Predictive Modeling of Pipetting Dynamics

Multivariate Regression Analysis: PLS and ANN for
Estimating Density and Volume from Pressure Recordings

Lisa Linárd Pedersen

2024

Master Thesis in
Biomedical Engineering

LUND
UNIVERSITY

Faculty of Engineering LTH
Department of Biomedical Engineering
In Collaboration with Thermo Fisher Scientific

Supervisors:
Frida Sandberg, Lund University
Jan Ybrahim, Thermo Fisher Scientific

Examiner:
Martin Stridh, Lund University

Contents

1 Abstract

Thermo Fisher Scientific manufacture automatic pipetting instruments for diagnostic tests. These tests are sensitive to abnormalities and changes in e.g. volume or density could potentially lead to less precision or other issues in the pipetting work flow. Utilizing data collected from a pressure sensor inside the pipette could be a way of automatically verifying different aspects related to the pipetting. Machine learning may be a powerful tool in continuously evaluating these aspects and keeping the handler notified of any changes.

This thesis aims to investigate the feasibility of extracting useful insights from pipetting pressure recordings. The initial objective was classifying error causes such as bubbles or foam in the pipette and data was collected with this in mind. This however was not successful as these errors were not detectable in the pressure recordings. Hence, the thesis focuses on the secondary objective, to estimate pipetted volume and density based on pressure sensor data.

The data collection was done using the Thermo Fisher pipetting instrument Phadia 200. Three different sets were collected. D1 data set consisting of 4 groups of 80 observations each. These were water, 5% glycerol, 10% glycerol and 40% glycerol. D2 data set consisting of 3 groups of 50 observations each. These were three different human samples. D3 data set consisting of 3 groups of 50 observations each. These were 2.5% glycerol, 7.5% glycerol and 20% glycerol. Pressure recordings as well as estimated volumes for each sample were collected. A partial least squares model (PLS) and an artificial neural network (ANN) model were used for the regression problem.

The results of the regressions were not satisfactory and it was concluded that the data was not ideal for the task. All models but the ones where all data sets were included in training yielded very poor R^2 scores, especially in the volume estimations. The best model was a PLS model which had an R^2 of 0.96 in volume predictions and 0.54 in density predictions. This model had an RMSE of 0.9660 in volume predictions and 0.0140 in density predictions. However, since this model was trained with all data and did not predict on any new densities, this does not say anything about generalizability to new and unseen data. The model that had the best results for predicting unseen data was a PLS model trained on D1 data set predicting the D3 data set. For these density predictions, R^2 was 0.80 and RMSE

0.0093. For the volume predictions however, R^2 was -40 and RMSE 2.1135.

The data was collected with the primary objective, a classification problem, in mind. Since the data was finally used for a regression task, it was concluded that shortcomings in the experimental design were a crucial aspect affecting the results. It is however not possible to say whether a better set up and data set would yield better results. There is a risk that the relationships between pressure, volume and density simply are not clear enough or are too easily affected by outside factors. The conclusion is therefore that further investigation needs to be done in order to evaluate the feasibility of the methods.

2 Dedications

I would like to dedicate this work to my family and loved ones. For always believing in me and encouraging me. During this work, through the entirety of my studies and in life. Special thanks to you Mom. To my chosen family Olli, Sakarias & Selma. And to my brothers Oscar & Adam. You are my support and inspiration. And Dad, I know you're watching proudly.

3 Acknowledgements

I would like to thank my supervisor Frida Sandberg of the department of Biomedical Engineering at Lunds Tekniska Högskola, for her continuous guidance during this process. Furthermore, Jan Ybrahim of Thermo Fisher Scientific, for providing his knowledge and insights.

4 Introduction

4.1 Background

The accuracy of in vitro diagnostic methods is contingent on precisely defined volumes for quantifying the concentration of substances in samples. To uphold the high quality of results in these diagnostic tests, the equipment utilized must incorporate features to validate the precision of sample pipetting.

Thermo Fisher specializes in immunodiagnostics, specifically in the diagnosis of allergies and autoimmunity through the detection of specific antibodies in blood plasma and serum. This is accomplished using Enzyme-Linked Immunosorbent Assay (ELISA). The measurement of antibody content involves a series of reactions, culminating in the utilization of a fluorescent molecule to quantify the amount of antibodies present in the sample. Here, the quantity information is correlated with the amount of fluorescent light emitted. In diagnostic procedures, the focus lies on concentration information rather than the intensity of emitted light, and therefore an essential step for accurate calculations is the employment of a calibration curve. However, the accuracy of this function is susceptible to volume inaccuracies, underscoring the critical importance of precise pipetting volumes.

Inaccuracies in aspiration and dispersion volumes may arise from various factors, such as blood clots obstructing the pipette or the presence of bubbles or foam impacting the actual pipetted volume. The handler is particularly interested in understanding both the pipetted volume and the reasons for faulty pipettings. An interesting aspect when pipetting is the difference in density in samples, as this may be useful information when analyzing and handling samples. Furthermore, when interpreting errors and differences that arise in the pipetting process. One potential method for determining these aspects involves analyzing the pressure inside the pipette.

Thermo Fisher manufacture a pipetting machine, Phadia 200, which allows the user to pipette predetermined volumes of liquid. The pipette incorporates a pressure sensor which can output information of pressure during aspiration and dispersion. As of today, this sensor is not used for controlling pipetting volumes or pipetting errors, but this is something the company would like to investigate.

Figure 1: Phadia 200 (Martin Danielson n.d.)

4.2 Aims and Objectives

The primary objective of this thesis was to detect and classify causes of error such as foam and bubbles based on pressure sensor data gathered by the Phadia 200. The secondary objective of this thesis was to estimate pipetted volume and density based on pressure sensor data gathered by the Phadia 200.

The primary objective encountered challenges as the sensors failed to detect the presence of foam or bubbles. Consequently, the secondary objective was necessitated, leading to the implementation of multivariate regression. This new purpose aimed to investigate the potential for predicting volume and density based on pipetting aspiration pressure curves. This thesis therefore handles the pressure data collection from pipetting samples using the Phadia 200 and predictive modeling of volume and density from this data. It investigates feasibility and discusses possible error sources. Furthermore, things to consider when gathering and processing data for the specific use.

The machine learning methods used are explained in section 5. Info on the data collection procedure as well as data processing steps can be found in section 6. The data and its rejected outliers, extracted features as well as prediction results are found in section 7. Comments and discussion regarding the results and possible error sources are found in section 8.

5 Theory

5.1 Multivariate Linear Regression

A multivariate linear regression from m independent variables X_i to n dependent variables Y_i can be expressed in matrix form as

$$Y = X\beta + \epsilon \tag{1}$$

where Y is an $n\times1$ vector of dependent variables, X is an $n\times(m+1)$ matrix of independent variables, β is a $(m+1)\times n$ matrix of coefficients, ϵ is an $n\times1$ vector of error terms.

The goal is to estimate the coefficients that minimize the errors.

5.2 Partial Least Squares Regression

The partial least square (PLS) algorithm originates from Herman Wolds work with the NIPALS algorithm (Wold 1966). It is a multivariate statistical method that is used for modeling relationships between sets of variables. It is particularly useful for predictive modeling in cases where there is multicollinearity between variables or high dimensional data. This is due to its robustness, deriving from the fact that the model parameters are not as sensitive to changes in the data as in other methods such as ordinary least squares (Geladi and Kowalski 1986). Because of this, a more stable model can be achieved, with better predictive and generalizing abilities.

In PLS regression the goal is to identify n components

$$K = [k_1...., k_n] \tag{2}$$

as linear combinations of inputs X

$$k_k = Xw_k \tag{3}$$

for $k = [1...., n]$, where w_k is some weight vector. We aim to derive components that serve as effective predictors not only for the response variable Y but also for the input variables X_j, where j=1,. . . ,p, fulfilling the relationships

$$X = KL^T + E \tag{4}$$

$$Y = Kb + e \tag{5}$$

where K is the matrix of scores (PLS components), L is a matrix of PLS loadings, E is a matrix of X- residuals, b is a vector of PLS regression coefficients and e is a vector of y- residuals.

By choosing n components (n<p), we approximate the model for the X-space

$$\hat{X} = \Xi\Gamma^T \tag{6}$$

and Y- space

$$\hat{Y} = \Omega\Delta^T \tag{7}$$

where

- $\Xi \in R^{n\times K}$, $\Omega \in R^{n\times K}$ contain the scores in their columns.
- $\Gamma^T \in R^{K\times d}$, $\Delta^T \in R^{K\times d}$ contain the loadings in their rows.

The components are obtained through the NIPALS algorithm (see 5.2.1)(Sanchez and Marzban 2020).

5.2.1 NIPALS: Nonlinear Iterative Partial Least Squares algorithm

Given matrices X and Y and a number of components

$$X \in R^{n\times d},$$
$$Y \in R^{n\times t},$$
$$\text{components } K$$

X_1 is set to X and Y_1 is set to Y and for each $k \in [1, K]$:

1. Compute the weights $u_k \in R^d$ and $v_k \in R^t$. These are the first left and right singular vectors of the cross-covariance matrix $C = X_k^T Y_k$. These are computed by the power method (5.2.2). By definition, u_k and v_k are chosen so that they maximize the covariance between the projected X_k and the projected target, that is $\text{Cov}\,(X_k u_k, Y_k v_k)$.
2. Find the scores ξ_k, ω_k, these are obtained by projecting X_k and Y_k on u_k and v_k:
   - $\xi_k = X_k u_k$
   - $\omega_k = Y_k v_k$
3. Find the loading vectors, γ_k and δ_k, these are obtained by

- γ_k: Finding a vector $\gamma_k \in R^d$ such that the rank-1 matrix $\xi_k \gamma_k^T$ is as close as possible to X_k.
- δ_k: Y_k are approximated using the projection of X_k (i.e. ξ_k) : finding a vector $\gamma_k \in R^d$ such that the rank-1 matrix $\xi_k \delta_k^T$ is as close as possible to Y_k.

4. deflate X_k and Y_k, i.e. subtract the rank-1 approximations:
   - $X_{k+1} = X_k - \xi_k \gamma_k^T$
   - $Y_{k+1} = Y_k - \xi_k \delta_k^T$.

At the end, we have approximated X as a sum of rank-1 matrices:

$$X = \Xi \Gamma^T \tag{8}$$

where

- $\Xi \in R^{n \times K}$ contains the scores in its columns
- $\Gamma^T \in R^{K \times d}$ contains the loadings in its rows.

Similarly for Y, we have

$$Y = \Omega \Delta^T \tag{9}$$

where

- $\Omega \in R^{n \times K}$ contains the scores in its columns
- $\Delta^T \in R^{K \times d}$ contains the loadings in its rows.

Note that the scores matrices Ξ and Ω correspond to the projections of the training data X and Y, respectively (scikit-learn developers 2023).

5.2.2 The power method

The PLS algorithm contains an inner loop, the power method, which computes the eigenvectors of C (see 5.2.1). Here, superscript k in u^k, v^k denote iteration in the inner loop and subscripts u_r, v_r denote iteration in outer loop (5.2.1) and are the columns of U and V. Superscripts on scalars indicate exponents. The power method algorithm as follows:

Let $\mathbf{C}$ be an $I \times J$ matrix. Set $r \leftarrow 0$.

1. Repeat

(a) Set $r \leftarrow r + 1$

(b) Choose $\mathbf{u}^0 \in R^I$

(c) Set $k \leftarrow 0$

(d) Repeat

$$\begin{aligned} k &\leftarrow k + 1 \\ \mathbf{v}^k &\leftarrow \mathrm{C}^T \mathbf{u}^{k-1} \\ \mathbf{v}^k &\leftarrow \mathbf{v}^k \\ \mathbf{u}^k &\leftarrow \mathbf{C}\mathbf{v}^k \\ \mathbf{u}^k &\leftarrow \mathbf{u}^k / \left\|\mathbf{u}^k\right\| \end{aligned}$$

until $\left\|\mathbf{u}^k - \mathbf{u}^{k-1}\right\|$ is less than some convergence criterion.

(e) Save

$$\begin{aligned} \mathbf{u}_r &\leftarrow \mathbf{u}^k \\ \mathbf{v}_r &\leftarrow \mathbf{v}^k \\ d_r &\leftarrow (\mathbf{u}_r)^T \mathbf{C}\mathbf{v}_r \end{aligned}$$

(f) (Set $\mathbf{C} \leftarrow \mathbf{C} - d_r \mathbf{u}_r \mathbf{v}_r^T$. until $\|\mathbf{C}\|$ is less than some criterion.

2. Reorder the d_r so that

$$d_1 > d_2 > \ldots > d_R$$

and reorder the $\mathbf{u}_r$ and $\mathbf{v}_r$ accordingly.

(Wegelin (2000). For proof of convergence see section 13, pg. 37-39 Wegelin (*ibid.*)).

5.2.3 Data assumptions

There are several assumptions made when attempting a linear regression. In this work, partial least squares estimation is used for the reason that it is a robust method that handles some aspects better than other regression methods. Nonetheless, it is good practice to have some things in mind when attempting a regression.

- Linear relationship: PLS assumes a linear relationship between the independent and dependent variables. Linearity can stem from a linear relationship but can also in some cases be achieved by transformation of the data.

- Independence of observations: PLS, like other regression techniques, assumes that observations are independent of each other.

- Multicollinearity: PLS is designed in a way that makes it robust to multicollinearity. When there are high correlations among the independent variables, PLS is particularly useful since the method not uses the observed variables directly, but constructs new latent variables which are designed to capture the shared variance among the predictors (El-Salam 2014). Dissimilar to the traditional regression approach, which generally focuses on predicting the dependent variable, it maximizes the covariance between the dependent variables and predictors while also finding components which have maximum correlation with the response (Johnsson and Kuhn 2013). This makes it less sensitive to multicollinearity.

- Outliers: PLS is generally less sensitive to the presence of outliers compared to some other regression techniques. since PLS performs dimensionality reduction by constructing latent variables, outliers may have less influence on these compared to the original predictors, yielding a more robust model. Still, extreme outliers can influence the results, and it is good practice to check for their presence.

- Homoscedasticity: If the variance of the residuals is not homogenous accross all levels of predictor variables, this can yield bad results and skewed predictions and this should therefore be taken into account. PLS is less sensitive to this than e.g. ordinary least squares, since it constructs latent variables that are linear combinations of the original predictors, and these latent variables are designed to capture the shared variance among the predictors, potentially reducing the impact of heteroscedasticity.

5.3 Artificial Neural Network

The Artificial neural network (ANN) used in this thesis is a feedforward neural network, i.e. there is only one way connections and no loop backs in the network. The primary objective of a feedforward network is to approximate a specific function f^*.

Here, we have a regression task incorporating two continuous output variables. Hence, the function

$$\mathbf{y} = f^*(\mathbf{x}) \tag{10}$$

maps an input x to a pair of outputs y_1 and y_2. By defining a mapping

$$\mathbf{y} = f^*(\mathbf{x}; \theta) \tag{11}$$

the feedforward network strives to learn the optimal values for the parameters θ that result in the most accurate function approximation (Goodfellow, Bengio and Courville 2016).

5.3.1 Architecture

The network consists of various layers, each employing computations to achieve an optimal mapping from the input to the output. While there are countless ways to structure the model, here a simple model with one hidden layer is used. The layers are fully connected (i.e. each neuron in every layer is connected to each neuron in the next).

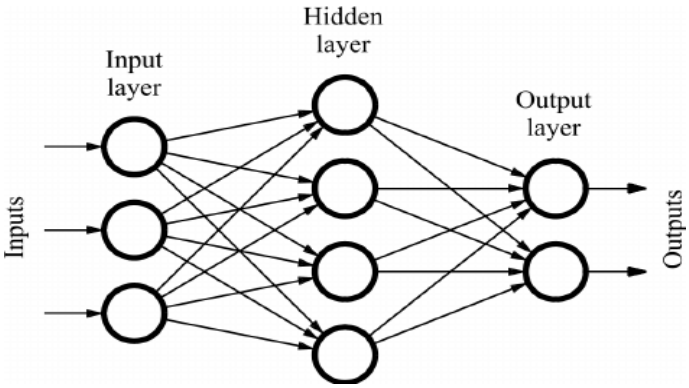


Figure 2: A feedforward neural network (Nguembang Fadja, Lamma and Riguzzi 2018)

Each layer is built up of a n of nodes. The input layer has n equal to the number of input variables, the output layer has n equal to the number of output variables and the number of nodes in the hidden layers are a design choice. Each node is associated with an activation function f. There are several choices of activation functions. Each input x is weighted by some weight w. The weights are adjusted during training in order to achieve the desired output, minimizing some error function. (Zou, Han and So 2009).

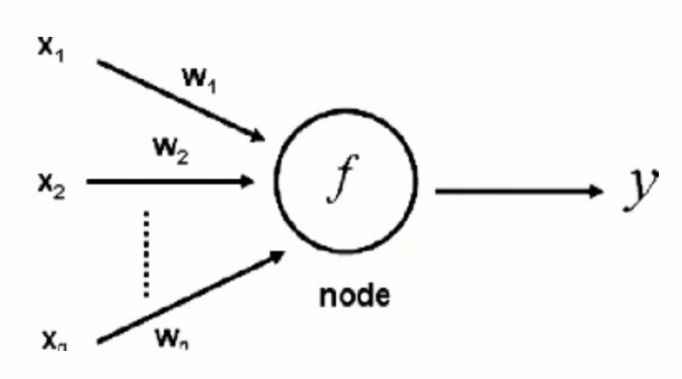


Figure 3: A node (Nguembang Fadja, Lamma and Riguzzi 2018)

5.3.2 Activation functions

In this work, two different activation functions are used.

A rectified linear unit function, defined as

$$ReLU(x) = max(0, x) \tag{12}$$

Here, the output is zero for any negative input and equal to the input for any non-negative input.

The output function is a linear activation function, defined as

$$F(x) = x \tag{13}$$

here, the network will output a linear variable, which is appropriate for a regression task (Siddharth Sharma 2020).

5.3.3 Hyperparameters

Several hyperparameters can be experimented with in an ANN. In this work the following hyperparameters will be tuned during training.

- Number of nodes in each layer. The number of nodes may affect generalization and capacity of fitting the data. In the hidden layers, this is a parameter that can be evaluated and optimized.

- Batch size. This determines how many training examples are considered together before the model's weights are updated. It is a trade-off between computational efficiency, memory requirements, and the quality of weight updates.
- Epochs. An epoch is one complete pass through the entire training dataset during the training phase. This hyperparameter determines how many iterations the the model gets for learning the data. Here, the trade-off lies between underfitting and overfitting the data.

5.4 Model training and testing

There are some things to consider and methods to utilize when training a model to achieve good results on both traning and test data. Overfitting is a common problem in machine learning, and here, good generalization to new data is the goal.

5.4.1 Model complexity and hyperparameters

Finding the right balance between model complexity and generalization is essential. In the case of partial least squares, choosing the right number of components is key. While many components may yield a good fit on training data, it may lead bad generalization and hence bad performance on future data. Too few may lead to underfitting. In the case of neural networks, the architecture plays a crucial role in its performance, and having a very complex architecture with many layers and nodes can lead to overfitting.

Hyperparameters play a crucial role in the performance of a machine learning model and can significantly impact its ability to generalize well to new, unseen data. Balancing model complexity and hyperparameters is an ongoing process that involves experimentation and tuning.

5.4.2 K-fold cross-validation

K-fold cross-validation is a widely used technique for assessing the performance and generalization ability of a predictive model. It provides a more robust estimate of the model's performance by averaging over multiple test sets. Furthermore, it ensures all data points are utilized for training and validating. The steps of K- fold cross validation are as follows:

1. Split the dataset into K subsets

2. The model is trained K times. In each iteration, K-1 folds are used for training, and the remaining fold is used for validating. The model's performance metrics are recorded for each iteration.

3. The performance metrics obtained from each iteration are averaged to provide an overall assessment of the model's performance.

5.4.3 Model testing

Model testing is a crucial step in the machine learning workflow that involves evaluating the performance of a trained model on new, unseen data. The primary goal of model testing is to assess how well the model generalizes to data it has not encountered during the training phase. The performance on the test data is a critical indicator of how well the model is likely to perform in real-world scenarios. It helps make informed decisions about the model's suitability for deployment or further refinement. Still, although the model is not fit on the test data, there is a danger of overfitting the model on the test data by tweaking it too much in the testing stage. This is something to take into account during the process as the test data should remain representative of unseen data in the future.

5.5 Evaluation Metrics

The choice of evaluation metrics is a crucial aspect in assessing the performance of a predictive model. As illustrated in figure 4, different metrics highlight different aspects of model performance, and the selection should be based on the specific characteristics of the problem at hand.

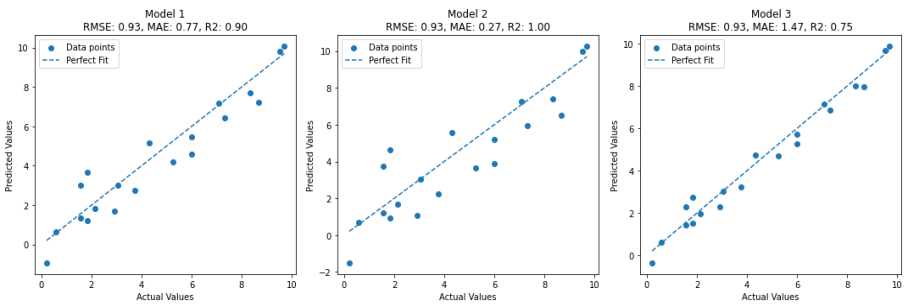


Figure 4: Evaluation Metrics

In figure 4, three different models have been used to predict values. Here, the root mean squared error (RMSE) is the same for all three models, meaning they have a similar overall accuracy in predicting values. However, mean absolute error (MAE) and R^2 scores differ. MAE reflects the average magnitude of errors, and R^2 reflects the goodness of fit, indicating how well the model explains the variance in the data. The variations in MAE and R^2 between the models illustrate how they perform differently in terms of the absolute errors and how well they explain the variance in the data.

5.5.1 Root Mean Squared Error

Root mean squared error is useful when the error measured in units is of interest. It measures the average magnitude of the errors between predicted values and actual values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

Where n is the number of observations or data points, y_i is the actual value of the dependent variable for observation i, $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

5.5.2 Mean Absolute Error

Mean absolute error measures the average absolute difference between predicted values and actual values. It is easy to interpret since it provides a straightforward measure of how far, on average, the predictions deviate from the actual values. Additionally, MAE is less sensitive to outliers compared to other metrics like Root Mean Square Error (5.5.1).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

Where n is the number of observations, y_i is the actual (observed) value, $\hat{y}_i$ is the predicted value.

5.5.3 R^2 score

The R^2 score, or coefficient of determination, indicates the proportion of the variance in the dependent variable that is predictable from the independent

variables. In other words, it measures the percentage of variation in the dependent variable that can be explained by the independent variables.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{16}$$

Here SS_{tot} represents the total variability in the dependent variable and SS_{res} represents the unexplained variability, n is the number of data points, y_i is the actual value of the dependent variable for the i-th data point, $\hat{y}_i$ is the predicted value of the dependent variable for the i-th data point, $\bar{y}$ is the mean of the actual values y.

6 Methodology

6.1 Data Collection

The data collection was done on two occasions. The thesis worker collected the D1 set and D2 set at the beginning of the process. A test set, D3, was later collected by an employee at Thermo Fisher. The collection methods are presented in this section.

6.1.1 Materials and Equipment

6.1.1.1 Phadia 200 Machine

The Phadia 200 machine, manufactured by Thermo Fisher, served as the instrument for automated pipetting. It is equipped with precision pumps and sensors to monitor and record pressure changes during liquid dispensing.

6.1.1.2 Pipetting wells

Glucose-stained sponges within wells were subjected to a washing and drying process. Subsequently, these sponges were transferred to empty wells and utilized in the pipetting procedure. This transfer was undertaken with the aim of minimizing evaporation as much as possible.

6.1.1.3 Manual pipette and pipette tips

A micropipette, from Thermo Fisher scientific, and disposable pipetting tips were utilized in the sample preparation and density measurements.

6.1.1.4 Scale

A Mettler Toledo AT261 deltarange was used for the weight measurements.

6.1.1.5 Camera

An RS Pro Wifi Digital Microscope was used to monitor the process.

6.1.1.6 Liquids

Glycerol and water were used to achieve different viscosities and densities.

6.1.2 Data sets

The three datasets collected with the measured densities can be seen in tables 1, 2 and 3.

liquid	water	5 % glycerol	10 % glycerol	40 % glycerol
measured density	0.99890 g/cm3	1.01014 g/cm3	1.02270 g/cm3	1.11240 g/cm3
observations	80	80	80	80

Table 1: D1 data set

liquid	plasma	serum 1	serum 2
measured density	1.01494 g/cm3	1.01668 g/cm3	1.01652 g/cm3
observations	50	50	50

Table 2: D2 data set

liquid	2.5 % glycerol	7.5 % glycerol	20 % glycerol
measured density	1.00330 g/cm3	1.01959 g/cm3	1.05526 g/cm3
observations	50	50	50

Table 3: D3 data set

6.1.3 D1: Water/ glycerol sample preparation

Various blends of water and glycerol were created to encompass a spectrum of compositions. Each mixture underwent thorough mixing to ensure homogeneity. These prepared mixtures were then categorized into distinct experimental groups, each corresponding to a specific water/glycerol ratio. The objective of the experiment was to span a range of concentrations, systematically exploring how differences in viscosity and density impact the pressure applied during pipetting.

At this stage, the primary goal was classification. Accordingly, the samples were distributed to investigate the potential scale for distinguishing between densities. For each mixture, 1000 μL of liquid was aspirated using a pipette, and the resulting liquid was subsequently weighed to estimate its density.

6.1.4 D2: Serum/ plasma sample preparation

Three different samples of serum/ plasma were prepared and used to collect data. These were provided by Thermo Fisher. The samples were thawed and for each mixture, 1000 μL of liquid was aspirated using a pipette, and the resulting liquid was subsequently weighed to estimate its density.

6.1.5 Pipetting Procedure

The D1 and D2 data sets were collected by the thesis worker. Prior to the main experiment, the Phadia 200 was calibrated according to the manufacturer's specifications. This calibration ensured accurate and consistent pipetting across all experimental conditions. The heat inside the instrument was turned of to minimize evaporation. For each water/ glycerol experimental group, 80 measurements of a fixed volume of 40 μL of the sample mixture was pipetted using the Phadia 200 machine. For each serum/ plasma group, 50 measurements were pipetted in the same manner. The wells were manually loaded into the machine. The machine's automated pipetting protocol was employed, and pressure curves were recorded in real-time during the dispensing process. Pressure curves generated during the pipetting process were recorded and stored for subsequent analysis. The wells were weighed before and after each pipetting and the results automatically put into an excel sheet, enabling calculations of actual pipetted volume. The complete pipetting process was monitored using a camera, ensuring subsequent verification that no errors occurred.

6.1.6 D3: Test data set

The D3 dataset was not gathered by the thesis worker, as this determination was deferred to a later stage following a shift in objectives. Instead, it was collected by a Thermo Fisher employee and provided to the thesis worker near the conclusion of the process. The test data was however collected using the same process. The same instrument, scale and pipette was used. The heat inside the instrument was turned off. The density was measured in the same manner. The only clear difference is that about a month passed before the test data was collected.

6.2 Data Preprocessing

To yield better results, data preprocessing and feature extraction are powerful tools. The goals are finding the most discriminating features and removing noise and unnecessary information.

6.2.1 Filtering and sampling by Phadia 200

In the pressure sensor, the data is filtered using a second order low pass Bessel filter with a cutoff frequency of 20 Hz. This is then sampled with a sampling frequency of 5 kHz and further filtered with a second order low pass Bessel filter with cutoff frequency 3 Hz.

6.2.2 Environment

The response data was collected to Excel and pressure data was extracted as log files from the Phadia 200. The pressure data was then initially visualized using Matlab. Scipy was used for the remaining programming work.

6.2.3 Outlier Rejection

Outliers can derive from different sources and hence contain more or less interesting information. Since they can greatly affect the results in a machine learning model, it is crucial that they are handled correctly. Outliers are in this work considered data points that greatly deviate from the data set. These are interpreted to originate from error in the data collection process and are assumed not to represent any relevant information for the analysis at hand, but instead disturb the model and yield incorrect and skewed results. The outlier rejection was done by visual inspection of the signals.

6.3 Feature Extraction

In this work, both the use of hand crafted features and using the entire pressure signals as predictor is attempted. This is due to interest from Thermo Fisher in understanding what basic aspects of the signals may serve as good predictors for the problem. The features are extracted and evaluated on the D1 data set.

The signal was evaluated and features were chosen by visually finding the discriminating factors for the different density clusters as these might serve as good features. The features used initially were

1. Mean value
2. Standard Deviation
3. Minimum value
4. Maximum value
5. Index of minimum value
6. Index of maximum value
7. Maximum positive derivative
8. Maximum negative derivative
9. Index of maximum positive derivative
10. Index of maximum negative derivative

The features were scaled using z-score normalization. This is a method used to rescale a distribution with a mean μ and standard deviation σ to a standard normal distribution with a mean of 0 and a standard deviation of 1. The z-score for a data point x is calculated using equation 11.

$$Z = \frac{x - \mu}{\sigma} \tag{17}$$

where Z is the z-score, x is the individual data point, μ is the mean of the distribution,σ is the standard deviation of the distribution.

The features were evaluated by looking at 10-fold cross-validation mean square error. The selection of features involved examining the consequences of excluding each feature. Ultimately, the reduced feature set yielding the best 10-fold cross-validation results was used.

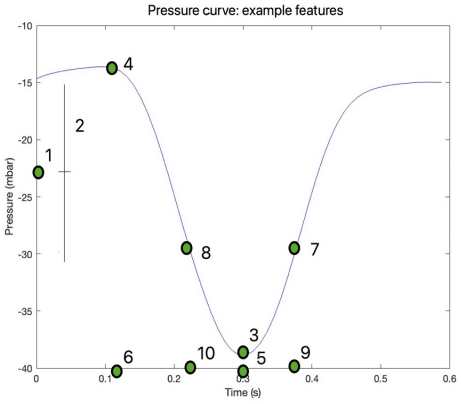


Figure 5: Example features

6.4 Prediction of volume and density

6.4.1 Partial Least Squares Regression

Partial Least Squares regression was performed using scikit function PLS-regression (Pedregosa *et al.* 2011) which follows the NIPALS algorithm (see section 5.2.1). The number of components to be used in the PLS model was optimized by minimization of the Mean Square Error as well as maximization of R^2 score. This was done in an iterative algorithm, evaluating the results of different choices. The PLS model was evaluated using 10-fold cross-validation.

First, the model was trained and tested using the entire signal as input. Examination of R^2 scores raised concerns, suggesting potential significant differences in the data. Consequently, a more in-depth analysis and visualization of the data was done to gain deeper insights. Since the model showed poor results when predicting the D2 and D3 sets, the effect of involving different parts of the D1, D2 and D3 data in training and testing of the model was explored. Secondly, the model was trained using the hand picked features as input. Here as well, different choices of data incorporation were investigated. Furthermore, the effect of log transforming the response variables was also examined.

6.4.2 Artificial Neural Network

A simple feedforward network was built using Keras sequential model (Martín Abadi *et al.* 2015). The Sequential model is a linear stack of layers, where you can add one layer at a time. The following architecture was used:

- Input layer: A fully connected layer with a reLU activation function and a varied number of nodes.
- Hidden layer: A fully connected layer with half the nodes compared to the first layer and a reLU activation function.
- Output layer: Two output nodes (yielding two predictions) and a linear activation function (predicting a continuous response).

Different batch sizes, epochs and numbers of nodes were tested for each training setup and the ones yielding the lowest MSE and highest R^2 scores were used in the model. The same choices of data in traning as for the PLS were investigated. The use of ANN was mainly investigated in order to evaluate the possibility of achieving higher R^2 scores and to inspect whether the residuals would distribute differently than for the PLS models.

7 Results

7.1 Data Preprocessing

All data collected in D1 data set is presented in figure 6.

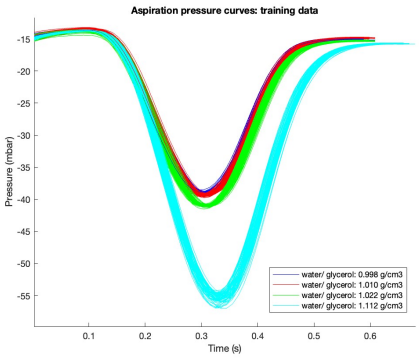


Figure 6: Aspiration pressure curves: D1

All data collected in D2 data set is presented in figure 7.

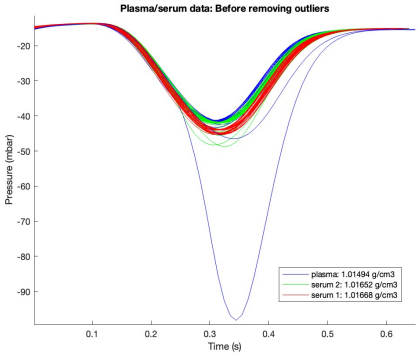


Figure 7: Aspiration pressure curves: D2

All data collected in D3 data set is presented in figure 8.

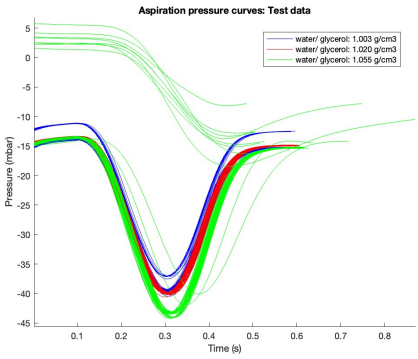
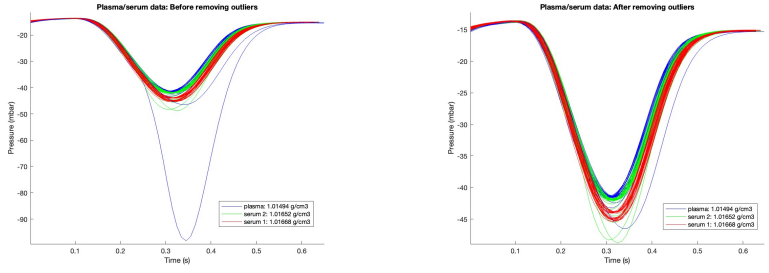


Figure 8: Aspiration pressure curves: D3

7.1.1 Outlier Rejection

One pressure recording in the D2 set is removed. This outlier (see figure 9) comes from a clog in the pipette and its value does not represent information on volume or density, but rather the clog caused by lumps in the liquid. It is therefore considered an incorrect measurement for the use in this study.

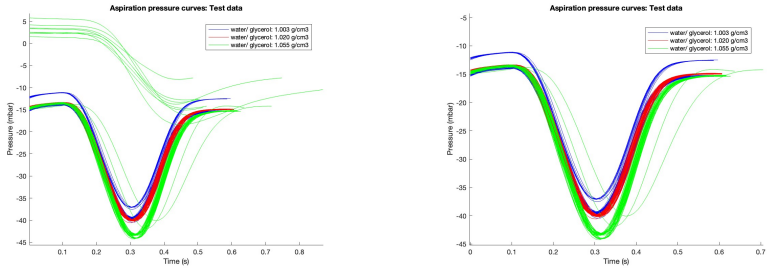


(a) Before removing outliers

(b) After removing outliers

Figure 9: D2 set: before and after removing outliers

The D3 data set has 8 outliers in the highest density measurements (see figure 10), these seem to come from some issue in the data collection as they differ from the rest of the curves to a great extent. There are some measurements that deviate slightly, these are kept as they may represent some true variation in the data.



(a) Before removing outliers (b) After removing outliers

Figure 10: D3: before and after removing outliers

Mean pressure curves for D1 and D2 data sets after removing outliers are presented in figure 11. Here, lower density is consistently related to higher mean pressure.

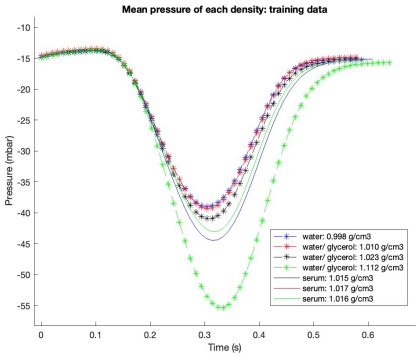


Figure 11: Mean pressure curves for D1 and D2 data sets, asterisk lines mark D1 set.

Mean pressure curves for D1 and D3 data sets after removing outliers are presented in figure 12. This comparison shows an unexpected result, as the lowest density is not related to the highest mean pressure.

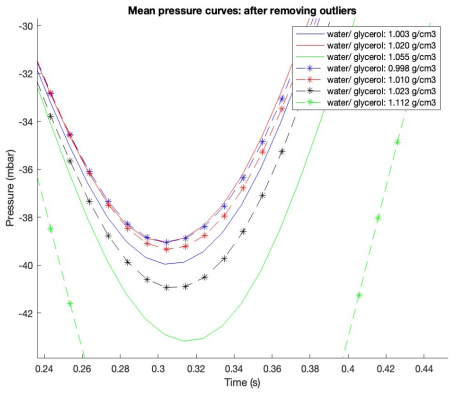


Figure 12: Mean pressure curve for D1 and D3 data sets, asterisk lines mark D1 set

7.2 Feature Extraction

In this section, the feature set yielding the best 10-fold cross-validation result on the D1 data set is presented. The reduced feature composition yielding the best results was

- Mean value
- Standard deviation
- Maximum absolute pressure
- Indices of maximum absolute pressure
- Maximum negative derivative
- Indices of maximum derivative

with the average errors of 10 iterations presented in table 4 and table 5.

volume	RMSE	MAE	R2
all features	0.6991	0.5402	0.9607
reduced set	0.6805	0.5309	0.9609

Table 4: Volume prediction errors: all features incorporated vs reduced feature set

density	RMSE	MAE	R2
all features	0.0049	0.0045	0.9863
reduced set	0.0046	0.0039	0.9871

Table 5: Density prediction errors: all features incorporated vs reduced feature set

Figure 13 illustrates the relationship between volume and density across all 3 data sets. D1 demonstrates expected correlation between volume and density. D3 however shows a completely different pattern. D2 looks similar to D1, however due to the narrow density range it is not possible to draw conclusions from this.

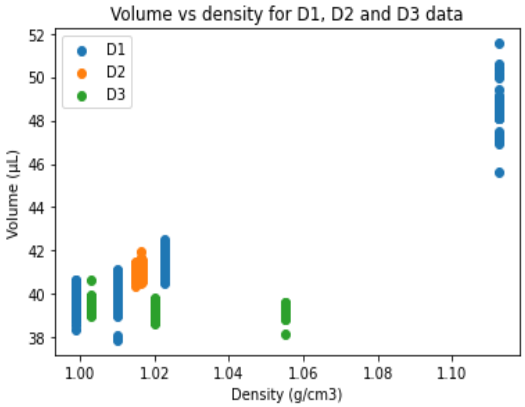
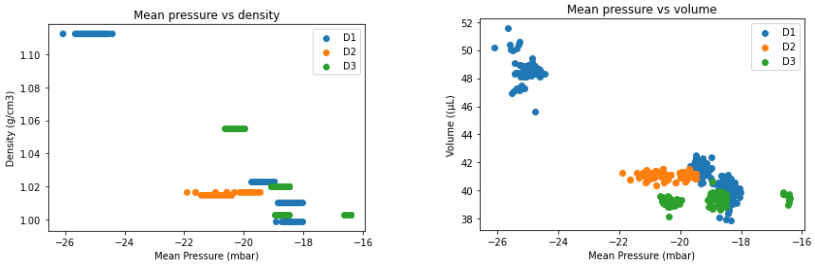


Figure 13: Volume vs density for D1, D2 and D3 data sets

All features for each dataset are visually represented in figures 14 through 19. The features are plotted against density in the (a) figures and volume in the (b) figures. Examining all features, patterns in the D1 dataset, from which the features were derived, appear discernible. However, this consistency is not observed in the other two datasets.

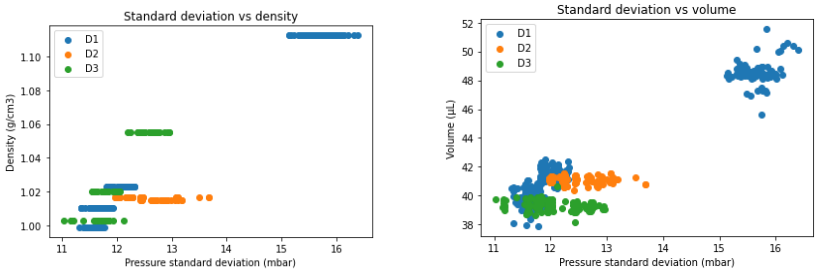
The mean pressure, in figure 14(a), shows some similarity between D1 and D3, however D2 shows a dissimilar pattern. In 14(b), D1 exhibits dissimilar behavior compared to both the D2 and D3 datasets.



(a) Density (b) Volume

Figure 14: Mean pressure vs volume and density

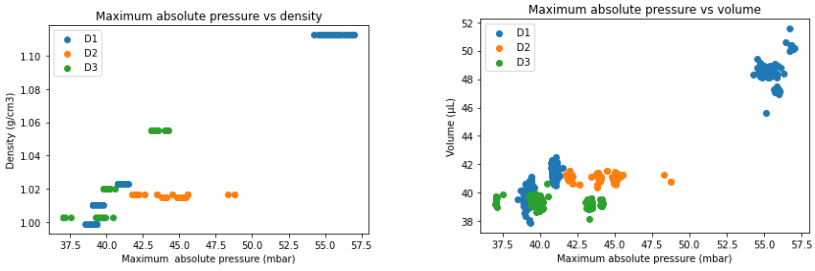
The standard deviation, presented in figures 15(a) and 15(b) behave comparable to the mean pressure. In 15(a), there is some similarity between D1 and D3 but not D2. In 15(b), D1 exhibits dissimilar behavior compared to both the D2 and D3 datasets.



(a) Density (b) Volume

Figure 15: Standard deviation vs volume and density

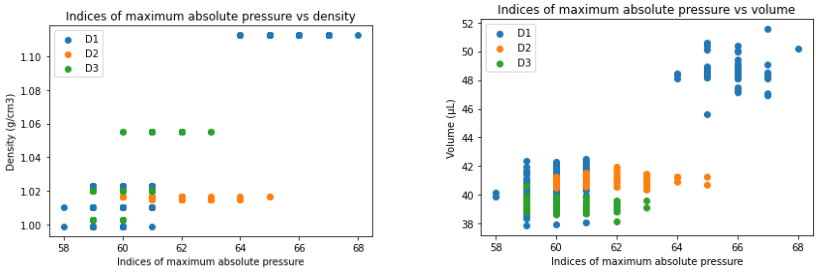
This pattern seems to continue in figures 16(a) and 16(b) with the only clear correlation between D1 and D3 in 16(a).



(a) Density (b) Volume

Figure 16: Maximum absolute pressure vs volume and density

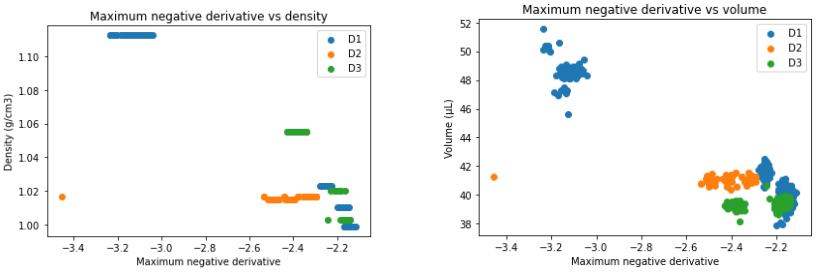
In figures 17(a) and 17(b), D2 still does not show any clear correlation with any of the other data sets. In this case, the only resemblance is still between D1 and D2 in figure 17(a).



(a) Density (b) Volume

Figure 17: Indices of maximum absolute pressure vs volume and density

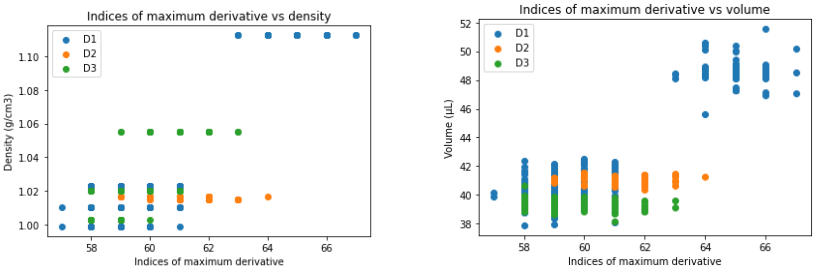
In figures 18(a) and 18(b), the relationship in D1 appears evident, while in D2 it is not apparent at all, and in D3, it is only noticeable for the density measurements.



(a) Density (b) Volume

Figure 18: Maximum negative derivative vs volume and density

In figures 19(a) and 19(b), this trend persists, although the correlation appears weaker in D1. D2 exhibits some resemblance to D1 in the density measurements, but neither D2 nor D3 show significant similarity in the volume measurements.



(a) Density (b) Volume

Figure 19: Indices of maximum derivative vs volume and density

7.3 Prediction of volume and density

The results of predicting volume and density are presented in this section. In the first part using PLS with feature input and in the second part using PLS and ANN and the entire signal as input.

7.3.1 PLS: Feature input

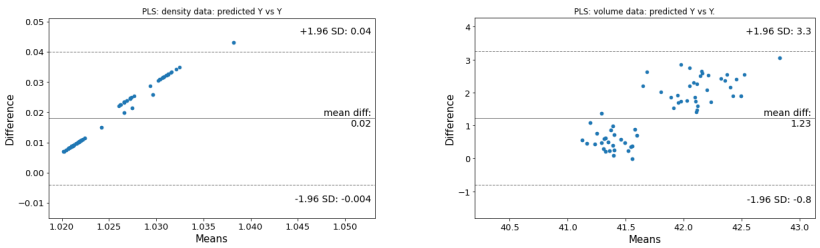
The results of attempting PLS regression using the feature set as input is presented in this section. Using features taken from the D1 set yielded the volume prediction errors presented in table 6 and density prediction errors in table 7. The tukey mean difference plots of the predictions can be seen in figures 20 and 21.

test set	RMSE	MAE	R^2
D1 (cross validation)	0.6613	0.5070	0.9631
D2	1.6036	0.8863	-42.19
D3	2.1431	1.1014	-41.73

Table 6: PLS volume predictions: D1 feature input

test set	RMSE	MAE	R^2
D1 (cross validation)	0.0040	0.0033	0.9906
D2	0.0212	0.0111	-660.3
D3	0.0125	0.0089	0.6425

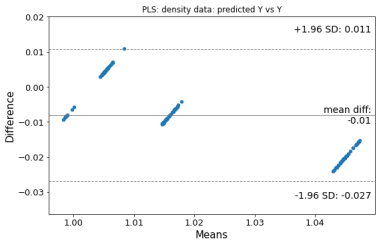
Table 7: PLS density predictions: D1 feature input



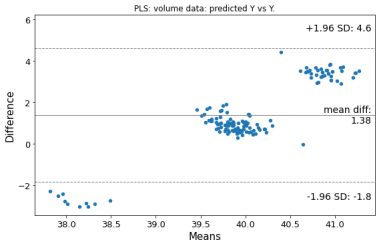
(a) Density

(b) Volume

Figure 20: Tukey mean-difference: training PLS model on D1 feature set and predicting D2 data set



(a) Density



(b) Volume

Figure 21: Tukey mean-difference: training PLS model on D1 feature set and predicting D3 data set

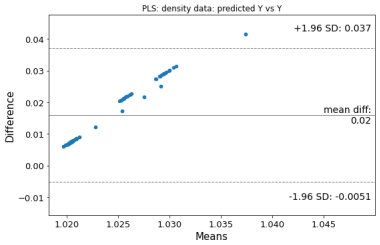
Using features taken from D1, D2 and D3 sets combined yielded the volume prediction errors presented in 8 and density prediction errors in 9. The tukey mean difference plots of the predictions can be seen in figures 22 and 23.

test set	RMSE	MAE	R^2
D1 (cross validation)	1.0727	0.8254	0.8757
D2	0.9892	0.5208	-15.4369
D3	1.1674	0.6136	-11.6811

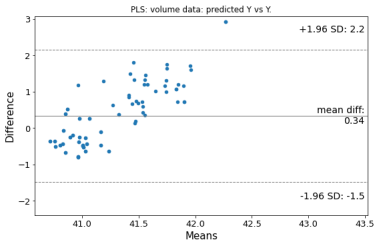
Table 8: PLS volume prediction: combined D1, D2 and D3 feature input

test set	RMSE	MAE	R^2
D1 (cross validation)	0.0110	0.0083	0.9051
D2	0.0192	0.0085	-540.5328
D3	0.0160	0.0095	0.4141

Table 9: PLS density predictions: combined D1, D2 and D3 feature input

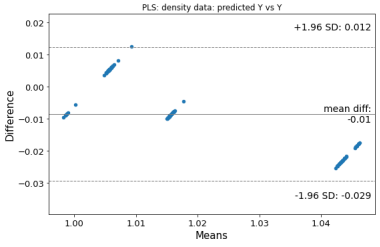


(a) Density

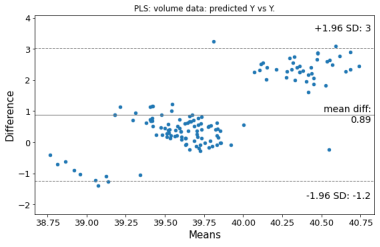


(b) Volume

Figure 22: Tukey mean-difference: training PLS model on combined D1,D2 and D3 feature set and predicting D2 data set



(a) Density



(b) Volume

Figure 23: Tukey mean-difference: training PLS model on combined D1,D2 and D3 feature set and predicting D3 data set

7.3.2 PLS vs ANN: Raw data input

In this section, the results when using raw input data when training a PLS model and an ANN model is presented. The ANN hyperparameters were optimized for each training input, these are presented in table 10.

training input	nodes in layer 1	nodes in layer 2	batch size	epochs
D1	10	5	32	10
D1, D2 and D3	12	6	64	20

Table 10: ANN hyperparameters

Training the models on the D1 data set and predicting the D2 data set

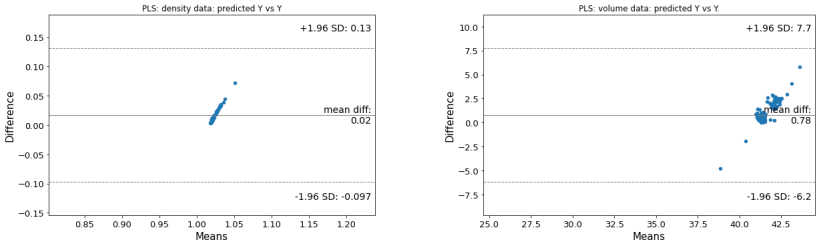
yielded the volume prediction errors presented in 11 and density prediction errors in 12. The tukey mean difference plots of the predictions can be seen for PLS in figure 24 and for ANN in figure 25.

volume	RMSE	MAE	R^2
PLS cross validation	0.6613	0.5070	0.9769
ANN cross validation	0.5282	0.5020	-50.14
PLS test	3.6311	0.7914	-170
ANN test	0.6126	1.7365	-0.5151

Table 11: PLS vs ANN volume predictions: Training models on D1 data set and predicting D2 data set.

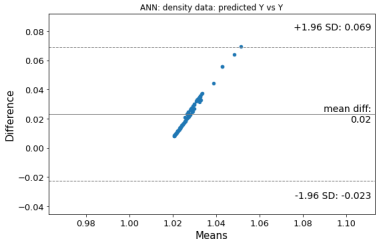
density	RMSE	MAE	R^2
PLS cross validation	0.0040	0.0033	0.9906
ANN cross validation	0.1892	0.1722	0.8066
PLS test	0.0605	0.0121	-6617
ANN test	0.7316	0.0251	-1751

Table 12: PLS vs ANN density predictions: Training models on D1 data set and predicting D2 data set.

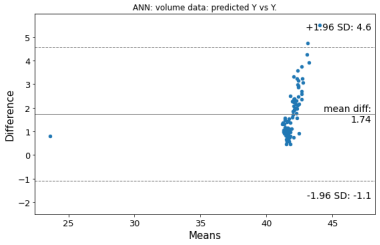


(a) Density (b) Volume

Figure 24: Tukey mean-difference: Training PLS model on D1 data set and predicting D2 data set.



(a) Density



(b) Volume

Figure 25: Tukey mean-difference: Training ANN model on D1 data set and predicting D2 data set.

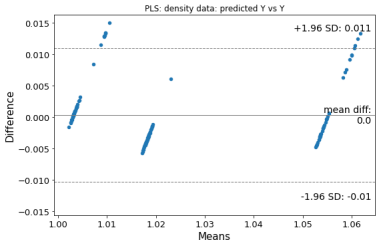
Training the models on the D1 data set and predicting the D3 data set yielded the volume prediction errors presented in 13 and density prediction errors in 14. The tukey mean difference plots of the predictions can be seen for PLS in figure 26 and for ANN in figure 27.

volume	RMSE	MAE	R^2
PLS cross validation	0.6613	0.5070	0.9769
ANN cross validation	0.5282	0.5020	-50.14
PLS test	2.1135	1.1786	-40.56
ANN test	1.792	1.140	-28.88

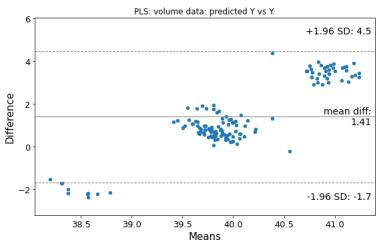
Table 13: PLS vs ANN volume predictions: Training models on D1 data set and predicting D3 data set.

density	RMSE	MAE	R^2
PLS cross validation	0.0040	0.0033	0.9906
ANN cross validation	0.1892	0.1722	0.8066
PLS test	0.0093	0.0051	0.80
ANN test	10.294	7.211	0.7595

Table 14: PLS vs ANN density predictions: Training models on D1 data set and predicting D3 data set.

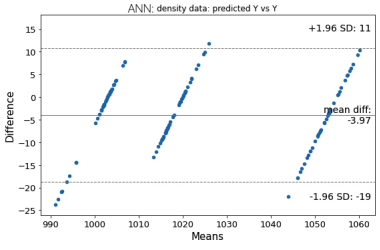


(a) Density

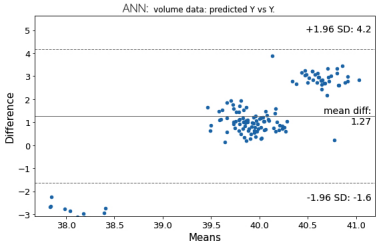


(b) Volume

Figure 26: Tukey mean-difference: Training PLS model on D1 data set and predicting D3 data set.



(a) Density



(b) Volume

Figure 27: Tukey mean-difference: Training ANN model on D1 data set and predicting D3 data set.

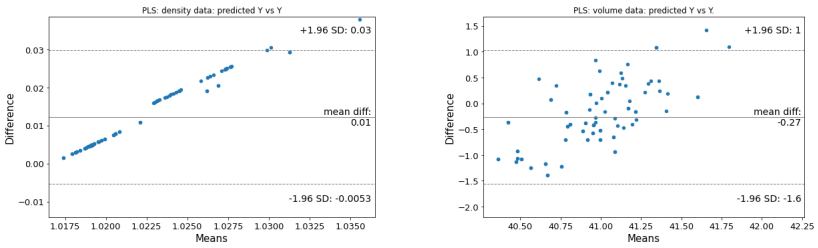
Training the models on the D1, D2 and D3 data set and predicting the D2 data set yielded the volume prediction errors presented in 15 and density prediction errors in 16. The tukey mean difference plots of the predictions can be seen for PLS in figure 28 and for ANN in figure 29.

volume	RMSE	MAE	R^2
PLS cross validation	0.9450	0.7655	0.9046
ANN cross validation	1.1481	0.8751	0.9039
PLS test	0.7548	0.5191	-8.5692
ANN test	0.4901	0.8060	0.3622

Table 15: PLS vs ANN volume predictions: Training models on D1, D2 and D3 data sets combined and predicting D2 data set.

density	RMSE	MAE	R^2
PLS cross validation	0.0092	0.0066	0.9318
ANN cross validation	0.0126	0.0086	0.8915
PLS test	0.0151	0.0082	0.9318
ANN test	0.4986	0.0122	-334.3

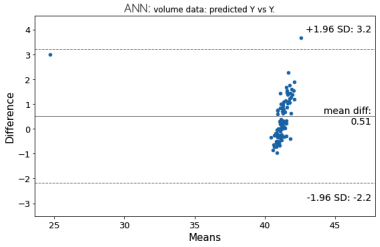
Table 16: PLS vs ANN density predictions: Training models on D1, D2 and D3 data sets combined and predicting D2 data set.



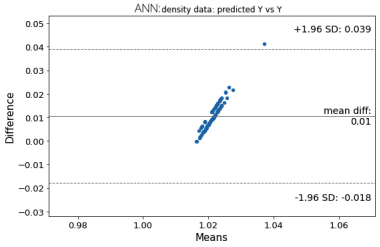
(a) Density

(b) Volume

Figure 28: Tukey mean-difference: Training PLS model on D1, D2 and D3 data sets combined and predicting D2 data set.



(a) Density



(b) Volume

Figure 29: Tukey mean-difference: Training ANN model on D1, D2 and D3 data sets combined and predicting D2 data set.

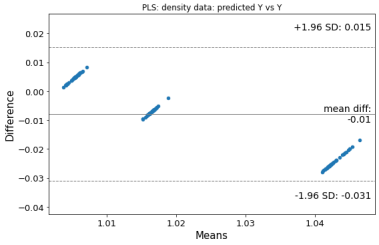
Training the models on the D1, D2 and D3 data set and predicting the D3 data set yielded the volume prediction errors presented in 17 and density prediction errors in 18. The tukey mean difference plots of the predictions can be seen for PLS in figure 30 and for ANN in figure 31.

volume	RMSE	MAE	R^2
PLS cross validation	0.9450	0.7655	0.9046
ANN cross validation	1.1477	0.8290	-11.25
PLS test	0.9660	0.5690	0.9642
ANN test	0.6100	1.9019	-45.95

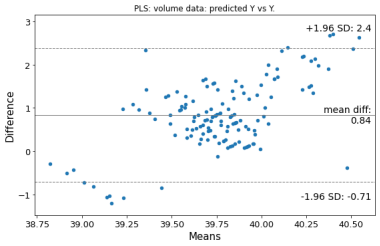
Table 17: PLS vs ANN volume predictions: Training models on D1, D2 and D3 data sets combined and predicting D3 data set.

density	RMSE	MAE	R^2
PLS cross validation	0.0092	0.0066	0.9318
ANN cross validation	0.0126	0.0086	0.8915
PLS test	0.0141	0.0069	0.5432
ANN test	0.2481	0.0080	0.7181

Table 18: PLS vs ANN density predictions: Training models on D1, D2 and D3 data sets combined and predicting D3 data set.

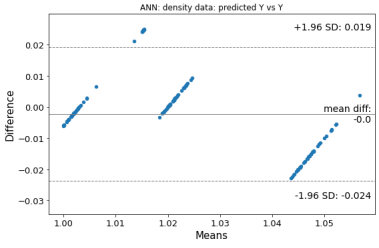


(a) Density

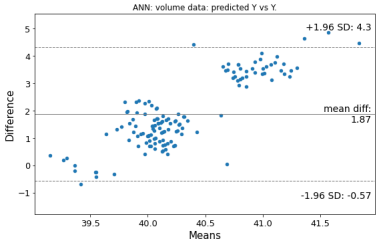


(b) Volume

Figure 30: Tukey mean-difference: Training PLS model on D1, D2 and D3 data sets combined and predicting D3 data set.



(a) Density



(b) Volume

Figure 31: Tukey mean-difference: Training ANN model on D1, D2 and D3 data sets combined and predicting D3 data set.

8 Discussion

There are two main aspects to consider when evaluating the feasibility of the methods in this thesis. The first is how well the model can predict data that has been incorporated in training the model. The second is the ability to generalize to and predict new, unseen data. It is expected that the model predicts data belonging to the same data set best, and here the issue lies in how representative the data is as well as how much data needs to be collected to ensure good results. When predicting new, unseen data, other factors may play a significant role. Generalizing the model to different instruments, handlers and new measurements is subject to many different insecurities and differences.

The only model that yielded sufficiently good R^2 scores in both volume and density predictions was the PLS model trained on all data sets (see tables 30 and 18). This suggests that there perhaps are some differences between data sets that are difficult to generalize to. This nonetheless, shows that the model can in fact predict data that has been incorporated in the training of it.

A fixed volume of 40 µL was used for the pipettings, and this or 90 µL is what is usually pipetted for the most common tests employed by the Phadia 200. In relation to this, the resulting RMSE scores were satisfactory for most cases presented. While the RMSE scores were acceptable, the R^2 scores observed across all models but the one presented in tables 30 and 18 were significantly below the desired level, revealing a notable shortfall in meeting the specified objectives. This underscores a substantial need for improvement in the employed methodology. The negative R^2 scores might indicate that there is such high variability in the data that the model is unable to capture underlying patterns. Significant differences in the data sets are clear in the visualizations of features (see figures 14- 19), correlations between volume and density (see figure 13) and mean pressure curves (see figure 12). It is probable that the volume predictions are simply some mean value, and that the differentiation of these small volumes is not possible. Notably, slightly more favorable results were obtained for density predictions. This could be attributed to the increased volume of input data available for each density response, potentially contributing to a better explanation of its variance. It could also be the case that the relationship between volume and pressure is more difficult to capture or perhaps not as clear.

The ANN performed equally as bad as PLS regarding R^2 score and the MSE scores were worse notably worse for many models. The difference in cross validation and test suggest over training and several different techniques were tested, without yielding better results. The addition of a dropout layer, experimenting with the model complexity (i.e. hidden layers), batches and epochs and the use of metrics when optimizing the model. The model might be more sensitive to learning the noise in the signal than the PLS model. ANN had more similar prediction residuals, when comparing high and low density predictions.

Mean pressure curve of each density cluster can be seen in figure 12. Here, it is clear that the either the assumed densities are incorrect or their effect on the pressure curves is ambiguous. The mean curve with the highest values is not that with the lowest measured density. The accuracy of the algorithm can not be expected to be higher than this inseparable difference, at least when features derived from these aspects are used. It suggests that there might have been either a calibration difference or a handling difference for the different sets as some of the curves were collected in a later stage, by a different technician. It could also be due to inaccuracies in density measurements.

Looking at figures 14 through 19, the features that effectively captured diverse volumes and densities in the training data (D1) exhibit dissimilar patterns in the test data (D2 and D3). While this discrepancy could be attributed to suboptimal feature choices, a noteworthy aspect is that these features demonstrated efficacy with the training data, as is apparent in figures 14 through 19. The densities appear to have slightly more consistent impact on pressure across the sets; perhaps further explaining the better density prediction results. Nevertheless, there are still notable differences in the patterns.

The models consistently demonstrated superior performance for values in the middle range compared to peripheral values across all cases. Examination of mean-difference plots (e.g. figure 22) suggests the need for a transformation. Log transformation was tested and it did not significantly improve results. However, it is observed that differences vary with choice of training and test data. A notable contrast is evident when comparing figures 28 and 30, revealing differences distributed in opposing patterns.

This suggests differing relationships in the sets used for training and test as training with the different sets makes the model predict high and low values in opposed patterns. One reason of the center values being predicted more accurately could be that there simply is more data available in this region. When the data collection was done, a balanced data set meant training on groups of data of similar size. Since the data now has been used in regression, also the choice of data points is of importance. Since more data was recorded in the center region, the model might perform better here. In any case, while the training dataset exhibited characteristics that suggested the viability of the methods, the consistency of the relationships observed in other datasets collected at different times remains uncertain.

There are several things to take into consideration when constructing a data set. How the collection process and methods affect the data and its validity as well as what data is best suited for the problem at hand are things that directly affect results and model generalization abilities. Deciding on what data to collect is an important step and different choices may be advantageous depending on the application. For a classification task, well clustered data points are favourable. For a regression task, accuracy and consistency in the data might be even more important. Furthermore, machine learning algorithms tend to be sensitive to imbalanced data. When trained on an imbalanced data set, the algorithm might get biased towards the majority group of data. It learns to favor the most common outcome. It is easy to misinterpret the quality of the results since this might give good accuracy but bad overall performance. A balanced dataset is therefore of great importance (Frederik Hvilshøj 2023).

Errors originating from the data collection process may occur due to a variety of reasons. Random errors from e.g. inconsistent mixing of liquids or incorrect measurements or systematic errors from e.g. miscalibrated machinery are some examples. When generating data one can argue that it is favorable to randomize the error by e.g. mixing the same solution several times in order to avoid systematic error in the data. Systematic error can be a big problem since it skews the data, yielding data centered around an incorrect mean value. Random error can yield better precision, especially when working with big data sets, as the data points tend to center around a more correct mean value. On the other hand, the data will have more variance (Bhandari, P. 2023).

The assumption when collecting the data was that the problem at hand

was a classification problem. It was therefore decided that low variance was favoured over high accuracy, and the data collection process was designed with this in mind. Small systematic errors were not expected to affect the ability of clustering data points. Despite this, as the aims changed after the collection process, and the data was instead used for a regression problem, it is clear that this approach was not ideal. The algorithms issues with accuracy could have been smaller if the error was distributed more randomly. Furthermore, the data was collected in big density clusters. While this approach was logical for a classification problem, it lacks applicability for a regression problem. In this context, a more diverse set of densities and spread measurements would be more suitable. Since the data was collected in order to investigate the scale in which the data points would be discernible, meaning clusters with decreasing density differences were collected, the data set was not balanced for a regression problem. While the data within each set was collected as clusters with equal size, these were not spread equally. A better balanced set for regression would incorporate more evenly spread observations. Furthermore, the D1 and D3 sets were not collected in equally as big clusters. This was due to the fact that these were initially meant to be used only for testing the model. When incorporating this data into the training, this too affects the balance in the training data.

This all further raises concerns of the accuracy in measurements. There are several measurement uncertainties involved in the experiment. The pipette used for pipetting the liquid in the density measurement, the effect of higher viscosity (and hence density) on the pipetting procedure and the scale accuracy all affect the density measurements. The liquid has been mixed once and density measured once for each density. This process makes the risk of measurement errors significant. Furthermore, the volume measurements could be inaccurate due to the scale, differences in temperature and process time affecting the risk for evaporation. While the thesis worker and the technician collecting the test set used the same methods in theory, there might be uncontrolled differences. One obvious aspect is the fact that the D3 set was collected about a month after the D1 and D2 sets.

8.1 Further Development

Given the sensitivity of the analysis to various factors, additional measurements must be undertaken and assessed to assess the feasibility of the methods. It is suggested that less clustered data is collected, i.e. more dens-

ities with fewer samples per density. Also data that better represents the entire spectrum of densities. Furthermore, collecting the data in a way that introduces random error rather than systematic. This can be done by mixing liquids several times, measuring the density for each mix and measuring the density using a hydrometer.

8.2 Ethical Considerations

It is essential to always consider ethical considerations and potential risks when working with human samples. The decision to utilize water/glycerol mixes instead of exclusively relying on human samples was driven partially by convenience and also stemmed from ethical considerations. Extracting the required amount of human blood for this study was deemed ethically impractical.

There were however a couple of human samples used for one of the data sets in this study. These were handled with caution, the appropriate protection gear and in a separate space of the laboratory. The samples used were provided by Thermo Fisher and were old samples that had been stored for some time. These samples did not have any specific preallocated purpose and were therefore utilized for this study when the aims were changed to further add to the analysis. The aim for using them was to investigate whether it was possible to train a model on water/ glycerol data and still use human samples in testing. Also, investigating if they could perhaps add useful information or would just lead to more noise if used in training.

Regarding the methods, there are ethical aspects to consider regarding these as well. As the diagnostic tests may be affected by e.g. hidden errors and wrongly estimated volumes or densities this poses a patient risk. Since these tests are sensitive to small differences in volume, it is crucial that the methods are precise and reliable in order to avoid misdiagnosis.

9 Conclusion

The primary objective was not successful as the pressure sensor in the pipette was unable to detect the errors of interest. Regarding the secondary objective, based on the data collected in this work, it is not possible to say that there is a clear enough relationship between the pressure sensor data and the volume and density.

An approach using a PLS regression model and an ANN model has been presented. PLS seems to have better generalization abilities, but the R^2 score remains unsatisfactory for most models. Some outliers were identified and dealt with but besides that no apparent issues were found in the data. The model still has issues generalizing to new data. The conclusion is that there might be two different sources for these issues. There might not be a clear enough relationship between the response variables and pressure data. Especially comparing the different data sets this correlation seems unclear. Perhaps the relationship is not similar enough in different density clusters, as the model only performed satisfactory when all densities were included in training. The second possibility is that the issues stem from measurement insecurities or calibration differences.

The data collection process is thought to have affected the results greatly. Using a more appropriate experimental design might yield better results with higher accuracy. It is likely that with a sufficient amount of well-representative data, calibration of a model could be possible. Nonetheless, the challenge might lie in addressing variations in instrument calibrations. Currently, it is challenging to determine the precision of such a model, and it may be necessary to calibrate the model individually for each instrument, and to re calibrate it routinely. This is clear from the differences in the data sets collected by the thesis worker and the one later collected by the Thermo Fisher employee, which showed significant differences despite having been collected in the same manner using the same instruments.

References

Bhandari, P. (2023). *Random vs. Systematic Error | Definition Examples. Scribbr.* URL: https://www.scribbr.com/methodology/random-vs-systematic-error/ (visited on 17/10/2023).

Frederik Hvilshøj (2023). *Introduction to Balanced and Imbalanced Datasets in Machine Learning.* URL: https://encord.com/blog/an-introduction-to-balanced-and-imbalanced-datasets-in-machine-learning/ (visited on 17/10/2023).

Geladi, Paul and Bruce R. Kowalski (1986). 'Partial least-squares regression: a tutorial'. In: *Analytica Chimica Acta* 185, pp. 1–17. ISSN: 0003-2670. DOI: https://doi.org/10.1016/0003-2670(86)80028-9. URL: https://www.sciencedirect.com/science/article/pii/0003267086800289.

Goodfellow, Ian, Yoshua Bengio and Aaron Courville (2016). *Deep Learning.* http://www.deeplearningbook.org. MIT Press.

Johnsson, K. and M. Kuhn (2013). *Applied Predictive Modeling.* Springer New York Heidelberg Dordrecht London. ISBN: 978-1-4614-6848-6. DOI: 10.1007/978-1-4614-6849-3.

Martín Abadi *et al.* (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.* Software available from tensorflow.org. URL: https://www.tensorflow.org/.

Martin Danielson, Thermo Fisher Scientific (n.d.). *New Phadia 200 Advances In-vitro Diagnostics in Europe for Allergy and Autoimmune Conditions.* URL: https://www.prnewswire.com/news-releases/new-phadia-200-advances-in-vitro-diagnostics-in-europe-for-allergy-and-autoimmune-conditions-300676762.html.

Nguembang Fadja, Arnaud, Evelina Lamma and Fabrizio Riguzzi (Dec. 2018). 'Vision Inspection with Neural Networks'. In.

Pedregosa, F. *et al.* (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

El-Salam, Moawad El-Fallah Abd (Jan. 2014). 'A Note on Partial Least Squares Regression for Multicollinearity (A Comparative Study)'. In: *International Journal of Applied Science and Technology* Vol. 4.1, pp. 164–165.

Sanchez, G. and E. Marzban (2020). 'All Models Are Wrong: Concepts of Statistical Learning.' In: URL: https://allmodelsarewrong.github.io (visited on 08/01/2024).

scikit-learn developers (2023). *Cross decomposition*. URL: https://scikit-learn.org/stable/modules/cross_decomposition.html#plsregression (visited on 25/12/2023).

Siddharth Sharma Simone Sharma, Anidhya Athaiya (2020). 'ACTIVATION FUNCTIONS IN NEURAL NETWORKS'. In: *International Journal of Engineering Applied Sciences and Technology* Vol. 4 (12), pp. 310–316. URL: http://www.ijeast.com.

Wegelin, Jacob A. (Mar. 2000). 'A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case'. In: *Technical report no 371, Department of statistics, University of Washington*, pp. 29–31.

Wold, H. (Mar. 1966). 'Estimation of Principal Components and Related Models by Iterative Least Squares.' In: *P Krishnaiah (ed.), "Multivariate Analyses," Academic Press, New York.*, pp. 391–420.

Zou, Jinming, Yi Han and Sung-Sau So (2009). 'Overview of Artificial Neural Networks'. In: *Artificial Neural Networks: Methods and Applications*. Ed. by David J. Livingstone. Totowa, NJ: Humana Press, pp. 14–22. ISBN: 978-1-60327-101-1. DOI: 10.1007/978-1-60327-101-1_2. URL: https://doi.org/10.1007/978-1-60327-101-1_2.