Using Machine Learning to Analyse Biscuit Structural Properties

M.T.D.R.Dolan
School of Chemistry, University of Bristol.
(Dated: April 30, 2024)

This project has been commissioned by McVitie's biscuit company to analyse various structural properties regarding several types of biscuit produced by them. It uses 'dunking' data provided for three biscuits, along with microscopy measurements of pore radius, and time-resolved measurements for each biscuit type. It will conclude, using machine learning methods and comparison to the Washburn equation, that results are consistent with capillary flow action for the absorption of tea.

1 Introduction

The aim of this task is to investigate the structural properties of different McVitie's biscuits when dunked in tea.

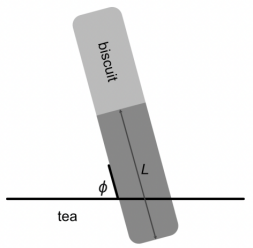


FIG. 1: Experimental Set-up [1]

The Washburn equation for capillary flow suggests the following relationship for the tea soaking up the biscuit:

$$L = \sqrt{\frac{\gamma r t \cos\phi}{2\eta}} \tag{1}$$

Where L is the length the tea soaks up the biscuit in time t, r is the radius of pore capillaries in the biscuit, ϕ the angle between the biscuit and tea, γ the tea surface tension, and η the tea dynamic viscosity. This can also be rearranged to:

$$L^2 = \frac{\gamma r \cos\phi}{2\eta} t \tag{2}$$

Errors have been propagated via partial differentiation so that:

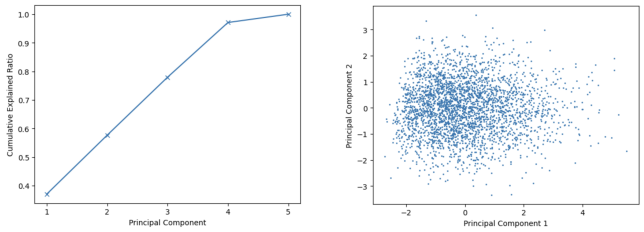
$$(\Delta L)^2 = (\Delta\gamma \frac{dL}{d\gamma} + \Delta r \frac{dL}{dr} + \Delta\phi \frac{dL}{d\phi} + ...)^2 \tag{3}$$

Three datasets have been provided. One containing 3000 recordings of one of three named biscuits being dunked, with all attributes except pore radius being named, a smaller subsection of these biscuits in which the pore radius was microscopically measured, and an additional three datasets which measured the tea flow rate through each of the three types of biscuit (although these are unidentified).

2 Analysis and Discussion

2.1 Dunk Data

The three biscuits were identified as Digestive, Hobnob or Rich Tea, with there being 1000 of each. A principle component analysis (PCA) was initially conducted.



(a) The cumulative variance (b) The first two against each other

FIG. 2: Exploration of the principal components for dunk data

Figure 2a shows that each principal component, except the last one, equates to a similar amount of explained variance. This means that although there could be some noise reduction in reducing the data to four principal components instead of four features, it is unlikely to make significant enough of a difference to warrant it.

This is echoed in figure 2b in which the first two principal components are plotted against each other. Any clustering in the data would indicated that biscuits would be able to be identified based on these first two dimensions, but there is none.

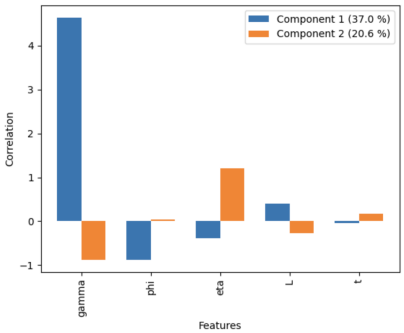


FIG. 3: The feature correlation of the first two principal components

Figure 3 shows that the feature with the largest correlation to the first two principal component makeups is the tea surface tension, this is peculiar considering that this does not depend on the biscuit chosen, although this is positive in the first component and negative and in the second. Interestingly, every feature has the reverse correlation in the second principal component as it does in the first.

The dunk data was then used to train two machine learning algorithms to test how well it could identify biscuits. Based on the PCA, the features themselves were just scaled to feed the model without any dimensionality reduction techniques put into place. Since this is classified data, a Random Forest (RF) and Special Vector Classification (SVC) method was tested. The total data was split into 75% training and 25% test data, stratifying so that each contains equal proportions of each biscuit.

model	biscuit	precision	recall	f1-score	accuracy
RF	Digestive	0.85	0.88	0.87	80.5%
	Hobnob	0.72	0.69	0.70	
	Rich Tea	0.83	0.84	0.84	
SVC	Digestive	0.94	0.89	0.81	86.1%
	Hobnob	0.83	0.78	0.81	
	Rich Tea	0.82	0.92	0.86	

TABLE I: Classification report on test data

The SVC was overall slightly more accurate than the RF model, although the it had a better f1-score for identifying Digestive biscuits. This could be because RF is more sensitive to noise, or it could just be due to the fundamental randomness within the model.

Figure 4 shows the feature importance for both models, with both of them having an equal order of feature importance. It should be noted that levels of importance cannot be compared between models as the RF is showing the percentage feature importance, whilst the SVC is showing the permutation importance which does not add up to 1.

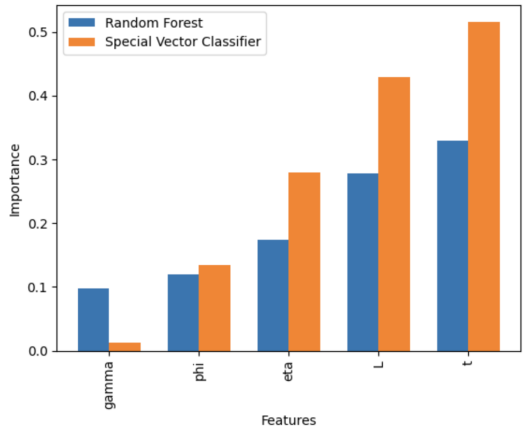
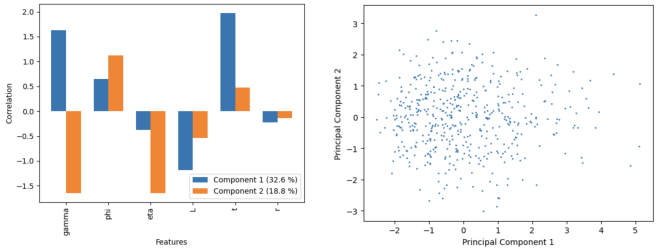


FIG. 4: feature importance for the models

2.2 Pore Radius



(a) The feature correlation (b) The first two against each other

FIG. 5: Exploration of the principal components for pore data

With the added pore radius information, an additional PCA was run. As shown in figure 5, the radius does not account for much of the variance and again no clustering can be seen when the first two principal components are put against each other.

Figure 6 shows the binned radius data, with three clear spikes in the data accounting for the three biscuit types. This is further demonstrated when the pore radius is shown just for each biscuit individually. Variances for the pore radius in each of the biscuits is shown in table II, each of them being of the same order and so they have similar spreads.

Biscuit	Mean (nm)	Variance
Digestive	804	3.36×10^{-6}
Hobnob	496	6.40×10^{-6}
Rich Tea	304	2.21×10^{-6}

TABLE II: The data spreads for the biscuit radii

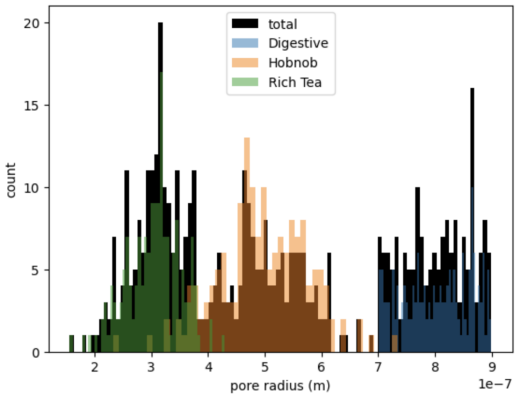


FIG. 6: Recorded pore radii

This data was again split into a train/test ration of 75/25 in order to compare radius predictions made by both an ML method and the Washburn equation. As this is not a classification problem, an SVC or traditional RF model cannot be

used, therefore a random forest regressor is used. The results of this are shown in figure 7, with the Washburn predictions being much more accurate. It has both a lower mean squared error and, in contrast to the RF regressor, correctly predicts the three different spikes (biscuit type was not used as a factor in the model training so that both would have equal input data).

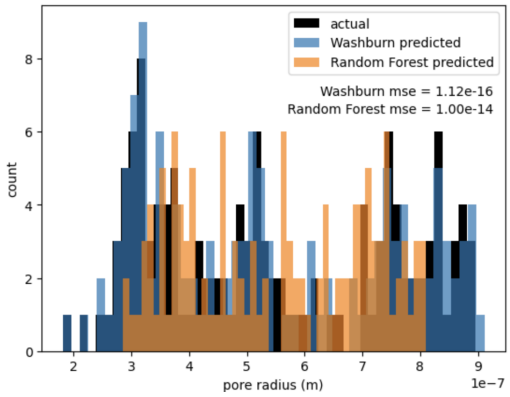
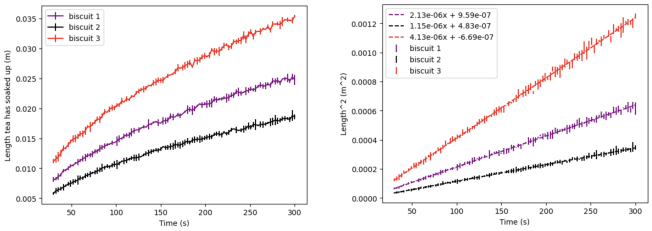


FIG. 7: Predicted pore radii

2.3 Soak Rate



(a) The length against time (b) The length^2 against time

FIG. 8: Soak length for the 3 unknown biscuits

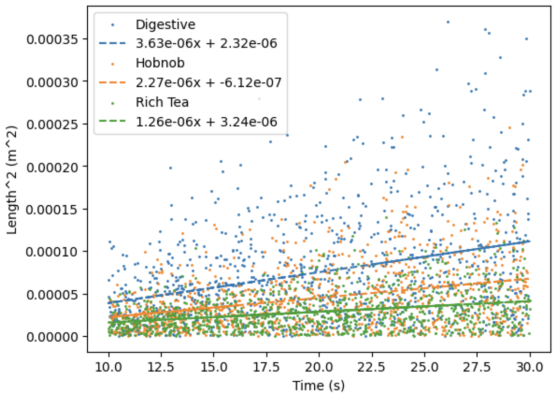


FIG. 9: Length^2 versus time for the pore data

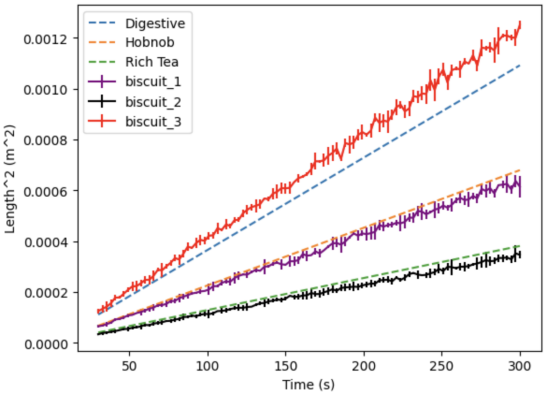


FIG. 10: Trend lines for known biscuits against unknown biscuit data

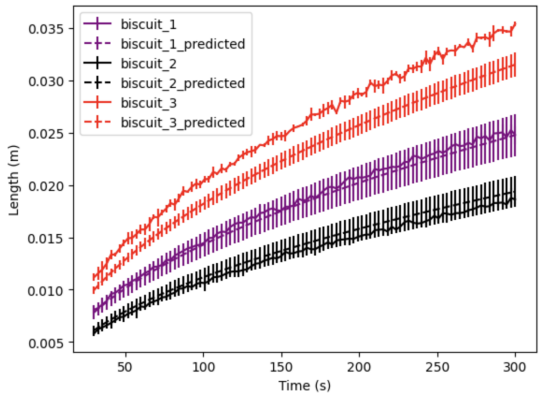


FIG. 11: Washburn predicted soak rate

Errors for L^2 and predicted L graphs have been calculated using:

$$\Delta L^2 = 2L\Delta L \tag{4}$$

$$\Delta L = \sigma_r \sqrt{\frac{\gamma t \cos\phi}{8r\eta}} \tag{5}$$

Where r is the average pore radius of a biscuit, and σ_r is its standard deviation. Equation 4 is used for figures 8b and 10, while equation 5 is used for figure 11.

The alternative Washburn equation, shown in equation 2, suggests a linear relationship between L^2 and t. This is demonstrated in figure 8b, with the data being almost perfectly predicted by a straight trend line.

Although the dunk data does not have time-resolved measurements for the same exact biscuit, it does have a comprehensive record of length versus time for many biscuits of the same type. Therefore it can be used to get a good estimation of the gradient (which depends on the r, which depends on the biscuit). When this is plotted against the unknown biscuit data, in figure 10, it can be extrapolated from the roughly

matching gradients which of the unknown biscuits is likely which type.

With the biscuits identified, the pore data can then be used to find estimations for the pore radius of each according to the average for its biscuit type. This can then be fed into the Washburn equation to find a predicted soak rate, as shown in figure 11.

3 Conclusion

A central narrative of this analysis is to show that while a ML method is often broadly successful, often it is not necessarily the best predictor. It performs excellently as a classifier, as shown in section 2.1. However, as a predictor of values, shown in the r predictions of section 2.2, it does not perform nearly as accurately as the Washburn equation does; the mean squared error (mse) was over 100 times larger. This also confirms that the absorption of tea is likely by capillary flow action.

Although, improvements can obviously be made to the ML method. A larger training data set would allow for increased generalization, and currently the method does not use biscuit type as a factor. Ultimately the aim would likely be to test if the soak length could be predicted by knowing the pore radius, and this has not been investigated.

There are a few other improvements that could be made within the study. Currently for the soak rate analysis, biscuits are manually identified by matching up the known biscuit trend line to the unknown biscuit data. This could likely be further automated by considering the mse for the data to the trend line, and then matching up biscuits via the lowest mse. In addition to this, a larger range of biscuit types would allow for a more complete investigation, as would the measurement of other likely contributory factors such as density, grain type etc.

4 References

[1] Programming Project 2 brief