# Modelling the Tea Absorption Ability of Different Biscuits using Machine Learning

Programming Project 2

Matthew Parker

**Abstract**

In this report, a machine learning model was trained using Support Vector Classification to distinguish between three biscuit types - Digestives, Hobnobs, and Rich Tea biscuits - producing an $F_1$-score of 0.90. Furthermore, the Washburn model of capillary flow action was improved by adding a correction factor, *k*, equal to 0.813 + 3.82e5 * r, to track the absorption of tea over time for each type of biscuit more accurately. The improved model produced a maximum log(B) factor of over 100,000. However, the reliability of the corrected model is very low due to the small amount of data used. With more data, this model could be used in tandem with customer feedback reports to improve customer satisfaction with McVitie's.

## Introduction

McVitie's was established in 1830 and is a staple of the British biscuit community.[1] For nearly two centuries, McVitie's biscuits have been dunked in glorious pools of hot, brown tea. The longevity of this relationship between tea and biscuit is due to the biscuits' unique properties which allow personal control of how soggy a biscuit gets.

Biscuits are made from an interlocking weave of glutenous fibres.[2] This weave of fibres produces some of the distinctive features of biscuits, such as their crisp snap and rigid structure. Furthermore, the fibres produce pores which allow the biscuits to effectively absorb fluids such as tea through capillary flow action.

The size of the pores in a biscuit affects the quantity and rate of tea absorption when dunked. The relationship between pore size and absorption rate is important to facilitate a better understanding of why some biscuits maintain their structural integrity while others crumple into a wet heap of biscuity mush.

It has been predicted that the capillary action of the biscuits will follow the Washburn relationship, given in equation 1:

$$\mathrm{L} = \sqrt{\frac{\gamma r t \cos(\phi)}{2\eta}} \qquad (1)$$

where *L* is the height of the tea up the biscuit (in m), $\gamma$ is the tea surface tension (in N $m^{-1}$), *r* is the radius of the pores within the biscuit (in m), *t* is the time the biscuit is dunked in the tea for (in s), $\phi$ is the angle between the biscuit and the tea (in rad), and finally $\eta$ is the dynamic viscosity of the tea (in Pa s). Clarification for how parameters $\phi$ and *L* are measured is shown visually in figure 1.

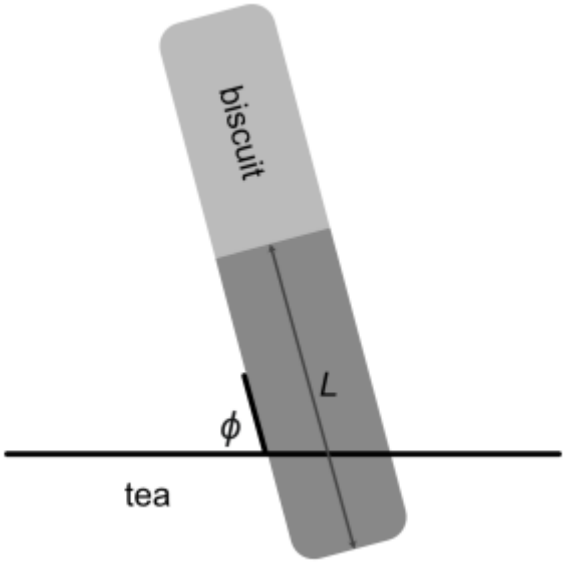


***Figure 1:*** Figure showing the angle, $\phi$, and distance, $L$, measurements taken in the provided dataset.

The Washburn relationship suggests that more porous biscuits will absorb greater quantities of tea, and at a faster rate. This relationship is important to test so the McVitie's team can produce biscuits which maintain the key biscuit characteristics while maximising absorption rates.

This report analyses a machine learning model used to effectively classify three different types of McVitie's biscuit: Digestives, Hobnobs, and Rich Tea biscuits. The pore size of these biscuits will be determined, with the pore radius of the biscuits used to determine the relationship between L and time for each biscuit type.

## Analysis

Two supervised machine learning models were trained to categorise biscuit type from a subset of a dataset containing values $\gamma$, $\phi$, $\eta$, $t$, $L$, and 'biscuit type'. The accuracy of each model was tested on the rest of the data not used in the training. Both a random forest classifier (RFC) and Support Vector Classifier (SVC) were trained and tested to find the best method. The SVC model was found to be more accurate, with an $F_1$-score of over 0.9, whereas the RFC model obtained an $F_1$-score of approximately 0.82. These exact results vary slightly depending on how the data is split for training and testing. The SVC model was used throughout the rest of the analysis due to its higher accuracy. Figure 2 shows how many of each type of biscuit were predicted for each biscuit type using the SVC model.

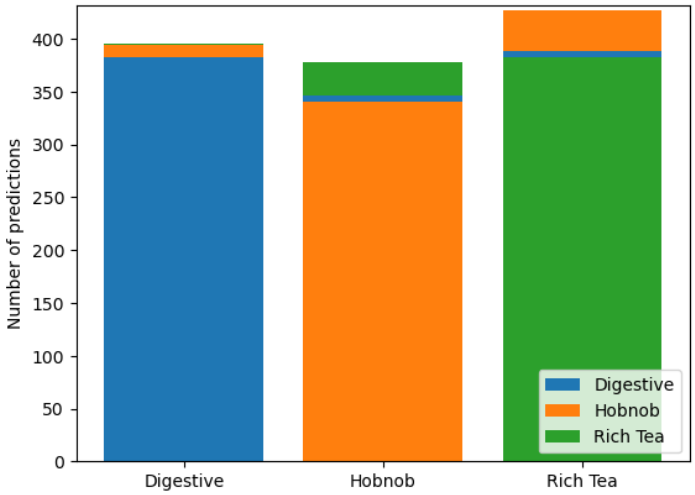


***Figure 2:*** Number of each biscuit type predicted for each biscuit type.

It can clearly be seen that the SVC model is effective at predicting each biscuit type, with a large majority of predictions being correct. The Digestives are classified most accurately, with very few Digestives classified as the other biscuits. The Hobnob and Rich Tea biscuits are more likely to be classified as each other, but are still classified correctly in most cases. Overall, the SVC model is very effective at classifying biscuit type from the provided data.

This model was then used to determine the biscuit type of each row of a second dataset which did not have the 'biscuit type' label. However, this second dataset included accurate measurements for the pore radius of each biscuit. This allowed a distribution of pore radii to be produced for each biscuit type, using the predicted biscuit labels. These distributions determined that Digestives have the largest pore size, then Hobnobs, and Rich Tea biscuits have the smallest pore size.

Furthermore, because the second dataset was actually a subset of the first, these predicted distributions could be compared to the real distributions obtained by merging the two datasets and comparing the true biscuit label with the pore radii. This allows a comparison of how effective the machine learning model is.

Figure 3 shows both the predicted and known distributions of radii for each predicted biscuit type. It can clearly be seen that the two models have very similar distributions, with the only minor discrepancy being the value for the standard deviation of the Rich Tea biscuit distribution.

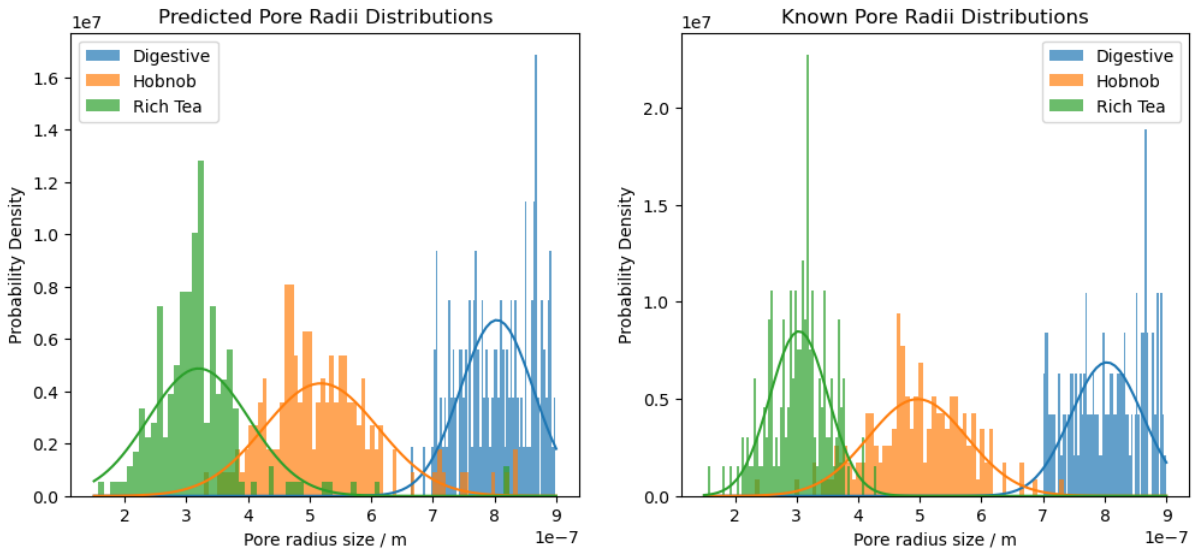


***Figure 3:*** Both predicted and know pore radii distributions for each biscuit type.

The final dataset provided gave information on how L changed over time for a single biscuit within each biscuit type. This relationship was predicted to follow the Washburn equation, given in equation 1. Figure 4 shows a comparison between the real data and the expected data from the Washburn equation when using the mean radii determined in figure 3. $\gamma$, $\phi$, $\eta$ were provided, with the value of each staying constant for each biscuit with values of 6.78e-2 N $m^{-1}$, 1.45 radians and 9.93e-4 Pa s respectively.

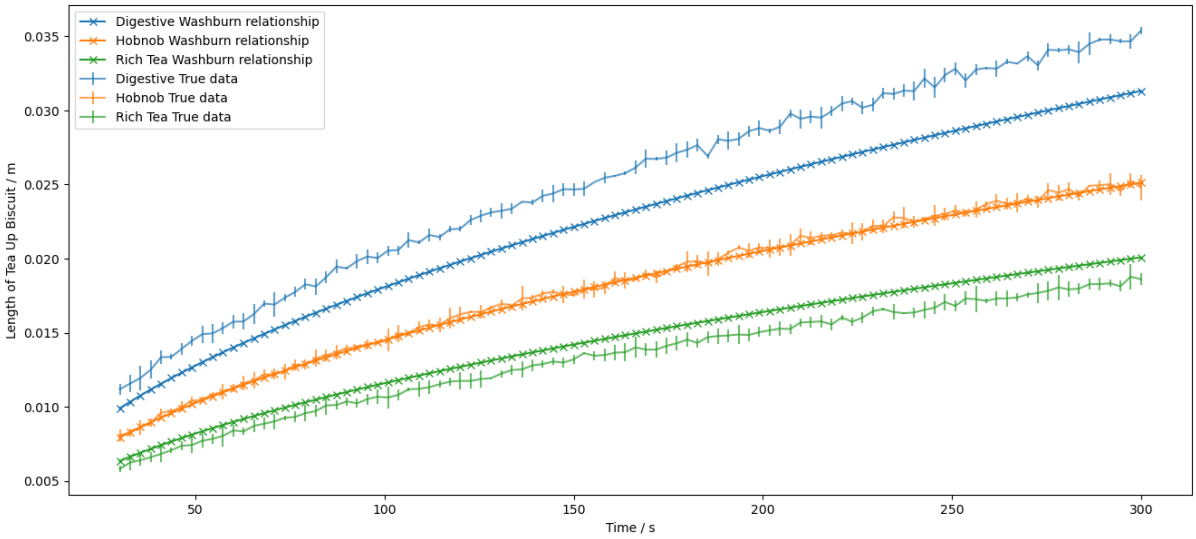


***Figure 4:*** Comparison of Washburn model and true data for each biscuit, assuming mean pore radius of each biscuit type.

As shown in Figure 4, the Washburn relationship effectively predicts how L changes over time for the Hobnob, with an MSE score of 5.21e-8. However, the model seems unsatisfactory for

Digestives and Rich Tea, with MSE of 7.30e-06 and 1.15e-06. This offset may be due to incorrect estimates for the radii of each biscuit, as the assumption that the radius of that specific biscuit was the mean may be inaccurate. However, more analysis must be done to determine if this is the case or if the model itself is inaccurate.

Markov chain Monte Carlo sampling was done to determine the pore radius of each biscuit required for the Washburn relationship to fit the true datasets shown in figure 4. Table 2 shows the calculated pore radius required for each biscuit to fit the data, and the probability of this pore radius occurring for each biscuit.

***Table 2:*** Summary statistics describing likelihood of the required radius occurring.

| ***Biscuit Type*** | **Mean** | **Standard deviation** | **Required radius value** | **Probability of required value occurring / %** | **Number of stds that value is from the mean** |
|---|---|---|---|---|---|
| *Digestive* | 8.04e-07 | 5.90e-08 | 1.00e-06 | 0.068 | 3.40 |
| *Hobnob* | 5.16e-07 | 9.26e-08 | 5.19e-07 | 97.9 | 0.0260 |
| *Rich Tea* | 3.23e-07 | 8.92e-08 | 2.80e-07 | 63.5 | 0.475 |

The table shows that the difference between the true data and the predicted Washburn relationship for the Rich Tea biscuit in figure 4 may have been due to an inaccurate estimate of the pore radius. However, the probability of this being the case for the Digestive biscuit is extremely low at 0.06%. This provides significant evidence that the deviation of the model is not due to an abnormally large pore radius, but rather due to the Washburn relationship being a poor model.

A new model was produced to improve upon the Washburn model, with a correction factor added to the model to account for this discrepancy. This is shown in equation 2:

$$\mathrm{L} = \mathrm{k}\sqrt{\frac{\gamma r t \cos(\phi)}{2\eta}} \tag{2}$$

where *k* is the correction constant, and the other parameters were defined in equation 1.

The correction constant was calculated for each model using nested sampling, again assuming the pore radius of each biscuit was the mean of its pore radii distribution. Figure 5 shows both the original models and the corrected models for each dataset. Clearly, the corrected version provides a better fit than the original version.

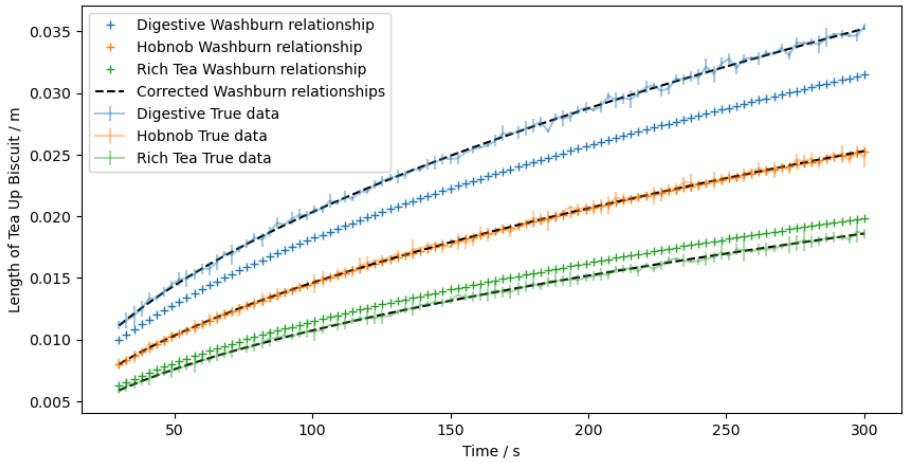


***Figure 5:*** Comparison of the Washburn and corrected Washburn models compared to the true time data.

The Bayesian evidence of the original model was compared to the corrected model. Table 3 shows the quantitative values obtained from the Bayesian evidence of the two models. This table shows that the model improves when adding the correction factor, in particular for modelling the Digestive data. The corrected model for the Digestive biscuit has a mean squared error (MSE) of 5.61e-8, much smaller than the value of 7.30e-6 for the original model.

***Table 3:*** Comparison of the effectiveness of the Washburn and corrected Washburn models.

| ***Biscuit Type*** | **Log(B)** | **Original model MSE** | **Corrected model MSE** |
|---|---|---|---|
| *Digestive* | 1.06e+05 | 7.30e-06 | 5.61e-08 |
| *Hobnob* | 17.7 | 5.21e-08 | 3.43e-08 |
| *Rich Tea* | 946 | 1.15e-06 | 1.73e-08 |

Linear regression was then used to determine a relationship between the correction factor and the radius used in the model for each biscuit, as shown in figure 6. The correction factor was determined to follow the relationship shown in equation 3:

$$k = 0.813 + 3.82e5 * r \quad (3)$$

with an $R^2$ value of 0.9975. However, this relationship is very unreliable due to the very low number of points used to determine this relationship.

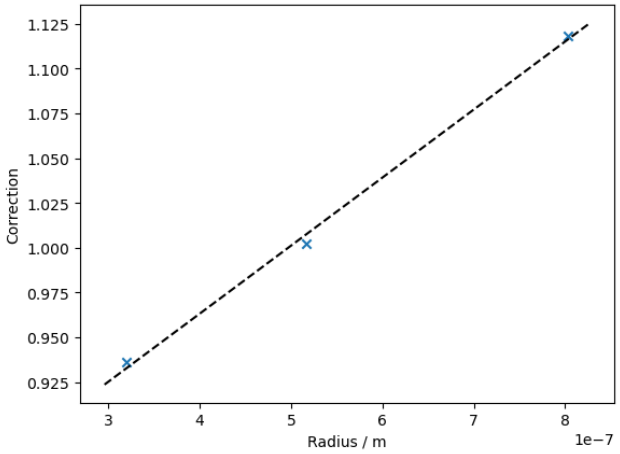


***Figure 6:*** Initial determination of relationship between correction factor and radius size. Equation of regression line calculated as k = 0.813 + 3.82e5 * r

**Conclusions**

This analysis shows that the SVC model is a very effective method to classify biscuit type given the required physical data. This classification method can be used in tandem with true pore radii sizes to determine the pore radius distribution of each type of biscuit.

Generally, biscuit type can be classified by sight. This visual classification could have been used with the microscopy data to determine the radius pore size of a given biscuit, eliminating the need for a machine learning model. However, the model produced would be useful in future data analysis to distinguish between two similar-looking biscuits, for example to explore how the properties of some McVitie's biscuits compare with a rival companies' biscuits.

The final part of the analysis is particularly relevant for quantifying how different biscuits interact with tea when dunked. The improved model allows an effective method to track tea absorption of biscuits over time, which can be directly related to customer experience.

Firstly, customer feedback should be obtained to determine if certain biscuit types are getting too soggy too quickly, or if they are not absorbing tea at a fast enough rate. If this occurrence is designated as a significant issue by McVitie's, the recipe could be changed to vary the pore size as required, using the analysis in this report to directly enhance customer experience.

Secondly, the improved model could be incorporated into the company's marketing strategy. For example, a tool could be added to the McVitie's website to allow customers to calculate how long they need to dunk a biscuit to produce their optimal biscuit sogginess. This would allow the consumer to reproduce biscuit sogginess across different biscuit types.

However, to make a commercial product such as this, more data must be taken to ensure a reliable relationship is determined. In this report, the analysis was conducted with just one biscuit of each type, significantly reducing the reliability of the model and results. Furthermore, the exact pore radius of each biscuit was not known. As it was predicted as the mean of the determined distribution, the relationship may be wrong due to the use of an incorrect value for biscuit pore radius. This risk would also be minimised if more of each biscuit type were used, or if the pore radius of the biscuits used in the dunking test were accurately measured.

In conclusion, this report provides analysis of an effective model to classify biscuits using Support Vector Classification with an $F_1$-score of over 0.90. It also allowed accurate predictions of the pore radii distributions of three different biscuit types. Furthermore, an improved version of the Washburn equation was determined which significantly improved the modelling of tea absorption over time for the three different biscuit types. This data should be used in tandem with customer feedback to improve customer satisfaction with the company's biscuits. Furthermore, it could potentially be used to create an effective marketing tool. However, before this is implemented commercially, more data must be taken to increase the reliability of the relationship.

## References

[1] About McVitie's , https://mcvities.co.uk/about (accessed 31/03/2025)
[2] D. J. Burt, T. Fearn, *Biosynthesis Nutrition Structure*, 1983, **35**, 351-354.