Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

Milk Solid-Not-Fat Prediction using Artificial Intelligence and Digital Image Processing

Abhinav Senthil Chinnaiyan
Independent
Chennai, Tamil Nadu, India

***Abstract*—The Supply chain of how milk is collected and processed is complicated. First, it starts with milk societies where farmers gather to sell their milk for a price. When the milk is given to the societies, a sample of milk is taken for testing of Milk Fat Percentages, 'Solid-Not-Fat' values, and the water content. Based on one of these values, the price is fixed. However, there are quite a few issues when the employees collect data. An issue is high costs of production. With the setup costs of data collecting machinery and balance scales, the cost will be significantly higher. Hence, there was an idea of creating a novel digital method of predicting milk SNF values, mainly through capturing it by a snap of an image from one's phone. The way to predict the milk SNF was through using the RGB values of the image, and then using models such as Linear Regression, Support Vector Regression and Artificial Neural Networks. By collecting primary data and feeding it through the models, it was found that a Linear Regression model was best suited, with an overall Mean Absolute Error of 0.12, and a Mean Squared Error of 0.2**

Keywords*—*Artificial Neural Network (ANN), Mean Squared Error (MSE), Mean Absolute Error (MAE) RGB, Solid-Not-Fat (SNF)

I. INTRODUCTION

India is the largest producer of milk on Earth. India, as of 2018-2019, has produced over 187749 tons of milk and contributes to about 5.3% of India's GDP as of 2019. When the supply chain of the dairy farming industry in India is analyzed, it could be seen that the main source of milk comes from small farmers (About 90%). Then, the milk is collected at designated centers called milk co-operative societies. The price per liter of milk a farmer is paid is determined by the Solid-Not-Fat value, the fat percentage and the water content of the milk itself. There are various methods used to test this, ranging from manual to chemical methods. This work tries to make a novel method of predicting some of the properties of milk: By using Machine Learning. The idea was to use images of milk samples, split it into RGB values, and then use Machine Learning models such as Artificial Neural Networks, Support Vector Regression, or Linear Regression to predict the milk properties through a given value of the red, blue or green pixel intensity

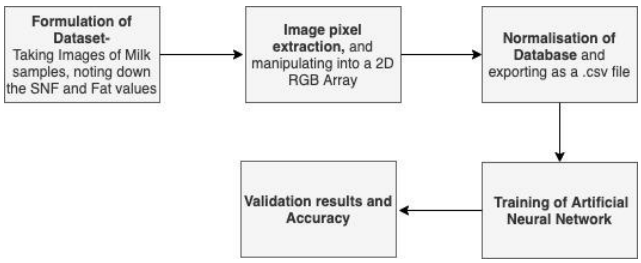
II. EXISTING SYSTEM

Although there is a plethora of methods available to predict milk fat and SNF, there are two main methods that farmers use in order to predict the respective yield. One is the ASE [2] (Accelerated Solvent Extraction) Method and the other is the Mojonnier method [1], and they have their own disadvantages:

A. *Disadvantages*

- The Accelerated Solvent Extraction (ASE) method takes place in a high temperature and pressure environment in order to extract the desired material, which in this case is the Milk Fat residue [2]. However, this could not be achieved in a rural setting too, as one cannot have easy access to chemicals or a laboratory in a village.
- The Mojonnier method uses organic solvents to extract fat of a milk sample, which will then be measured in a flask and subsequently, the fat per liter will be determined [1]. Since this process requires the use of organic solvents and man power to evaporate and extract dry fat, the costs required to determine the milk fat are high, especially in a rural setting. Furthermore, it is estimated that the Mojonnier method takes 2 to 3 hours to completely determine the milk fat yield.

III. METHODOLOGY



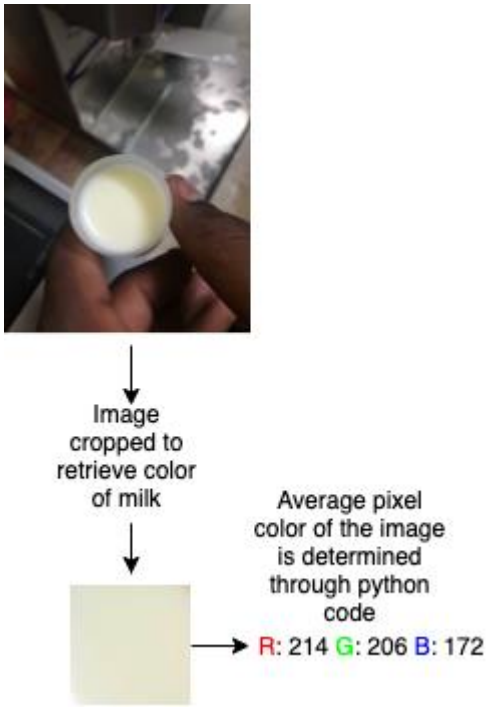
A. *Formulation of Dataset*

Primary data was collected in a local Village called 'Kadayampatti'; with over 100 samples of milk images taken, the actual SNF and Fat Percentage values taken into account.

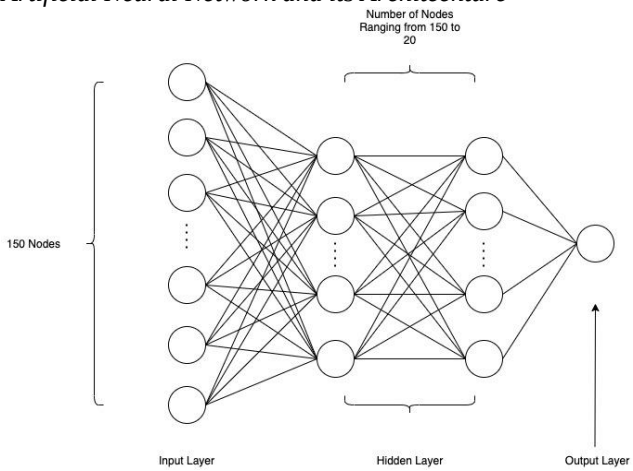
B. *Image Pixel Intensity Extraction*

After data collection, the average RGB pixel value was taken for each of the photos using the python library 'OpenCV'. Then, in a .csv file, the red, blue and green pixel values were uploaded separately including its respective measured Fat percentage and the Solid-Not-Fat value. Then, using the python library 'Pandas', data were being analyzed and the correlation was found using the Chi-square goodness of fit test. It was found that the blue pixel value of the image of a milk sample had the greatest correlation with the SNF value of that respective milk sample. Therefore, the only two variables that were taken into account for the prediction were the Solid Not Fat value and the Average Blue Pixel Value.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021



C. Artificial Neural Network and its Architechture



Artificial Neural networks (ANNs) have become very popular in using it to predict linear and nonlinear relationships. An ANN has weights and biases, to which its timely updates after each iteration as the dataset of information is fed through. With its optimization functions, the ANN effectively updates its weights and biases to predict the input data. An Artificial Neural Network has nodes, from which there can be different Neural Net Architectures. The Architecture of the Neural Network intended for this research is shown above. There are about 150 input neurons/nodes wherein the multidimensional RGB pixel value array is fed forward. Then, there are 6 different hidden layers in the middle (shown together in diagram) that have 100, 85, 70, 25, 20, and 20 nodes respectively. Finally, there is a singular output node that gives a computed fat value in a 1D array.

D. Using Support Vector Regression

Support Vector Regression is a supervised learning model which works similarly to support vector machines. A support vector machine (SVM) is used for classification, where it analyses data points, creates boundaries and a line of best fit given the existing parameters. But the key difference is, in Support Vector regression, the line of best fit is used to predict continuous data, whereas in an SVM it is only used as a decision boundary to classify data into two types.

E. Using Linear Regression

Linear regression is a statistical tool used to effectively model the correlation between two different variables [3]. In this case, the independent variable is the blue pixel value, and the Solid Not Fat value as the dependent variable. After plotting a regression line, one could extrapolate and interpolate Solid Not Fat values with an input of the value of the blue pixel.

IV. HYPER PARAMETER OPTIMIZATION

A. Artificial Neural Network

- *Learning Rate:* The learning rate is one of the parameters used in an ANN, wherein it controls how quickly the model adapts to predicting the given data. Choosing the learning rate is a trade off between accuracy and time. A larger learning rate will be faster but less accurate predictions and vice versa. The other parameters were kept constant, with (epochs = 200, batch size = 256, activation function is 'softmax', optimizer is Adam). Different learning rates were tested on the validation dataset, and as shown below, it seems that any learning rate from 1 - 0.1 seems to give the lowest Mean Absolute Error.

TABLE I. LEARNING RATE AND MAE

S.no.	Table 1	
	Learning Rate	***Mean Absolute Error***
1	10	2.4115
2	1	0.56
3	0.1	0.56
4	0.01	0.56

- *Batch Size:* The Batch Size is a number which indicates the number of samples to go through during each epoch or iteration to update the weights of the Neural Network. In order to find which batch size fits the model best, different batch sizes were tested using the validation dataset, with other parameters being constant (epochs=200, learning rate=0.1, activation function is SoftMax, Adam optimizer). The batch size of 75 and 100 performed the best, with MAE = 0.56

TABLE II. BATCH SIZE AND MAE

S.no.	Table 2	
	Batch Size	***Mean Absolute Error***
1	1	0.5924
2	5	0.6214
3	20	0.5728
4	40	0.5601
5	75	0.56
6	100	0.56

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

- *Epochs:* The number of epochs is the number of iterations a model goes through. During each epoch, the model is fed through the training dataset each time and internal weights are updated accordingly. In order to find which number of epochs fits the model best, different epochs were tested using the validation dataset, with other parameters being constant (batch size = 20, learning rate=0.1, activation function is SoftMax, and has an Adam optimizer). It could be seen that it performed the best when there were 50 epochs, with MAE = 0.5733. The reason higher epochs did not work were because the model might have over-fit the training dataset, that it performs in a worse accuracy during the validation dataset

TABLE III. EPOCHS AND MAE

S.no.	Table 3	
	Batch Size	***Mean Absolute Error***
1	1	5.715
2	10	0.9901
3	50	0.5733
4	100	0.7103
5	200	0.6046

- *Choosing Optimizer:* Optimizers are algorithms used to change attributes such as weights and biases in order to achieve the lowest possible loss, or error. To find which optimizer fits the model best, different optimizers were trained and tested, with other parameters being constant (batch size = 20, learning rate = 0.1, activation function is SoftMax, 50 epochs). It could be seen that an SGD (Stochastic Gradient Descent) optimizer performed the best, with MAE = 0.56.

TABLE IV. OPTIMIZER AND MAE

S.no.	Table 4	
	Optimizer	***Mean Absolute Error***
1	RMSProp	1.1758
2	Adam	0.571
3	SGD	0.56

B. *Support Vector Regression*

- *Kernel:* Kernels are functions that process the given input data to an SVR model. The types of kernels that could be used in the Sci-Kit Learn Library are Linear, Poly, RBF (Radial Basis Function), Sigmoid and Precomputed. All these parameters, with the exception of precomputed as it does not fit the dimension of the matrix, will be tested in order to see which gives the lowest Mean Squared Error. The parameter which does not change throughout changing kernels is the Regularization parameter C and another parameter called gamma.

TABLE V. KERNAL AND MSE

S.no.	Table 5	
	Kernel	***Mean Squared Error***
1	Linear	0.172
2	Poly	3007375041.331
3	RBF	0.54
4	Sigmoid	0.018

As shown in the table, the kernel Linear had the lowest Mean Squared Error while Poly had the highest. Therefore, the Linear kernel was chosen as the parameter for the final tuning of the Support Vector Regression Model.

- *C -Regularization parameter:* An SVR tries to optimize the function, where it tries to maximize the number of points that are correctly classified in the training set. However, it is highly likely that the model might overfit. Overfitting occurs when the model gains a high training accuracy in the training data set, but fails to predict poorly on the validation set or whenever new data is fed through. The C parameter [4], here, focuses on the extent to which the model could be overfit.

A small parameter of C increases the length of the decision boundaries, while a larger parameter of C encourages a smaller margin. In order to find what value of C suits the model the best, trial and error needs to be taken place, with the mean squared error of the validation data (a completely unseen dataset apart from the training dataset) taken into account. Furthermore, it is important to keep the other parameters constant in order to ensure that C is the only factor/variable affecting the obtained MSE value. Hence, the kernel was fixed to Linear for all the tested trials, and the gamma was fixed to 0.1 as well.

TABLE VI. C AND MSE

S.no.	Table 6	
	C	***Mean Squared Error***
1	0.001	1.80
2	0.01	1.80
3	0.1	1.81
4	1	1.80
5	10	1.72
6	100	3.61
7	1000	71.03
8	10000	5442.26

Analyzing the trend of the Mean Squared Error with respect to an increase in the value of C, it could be seen that, generally, from C = 0.001 to C = 10, the MSE is approximately 1.8, and then as C increases, the value of the MSE increases from 3.61 to 5442.26. The value of C as 10 performed the best, to conclude, because it had garnered the least mean squared error of all.

V. ANALYSIS AND DISCUSSION

Metrics used: *MSE, MAE, R^2*

MSE, or Mean Squared Error, has the following formula:

$$MSE = \left(\frac{\sum_{i=1}^{n}(y - y_i)^2}{n}\right)$$

Where y is the actual value, y_i is the predicted value, n is the number of data points.

Mean Absolute Error is another metric, wherein the absolute value of the difference of the observed and predicted value is taken, instead of squaring it. This could be expressed by the following:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 07, July-2021

$$MAE = \frac{\sum_{i=1}^{n} |y - y_i|}{n} \times 100$$

Again, y is the observed value, y_i is the predicted value of the statistical model, and n is the number of data points Lastly, there is R^2, which is the coefficient of determination [7], where SS_{RES} is the Residual sum of squares and SS_{TOT} is the total sum of squares. R^2 shows the extent to which there is correlation between the 2 variables

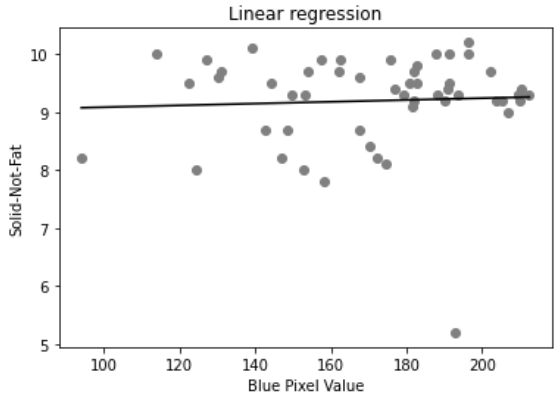
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \tilde{y})^2}$$

A. Linear Regression

Here is the Table of metrics and the graph obtained by using the linear regression model.

TABLE VII. LINEAR REGRESSION RESULTS

S.no.	Table 7	
	Metrics	***Result***
1	R Squared	0.60
2	MAE	0.1
3	MSE	0.2



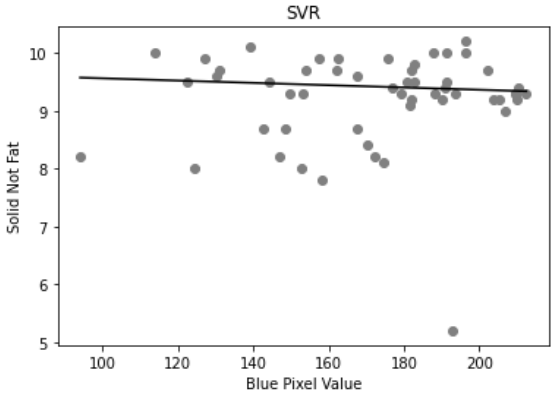
B. Support Vector Regression

From the Hyper Parameter tuning section of this paper, the parameters selected are C = 10, Kernel = Linear. And now, to look at the metrics of the model:

TABLE VIII. SUPPORT VECTOR REGRESSION

S.no.	Table 8	
	Metrics	***Result***
1	R Squared	0.34
2	MAE	0.12
3	MSE	0.2

Here's how the final model of SVR is, with parameters tuned for a high degree of accuracy



C. Artificial Neural Network

After choosing the parameters in which the Mean Absolute Error of the validation dataset is the lowest (50 epochs, 75 Batch Size, 0.01 Learning Rate, and SGD Optimizer), the MSE and the MAE was found to be 0.4891 and 0.5602 respectively.

VI. CONCLUSION

To conclude, if there exists a need to make a digitized system to predict the Solid Not Fat value, different from the traditional methodologies, one could opt for the Linear regression method, than using an Artificial Neural Network, or a Support Vector regression model, as it obtained relatively lower MAE, MSE and higher R^2 values.

ACKNOWLEDGMENT

The author would like to thank Dr.E.Mariappan.B.V.Sc., Assistant Director of Animal Husbandry Veterinary Hospital at Perundurai, Erode, Tamil Nadu, India for his guidance throughout the project work.

REFERENCES

[1] Dionex, Rapid Determination of Total Fat from Dairy Products, https://assets.fishersci.com/TFS-Assets/CMD/Application-Notes/109982- AN364-ASE-TotalFat-DairyProducts-09Feb2011-LPN2713.pdf
[2] Accelerated Solvent Extraction: A Technique for Sample Preparation, Bruce E. Richter, Brian A. Jones, John L. Ezzell, Nathan L. Porter, Nebojsa Avdalovic, and Chris Pohl, Analytical Chemistry 1996 68 (6), 1033-1039, DOI: 10.1021/ac9508199
[3] Linear Regression, Department of Statistics and Data Science, Yale University, www.stat.yale.edu/Courses/1997-98/101/linreg.htm.
[4] Hsia Jui-Yang, Parameter Selection for Linear Support Vector Regression
[5] Reserve Bank of India - Publications, m.rbi.org.in/Scripts/PublicationsView.aspx?id=20072.
[6] An Outlook on Growth of Dairy Sector and its Contribution to Indian Economy, J. Shilpa Shree and M. Prabu,International Journal of Livestock Research.
[7] sklearn.metrics.r2 score — scikit-learn 0.24.2 documentation. (2007). Scikit Learn. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2 score.htm