

pHase - Fermented Raw Milk Stage Prediction Model

Nate Oberdier

noberdier@oakland.edu

ABSTRACT

Raw milk goes through a gradual and unique fermentation process from its naturally occurring bacteria. As raw milk sits at room temperature, its pH, texture, and flavor shift in a relatively consistent order, beginning with fresh milk and moving through phases such as sour milk, yogurt/kefir, and curds before finally settling into a soft, cheese state. Although this progression has been recognized in many food practices, there has been little attempt to model it technically or artificially. The goal of this project, pHase, was to determine whether machine learning techniques could accurately estimate the current stage of fermentation and the expected pH value of raw milk over time. Before building the model, I collected data from 12 jars of raw milk over a period of 45 days, recording 50 separate time points per jar. Each entry included pH measurements along with the time of day and the current date. After organizing the data into a usable format and applying smoothing to remove measurement noise, I trained two models, a Random Forest Regressor to predict pH and a Random Forest Classifier to estimate the fermentation stage. The regression model performed extremely well, reaching an RMSE value of 0.0375 and an R² value of 0.9976. The classification model achieved perfect accuracy on the test set, correctly identifying every stage without error. An interactive interface was then created to allow users to input a start time, temperature, and culture preference to receive real-time predictions. Overall, pHase demonstrates that raw milk fermentation follows a trackable pattern that is recognizable enough to be able to predict fermentation stages along with pH values with great accuracy and reliability.

1. INTRODUCTION

Raw milk goes through a natural fermentation process due to the activity of its natural lactic acid producing bacteria. When stored at room temperature, the milk gradually acidifies and transitions through stages often described as fresh, sour, yogurt/kefir, curds, and finally cheese. These transformations are important both scientifically and in modern fermentation practices, yet little computational work has been done to model or predict this progression. Understanding the process and specific changes within these stages are important for both consumers and producers such that each stage has unique health benefits and can be used for many different reasons.

Problem Statement: Although raw milk fermentation follows a consistent pattern, there is currently no machine learning model that predicts the fermentation stage based on simple variables such as time, temperature, and culture addition due to lack of interest for scientists and lack of resources for consumers. Existing research largely covers pasteurized milk, commercial cultures, or industrial fermentation processes, leaving naturally fermenting raw milk minimally researched.

Existing Solutions: Most existing studies focus on yogurt or kefir made using standardized starter cultures under strict laboratory conditions. These datasets do not reflect the microbial diversity or natural pH fluctuations seen in raw milk, and discontinue experimentation after the first few days. Spoilage prediction research also exists, but these models classify food as fresh or spoiled rather than identifying the current fermentation process. No accessible consumer oriented tools exist to estimate pH or the complete stage progression in raw milk.

Proposed Solution: This project introduces pHase, a machine learning algorithm designed to model raw milk fermentation using an accurate 45 day dataset collected from 12 jars of raw milk. The project makes the following contributions:

- It presents the first raw milk fermentation dataset collected at household room temperatures with both cultured and non-cultured jars.
- It develops an incremental stage labeling method reflecting the natural occurring process of fermentation.
- It trains two Random Forest models, one for pH regression and one for stage classification, that demonstrate high accuracy ($\text{RMSE} \approx 0.0375$; stage accuracy 100%).
- It introduces an interactive tool that allows users to input start time, temperature, and culture information to receive predictions about the current stage, pH, and upcoming transitions.

2. RELATED WORK

Research in dairy fermentation typically focuses on commercial or laboratory settings. Yogurt fermentation is often monitored using spectroscopy, pH curves, or optical sensors. These studies use standardized cultures and pasteurized milk, which produce consistent and predictable fermentation curves. Kefir research follows similar patterns and relies on inoculation with known microbial communities. Although these findings are useful for industrial control, they are not applicable to raw milk, which contains a wider range of microorganisms and follows a slower and more variable fermentation path. Food spoilage detection has also been analyzed with machine learning. These studies aim to identify whether food is fresh or spoiled based on gas formation, temperature exposure, or microbial load. Spoilage research does not take into account the fermentation process of foods, which scientifically examines how food transforms rather than the natural process it changes its bioavailability. Compared to these areas, pHase focuses on the unique behavior of unpasteurized milk. It relies on data collected under real household conditions, and it models multiple stages rather than a simple fresh versus spoiled classification. This makes the project unique compared to most work previously done in dairy science and from most machine learning applications in food research.

3. A MOTIVATING EXAMPLE

Imagine someone who places a jar of raw milk on the counter with the goal of making clabbered milk, kefir-like fermented milk, or a soft cheese. They might smell the milk or look at its texture to guess where it is in the process, but these observations can be misleading, especially for someone without experience. Even small differences in room temperature or culture addition can change the rate at which the milk transforms. With pHase, that person can enter the date and time when fermentation began, the approximate temperature of the room, and whether any culture was added. The tool then predicts the milk's current stage, estimates its pH value, and even provides a rough timeline for when it will reach the next major transformations. This allows individuals to make informed decisions rather than relying on guesswork.

4. Approach

4.1 Data Collection and Preprocessing

Data Sources: The dataset used in this project was generated entirely from my own raw milk fermentation experiment. Twelve 16 ounce jars of fresh raw milk, straight from a farm, were prepared at the beginning of the study. Each jar contained 10 ounces of milk. Jars 1 through 6 contained only raw milk, while jars 7 through 12 included one tablespoon of culture of previously naturally fermented milk. All jars were kept in a kitchen cabinet for 45 consecutive days. The room temperature fluctuated naturally between 67 and 69 degrees Fahrenheit, averaging around 68 degrees. Each jar was measured roughly fifty times during the experiment, resulting in approximately 600 total pH datapoints. All jars were measured together at each entry, ensuring consistent sampling conditions. I began collecting data multiple times per day to evaluate the rapid pH changes near the beginning. After a few days, I began collecting the pH values once a day. After 30 days, I collected the values every other day and progressively bigger time gaps due to minimal changes. All data was originally collected in a notebook and then later transferred into an excel spreadsheet for technical analysis.

Preprocessing Steps: The collected data was organized into long format, and each entry's date and time were merged into a single piece of information. This timestamp was then converted into the number of hours since the beginning of fermentation, providing a smooth temporal index for modeling. Because manual pH readings can show small fluctuations between measurements, an exponential smoothing method was applied to reduce noise in the curves. Additional features such as the rate of change in pH, squared time values, and pH time were taken into account to accurately estimate the small inaccuracies of my measurements. Fermentation stages were not directly observed, so a stage labeling method based on published pH ranges and the known biological order of fermentation was used. An incremental constraint ensured that stages progressed forward over time.

4.2 Model Selection and Training

Algorithms Used: Two machine learning models were applied to the dataset. I first used a Random Forest Regressor for predicting pH and then a Random Forest Classifier for predicting the fermentation stage. Random Forest was chosen because it is accurate with small datasets and would be best for my relatively minuscule dataset. The structure of decision trees also allows the model to capture complex interactions which are best suited for the unique occurrences along the path of fermentation.

Training Process: To prevent inaccuracies between jars, the dataset was split into 2 groups, one for culture added and the other without culture added. Within each group, four jars were used for training, one for validation, and one for testing. This ensured that each model was evaluated on jars it had never encountered. Both the regression and classifier models were trained using four hundred trees with no maximum depth limit, allowing the trees to grow fully and capture the detailed fermentation curves. Training involved fitting the models on through the hyperparameters mentioned earlier with the validation set to ensure a smooth transition and minimal inaccuracies. After training, both models were tested on the final withheld jars.

5. Experimental Evaluation

5.1 Methodology

The main goal of this project is to determine whether the fermentation stage and pH of raw milk can be predicted using only time, temperature, and whether a culture was added. The hypothesis behind the experiment is that raw milk fermentation follows a predictable pattern when kept at a steady room temperature, and that a machine learning model can learn this pattern well enough to make accurate estimates. The method used for the experiment begins with the dataset of 600 pH measurements taken from 12 jars over 45 days. The independent variables include the number of hours since fermentation began, if culture was added, and the engineered time based features. The dependent variables are the pH, for regression, and the fermentation stage, for classification. Because all jars were measured together and went through the same temperature changes, the dataset reflects a realistic and natural fermentation process. The training and test data were created by splitting the jars into groups. 4 jars from each group were used for training, 1 was used for validation, and 1 was used for testing. This prevents the models from memorizing the behavior of individual jars. The performance data collected included predicted pH values, predicted stages, and the errors between the predictions and the real values. These results were compared visually through graphs and also through numerical evaluation measures. The models were implemented in Python using the scikit-learn library. The libraries ZoneInfo, ipywidgets, and datetime were also used to construct the interactive tool so users can rely on accurate time measurements using Eastern Standard Time. The main research question this project addresses is whether raw milk fermentation can be modeled with enough accuracy to provide practical and reliable predictions for users.

5.2 Evaluation Metrics

Two types of metrics were used to evaluate the models. For the regression model, RMSE and R^2 were used. RMSE measures the average difference between the predicted and real pH values, so a lower value indicates better accuracy. R^2 measures how much of the variation in the data the model is able to explain. A higher value means the model fits the real fermentation curve well. For the classification model, accuracy and F1 score were used. Accuracy measures how often the model predicted the correct stage, while the F1 score treats each stage equally and checks how well the model performs across all classes. These metrics were chosen because they give a precise estimate of both the overall performance and the performance across different fermentation stages.

5.3 Results

The Random Forest Regressor performed extremely well. It achieved an RMSE value of about 0.0375 and an R^2 value of 0.9976. These results show that the model was able to mimic the natural pH curve very closely and could track fermentation precisely. The classification model achieved perfect accuracy on the 2 test jars. Every predicted stage matched the labeled stage, and the F1 score was also perfect. This means the model accurately matched each stage very well. When the data is viewed on graphs, the pH data show a minimal difference between jars with culture and those without culture. The cultured jars lower in pH much faster near the beginning, while the non-culture jars rise in pH sooner near the end of fermentation but the added culture jars still end with a higher pH value. These differences are well evaluated and visualized in the AI tool of pHase. These results show that the models can imitate the natural process of fermentation quite well with minimal error.

5.4 Discussion

The results support the initial hypothesis that raw milk fermentation can be predicted using machine learning. The performance of the regression model shows that the acidity is strongly tied to time and culture conditions. The classification model's perfect accuracy suggests that the incremental labeling of the stages captured the natural order of fermentation reliably. Random Forest models do not require complex tuning, yet they were able to model the fermentation process with nearly perfect accuracy. The structure of the data also helped, since the fermentation curves were smooth and followed a consistent trend across jars. Naturally, during fermentation, there are differences when using culture and not using culture, and this project accurately displayed the differences with each process. Another interesting aspect to acknowledge is that the content within each of the 12 jars reduced by 2 ounces, resulting in a 20% decrease of total milk volume. This process occurs due to carbon dioxide forming within the milk, mixed with minimal amounts of evaporation, resulting in a loss of volume.

6. LIMITATIONS

There are many limitations in this project, mainly revolving around data collection. I was only able to use 1 gallon of milk to collect data. If I had more jars, it would take significantly longer to collect data each day which would make it much more difficult. There was also a limited amount of time relevant to milk fermentation to conclude the project. I was not able to accurately collect pH values during or after cheese formation, leaving this project unfinished. I was also only able to use 1 source of milk. I already had a herdshare with this farm so obtaining another herdshare from another farm would require \$100+ and 2 hours of driving. I would also have to discontinue the other herdshare the same day I receive the milk which would be unideal. When collecting the pH values, I had to wait a minute or two for each jar until the pH value would settle. This would cause inaccuracies and waste a lot of time, sometimes taking 20+ minutes per data collection each day. I was also only able to store the milk at one temperature the entire time. This made it difficult to calculate different temperature fermentations when calculating pH values and stages within the interactive tool.

7. CONCLUSIONS AND FUTURE WORK

This project shows that modeling the fermentation timeline of raw milk is possible with the help of machine learning. The pHase model can accurately estimate pH and fermentation stages from simple inputs such as date, time, and indicating if a culture was added. The high accuracy of my models show that raw milk fermentation follows a trend that can be recreated artificially. For future experiments, there should be a wider variety of milk sources, longer and more often pH observations, more accurate stage collection, and bacteria examination. I hope to see this project inspire others to further experiment on raw milk and the fascinating fermentation process it endures.

8. DATA AVAILABILITY

The entire project was coded within Google Colab. It can be accessed via Github using the following link:
<https://github.com/NateOberdier/pHase>

This work is licensed under CC BY 4.0. To view a copy of this license, visit
<https://creativecommons.org/licenses/by/4.0/>

9. ACKNOWLEDGEMENTS

I would like to thank British Family Farms for providing the herdshare program that allowed me to collect data for this project. Without them, I wouldn't be able to develop a love for raw milk and have the inspiration to investigate and pursue this project.

10. REFERENCES

- Aydogdu, M., Altuntas, A., & Kocak, P. (2023). pH: The fundamentals for milk and dairy processing – A review. *Journal of Food Science and Nutrition*, 11(7), 4567–4583. Retrieved from <https://www.researchgate.net/publication/371901325>
- D'Amico, D. J., & Donnelly, C. W. (2010). Microbiological and chemical changes in raw milk during storage. *Journal of Dairy Science*, 93(3), 910–921. <https://doi.org/10.3168/jds.2009-2761>
- Gagara, M., Nonga, H. E., & Ijumba, J. N. (2022). Hygienic quality of raw and fermented cow milk in the local market in Tanzania. *Veterinary Medicine International*, 2022, Article 9375221. <https://doi.org/10.1155/2022/9375221>
- Tomovska, J., Gjorgjevska, B., & Kuzmanova, S. (2016). Examination of pH, titratable acidity and antioxidant activity in fermented milk. *Journal of Food and Nutrition Sciences*, 4(3), 51–57. Retrieved from <https://www.davidpublisher.com/Public/uploads/Contribute/58ca5d728d24d.pdf>
- Yang, H., Choi, J., Lee, S., & Kim, D. (2023). Investigation on bacterial growth and pH in milk after the introduction of a model spoilage bacterium. *Microorganisms*, 11(10), 2259. <https://doi.org/10.3390/microorganisms11102259>
- SafeMilk Labs. (2022). Understanding the pH of milk and milk products: A key indicator of quality and freshness. Retrieved from <https://safemilklabs.com/understanding-the-ph-of-milk-and-milk-products-a-key-indicatorof-quality-and-freshness/>
- University of Guelph. (2015). Raw milk quality. In Cheese-Making Technology eBook. Retrieved from <https://books.lib.uoguelph.ca/cheesemakingtechnologybook/chapter/raw-milk-quality/>
- Tetra Pak. (2023). Dairy processing handbook. Retrieved from <https://dairyprocessinghandbook.tetrapak.com/chapter/chemistry-milk>
- SafeMilk Labs. (2022). Milk spoilage: Methods and practices of detecting milk quality. Retrieved from [https://www.researchgate.net/publication/269963794 Milk Spoilage Methods and Practices of Detecting Milk Quality](https://www.researchgate.net/publication/269963794_Milk_Spoilage_Methods_and_Practices_of_Detecting_Milk_Quality)
- Utah State University Extension. (2021). Tips to safely ferment at home. Retrieved from <https://extension.usu.edu/preserve-the-harvest/research/tips-to-safely-ferment-at-home>