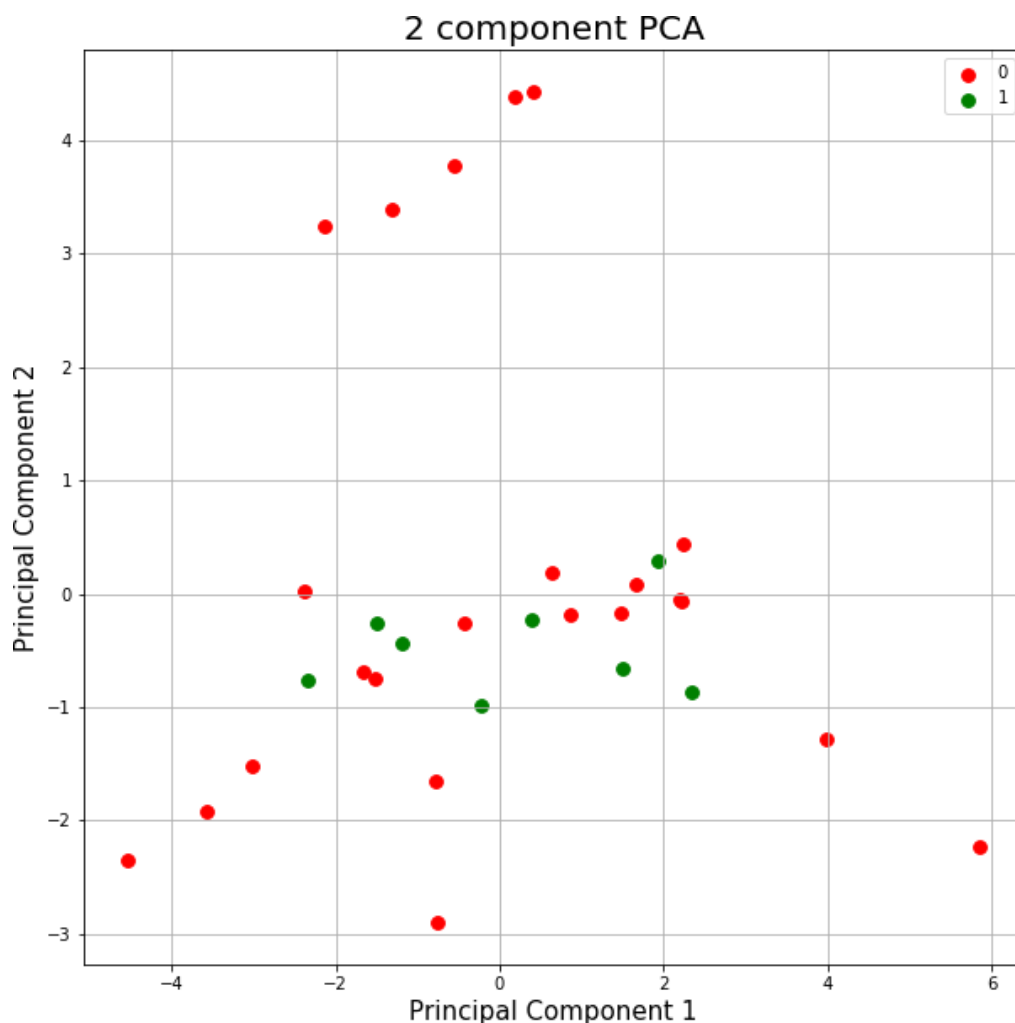


Figure 7-21: PCA of surface area. Low surface area powders (class 0) are represented in red and high surface area powders (class 1) are represented in green.

7.3.2.3. Machine learning classification models for the prediction of surface area

The same algorithms used in section 3.1. were trained and compare for the classification of surface area. RF and NB outperformed the rest of the algorithms, achieving an AUC of 0.82 and 0.84, respectively (see Fig 7-22). SHAP methods are only available for certain algorithms, and NB is currently not among them. Hence, RF was chosen to perform the interpretability analysis.

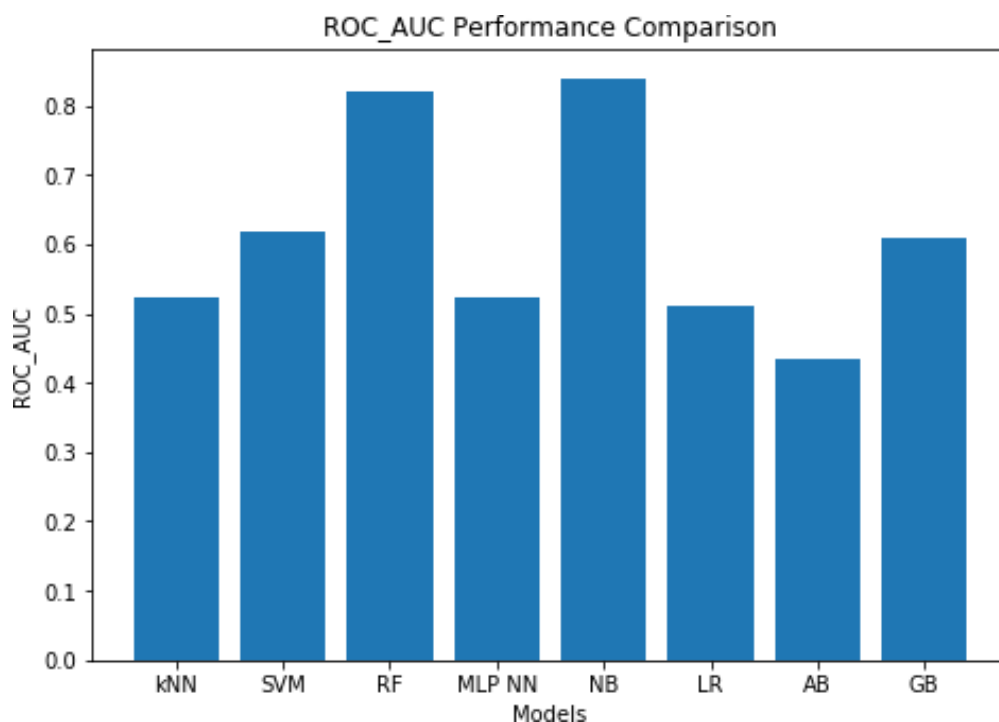


Figure 7-22: Comparison of the performance of the classification models for the prediction of the surface area, using a binary classification system.

Figs 7-23 and 7-24 present the SHAP analysis for the RF classification models for surface area. Fig 7-23(a) shows that bulk density was the most important variable for the RF model for the classification of surface area of pharmaceutical materials, and Fig 7-23(b) shows the negative correlation between bulk density and the surface area in the experimental data. Figs 7-24(a) and 7-24(b) show that as surface area decreases, bulk density increases. The RF model dependence plot (Fig 7-24(b)) shows that as the value of bulk density increases, the surface area decreases. Particularly, values of bulk density smaller than 0.5 g/ml have a negative impact on the surface area. The negative values of SHAP indicate that higher values of bulk density reduce the predicted value of surface area, increasing the likelihood of the powder to be classified as a low surface area (smaller than 0.65 m²/g).

Interrogation of particle and bulk property descriptors in the context of machine learning and prediction of pharmaceutical materials.

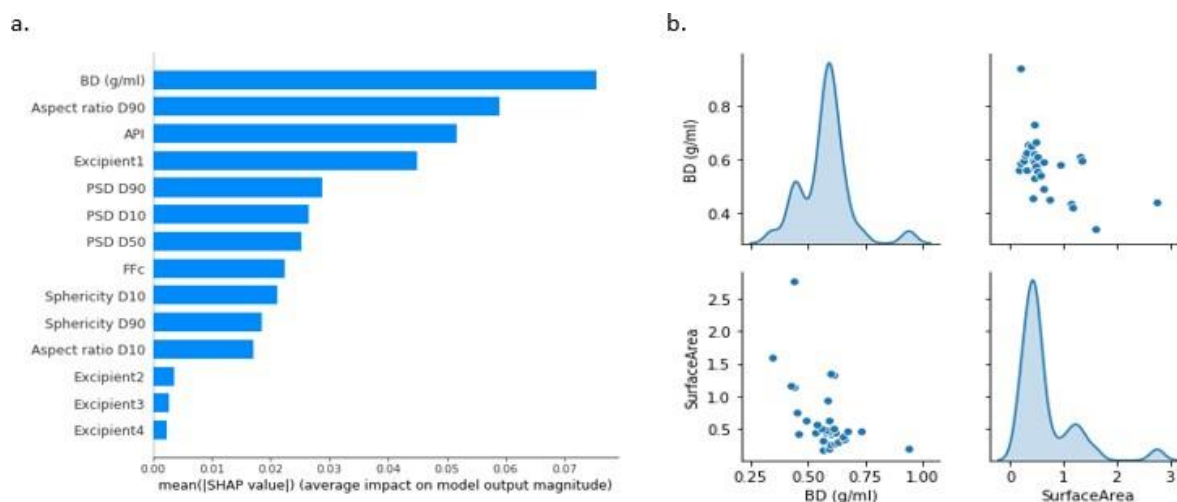


Figure 7-23: a) RF model SHAP summary plot of the ranking of the most important variables for the classification of surface area; b) Pairplot of surface area and bulk density to study the correlation between the actual data of the two variables.

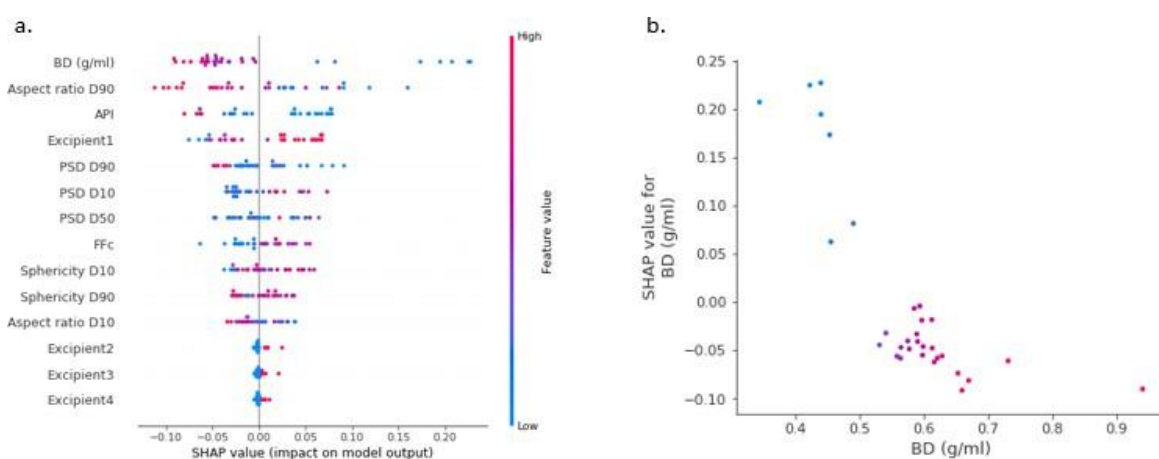


Figure 7-24: a) RF model SHAP values summary plot that shows the direction in which each variable contributes to the prediction. The further to the right the data points appear, the more their contribution to a higher surface area; b) RF model SHAP dependence plot. Bulk density was plotted on the x-axis and the SHAP values for bulk density (the contribution of the variable to classification of surface area) were plotted on the y-axis.

The results of the correlation between bulk density and surface area tie in with previous studies in literature. Angelos *et al.* (2009) reported porous materials to be useful to increase drug bioavailability (Angelos, Liong, Choi, & Zink, 2008; Wang, 2009). As mentioned in the introduction section, porosity

is the proportion of the volume of the sample that does not contain solid, and thus, porosity is inversely correlated with bulk density. The bioavailability of a drug depends on its dissolution rate and therefore on its ability to be absorbed in the intended biological destinations. Following Noyes-Whitney equation, the dissolution rate of a drug depends on surface area, and, therefore, the results obtained by the machine learning model agree with what would be expected: a decrease in bulk density (increase in porosity) will result in an increase of surface area available for dissolution of the solid and hence, an increase in the dissolution rate of the drug and potentially its bioavailability. This study links the prediction of surface area from particle size and shape to the biopharmaceutical classification system (BCS).

Moreover, the results obtained from the surface area classification models can be linked to the results obtained in Chapter 4 (section 4.3.4.). The results in Chapter 4 showed that to ensure suitability for direct compression based on the FFC, the PSD D90 of the powder should be between 300 and 700 μm . In contrast, the results obtained in the surface-area model showed that PSD D90 greater than 200 μm would have a detrimental impact on the surface area, and therefore on the drug bioavailability (see Fig 7-25). Hence, the guidelines that were proposed in Chapter 4 to ensure suitability for direct compression based on powder flow did not include a consideration regarding the bioavailability of the API. Therefore, these guidelines should be further expanded into unified guidelines that not only ensure manufacturability but also drug bioavailability, based on the targeting desired powder flow and surface properties of new pharmaceutical materials.

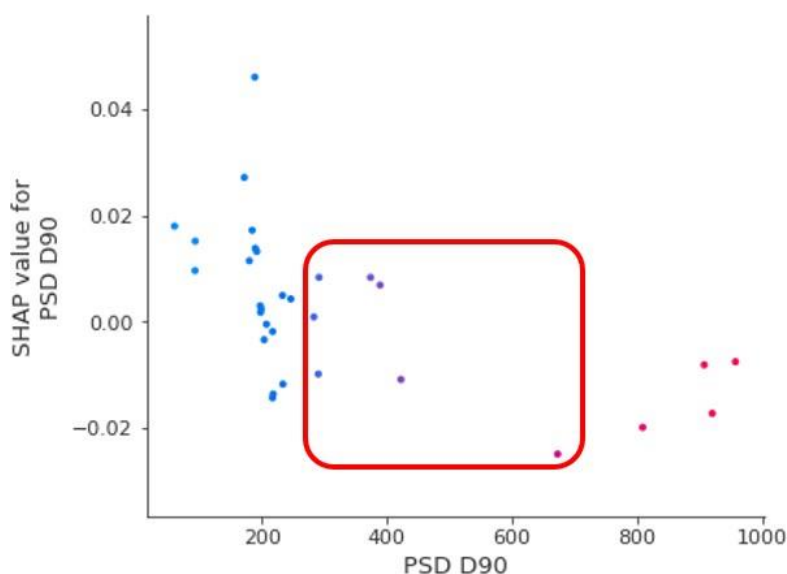


Figure 7-25: RF model SHAP dependence plot shows the impact of PSD D90 on the surface area class. As the PSD D90 increases, the surface area decreased, and therefore, a PSD D90 greater than 300 μm would have a negative impact on the surface area class. The red rectangle shows the target PSD D90 to ensure suitability for direct compression based on achieving desired powder flow.

While more data are needed to improve the model performance, this approach explores the feasibility of defining guidelines of “ideal” particle properties to design new pharmaceutical powders that are suitable for direct compression while ensuring drug bioavailability.

7.3.3. Surface energy measurement and associated ML models

7.3.3.1. Surface energy experimental measurements

To establish the dataset of surface energy for training ML models a total of 31 pharmaceutical powders analysed with the SEA. The distribution of the surface energy across the dataset is shown in Fig 7-26. The surface energy ranged between 3.15 and 16.81 mJ/m^2 , but most of the materials had a value between 4 and 8 mJ/m^2 .

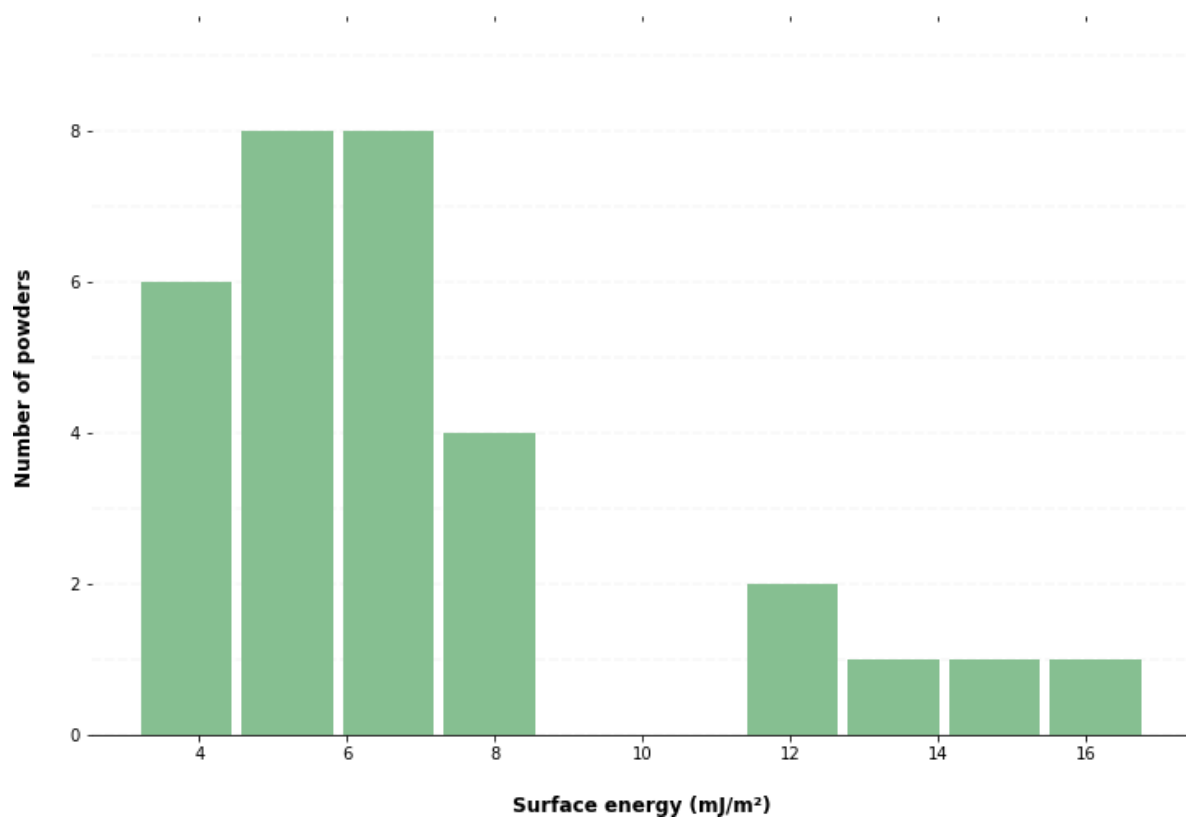


Figure 7-26: The distribution of surface energy values across the powders included in the training dataset. Most of the materials had a surface energy between 4 and 8 mJ/m², and the mean value was 6.97 mJ/m².

For the classification of surface energy, the powders included in the training data were divided into two categories. Powders that had specific surface energy less than 6.97 mJ/m² were considered low surface energy powders (class 0), and powders that had a surface energy greater than 6.97 mJ/m² were considered high surface energy powders (class 1), as shown in Table 7-12. The mean of the surface energy values of the training dataset was taken as the threshold to divide the classes. The variables included in the surface energy models were the same as the variable included in the surface area models (see Table 7-11, section 7.3.2.1.).

Table 7-12: number of powders that belong to each class in the training dataset.

Class	Range of values	Number of powders
0	Surface energy < 6.97 mJ/m ²	21
1	Surface energy > 6.97 mJ/m ²	10

7.3.3.2. Data analytics

As in section 7.3.2.2., a 2-component PCA was performed for clustering purposes. The colour of the data points correspond to their surface energy values. Thus, low surface energy powders are plotted in red (class 0), and high surface energy powders are plotted in green (class 1) (see Fig 7-27). Equally to the surface area PCA results, the red and green data points overlapped and therefore, PCA was not useful to cluster the data based on surface-energy classes. The same five low-surface-energy formulations were found in a different cluster than the rest of the materials included in the training dataset (see Table 7-13). These results showed that while PCA was not useful to classify the powders based on their surface energy, but this analysis was able to differentiate the individual materials from the blends.

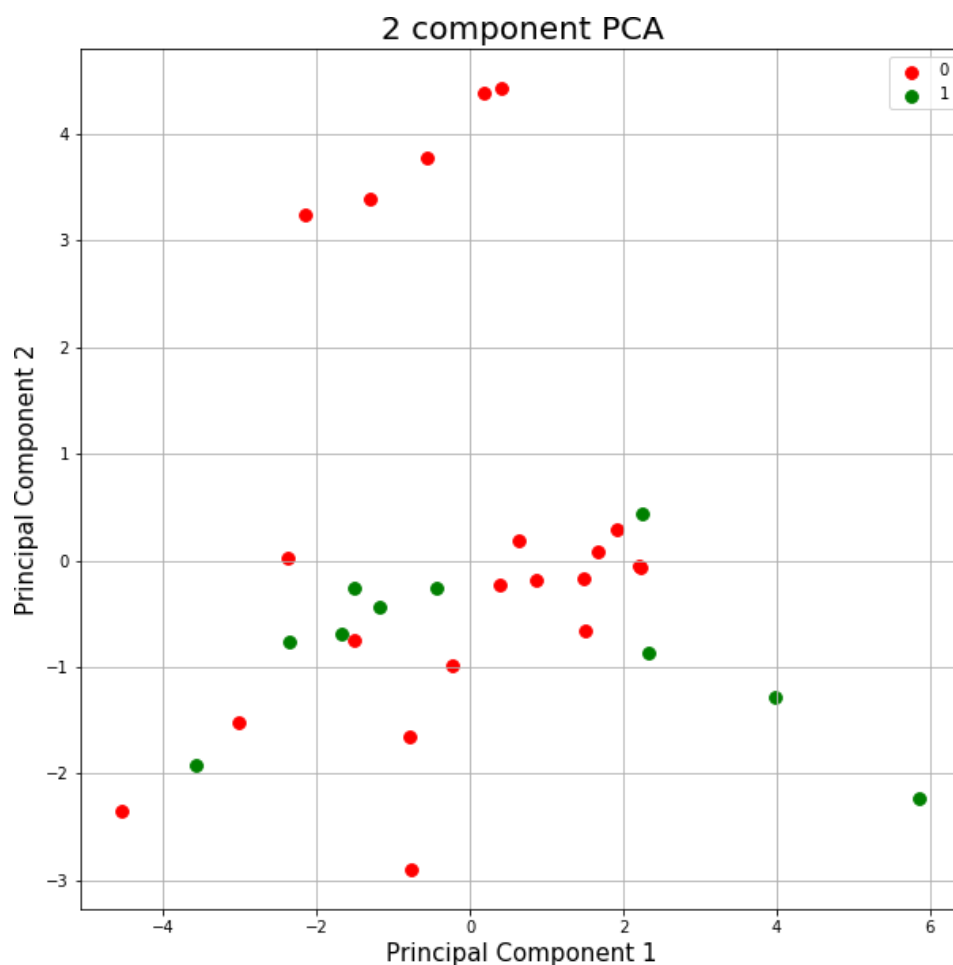


Figure 7-27: PCA for the classification of surface energy.

Table 7-13: The five high-surface energy powders that were clustered in a different group using PCA. These five powders were pharmaceutical blends, as opposed to the rest of the powders grouped in the main cluster that were individual materials.

Name	Surface energy (mJ/m ²)
Paracetamol Powder (5%) - MC	3.67
Paracetamol Powder (20%) - MC	6.15
Paracetamol Powder (40%) - MC	6.10
Ibuprofen 50 (5%) - multicomponent	6.18
Ibuprofen 50 (20%) - multicomponent	4.83

7.3.3.3. Machine learning classification models for the prediction of surface energy

Classification algorithms were trained to classify powders based on their surface energy. SVM (AUC-ROC = 0.63), RF (AUC-ROC = 0.64), and LR (AUC-ROC = 0.62) achieved the highest performance among the learners (shown in Fig 7-28). These results are poorer than the results achieved in the previous models (bulk-density and surface-area models). Therefore, the surface-energy classification models were used to improve the understanding of the relationship between the surface energy and the other properties included in the training dataset (see Table 7-10), rather than aiming to classify new powders based on their surface energy. This improvement of the understanding of the correlation between variables was achieved by the study of the SAHP values for the RF classification model.

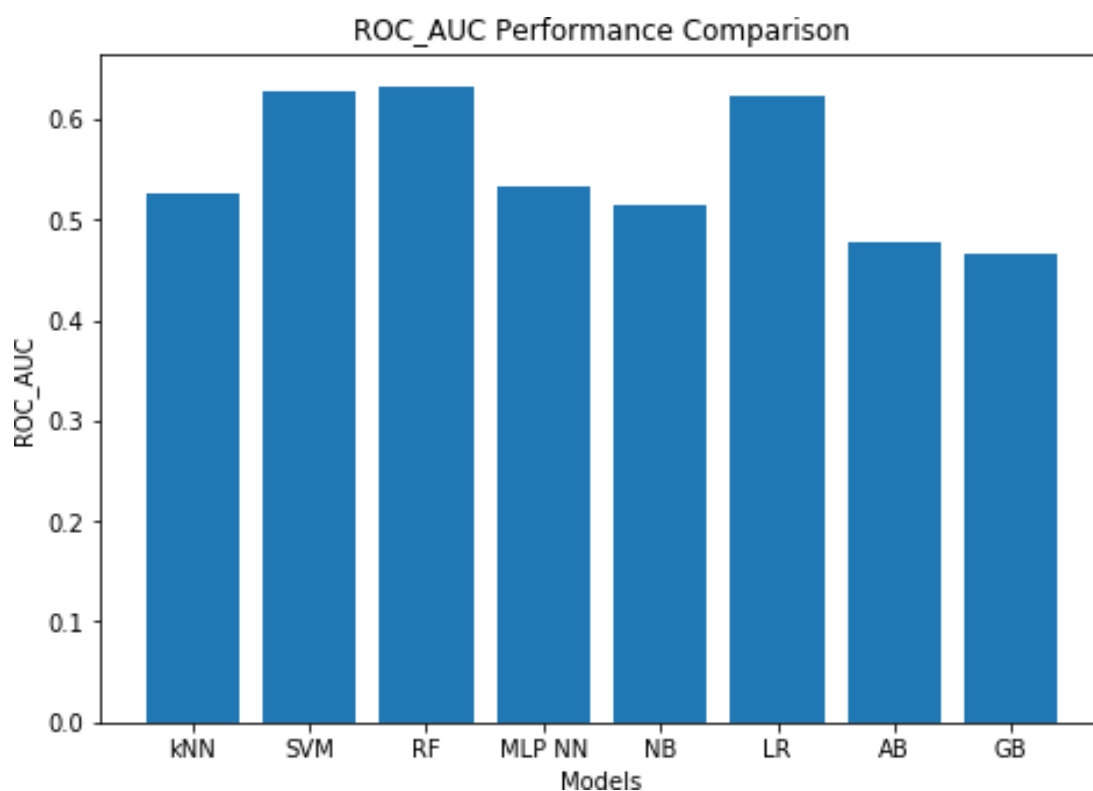


Figure 7-28: The comparison of the results of the performance of the classification algorithms trained for the prediction of surface energy.

The SHAP feature importance analysis was calculated for the RF surface-energy model. Fig 7-29 shows that sphericity D90 was the most important variable for the classification of surface energy, followed

by the particle size distribution D50. Fig 7-29(b) shows that high values of sphericity D90, represented in pink, led to low values of surface energy (left side of the SHAP plot).

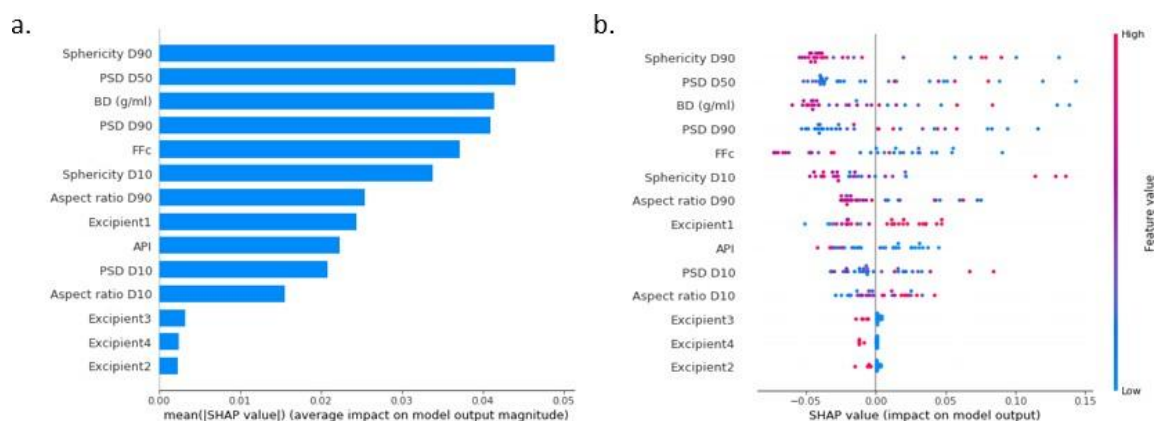


Figure 7-29: a) Feature importance analysis using the RF SHAP methods for model interpretability. The SHAP values are assigned to the variables depending on their impact on the model target (surface energy); b) RF model SHAP summary plot of the direction in which each variable contributes towards the prediction of surface energy. High values of the variables are represented in pink and low values in blue.

To explore the correlation between sphericity D90 and the surface energy, the RF SHAP dependence plots were analysed. As the sphericity D90 increases, the surface energy decreases (see Fig 7-30(a)). Values of sphericity D90 greater than 0.75 had negative impact on surface energy, decreasing the predicted value. The SHAP dependence plot showed that three powders that had a sphericity D90 greater than 0.9 had a positive impact on surface energy, increasing the predicted value of surface energy. This trend could be further explored by adding more data to see whether very high values of sphericity D90 lead to high surface energy, or these three powders were outliers.

The impact of particle size on surface energy was also explored using SHAP dependence plots and compared with previous publications. Ali et al. demonstrated the dependence of surface energy to particle size in nanoparticles of gold. Their calculations showed that the surface energy decreases with increasing the particle size (Ali, Myasnichenko, & Neyts, 2016). The results of the SHAP dependence plot for the RF classification model showed that the impact of particle size on surface energy depends on the range of particle size (see 7-Fig 30(b)):

- (i) PSD D50 greater than 100 μm had a positive impact on surface energy.
- (ii) PSD D50 between 100 and 200 μm had a negative impact on the surface energy.
- (iii) PSD D50 greater than 200 μm did not have an impact on the predicted values of surface energy.

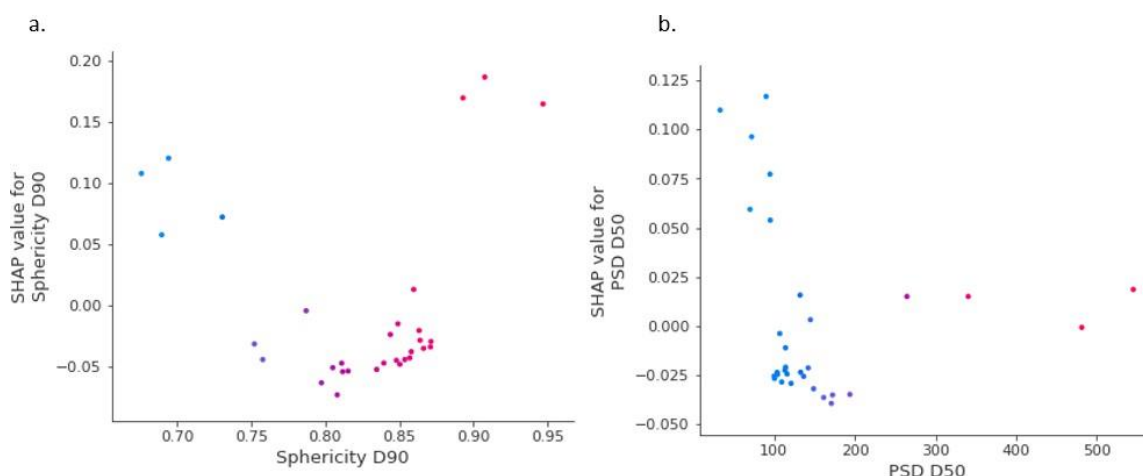


Figure 7-30: a) RF model SHAP dependence plot between sphericity D90 and its impact on the surface energy; b) RF SHAP dependence plot of PSD D50.

A final route was explored to study the impact of surface area on the surface energy. So far, the prediction of surface energy has been attempted by considering particle size and shape, but in this case, the surface area data were also included as an input of the RF model. As described in the methods section, both surface area and surface energy are measured using the same instrument (iGC). The first step of the measurement is the analysis of the surface area of the samples. This measurement takes approximately 8 hours. Once completed, the user must initialise the measurement of the surface energy, which takes approximately 24 hours. The addition of the surface area data for the classification of surface energy would reduce the time of the iGC measurement from two days to eight hours. Moreover, the iGC is a very sensitive instrument likely to shut down if there is any change in the environment, such as a change in pressure. By reducing the time required to run the measurement, the risks of measurement failure are also reduced.

Supervised ML algorithms were trained for the classification of surface energy from particle size, particle shape and surface area. The performance of the ML algorithms is shown in Fig 7-31. Overall, the performance improved compared to the results of the previous surface energy models when the surface area was not included. SVM, RF, and LR achieved a performance of over 0.7.

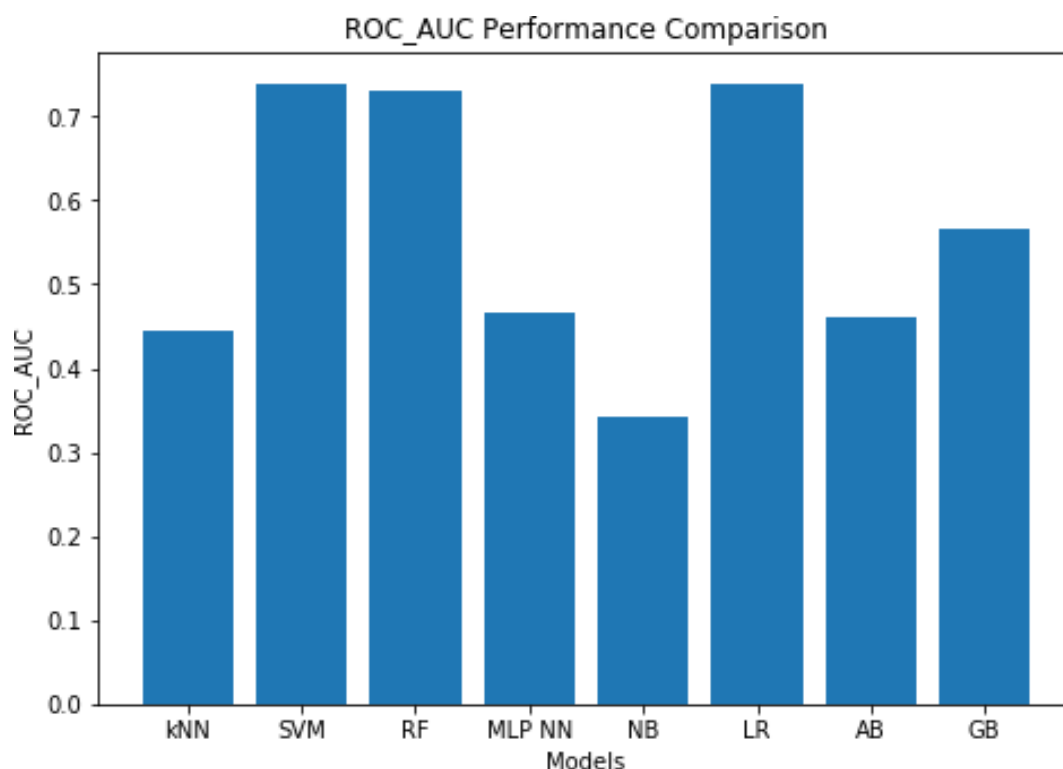


Figure 7-31: The comparison of the performance of classification learning algorithms for the prediction of surface energy from particle size, particle shape and surface area.

Since the addition of the surface area data improved the performance of the RF surface-energy classification model, the impact of the surface area on the surface energy was investigated using RF SHAP dependence plots (Fig 7-32). The plot shows that for values of surface area smaller than $0.6 \text{ m}^2/\text{g}$, as the surface area increases, the surface energy decreases. When the surface area is greater than $0.6 \text{ m}^2/\text{g}$, the increase of surface area has a positive impact on surface energy. Furthermore, values of surface area between 0.25 and $1.2 \text{ m}^2/\text{g}$ had a negative impact on the predicted value of surface energy.

While more data are still needed to improve the model performance, this approach explores the feasibility of the classification of surface energy from physical particle properties and surface area data, decreasing the total time needed for the measurement of surface energy.

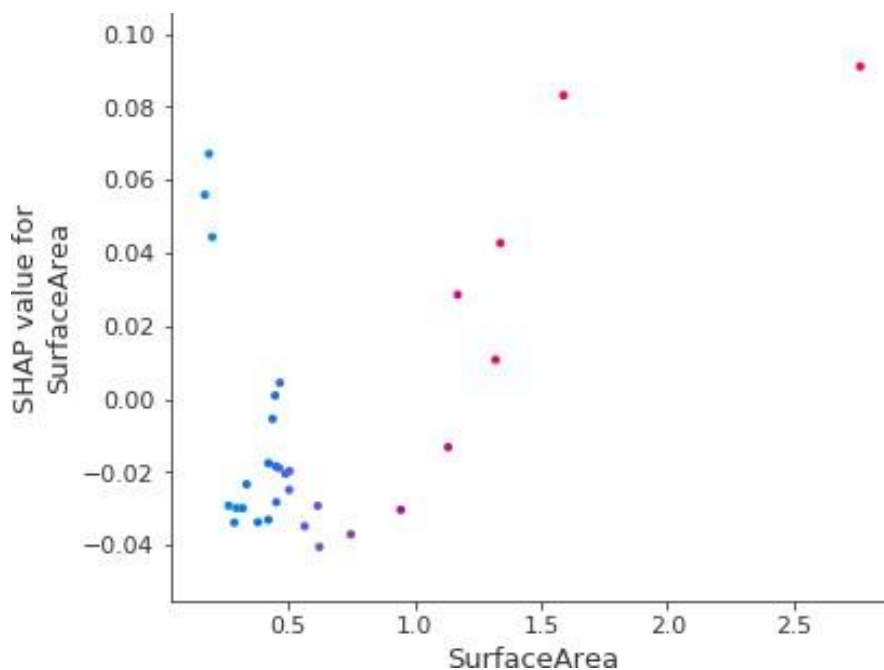


Figure 7-32: RF model SHAP dependence plot of surface area values, plotted on the x-axis, and its SHAP values, plotted on the y-axis, indicating the impact of surface area on surface energy.

7.4. Conclusions

In the work presented in this chapter, bulk density, surface area and surface energy properties were predicted from particle size and shape descriptors. A total of 112 materials with a wide range of particle properties were included in the training dataset for the classification and prediction of bulk density. Due to time constraints associated with measuring surface area and energy, only 31 materials of the 112 materials were included in the surface area and surface energy models.


The studies show that particle size and shape distribution measured with dynamic image analysis are sufficient to enable the prediction of key bulk properties relevant to pharmaceutical processing. The best performing model for the classification of bulk density was RF, with over 80% probability of distinguishing between classes. Regarding the bulk density regression models, both RF and GB achieved the lowest MAE (0.9 g/ml). To reduce class imbalance, the model performance would likely be improved by the addition of data for materials with a bulk density lower than 0.5 g/ml. The applications of these models could theoretically be extended by adjusting the threshold at which the classes were divided (0.5 g/ml) to suit other processes or pieces of equipment.

The work in this chapter shows that surface area could be also predicted from particle size and shape descriptors, measured with dynamic image analysis. The best performing model for the classification of bulk density was RF with over 80% probability of distinguishing between classes. It would be interesting to investigate the relationship between surface area and dissolution rate to establish applicable threshold between the classes. Finally, the results achieved in the surface energy showed that the correlation between surface area and surface energy enabled the possibility of eliminating the additional time required for the characterisation of surface energy.

The further development of the models presented in this chapter would help link the manufacturability requirements to the bioavailability requirements, working towards unified guidelines for product and process design. The implementation for the bulk density model could save time and resources when developing a new API as this model allows early-stage prediction of a powder's suitability for direct compression from particle shape and size distribution alone, following the MCS guidelines. Similarly tying the surface area prediction model presented here to the biopharmaceutical classification system would help not only assess but optimise the bioavailability of a new API in early-stage process development by reverse engineering the optimum properties that ensure both manufacturability and drug availability. Furthermore, the consideration of the analytical measurement error of the pieces of equipment considered in this chapter should be included to improve the models performance. With further development these models can allow us to extend the application of digital design and ML in particular to connect particle properties to bulk powder properties to manufacturability and drug bioavailability in the patient.

7.5. References

- Abdullah, E. C., & Geldart, D. (1999). The use of bulk density measurements as flowability indicators. *Powder Technology*, *102*(2), 151-165.
- Ali, S., Myasnichenko, V., & Neyts, E. (2016). Size-dependent strain and surface energies of gold nanoclusters. *Physical Chemistry Chemical Physics*, *18*(2), 792-800.
- Angelos, S., Liong, M., Choi, E., & Zink, J. I. (2008). Mesoporous silicate materials as substrates for molecular machines and drug delivery. *Chemical Engineering Journal*, *137*(1), 4-13.
- Barjat, H., Checkley, S., Chitu, T., Dawson, N., Farshchi, A., Ferreira, A., . . . Tobyn, M. (2021). Demonstration of the Feasibility of Predicting the Flow of Pharmaceutically Relevant Powders from Particle and Bulk Physical Properties. *Journal of pharmaceutical innovation*, *16*(1), 181-196. doi:10.1007/s12247-020-09433-5
- Benet, L. Z. (2013). The role of BCS (biopharmaceutics classification system) and BDDCS (biopharmaceutics drug disposition classification system) in drug development. *Journal of pharmaceutical sciences*, *102*(1), 34-42.
- Darst, B. F., Malecki, K. C., & Engelman, C. D. (2018). Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC genetics*, *19*(1), 1-6.
- Dressman, J., Amidon, G., & Fleisher, D. (1985). Absorption potential: estimating the fraction absorbed for orally administered compounds. *Journal of pharmaceutical sciences*, *74*(5), 588-589.
- Felton, L. A. (2013). *Remington-essentials of pharmaceuticals*: Pharmaceutical press.
- Fichtner, F., Mahlin, D., Welch, K., Gaisford, S., & Alderborn, G. (2008). Effect of surface energy on powder compactibility. *Pharmaceutical research*, *25*(12), 2750-2759.
- Fitzpatrick, J. (2013). Powder properties in food production systems. In *Handbook of food powders* (pp. 285-308): Elsevier.
- German, R. M. (1989). Particle packing characteristics.
- Gibbs, J. (1875). Trans. Conn. Acad. Arts Sci.
- Gohel, M. C., & Jogani, P. D. (2002). Functionality testing of a multifunctional directly compressible adjuvant containing lactose, polyvinylpyrrolidone, and croscarmellose sodium. *Pharmaceutical technology*, *26*(3), 64-64.
- Gorle, A. P., & Chopade, S. S. (2020). Lquisolid Technology: Preparation, Characterization and Applications. *Journal of Drug Delivery and Therapeutics*, *10*(3-s), 295-307.
- Grey, R., & Beddow, J. (1969). On the Hausner ratio and its relationship to some properties of metal powders. *Powder Technology*, *2*(6), 323-326.
- Hausner, H. H. (1967). *Friction conditions in a mass of metal powder*. Retrieved from
- Kalman, H. (2021). Quantification of mechanisms governing the angle of repose, angle of tilting, and Hausner ratio to estimate the flowability of particulate materials. *Powder Technology*, *382*, 573-593.
- Karehill, P., Glazer, M., & Nyström, C. (1990). Studies on direct compression of tablets. XXIII. The importance of surface roughness for the compactability of some directly compressible materials with different bonding and volume reduction properties. *International journal of pharmaceuticals*, *64*(1), 35-43.
- Kumar, D., Chirravuri, S. S., & Shastri, N. R. (2014). Impact of surface area of silica particles on dissolution rate and oral bioavailability of poorly water soluble drugs: A case study with aceclofenac. *International journal of pharmaceuticals*, *461*(1-2), 459-468.
- Kunnath, K., Chen, L., Zheng, K., & Davé, R. N. (2021). Assessing predictability of packing porosity and bulk density enhancements after dry coating of pharmaceutical powders. *Powder Technology*, *377*, 709-722.
- Leane, M., Pitt, K., Reynolds, G., & Group, M. C. S. W. (2015). A proposal for a drug product Manufacturing Classification System (MCS) for oral solid dosage forms. *Pharmaceutical development and technology*, *20*(1), 12-21.

- Liversidge, G. G., & Cundy, K. C. (1995). Particle size reduction for improvement of oral bioavailability of hydrophobic drugs: I. Absolute oral bioavailability of nanocrystalline danazol in beagle dogs. *International journal of pharmaceutics*, 125(1), 91-97.
- Lundberg, S. (2018). Welcome to the SHAP documentation .
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Mase, M., Owen, A. B., & Seiler, B. (2019). Explaining black box decisions by shapley cohort refinement. *arXiv preprint arXiv:1911.00467*.
- Moore, R. (1991). The chemical and mineralogical controls upon the residual strength of pure and natural clays. *Geotechnique*, 41(1), 35-47.
- Noyes, A. A., & Whitney, W. R. (1897). The rate of solution of solid substances in their own solutions. *Journal of the American Chemical Society*, 19(12), 930-934.
- Nyström, C., Alderborn, G., Duberg, M., & Karehill, P.-G. (1993). Bonding surface area and bonding mechanism—two important factors for the understanding of powder comparability. *Drug Development and Industrial Pharmacy*, 19(17-18), 2143-2196.
- Pharmacopeia, U. (2012). Bulk density and tapped density of powders. *USP*, 36, 265-268.
- R Williams, D. (2015). Particle engineering in pharmaceutical solids processing: surface energy considerations. *Current Pharmaceutical Design*, 21(19), 2677-2694.
- Rowe, R. (1988). Binder-substrate interactions in tablets: a theoretical approach based on solubility parameters. *Acta pharmaceutica technologica*, 34(3), 144-146.
- Saleem, I., Smyth, H., & Telko, M. (2008). Prediction of dry powder inhaler formulation performance from surface energetics and blending dynamics. *Drug Development and Industrial Pharmacy*, 34(9), 1002-1010.
- Samiei, L., Kelly, K., Taylor, L., Forbes, B., Collins, E., & Rowland, M. (2017). The influence of electrostatic properties on the punch sticking propensity of pharmaceutical blends. *Powder Technology*, 305, 509-517.
- Saw, H. Y., Davies, C. E., Paterson, A. H., & Jones, J. R. (2015). Correlation between powder flow properties measured by shear testing and Hausner ratio. *Procedia engineering*, 102, 218-225.
- Shi, B., Wang, Y., & Jia, L. (2011). Comparison of Dorris–Gray and Schultz methods for the calculation of surface dispersive free energy by inverse gas chromatography. *Journal of Chromatography A*, 1218(6), 860-862.
- Sridharan, A., Rao, S., & Murthy, N. (1986). Compressibility behaviour of homoionized bentonites. *Geotechnique*, 36(4), 551-564.
- Wang, S. (2009). Ordered mesoporous materials for drug delivery. *Microporous and mesoporous materials*, 117(1-2), 1-9.
- Wells, J. I., & Walker, C. V. (1983). The influence of granulating fluids upon granule and tablet properties: the role of secondary binding. *International journal of pharmaceutics*, 15(1), 97-111.
- Yu, A.-B., Zou, R., & Standish, N. (1996). Modifying the linear packing model for predicting the porosity of nonspherical particle mixtures. *Industrial & engineering chemistry research*, 35(10), 3730-3741.
- Zajic, L., & Buckton, G. (1990). The use of surface energy values to predict optimum binder selection for granulations. *International journal of pharmaceutics*, 59(2), 155-164.

8. Conclusions and future work

The work presented in this thesis investigates aspects of the correlation between particle attributes and bulk powder properties to inform enhanced digital design methods for the development and manufacture of tablets while ensuring the quality and safety of medicines to patients. A key area of focus has been understanding how particle properties impact bulk material and formulation properties, such as powder flow and bulk density, and how the learnings from this investigation can help minimise cost, time, resources, and waste in Chemistry, Manufacturing and Control processes for new pharmaceutical ingredients. The application of machine learning and deep learning methods has enabled the development of predictive models that allowed direct linkage of drug substance properties to decision-making in drug product manufacturing. This is key to enabling closer integration of primary and secondary manufacturing through to achieving integrated end-to-end continuous manufacturing process (Quon et al., 2012). The data-driven model workflow outlined in Chapter 4 (see Fig 4-14) was used on a wide variety of tasks throughout the research, from the prediction of powder flow of individual materials and the assessment of the viability of formulations for continuous direct compression (cDC) to the interrogation of the importance of bulk density and surface properties. This generic machine learning workflow could be potentially applied to any stage of the pharmaceutical development pipeline (Gaudelet et al., 2021; Réda, Kaufmann, & Delahaye-Duriez, 2020; Vamathevan et al., 2019). The methodologies captured in this data-driven model development workflow aim to provide guidelines for the development of reliable, useful machine learning models in the context of pharmaceutical development and manufacturing. One particular aspect that has received particular attention is the interpretation of the models' results using SHAP parameters to further enhance the utility of these approaches.

Machine learning models were built for the classification and prediction of the powder flow of individual powders (Chapter 4) and pharmaceutical formulations (Chapter 5). Particle size and shape descriptors extracted from widely used measurement instruments, QICPIC®, in Chapter 4, or Mastersizer® 3000 in combination with Morphologi® G3, in Chapter 5, were used to train the models.

For the classification of powder flow of individual powders, several algorithms were trained, tuned, and compared to find the best-performing model. Hence, Random Forest (RF) was trained on a library of 112 pharmaceutical powders, with a wide range of particle size, i.e., PSD D10 values between 9 and 225 µm, PSD D50 values between 25 and 644 µm, PSD D90 values between 53 and 1892 µm; particle shape, i.e., aspect ratio values between 0.4 and 0.95 and sphericity values between 0.1 and 0.9; bulk density, with a range of values between 0.2 and 1.2; and flow function coefficient, with a range of

values between 0.95 and 107, was used. The model was assessed using 10-fold cross-validation and an external dataset, achieving a performance of over 80% probability of distinguishing between classes.

SHAP interpretation methods were applied to understand how the models made the predictions. This technique also allowed the development of manufacturing recommendations to reverse engineer particles with the “ideal” properties for the desired manufacturing route, i.e., direct compression. Targets of particle size and shape were presented to ensure direct compressibility of new pharmaceutical materials. The SHAP method used to propose the manufacturing recommendations is heavily reliant on the training dataset, and therefore the addition of more data, particularly cohesive powders to minimise the effect of class imbalance and powders with a wider combination between particle descriptors and bulk properties, would help overcome this challenge. Another limitation of the methodology is that SHAP values are calculated assuming feature independence, which is not realistic for this application, as particle size and shape properties are not independent. Reverse engineering has been applied to several fields related to the pharmaceutical industry. For instance, several reverse engineering methods, i.e., linear regression, graphical Gaussian model, and dynamic Bayesian network, have been applied to gene networks (Camacho, VERA LICONA, Mendes, & Laubenbacher, 2007; He, Balling, & Zeng, 2009). These reverse engineer methods presented some limitations, namely biological assumptions that made the application of the methods to gene networks challenging. In the field of pharmaceutical manufacturing, Čapková *et al.* used Raman mapping and chemometrics coupled with statistical analysis to distinguish tablets based on the manufacturing technology used, the particle size of the API, or the composition of the components, and thereby reverse engineer pharmaceutical tablets for generic product development. The challenge presented by this method was the processability of the spectra contained in each map for chemometric evaluation (Čapková, Pekárek, Hanulíková, & Matějka, 2022). In this thesis, we presented a novel method for reverse engineering particle properties. This methodology could be further evaluated by testing the manufacturing recommendations in the design of a new pharmaceutical powder.

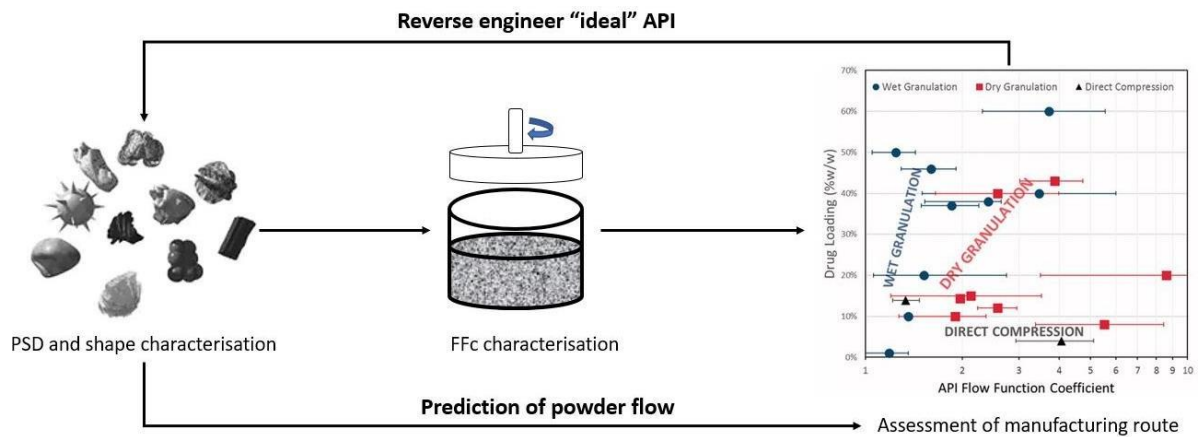


Figure 8-1: A diagram of the reverse engineering approach to obtain the “ideal” API for the desired pharmaceutical manufacturing route, including the flow map showing the Ffc of the API and the drug loading for direct compression, wet granulation and dry granulation presented by the MCS (Leane et al., 2018).

The data-driven model workflow described in Chapter 5 was applied for the assessment of the viability of pharmaceutical formulations for cDC. An RF model was built using particle size, particle shape, and composition information for 99 formulations and the model was assessed using 10-fold cross-validation, bootstrap sampling, and an external dataset, achieving a precision of 90%. Similar results were obtained in the different validation methods, suggesting that the model was robust and generating trust for the implementation of the model in an industrial setting. Furthermore, the results of the classification of new formulations were reported using a range of probabilities describing their suitability for cDC, i.e., whether they are viable for cDC. These probabilities were calculated using Monte Carlo simulation, accounting for the analytical error of the instruments, increasing the confidence in the prediction. The performance of the models could be improved by the addition of more viable formulations for cDC to the dataset, as these formulations were underrepresented. However despite this uneven representation in the training data set used in this study, the results obtained were satisfactory, allowing to proceed with the translation of these models to an industrial group in Roche by building a user-friendly app in using python and Heroku (Middleton & Schneeman, 2013) (see Fig 8-2).

The interface shows a form for configuring a pharmaceutical formulation. At the top, there is a field for 'API batch number' and a slider for 'API Concentration' set to 20. Below this are three sections for 'Excipient 1', 'Excipient 2', and 'Excipient 3'. Each section includes a dropdown menu (all set to 'Aerosil 200') and a slider for concentration (set to 40, 20, and 20 respectively). A green progress bar at the bottom indicates 'Total Concentration: 100'. A 'Submit' button is located at the very bottom.

Figure 8-2: The interface of the app implemented on-site. The model can predict the viability of a pharmaceutical formulation for cDC based on the physical properties of the API and the concentration of the components.

The prediction of powder flow was expanded and complemented with assessing the ability to predict bulk properties using deep learning (Chapter 6). A CNN architecture was trained on images of particles, achieving a similar result as the RF model of comparison. The RF and the CNN were compared based on their performance on the classification of nine powders included in the external dataset. Even though the accuracy of both algorithms was equal, the AUC – ROC value achieved by CNN was higher, indicating a higher reliability of the CNN model for the classification of powder flow. It would be expected that with the addition of more data to both training dataset, the CNN model outperforms the RF model. Moreover, in this chapter, a novel approach for the prediction of powder flow using images was introduced, opening the possibility of using images of samples to assess the behaviour of powders.

Lastly, the interrogation of particle and bulk properties in the context of machine learning was presented in Chapter 7. A threshold of 0.5 g/ml was defined for the bulk-density models to assess the suitability of new materials for direct compression. This threshold was adopted based on the results from the manufacturing classification system (MCS) (Leane, Pitt, Reynolds, & Group, 2015). The results

of the bulk-density model showed that as the aspect ratio increases, the bulk density increases too. This finding was in good agreement with the studies published by H. Ding *et al.* and Kunnath *et al.* regarding the investigation of the correlation between the particle sphericity and the bulk porosity (inversely correlated to bulk density) (H. Ding *et al.*, 2020; Kunnath, Chen, Zheng, & Davé, 2021). Another interesting finding was the impact of the FFC on the bulk density: for values of FFC smaller than 10, as the FFC increases, the positive impact on the model outcome (classification of bulk density) increases too; for values of FFC equal or greater than 10, the value of FFC does not impact the classification of bulk density. Since the correlation between FFC and bulk density was unclear, and this property was not the most important property for the classification of bulk density, the FFC was removed from the dataset. The bulk-density model was rebuilt using particle size and shape data exclusively, and the performance of the model did not decrease. The main advantage of predicting bulk density based on particle size and shape is that the time and amount of material required to generate the data to run the model for a new material is decreased from 30 grams and 2 hours to 2 grams and 5 minutes.

The results of the surface-area models showed a link between the bulk properties and the bulk performance in the patient, by understanding the correlation between surface area and drug bioavailability. While more data could be added to improve the performance of this model, the approach explored the feasibility of predicting surface area to assess the bioavailability of a new drug in the gastrointestinal tract. The bioavailability depends on the dissolution rate of the drug, which is calculated from the surface area using the Noyes-Whitney equation (Noyes & Whitney, 1897). The results obtained from the surface-area classification model enable the prediction of the surface area of a new material based on its particle size and shape properties, which can be used to calculate the bioavailability of the drug. The model also shows that as the particle size and aspect ratio increase, the surface area decreases, potentially jeopardizing the dissolution rate of the drug. For tablet manufacturing, particularly for direct compression, the efforts of the particle design scientist are in increasing the particle size and make the particles more spherical. Further development of the surface-area models would help link the manufacturability requirements to the bioavailability requirements, working towards unified recommendations for manufacturing, accelerating the digitalisation of pharmaceutical manufacturing, thereby reducing costs, improving efficiency and productivity (Hole, Hole, & McFalone-Shaw, 2021).

The implementation and understanding of machine learning models in early stage of pharmaceutical development can reduce costs of manufacturing new APIs while assuring drug efficacy, by reducing the amount of materials and the number of measurements required. The results outlined in the scientific chapters of this thesis show how data-driven models have the

potential to provide useful, reliable, and interpretable predictions that can facilitate the development of a new pharmaceutical product or formulation.

8.1. Future work

The manufacturing recommendations obtained by the reverse engineer analysis presented in Chapter 4, coupled with the guidelines proposed by Leane *et al.* (Leane et al., 2015) can guide pharmaceutical scientists to develop direct-compressible materials, based on the particle size and shape of the API. The addition of more API properties (e.g., crystal structure descriptors, or solubility), and how these properties affect manufacturability could be further explored to develop a more complete guide for the reverse engineering of pharmaceutical powders. The developability of APIs in other manufacturing techniques, i.e., roller compaction or wet granulation, could also be investigated by the prediction of FFC to, for example, inform spherical agglomeration or wet granulation route selection. Furthermore, generating a model that can transform particle size and shape data measured by different instruments would make the proposed manufacturing recommendations instrument agnostic, allowing their implementation regardless of the particle size characterisation instrument available.

The prediction of the viability of pharmaceutical formulations for cDC serves as a showcase of the potential applications of the development and implementation of predictive models in the pharmaceutical industry. The implementation of this model contributes to the increasing drive for digital transformation across the industry to improve material efficiency, R&D productivity, improve sustainability and reduce cost. Moreover, the excipients included in the training dataset could be grouped based on their properties, removing the need of retraining the model if different excipients were to be used in a new formulation. If well-trained and well-validated models were able to be more widely implemented, only formulations that obtained a positive result *in silico* would undergo further development, reaching the patient faster and with less material waste.

The further development of the deep learning models presented in Chapter 6 could facilitate their implementation either offline, as “analytical” measurements to predict bulk properties in early-stage development, or online, coupled with the current PAT tools. An example of a potential online implementation could be the use of the deep learning models in combination with focussed beam reflectance measurement (FBRM) particle size monitoring in crystallisation. If further work was developed on the training and implementation of the model, powder flowability could be predicted before the crystals left the vessel, and the assessment of the suitability of the material for the intended manufacturing operation could be done *in situ*, moving forward towards an end-to-end integrated continuous manufacturing process. This adaptive real time control would revolutionise how materials are developed and produced in pharmaceutical industry.

Finally, the work presented in Chapter 7 could be further enhanced by adding more data to the surface area and surface energy models. Particularly, the surface area models could be expanded by finding a threshold to divide the dataset into classes that are representative of an adequate dissolution rate, and therefore, the classification of the surface area would provide a better estimation of the drug bioavailability. Finding this threshold would allow to expand on the manufacturing recommendations proposed in Chapter 4, to design manufacturability while ensuring bioavailability of the drug in the patient.

8.2. References

- Camacho, D., VERA LICONA, P., Mendes, P., & Laubenbacher, R. (2007). Comparison of reverse-engineering methods using an in silico network. *Annals of the New York Academy of Sciences*, 1115(1), 73-89.
- Candanedo, I. S., Nieves, E. H., González, S. R., Martín, M., & Briones, A. G. (2018). *Machine learning predictive model for industry 4.0*. Paper presented at the International Conference on Knowledge Management in Organizations.
- Čapková, T., Pekárek, T., Hanulíková, B., & Matějka, P. (2022). Application of reverse engineering in the field of pharmaceutical tablets using Raman mapping and chemometrics. *Journal of Pharmaceutical and Biomedical Analysis*, 209, 114496.
- Ding, B. (2018). Pharma Industry 4.0: Literature review and research opportunities in sustainable pharmaceutical supply chains. *Process Safety and Environmental Protection*, 119, 115-130.
- Ding, H., Li, B., Boiarkina, I., Wilson, D. I., Yu, W., & Young, B. R. (2020). Effects of morphology on the bulk density of instant whole milk powder. *Foods*, 9(8), 1024.
- Gaudelet, T., Day, B., Jamasb, A. R., Soman, J., Regep, C., Liu, G., . . . Tang, J. (2021). Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 22(6), bbab159.
- He, F., Balling, R., & Zeng, A.-P. (2009). Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. *Journal of biotechnology*, 144(3), 190-203.
- Hole, G., Hole, A. S., & McFalone-Shaw, I. (2021). Digitalization in pharmaceutical industry: What to focus on under the digital implementation process? *International Journal of Pharmaceutics: X*, 3, 100095.
- Kunnath, K., Chen, L., Zheng, K., & Davé, R. N. (2021). Assessing predictability of packing porosity and bulk density enhancements after dry coating of pharmaceutical powders. *Powder Technology*, 377, 709-722.
- Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & information systems engineering*, 6(4), 239-242.
- Leane, M., Pitt, K., Reynolds, G., & Group, M. C. S. W. (2015). A proposal for a drug product Manufacturing Classification System (MCS) for oral solid dosage forms. *Pharmaceutical development and technology*, 20(1), 12-21.
- Leane, M., Pitt, K., Reynolds, G. K., Dawson, N., Ziegler, I., Szepes, A., . . . Group, M. C. S. W. (2018). Manufacturing classification system in the real world: factors influencing manufacturing process choices for filed commercial oral solid dosage formulations, case studies from industry and considerations for continuous processing. *Pharmaceutical development and technology*, 23(10), 964-977.
- Middleton, N., & Schneeman, R. (2013). *Heroku: up and running: effortless application deployment and scaling*: " O'Reilly Media, Inc."
- Noyes, A. A., & Whitney, W. R. (1897). The rate of solution of solid substances in their own solutions. *Journal of the American Chemical Society*, 19(12), 930-934.
- Quon, J. L., Zhang, H., Alvarez, A., Evans, J., Myerson, A. S., & Trout, B. L. (2012). Continuous crystallization of aliskiren hemifumarate. *Crystal growth & design*, 12(6), 3036-3044.
- Réda, C., Kaufmann, E., & Delahaye-Duriez, A. (2020). Machine learning applications in drug development. *Computational and structural biotechnology journal*, 18, 241-252.
- Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., . . . Spitzer, M. (2019). Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6), 463-477.