Udderly Advanced: AI's Leap Into Milk Analysis

Annie Cho

***Abstract*—Milk adulteration remains a significant challenge in the global effort to combat food fraud, particularly in developing nations where limited screening infrastructure exposes consumers to serious health risks. While developed countries often employ chromatography and spectrometry techniques for detection, these methods are often impractical in resource-constrained regions due to financial limitations and lack of specialized equipment. 'Udderly Advanced' addresses this issue by introducing a novel AI-based solution to detect water—the most prevalent milk adulterant—through the analysis of droplet evaporation images and use of an enhanced feature extraction pipeline. ML models capable of detecting dilution with up to 98.6% accuracy and 0.985 ROC-AUC scores have been produced, bridging the gap in comprehensiveness, accessibility, cost-effectiveness, efficiency, and scalability left by traditional and alternative AI methods. Rigorous evaluation, combining both qualitative and quantitative measures, ensures the models' effectiveness in real-world settings. Ultimately, 'Udderly Advanced' offers a promising solution for improving milk quality assessment in areas where conventional methods are not feasible.**

***Index Terms*—milk, adulteration, machine learning, droplet analysis, accessibility, public health, developing countries.**

GLOSSARY

AI	Artificial Intelligence
CV	Computer Vision
DL	Deep Learning
DT	Decision Tree
FCM	Full Cream Milk
KNN	K Nearest Neighbours
LOO	Leave-One-Out Cross Validation
LR	Logistic Regression
MC	Monte Carlo Cross Validation
ML	Machine Learning
MLP	Multi-Layer Perceptron
NBC	Naïve Bayes Classifier
P+	Protein Plus Milk
RFM	Reduced Fat Milk
RL	Reflection Line
ROC-AUC	Receiver Operating Characteristic Area Under the Curve
SP	Stout's Pipeline
SVM	Support Vector Machine
UIP	Udderly Improved Pipeline

I. INTRODUCTION

A. Motivation

THE dairy industry has experienced significant global growth over the decades, with cow's milk emerging as a versatile, nutritious, and highly sought-after commodity [1]. However, the dairy industry faces challenges concerning adulteration, the deliberate contamination of foreign substances into milk. Adulteration can occur at various stages of the supply chain system—often by unscrupulous providers seeking to maximise profits, but also from misguided efforts to improve hygiene or ignorance regarding appropriate drug administration practices [2], [3]. Substances range from seemingly innocuous water and whey to harmful additives like melamine, urea, soap, and hydrogen peroxide. This compromises the quality and safety of milk, posing substantial risks to consumer health and eroding trust in dairy products. One notorious case is the 2008 Chinese milk scandal, where nitrogen-rich melamine was added to milk to artificially inflate protein levels and deceive the nitrogen-based checks used to indicate milk quality. This malpractice resulted in 300,000 hospitalisations and 6 deaths among infants who ingested contaminated milk formula [2], [4], highlighting the severe health consequences of consuming adulterated products and the need for robust detection methods.

TABLE I: Potential Contaminants in Milk [2]

Contaminant	Motivation	Potential Health Risks
Water	Increases profits and milk volume.	Acute or severe malnutrition, diarrhoea, typhoid.
Melamine	Increases profits, falsely elevates protein content.	Kidney stones and failure, infant death, urinary stones, crystalluria, bladder cancer, toxic poisoning.
Urea	Increases profits, falsely elevates protein content, extends shelf life.	Digestive issues, kidney damage, diarrhoea, ulcers, impaired vision.
Hydrogen Peroxide	Extends shelf life without refrigeration.	Gastritis, intestinal inflammation, diarrhoea, vomiting, nausea.
Soap/ Detergent	Increases profits, improves appearance.	Gastro-intestinal and respiratory complications, hypotension, cancer.

B. Problem Statement

Adulteration is widely condemned as unethical and illegal in many jurisdictions [5]. Regulatory bodies like New Zealand Food Safety impose penalties to deter these practices; however, consumers in developing countries (e.g., China, Sudan, India) remain vulnerable due to poor access to high-quality products, limited education about food safety, and inadequate regulatory oversight [2], [6].

Detecting adulteration in developing countries poses numerous challenges due to the vast scale of milk production, milk's complex composition, and the extensive range of potential contaminants seen in Table I. Cow's milk contains both nonvolatile milk solids and volatile mediums that complicate the detection of foreign substances—approximately 87% water and 13% protein, fat, carbohydrates, and minerals [7], [8].

This project was supervised by Andrew Lensen (Primary) and Gideon Gouws (Secondary).

Purely motivated by financial greed, water emerges as the most pervasive adulterant due to its free, unregulated, and abundant availability [3]. Typically sourced from taps or ponds, water dilution not only reduces the milk's nutritional value but also heightens the risk of ingesting waterborne pathogens [2]. A 2011 survey by the Food Safety and Standards Authority of India exposed water as the leading adulterant among the staggering 68.7% of milk products failing to meet quality standards [2], [3]. Similar studies revealed water adulteration in 73% of milk samples analysed in Pakistan and over 95% in Sudan [2], [6]. The addition of sophisticated adulterants like rice flour, melamine, and detergent can subsequently restore the viscosity of diluted milk, deceive nitrogen-based protein tests, and conceal contamination.

Notably, the lack of standardised testing and equipment in developing countries significantly hinders detection efforts. Traditional analytical methods (e.g., chromatographic separation, spectroscopic analysis) provide high quantitative accuracy [9] but are time-consuming, expensive, and demand advanced instrumentation and skilled personnel—limiting their use in resource-constrained environments [3]. For example, amidst the 2007 Brazilian milk scandal, where products adulterated with oxygenated water, hydrogen peroxide, and caustic soda were sold to unsuspecting consumers, only one-third of products were inspected by national consumer health programme authorities [2]. This incident evidences the inefficiency and incapacity of government inspection systems in resource-constrained regions and the critical need for *accessible* detection methods to effectively regulate milk quality.

C. Solution and Deliverables

As such, Artificial Intelligence (AI) emerges as a solution that minimises the need for manual labour, time, and specialised equipment [10]. 'Udderly Advanced' introduces a novel machine learning (ML) approach to classify adulterated milk. Through the experimentation and development of six ML algorithms—K Nearest Neighbours, Decision Tree, Naïve Bayes Classifier, Logistic Regression, Support Vector Machine, and Multi-Layer Perceptron—a highly accurate model has been produced, capable of distinguishing untampered full cream milk from water-contaminated samples.

This process begins by capturing the spatio-temporal characteristics of evaporating droplets through a sequence of profile images. Using a solution developed by a previous researcher as a foundation, these images are then processed through a feature extraction pipeline, specifically reformed for detecting adulteration. This transformation generates a singular feature set by measuring the pixel heights of the droplet at regular intervals across its width. Subsequently, the refined data is utilised by the ML models to differentiate between diluted and pure samples.

Bridging the gaps left by both traditional and alternative AI methods, the final solution satisfies its five non-functional project requirements. Focusing on practical use in developing countries, these requirements are:

- **Comprehensiveness:** Effectively addresses key aspects of adulteration detection itself. This includes detailed capture of droplet dynamics, development of traceable systems resilient to variability, and precise classification.
- **Accessibility:** Utilises broadly available equipment and technology, ensuring ease of use for individuals with minimal technical expertise in resource-constrained environments.
- **Cost-Effectiveness:** Offers an affordable alternative to costly traditional methods and AI solutions, minimising expenses in apparatus and technological infrastructure.
- **Efficiency:** Delivers rapid results through lightweight, computationally-efficient ML algorithms. Suitable for deployment on accessible but lower-capability devices like smartphones.
- **Scalability:** Adaptable to accommodate increasing volumes of milk production.

From these project requirements, the following applicable measurable performance specifications are addressed:

- Development of the droplet dataset only requires a basic camera, substrate, and milk droplets while the ML task utilises free, open-source software libraries.
- The models are runnable through a Python command-line interface that outputs relevant performance metrics to an external CSV file, along with logging of system status and error messages.
- Six ML models were trained, achieving up to 98.6% accuracy and 0.985 ROC-AUC for dilution detection, along with 89.5% accuracy and 0.978 ROC-AUC for classifying diluted droplets among three legitimate milk types. This meets the project's target of over 90% accuracy and demonstrates competitive performance to benchmarks set by comparable AI-driven approaches.
- High-performing models are as compact as 1.56 KB and can operate in as little as 54.1 ms (7.7 ms per sample prediction) on standard hardware, enabling potential deployment on resource-limited devices, like smartphones.
- Lastly, optimisation upgrades to a baseline computer vision pipeline significantly improved feature extraction for varied droplets and conditions, boosting processing success rates from an initial 50% to 86%.

D. Environmental and Sustainability Issues

The project addresses three United Nations Sustainable Development Goals [11]: SDG2 *Zero Hunger*, SDG3 *Good Health and Wellbeing*, and SDG9 *Industry, Innovation, and Infrastructure*. Firstly, by developing robust adulteration detection methods, the project partly ensures food safety, a critical component of food security outlined by SDG2. With 1/3 of people worldwide facing moderate to severe food insecurity, protecting milk—a growing and essential nutritional resource in developing countries [2]—from adulterants secures the health and nutrition of populations with less stringent food safety standards. Secondly, in line with SDG3, the project aims to wholly prevent health hazards linked to consuming tainted milk, as seen in the 2008 Chinese milk scandal. Early detection safeguards public health, preventing direct effects experienced by 57% of people who consume adulterants [12] such as illnesses, hospitalizations, and deaths. Lastly, moving

beyond traditional methods to more comprehensive, accessible, cost-effective, efficient, and scalable AI solutions partly supports SDG9. AI solutions that cover the weaknesses left by traditional analytical methods contribute to building resilient infrastructure in the milk industry.

II. Background Research

A. Relevant Terms

While often conflated, AI broadly refers to machines' ability to emulate human-like intelligence via algorithms, while the ML subfield enables algorithms to learn iteratively and generalise to new data without explicit programming [13]. Deep learning (DL), a subset of ML, utilises multi-layered neural networks to identify and learn complex patterns in vast, complex datasets. In this context, a model refers to an instance of an algorithm. Models make predictions after training on datasets with distinctive features, often including target variables for supervised learning tasks. Lastly, computer vision (CV) enables machines to interpret information from visual media, such as digital images or videos [13].

B. Literature Review

The organised patterns arising from droplet evaporation yields valuable insights into the droplet's internal properties [14]. Proteins, fats, and other substances drastically influence resulting pattern and drying process, as pictured in Figure 1. As such, the droplet's shape over time (i.e., spatio-temporal data) can indicate potential adulterants, motivating the use of CV and AI to detect irregularities in a comprehensive, accessible, cost-effective, efficient, and scalable manner. An evaluation matrix of the related works against these five dimensions can be found at Appendix A.

Fig. 1: Resulting pattern of a full cream milk droplet. The evaporation process has left behind a residue of fats and proteins on the substrate.

1) ML/DL on Spectral Data: While our project centres on evaporation-based methods, spectral research provides valuable benchmarks for assessing the efficacy of our more accessible, cost-effective, and scalable approach. For example, Neto et al. integrated spectrometry with ML/DL to detect adulterants [15], achieving promising performance on a binary classification problem with Linear Regression (79.62% accuracy) and Convolutional Neural Network models (96.76% accuracy and ROC-AUC score of 0.9985). Despite this, the cost and limited availability of spectrometers and DL, coupled with the need for specialised knowledge, hinder its use in developing countries.

2) Identifying Static Droplet Patterns: Investigating milk stains to detect adulterants, Kumar et al. [8] observed distinct patterns during the evaporation of normal, diluted, and urea-contaminated milk. They revealed that adulterants lead to chemical crystallisations and the disruption of "coffee-ring" depositions in unadulterated stains. Requiring only a substrate and a smartphone, their method can visually detect adulteration without sophisticated infrastructure, making it a crucial starting point for designing accessible and inexpensive solutions for developing countries. However, relying on manual inspection rather than CV and ML—which the proposed solution does—introduces variability, human error, inefficiency, and scalability limitations. While the study provides a foundational step, it is an incomplete solution for practical application.

3) Imagery and ML on Static Droplet Patterns: Replacing the manual inspections used in Kumar et al.'s solution, Harindran et al. developed a feed-forward neural network to classify diluted milk droplets based on their drying patterns, achieving ~85% accuracy on bird's-eye images [16]. Similarly, Pérez-Calabuig et al. applied DL to categorise cow's milk adulterated with water, goat's milk, and sheep's milk [17]. Their ResNet50 model, trained on 10,400 overhead images under varying light conditions, reached accuracies of 93.5% and 93.2% under light and dark conditions, respectively. Both approaches boost efficiency and scalability through automation, while open-source tools like Python and TensorFlow minimise monetary costs. However, ResNet50's 50-layer architecture demands substantial resources, making it impractical compared to more lightweight models like those employed in this project. Additionally, static bird's-eye imagery fails to capture temporal changes in droplet patterns, necessitating improved feature extraction techniques to prioritise comprehensiveness.

4) Feature Extraction and ML on Static Droplet Patterns: Andalib et al. approach this similar CV task using point-by-point analysis of the droplet profile [18]. Naïve Bayes Classifier and Bagged Decision Tree models were trained on measurements of the droplet's diameter and contact angle, captured at a single time point t within a predefined period T. The models achieved accuracies of 75% and 96%, respectively. Andlib et al.'s method offers notable advantages. Translating the visual droplet data into a format suitable for computer analysis enables accessible and cost-effective processing, mitigating the risk of human error. ML also ensures consistency, efficiency, and scalability in classification. However, similar to Handridan et al. and Pérez-Calabuig et al.'s image capture approach, a disadvantage lies in the fragmentation of the droplet into discrete states rather than viewing it as an evolving entity, as proposed in the project.

5) Identifying Temporal Droplet Patterns: Suh et al. present an alternative CV approach [19]. Their method diverges from the project's solution in two key facets: droplets are again observed from a birds-eye perspective rather than in profile, and a limited set of spatio-temporal features are extracted using a DL intermediary, instead of generating feature sets from individual droplets. Despite reducing time and manual effort in analysing dynamic droplet behaviour, this approach has two disadvantages. Firstly, it fails to fully depict the evolving shape from a birds-eye angle, in which profile would

provide more intricate insights. Secondly, the reliance on DL incurs significant expenses that diminish its accessibility and cost-effectiveness, particularly in computational resources and necessary expertise. Nevertheless, Suh et al.'s research remains pertinent as they explicitly monitor spatio-temporal features, a fundamental aspect of the project's solution.

6) Feature Extraction and ML on Temporal Droplet Patterns: Lastly, expanding upon prior research into temporal data, Stout et al. investigated ML to classify milk types [20], providing a locally-developed foundation for this project to improve upon. Their method involves placing droplets on a substrate and capturing a sequence of profile images documenting the evaporation process over T frames. A feature extraction pipeline then records the pixel height of the droplet at k regular intervals, generating a matrix of the droplet's spatio-temporal features. The vector representation of the matrix is subsequently used for training.

Among six models trained on the extracted data, the Logistic Regression model notably achieved up to 96.3% accuracy and a ROC-AUC score of 0.999. Leveraging profile images offers a more precise representation of the droplet's evolution, while the use of non-invasive image capture techniques eliminates the need for expensive and sophisticated infrastructure, enhancing practicality for widespread adoption. While Stout et al.'s research addresses a simpler problem, their CV method demonstrates effectiveness within their research context and shows promise for this project's more complex task of detecting adulterants. As such, their findings inform this project's development and their performance metrics will serve as valuable benchmarks for evaluating this project's improvements in addressing a more challenging problem. A complete results table from Stout et al.'s study can be found in Appendix B. However, their method exhibits three main weaknesses in both its inherent design and application for adulteration detection:

1) Their use of up to 1200 images per sample necessitates substantial processing power and storage, an issue that falls outside the scope of this project.
2) The feature extraction pipeline has been tested exclusively on standard off-the-shelf milk droplets in highly-controlled environments, which is difficult to reproduce precisely. To uphold accuracy and reliability, the pipeline must be resilient to environmental factors such as variations in lighting, distance, and angle—requirements currently unmet. Furthermore, there are additional complexities specific to detecting adulterants that are unaddressed in the study, such as the quicker evaporation rates of diluted droplets, which could impact processing.
3) The feature extraction pipeline lacks effective error handling and reporting capabilities.

Therefore, to fully address these last two limitations, the project also focuses on: a) further developing their pipeline to accommodate variations in environment and droplet composition, and b) implementing improved error handling and reporting functionality.

7) Final Benchmarks: The benchmarks used to evaluate the project have been summarised in Table II.

TABLE II: Final Benchmarks [15], [20]

Author	Model	Accuracy	ROC-AUC	Limitation
Neto	Linear Regression	79.62%	N/A	Use of spectrometry
Neto	Convolutional Neural Network	96.76%	0.9985	Use of spectrometry and DL
Stout	Logistic Regression	96.30%	0.999	Not for adulteration

C. Tools and Methodology

1) Methodology and Development Process: This project adopts a hybrid methodology that combines the strengths of the Waterfall methodology and Agile practices, accommodating to the project's well-defined requirements while remaining flexible to evolving needs. Comprising five main development stages—planning and research, data collection and pre-processing, model selection and design, training and evaluation, and refinement—the project progresses sequentially with clear milestones that the Waterfall methodology encapsulates well, facilitating periodic reviews to monitor progress against predefined goals. Emphasis is placed on formal documentation, including meeting minutes with summaries and action items, and a risk register outlining risks and mitigation strategies. Waterfall also ensures a structured and smooth project closure, producing final documents and formally concluding project activities. Agile principles are integrated within each stage, promoting iterative development, continuous feedback, and adaptive planning. While not all Agile practices (e.g., stand-up meetings) have been adopted, the project still adopts iterative cycles of planning, execution, and review. Larger components are decomposed into smaller units, aimed to be completed within two-week iterations and tracked as Git issues. To align with project goals and uphold clear communication, regular meetings with project supervisors and Stout, when available, are held at the end of each iteration to discuss challenges and plan the next set of tasks.

2) Git: Git is an indispensable development tool used to track changes in source code. In an Agile context, Git's flexible branching and merging features enable experimentation without disrupting the main codebase, aligning with the principle of iterative development. By representing tasks as granular Issues and tagging them to different project stages, Git also enhances project organisation. In a Waterfall context, Git effectively manages the sequential development process by controlling versions for each stage, and serves as a centralised location for easily accessible documentation.

3) Programming Language: Python was selected due to its extensive libraries tailored for developing effective ML and CV solutions. Python's simplicity and strong community support make it easy to learn, enabling developers to overcome challenges more quickly. Compared to other languages, Python strikes an ideal balance with its performance and library support. This marks its superiority over R, which excels in statistical analysis and offers ML packages like *caret* but performs substantially slower. Similarly, while Julia stands out for its intuitive syntax and high-speed numerical computing, it lacks Python's popularity and extensive resources. Comparisons of the languages can be found at Appendix C.

TABLE III: Python Libraries (used to benefit the development process)

Library	Usage	Advantages	Disadvantages	Alternatives
Scikit-Learn	ML algorithms and tools. To be used in the model selection, design, training, and evaluation stages.	Intuitive, extensive documentation, easy to use, wide range of pre-built algorithms.	May not be as efficient for huge datasets.	TensorFlow and PyTorch (more complex).
OpenCV	CV and image processing. Used in the predefined feature extraction pipeline.	Comprehensive image processing and feature extraction functions, fast, easy to handle large image datasets.	Steeper learning curve, complex for simple tasks.	Scikit-Image (less comprehensive functionality); PIL (simpler but less powerful).
NumPy	Numerical computing and array operations.	Efficient array operations, broad functionality, widely used.	Less intuitive syntax for complete beginners.	SciPy (built on NumPy and not standalone).
Pandas	Data manipulation and analysis. To be used throughout most of the development process, especially in the data collection and preprocessing stages.	Powerful DataFrame object, easy data manipulation, cleaning, and exploratory data analysis.	Can be slow with huge datasets.	Dask (more complex); Vaex (newer and less mature).
Matplotlib	Data visualisation.	Highly customisable, extensive plotting capabilities	Verbose syntax, can be slow with huge datasets.	Seaborn (used sparingly, built on Matplotlib, and lacks some flexibility).

4) Python Libraries: Python libraries, seen in Table III, streamline the development process and are invaluable for this project's CV and ML tasks due to their extensive functionality and seamless integration. Notably, the chosen libraries are open-source, free, and offer extensive documentation; this ensures developers, particularly those in resource-constrained regions, can readily utilise them without financial constraints or excessive learning curves. Additionally, the large online community and open-source nature builds upon the solution's effectiveness and reliability through collaboration and continual improvement.

III. Design

A. System Model

Based on the literature review and supervisor consultations, five requirements guide the design of our AI-driven adulteration detection system for practical use in developing countries. These requirements are: comprehensiveness, accessibility, cost-effectiveness, efficiency, and scalability. To address the issue of adulteration while satisfying these five project requirements, the proposed system model comprises three structured phases, as illustrated in Figure 2: 1) Evaporation Capture, 2) Feature Extraction, and 3) Machine Learning.

Fig. 2: Flow chart of the three stages in the system model.

1) Phase One: Evaporation Capture: In the initial data collection phase, high-resolution profile images of evaporating droplets are captured over time, generating an image sequence per sample. Aligning with the project principles of comprehensiveness, accessibility, cost-effectiveness, this phase focuses on employing readily available, simple, and inexpensive equipment to thoroughly capture temporal droplet behaviour. This process is coordinated by the project's secondary supervisor.

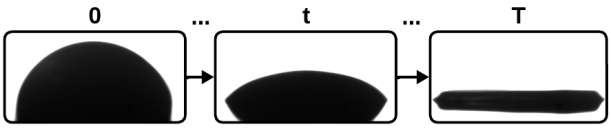

Fig. 3: Droplets captured as they evaporate over time T.

2) Phase Two: Feature Extraction: This extensive imagery forms the basis for the second phase, where the image sequences are subsequently processed through a pre-established feature extraction pipeline. This transforms the images into a singular, quantitative feature set that characterises the evolving shape of the droplets. In response to the limitations identified in existing literature—including the unscalable nature of manual detection [8], dependence on incomprehensive static data [16]–[18], and the inaccessibility, high costs, and inefficiency of spectrometers and DL [15], [19]—this phase prioritises all five project dimensions.

Stout et al.'s CV approach, which demonstrated success in a related problem, has been adapted to address the unique challenge of adulteration detection instead. Utilising an enhanced adulteration-focused feature extraction pipeline, the images are cropped, the reflection line (RL) where the droplet meets the substrate is located, and droplet heights are recorded at $k = 30$ regular intervals. Intervals towards the droplet's edges are taken more frequently to capture any nuanced fluctuations that may provide valuable insights [20]. A matrix of the droplet's outline over time is saved as a CSV file.

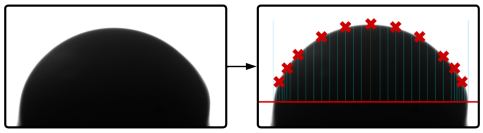

Fig. 4: Feature extraction process. The RL is first located. In intervals across the droplet's width, heights are then measured from the RL. Each image forms a row of quantitative data, creating a matrix.

3) Phase Three: Machine Learning: Finally, this time-series data serves as input for a selection of supervised ML classification algorithms to assess their effectiveness in distinguishing between pure and diluted droplets. This third phase is essential for developing an accessible, cost-effective, efficient, and scalable adulteration detection system—leveraging free, widely available software and lightweight ML algorithms.

Following Stout et al.'s approach, minor pre-processing steps are applied before training to the CSV data. To mitigate variance in observed heights, the droplet heights are normalised to a range of (0.0, 1.0). This ensures each sample's data is relative to its maximum height, which is particularly useful when relative changes are more significant than absolute physical measurements. Normalisation also improves the performance of distance-based algorithms [21]. Irrelevant timesteps after a specified time point t within T, are trimmed, optimally between 600-900, as negligible changes to droplet topology are observed beyond this point [20]. Lastly, the 2D matrix is flattened to to meet the ML algorithms' requirement for vector inputs.

Next, the Machine Learning phase involves constructing a flexible Python system capable of executing multiple models, conducting experiments and analyses, and selecting the model that best aligns with the project requirements. Two classification tasks have been selected to represent real-world adulteration problems.

The primary ML task focuses on classifying between pure and diluted samples, closely simulating the common adulteration practice in which standard full cream homogenised milk (FCM) is diluted with water to maximize profits. This fundamental problem assesses the models' ability to differentiate between the most basic categories before progressing to more complex analyses, such as detecting incremental dilutions or the presence of sophisticated adulterants often introduced in later stages of food fraud. To establish two distinct classes, the following samples were prepared: (1) Pure FCM and (2) 50:50 FCM/Water.

A secondary four-class problem is also explored to distinguish between a broader range of legitimate, market-relevant milk varieties, adding complexity and aligning the system with more diverse real-world conditions. This classification task involves: (1) Pure FCM, (2) 50:50 FCM/Water, (3) Reduced Fat Milk (RFM), and (4) Protein Plus (P+).

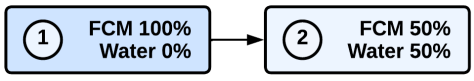

Fig. 5: Two classes for detecting diluted droplets from FCM.

B. Design and Analysis

1) Stout's Pipeline: Despite Stout et al.'s success in classifying milk types, the original pipeline—hereafter referred to as Stout's Pipeline (SP)—struggles with two fundamental weaknesses. Firstly, SP lacks robustness to environmental variations and changes in evaporation patterns caused by dilution, as noted in its review. Variations in camera angle can obscure the RL, changes in distance may blur edges, inconsistent lighting can impact contrast, and unknown artefacts can disrupt image clarity. Additionally, as seen in Figure 7, water introduces volatile elements that accelerate evaporation [8], leaving minimal discernable patterns in some diluted droplets that SP cannot effectively handle.

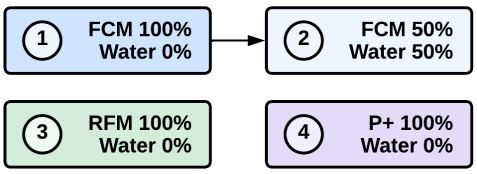

Fig. 6: Four classes for detecting diluted droplets from FCM, RFM, and P+.

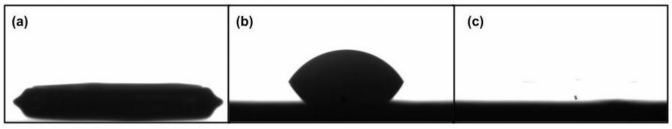

Fig. 7: Pure FCM droplet vs. diluted droplet. (a) Pure FCM droplet at the end of its evaporation process, showing notable residue. (b) Diluted droplet at the beginning of its evaporation process. (c) Diluted droplet at the end of its evaporation process, with no residue remaining.

When processing droplets differing in both environment and composition from Stout et al.'s original samples, several observable errors frequently arise, including the RL being detected at the bottom of the image and inaccurate width calculations. Furthermore, an uninformative "*Unable to find bounds*" error occurs wherein the leftmost and rightmost boundaries cannot be detected, arising when a) the RL cannot be located (and there are no "bounds" to locate) or b) when the droplet has faded beyond detection.

The second fundamental weakness is poor error handling and reporting. With hundreds of images per sample, pinpointing failures is demanding, and errors lack the sufficient detail needed for effective troubleshooting. Moreover, SP terminates prematurely upon encountering errors, and all reporting is confined to the terminal, which lacks persistence and useful logging analysis capabilities.

It is evident that an upgrade to SP is necessary for effective use in adulteration detection, prompting the design of the UIP—the Udderly Improved Pipeline.

2) The Udderly Improved Pipeline: To address SP's weaknesses, a combination of strategic design approaches is essential to optimise Phase Two: Feature Extraction. The primary issue, the lack of robustness, can be further subdivided into two problems: environmental factors and evaporated droplets.

For environmental factors, image augmentation—like adjusting brightness, contrast, and sharpness—can maintain consistent image quality regardless of variability. This is a common principle in signal processing to reduce noise and enhance the reliability of image signals [22]. For instance, histogram equalisation, a statistical imaging processing technique, can improve the contrast of an image by redistributing its pixel

intensity values to achieve a more uniform histogram [23]. However, applying such augmentation introduces additional computational overhead, significantly increasing processing time and hindering scalability as the dataset grows. Given n images and a time complexity of $O(m)$ per image adjustment, where m reflects the complexity of the augmentations, the total complexity for this step becomes $O(nm)$.

An alternative is optimising SP's hyperparameters for diverse environmental conditions. Two independent settings regulate the droplet-to-background thresholds, sensitive to lighting and distance, and two inderdependent hyperparameters govern the detection of the RL, which can trigger an "*Unable to find bounds*" error if the image angle obscures the point of symmetry. Therefore, tuning incurs a one-time cost with a complexity of $\Theta(k^2 + 2k)$, where k is the size of the search window. This constant time complexity minimises the impact on pipeline performance and resource demands compared to the cumulative costs of image augmentation.

In the second scenario leading to the "*Unable to find bounds*" error, SP struggles to handle evaporated droplets, necessitating a method to determine when to halt processing. One approach involves continuing processing until the droplet has visibly disappeared, by reducing the sensitivity of the hyperparameters that distinguish droplet from background. However, this introduces the risks of false positives, where background noise is mistaken for the droplet. Moreover, as a droplet becomes negligibly small or fragmented, its value to ML models diminishes, potentially decreasing accuracy and inflating computational costs without improving data quality. Instead, to enhance robustness beyond the capabilities of SP, the UIP will integrate threshold-based checks that monitor changes in droplet height and width, extending upon the existing bounds detection. Using the concept of 'events' in redundancy in system design ("a particular condition or change in condition" [24]), the system can deem the droplet evaporated if any prohibited events occur. This approach ensures greater reliability through the use of multiple indicators and allows the process to move forward efficiently.

This brings us to the second weakness: inadequate error handling and reporting. Four main methods were considered when encountering an error state triggered by evaporation:

1) Termination

- *Method*: Terminate the process, as employed by SP.
- *Benefit*: Avoids unnecessary overhead, prevents potential downstream errors, optimal for early errors.
- *Drawback*: Poor data retention, early errors are rare.

2) Graceful Degradation

- *Method*: Continue to process remaining images and log errors for later review.
- *Benefit*: Prioritises data retention.
- *Drawback*: Remaining images are also unlikely to process successfully, especially when evaporation leads to increasingly faint images.

3) Dynamic Adjustment

- *Method*: Retry failures with adjusted hyperparameters.
- *Benefit*: Flexible, prioritises data retention.

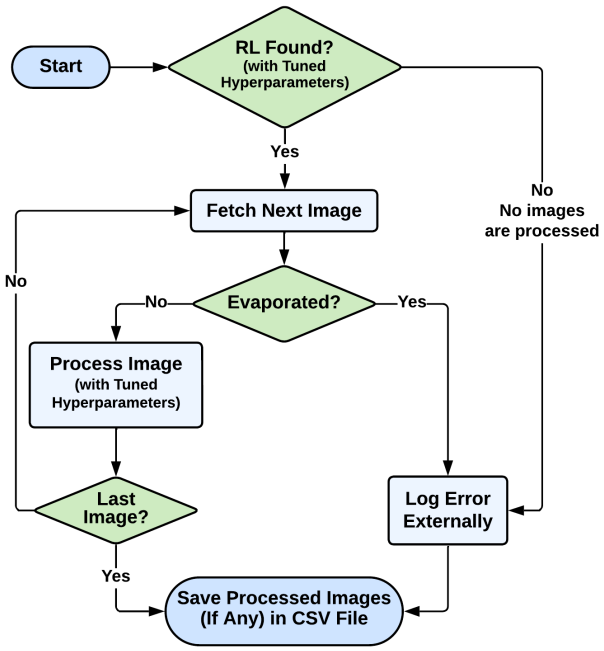

Fig. 8: Flow chart of the UIP used in Phase Two: Feature Extraction.

- *Drawback*: Extends processing times, introduces inconsistencies, creates performance bottlenecks when errors persist across multiple images.

4) Save Images

- *Method*: Halt processing at the error state and save/export only the successfully processed images. Log errors.
- *Benefit*: Balances data retention with clear failure boundaries, prevents cascading errors, minimal performance bottlenecks, ensures data consistency, reduces unnecessary resource consumption.
- *Drawback*: Not suitable for early errors.

For its adherence to efficiency, fault tolerance, and resource optimisation engineering principles, the Save Images design has been opted over the proposed alternatives. The final pipeline logic can be seen in Figure 8.

Finally, to improve reporting, an external logging system that saves detailed logs outside the terminal can allow logs to be stored, reviewed, and analysed after processing.

3) Model Selection: Selecting models offers a wide array of algorithms, each varying in complexity, computational demands, and suitability for different types of data. To prioritise accessibility, cost-effectiveness, and efficiency in Phase Three: Machine Learning, traditional supervised ML techniques are preferred over DL models.

DL perform exceptionally well in domains with large and high-dimensional data, excelling in applications involving imagery [25]; however, the complexity of DL architectures exceed this project's needs and contradicts with its key principles of accessibility, cost-effectiveness, and efficiency. ML algorithms, easier to comprehend and implement, benefit from tools like Scikit-Learn, which provide extensive documentation and community support, particularly valuable in

resource constrained regions. Moreover, ML demands less computational costs compared to DL methods, which often necessitate high-performance hardware, memory requirements, and inference time to execute a model on new data [25]. Often occupying gigabytes of storage, this makes DL models significantly less practical if deployed on less computationally capable devices.

ML algorithms also execute rapidly and offer improved interpretability, enabling understanding of prediction rationale and building trust in their outcomes [25]. Conversely, the opacity of DL may inhibit their applicability in contexts requiring human oversight, ethical considerations, bias management, transparency, and establishment of trust with stakeholders. As such, I propose six ML algorithms for Phase Three: K-Nearest Neighbours (KNN), Decision Tree (DT), Naïve Bayes Classifier (NBC), Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP).

4) Model Training: Performance metrics will be aggregated from multiple models to accurately represent the algorithms' performance. To ensure a comprehensive evaluation across diverse data partitions, three cross-validation resampling methods have been selected to be included in the experiments. This includes: Monte Carlo (MC), which randomly divides the dataset N times; K-Fold, which divides the data into N folds iteratively, training each model on $N-1$ folds while testing on the remaining fold; and Leave-One-Out (LOO), which trains on all but one sample, repeating this process for each data point [26], [27]. In contrast to holdout evaluation, which splits the dataset once into a single training and test set, these methods mitigate overfitting introduced by a single split, optimise smaller datasets, and yield more reliable results [27]. The models will also undergo a series of fine-tuning experiments, exclusively on the training set, with various hyperparameter values to determine appropriate settings.

C. Sustainability Considerations

1) Environmental: The selection of lightweight, efficient ML models mitigates environmental concerns by demanding fewer computational resources and consuming less energy than their DL counterparts, which, while effective in classification tasks, are deemed excessive for this task.

2) Social: The system design fundamentally prioritises equitable access in developing countries, where high-end resources are limited and consumers are most vulnerable due to inadequate detection infrastructure. By utilising simple equipment in Phase One, coupled with accessible and robust software in Phases Two and Three, this solution provides a comprehensive method to detect adulteration without the prohibitive nature of spectrometers and DL.

3) Technical: The system's modular design supports long-term maintenance and scalability. By structuring the process into three well-defined components, each phase can be independently updated or improved without requiring major overhauls to the entire system. Additionally, system updates can be implemented through software improvements and model retraining, avoiding costly and time-consuming hardware recalibration or replacement. The ability to adapt the model to new adulterants or classification tasks by training on new data, also ensures the system remains technically sustainable and relevant over time as adulteration practices evolve.

IV. Implementation

The implementation involves constructing the model in accordance with the phases outlined in the system model. The code can be accessed through its GitLab repository.

A. Phase One: Evaporation Capture

Following the preparation of the various milk mixtures, droplets were transferred to the substrate using a 5-microlitre pipette and captured in profile at one-second intervals by a fixed camera as they evaporated. This process could yield up to 1200 images per sample and take as long as 20 minutes. The resulting image sequences were subsequently utilised in the following phases of the project. It is important to note that this phase was entirely managed by the project's secondary supervisor, and I was not directly involved in the evaporation capture process.

B. Phase Two: Feature Extraction

As detailed in the *Design and Analysis* section and seen in Table IV, the UIP enhances the robustness, error handling, and reporting capabilities of SP.

TABLE IV: Pipeline Upgrades Overview—From SP to UIP

SP Weakness	Observable Issue	UIP Solution
Lack of Robustness: Environment	Inaccurate width/RL calculations, "*Unable to find bounds*" error if RL cannot be found.	Tune hyperparameters to be robust to environmental factors, e.g., angle, distance, lighting, unknown artefacts.
Lack of Robustness: Evaporation	"*Unable to find bounds*" error if droplet has evaporated.	Define a criterion that determines when a droplet has evaporated.
Inadequate: Error Handling	Process terminates when an error occurs.	Halt processing and export the successfully processed images.
Inadequate: Reporting Capabilities	Uninformative errors, terminal only.	External logging system, detailed messages.

1) Upgraded Reporting Capabilities: This first upgrade involved integrating an external logging system, prioritised to facilitate subsequent upgrades through detailed log analysis. Using Python's *logging* library, logs are generated for each run of the UIP and stored locally, capturing key information such as image names, timestamps, log levels, and error specifics.

2) Upgraded Robustness for Evaporation: Based on experimental results, I established criteria for determining when a droplet can be deemed evaporated:

- **Its width decreases by more than 15% from its original size.** In successfully processed diluted samples, the width of the resulting residue typically remains above 97% of its initial measurement at $t = 0$. A significant width reduction indicates rapid fading.

- **Its height falls below 20 pixels.** The data becomes negligible at this size and further processing is unjustified.
- **The droplet bounds are not detectable.**

These thresholds can be further refined to better suit the dilution problem or adapted for future applications.

3) Upgraded Error Handling: Based on the criterion above, the pipeline logic was rewritten to halt further image processing and complete the export process, thereby preserving valuable data.

4) Upgraded Robustness for Environmental Factors: Lastly, four hyperparameters were tuned to ensure accurate droplet measurements. A search window of 5 was used.

- **DROP_PXL_BORDER:** The maximum value of a black-and-white pixel for it to be considered part of the droplet when determining its height. *Default=152*
- **NONDROP:** The minimum value of a black-and-white pixel for it to be considered *not* part of the droplet when determining the side bounds. *Default=225*
- **RADIUS:** The radius of the search area (above and below a given row) used to identify the RL. *Default=10*
- **THRESH:** The maximum allowable difference between each side of the radius for a row to be considered the RL. *Default=2*

The RL is detected by scanning upwards from the bottom of a reference image and comparing RADIUS pixels above and below each row. When the difference between these pixels falls below the similarity threshold, THRESH, that row is marked as the RL. A small RADIUS risks false positives, while a larger RADIUS can potentially miss the RL altogether. Similarly, a higher THRESH may be overly permissive, whereas lower values demand a greater degree of symmetry, risking missed detection if such symmetry is absent. Given their interdependence, a grid search was conducted for RADIUS and THRESH:

1) Five values for each hyperparameter were defined based on prior knowledge and experimentation.
2) While controlling the remaining hyperparameters, the 25 combinations were each tested on six diluted samples and visually evaluated using qualitative assessments: Excellent, OK, and Poor.

The grid search results can be seen in Appendix D, in which RADIUS=30 and THRESH=2 emerged as one of the most robust combination, yielding a 100% Excellent rate.

The original DROP_PXL_BORDER value of 152 was left unchanged, but was found to be exclusively employed in the seldom-used 'bottom-up' detection method. The default 'top-down' method, which was hard-coded value to 250, was revised to apply DROP_PXL_BORDER for consistency and maintainability. A new NONDROP value of 235 also improved boundary detection for blurrier samples, while maintaining high performance on well-defined counterparts.

C. Phase Three: Machine Learning

1) Interface: A Python program was implemented to run with a versatile command-line interface (CLI) utilising Python's *Argparse* functionality. The CLI streamlines the configuration and execution of various functionalities without necessitating changes to the codebase, such as setting models, run counts, seeds, and resampling techniques. Additionally, logging to external files was set up to record telemetry data, tracking the models' performance over time.

2) Preprocessing: To overcome limitations in Stout et al.'s existing four-class hard-coded solution, a major focus involved enhancing the system to accommodate classification tasks with varying number of classes and samples with varying number of frames. Firstly, samples are dynamically loaded in from a chosen data directory, treating each subdirectory as a distinct class. The system processes the droplet CSV files by selecting relevant timesteps, flattening, normalising, and imputing missing data with zeroes to prepare them for the ML models. A notable divergence from Stout et al.'s approach lies in the enhanced preprocessing flexibility. While the original method is hard-coded to trim excess timesteps after $T* = 900$, the proposed solution enables users to define $T*$ via the command-line, which is essential for accommodating different T values between pure and adulterated droplets. If any droplet's T is less than $T*$, the droplet is padded with its last observed heights. Additionally, a label-encoding mapping class addresses Stout et al.'s hard-coded class names (especially useful when generating visualisations) and facilitates pickling into byte streams for future use.

3) Models: A Model class was constructed to accommodate the six models. The fine-tuned settings for the models can be seen in Appendix E. Models can be run sequentially for N iterations using the following cross-validation techniques: MC with a 75% train and 25% test split, K-Fold, or LOO.

4) Performance Metrics: Performance metrics are crucial for evaluating ML models, offering insights into their predictive capabilities. Outlined in Table V, the model outputs key metrics in a CSV file after training, including accuracy, Area under the Receiver Operating Characteristic Curve (ROC-AUC), precision, and recall. Confusion matrices are also provided. Among these, accuracy and ROC-AUC serve as primary indicators of model effectiveness in the experiments.

Accuracy, which measures the proportion of correct predictions, provides an intuitive gauge of overall performance but can be misleading for imbalanced datasets. ROC-AUC, ranging from 0 to 1, assesses the model's ability to discriminate between classes by evaluating the trade-off between true positive (TPR) and false positive rates (FPR). This makes it particularly valuable for datasets with uneven class distributions, providing a complementary perspective that balances the inherent weakness of accuracy.

While precision and recall offer insights into the model's performance for positive classes, they do not account for overall predictive accuracy across all classes. Consequently, accuracy and ROC-AUC serve as more comprehensive metrics, offering a balanced evaluation of the model's performance across diverse scenarios.

Lastly, inference times and model sizes on disk provide insights into the efficiency and scalability of the models.

5) Experiments: As highlighted in the literature review, diluted droplets exhibit unique evaporation processes, leading to distinct drying patterns detectable by ML models [20]. Initially, the ML task focused on binary classification between

TABLE V: Performance Metrics (TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative)

Metric	Definition	Advantages	Disadvantages
Accuracy	$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$	Intuitive; quick overall performance measure.	Misleading in imbalanced datasets; lacks error type insights (e.g., FP, FN).
ROC-AUC	$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$	Informative for imbalanced datasets.	More complex to interpret; may not reflect specific threshold performance.
Precision	Proportion of TPs among positive predictions. $\text{Precision} = \frac{TP}{TP + FP}$	Important when FPs are costly.	Misleading without recall; may overlook false positives in imbalanced datasets.
Recall	Proportion of TPs among actual positives. $\text{Recall} = \frac{TP}{TP + FN}$	Important when FNs are costly.	Misleading without precision; may ignore false positives.
Confusion Matrix	Breakdown of TPs, TNs, FPs, and FNs.	Offers detailed insights; identifies specific error types.	No single performance measure; complex for multi-class problems.

TABLE VI: Sample Counts for Each Class

Sample Count	FCM	Diluted	RFM	P+
Count	13	14	18	15

FCM and 50% diluted droplets before expanding to a secondary four-class problem that included RFM and P+. Sample counts for each class are summarised in Table VI.

Stout et al. observed that the variations in heights between consectutive frames were minimal, suggesting that not all timesteps are necessary to accurately model the evaporation process [20]. Timestep selection schemes were proposed, reducing sample size by extending the time interval between observations and focusing on the most informative timesteps.

To evaluate the effect of these schemes on the adulteration detection models, experiments were conducted using two 'Domain-Guided' (DG) approaches from Stout et al.'s study. Given the rapid evaporation of diluted droplets, a custom Udderly Different method was also developed, capturing more timesteps at the beginning of the sequence, where changes in observed heights are the most pronounced. These were compared with two schemeless approaches, trimming the data to 900 and 600, respectively.

- **DG Scheme:** First 200 timesteps, every 5th from 200-400, and every 20th from 400-900. *Total=265*
- **DG (Compounded) Scheme:** Every 2nd from 0-200, 10th from 200-400, and 40th from 400-900. *Total=132*
- **Udderly Different Scheme:** Every timestep from 0-300, and every 5th from 300-600. *Total=360*
- **Trimmed to 900:** First 900 timesteps. *Total=900*
- **Trimmed to 600:** First 600 timesteps. *Total=600*

Lastly, suitable model parameters were determined through performance evaluations across a range of configurations, utilising a search window of 5.

V. Evaluation

A. Results

1) The UIP Results: A performance analysis conducted on 14 diluted droplets revealed that SP and the UIP achieved success rates of 50.0% (7/14) and 85.7% (12/14), respectively. All seven SP failures resulted in an *"Unable to find bounds"* error—five caused by the inability to detect the RLs and two due to evaporating droplets. Further examination of the two UIP failures indicated premature detection of the RLs at the bottom of the images. These false positives stem from excessive sensitivity in the THRESHOLD x RADIUS settings, highlighting the inherent compromise between specificity and adaptability. Given the unavoidable variability of droplets due to environmental factors, these results are deemed acceptable. The UIP was further validated using two pure samples from Stout et al.'s initial dataset.

Moreover, the modifications implemented in the UIP had a negligible impact on processing times, with the observed speed difference between SP and the UIP falling within the acceptable variability of the hardware (207.97 vs. 217.74 seconds for a sample containing 1203 images, resulting in an approximate difference of 0.008 seconds per image).

Overall, the UIP's superior performance and its ability to address SP's key weaknesses demonstrate its enhanced suitability for tackling adulteration tasks.

2) FCM-Diluted Experiment Results: Performance metrics were obtained using MC Cross Validation, averaged across 50 independently trained models with a 75%/25% train-test split. Compared to LOO, which is overly sensitive to individual observations, and K-Fold, which may introduce bias from uneven class distributions, MC provides more stable performance estimate by reducing single-sample impacts and minimising bias through repeated sampling. This ensures a reliable, representative basis for comparing the following results.

The experimental results from the FCM-Diluted task, summarised in Table VII, revealed that NBC achieved the highest accuracy of 98.6% and a corresponding 0.985 ROC-AUC score when trimmed to 900 timesteps. The SVM model also demonstrated excellent discriminatory ability, reaching a 0.999 ROC-AUC under the same conditions.

These results show competitive performance relative to established benchmarks—particularly Neto et al.'s Linear Regression model (79.62% accuracy) and their Convolu-

TABLE VII: FCM-Diluted Accuracies and ROC-AUC Scores

	DG Scheme		DG Compounded Scheme		Udderly Different Scheme		Trimmed to 900		Trimmed to 600	
Model	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**
KNN	91.4	0.939	89.9	0.942	93.0	0.953	95.1	0.962	95.6	0.967
NBC	89.0	0.893	89.7	0.902	90.7	0.906	98.6	0.985	95.7	0.951
LR	89.6	0.916	88.9	0.926	90.0	0.917	94.3	0.973	90.3	0.912
SVM	78.0	0.949	75.4	0.972	81.1	0.988	90.6	0.999	86.3	0.998
DT	88.9	0.891	89.7	0.905	89.0	0.894	91.9	0.925	91.4	0.92
MLP	92.0	0.908	93.9	0.932	93.7	0.918	91.6	0.93	91.3	0.898

TABLE IX: 4 Class Accuracies and ROC-AUC Scores

	DG Scheme		DG Compounded Scheme		Udderly Different Scheme		Trimmed to 900		Trimmed to 600	
Model	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**	**Accuracy**	**ROC**
KNN	79.1	0.933	80.3	0.935	81.2	0.933	86.0	0.937	85.3	0.956
NBC	75.2	0.848	80.0	0.901	74.3	0.832	78.5	0.858	82.0	0.884
LR	77.6	0.924	79.2	0.958	78.0	0.931	87.2	0.966	89.5	0.978
SVM	78.7	0.956	80.8	0.962	78.4	0.958	83.2	0.964	88.1	0.98
DT	74.7	0.831	74.1	0.83	76.4	0.842	78.7	0.857	80.5	0.867
MLP	87.3	0.944	87.7	0.96	84.4	0.942	79.7	0.905	84.4	0.945

tional Neural Network (96.76% accuracy, 0.9985 ROC-AUC). In comparison to the Convolutional Neural Network, the slight edge in accuracy and comparable discriminatory power achieved by the NBC model is especially noteworthy, as it accomplishes this without costly spectrometry equipment or resource-intensive DL architectures.

Stout et al.'s LR model, although designed for a simpler problem involving milk types, provides another interesting point of comparison with its 96.30% accuracy and 0.999 ROC-AUC. While the NBC slightly outperformed this benchmark in accuracy, it fell short in discriminative ability. Regardless, the top-performing SVM model (under the same 900-trimmed conditions) matched Stout et al.'s ROC-AUC equally.

TABLE VIII: Average Inference Times and Model Sizes (Trimmed to 900)

	DT	**LR**	**NBC**	**SVM**	**KNN**	**MLP**
Size (KB)	1.56	113	451	737	2551	10826
Time (ms)	54.1	87.8	59.1	57.3	66.0	93.5
Time per Sample (ms)	7.7	12.5	8.4	8.2	9.4	13.4

The top-performing NBC model also exhibited a relatively quick inference time (59.1 ms for a test set of size $n = 7$, 8.4 ms per sample) and compact size (451 KB), making it a practical choice for real-world applications in developing countries. Small, fast models are preferred in these settings due to their lower resource requirements, allowing for deployment on accessible devices, reducing operational costs, and facilitating real-time decision-making. In comparison, the DT model proved even faster (54.1 ms, 7.7 ms per sample) and *exceptionally* smaller (1.56 KB) while maintaining high performance (91.9% accuracy and 0.925 ROC-AUC). Compared to larger models like MLP, which stores hundreds of weights, and KNN, which retains training data for its predictions, DT presents another viable option, especially as dataset sizes increase.

A comparison of the two trimmed approaches with the timestep selection schemes (see Appendix F) revealed that the application of schemes, particularly DG-Compounded, significantly reduced inference times and model sizes. For instance, 50 independent MLP models employing the 900-trimmed schemeless approach averaged an inference time of 93.5 ms (13.4 ms per sample) and a model size of 10826 KB. In contrast, the DG-Compounded scheme recorded 34.3 ms (4.9 ms per sample) and 1623 KB—a 63.3% and 85% decrease, respectively.

However, despite these notable savings, all models except the MLP experienced minor performance decreases under schemed conditions. Interestingly, the MLP classifier excelled, achieving 93.9% accuracy and a 0.932 ROC-AUC with DG-Compounded, and 93.7% accuracy with a 0.918 ROC-AUC under the Udderly Different scheme. Nevertheless, the minor performance deficit in the other models suggests that the transformed data still mostly retains the essential characteristics of the original dataset.

The reduction in computational demands from the schemes aligns with the project's focus on developing lightweight, high-performing models suitable for resource-constrained developing countries. Faster and smaller models are essential for practical applications, particularly on accessible but less capable devices like smartphones. However, further refinement is necessary to determine the optimal timestep selection scheme for diluted data, which could potentially minimise or overcome the gap between schemed and schemeless approaches.

3) FCM-Diluted-RFM-P+ Experiment Results: Next, the experiments were performed on the four-class classification problem, distinguishing between FCM, Diluted, RFM, and P+. The results can be seen in Table VII, with supplementary inference times and model sizes found in Appendix G.

The LR classifier under the 600-trimmed schemeless approach emerged as the most accurate model, achieving 89.5% accuracy and 0.978 ROC-AUC. The SVM model, under the same conditions, achieved a lower accuracy of 88.1% but demonstrated the highest ROC-AUC score of 0.98, identifying the positive class correctly across different thresholds. It is evident that this problem, which aims to align the system with more diverse real-world conditions, is more complex than the primary binary classification problem.

Investigating into this slightly poorer performance revealed that several droplets consistently exhibited high misclassification rates, ranging from 50% to 100% across multiple models. These are seen in Table X, with its average and maximum misclassification rate summarised across the six models.

Principal Component Analysis (PCA), a dimensionality reduction technique that identifies features with the highest variance, was employed to investigate these misclassifications. In Figure 9a, the 2D representations of droplet data revealed two points—Sample 13 and 55—significantly distant from their respective class majority, indicating a deviation from the primary trends captured by the principal components. Further investigation found that Sample 55 contained substantial

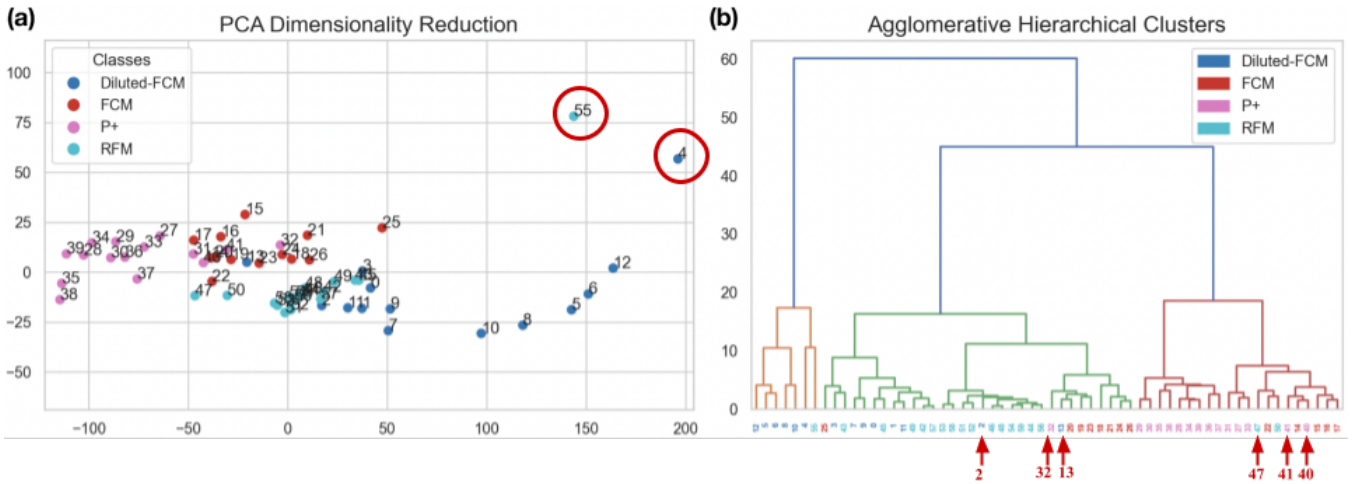

Fig. 9: PCA and Agglomerative Clustering Dendrogram of FCM, Diluted, RFM, and P+ droplets.
(a) Outlier Samples 55 and 4 have circled in red. (b) Outlier Samples 2, 13, 47, 32, 40, and 41 have been marked in red.

TABLE X: Persistently Misclassified Samples

ID	Name	Class	Avg. Rate	Max. Rate
2	240404D11	Diluted	83.17%	100%
3	240404D12	Diluted	85.57%	100%
13	240423E	Diluted	100%	100%
32	221128CP2	P+	97.92%	100%
40	221217F	P+	51.8%	93.75%
41	221217G	P+	76.58%	100%
47	221128BP	RFM	82.98%	100%
55	221219E	RFM	85.42%	100%

numerical errors in its initial timestep; however, removing the erroneous data repositioned the sample within its class boundaries and notably decreased its misclassification rate to an acceptable level.

To validate potential outliers, Agglomerative Clustering was also utilised. This hierarchical clustering technique progressively merges individual data points into clusters based on similarity, forming a dendrogram seen in Figure 9b. The analysis revealed that frequently misclassified samples exhibited noticeable distance from their main class clusters. Samples 2, 13, and 47 were considerably distant from their Diluted and RFM counterparts, while Samples 32, 40, and 41 also showed similar, but less significant, separation from other P+ instances.

These observations prompted a deeper investigation into the causes of outlier status and misclassification, particularly concerning physical characteristics not captured by the spatio-temporal data, such as volume, contact angle, and wetting area.

While the Evaporation Capture phase aims to transfer 5 microlitres (µL) droplets to the substrate, this imperfect process can lead to volume variations from 3µL to 6µL. Although droplet measurements are normalised based on maximum height during evaporation, the droplet's physical characteristics distinctly affect the *rate* of evaporation. Droplets with larger volumes typically evaporate more gradually than smaller ones due to greater mass, meaning two droplets within the same class may exhibit similar spatio-temporal patterns while differing in evaporation rate.

Illustrated in Figure 10, the contact angle at which the droplet meets the substrate further impacts evaporation dynamics. Greater contact angles indicate a more spherical shape and less spreading, while lower angles allow for a greater wetting area, accelerating evaporation by increasing the surface area interacting with the substrate. Volume and contact angle/wetting area interact in complex ways that can be unpredictable, introducing variability that inhibits a model from fully grasping the evaporation process using spatio-temporal data alone.

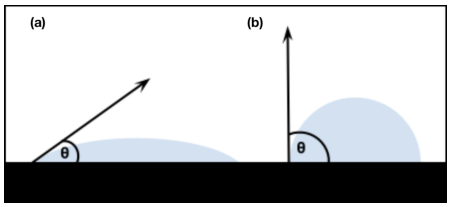

Fig. 10: Low vs. high contact angle. Despite having the same volume (not to scale), droplet in (a) has a larger wetting area compared to the droplet in (b).

To visualise these factors, three Box-and-Whiskers graphs (see Figure 11) were produced for volume, contact angle, and wetting area, providing a visual summary of central tendency and variability within the droplet measurements. Spanning from the 25th percentile to the 75th percentile, the box represents the interquartile range (IQR), with the 50th percentile (median) indicated by a line within the box. Whiskers extend to the minimum and maximum values within 1.5 times the IQR from the quartiles, while outliers beyond this range are displayed as individual points.

Across the three plots, numerous misclassified instances were positioned on the whiskers, either above or below the central 50%. Notably, the wetting area graph revealed that most misclassified instances fall on the lower whisker of their respective classes. Although these samples remain within the acceptable variability for their class, the relationship between their classification and positioning is particularly revealing. For frequently misclassified P+ samples, the average misclassifi-

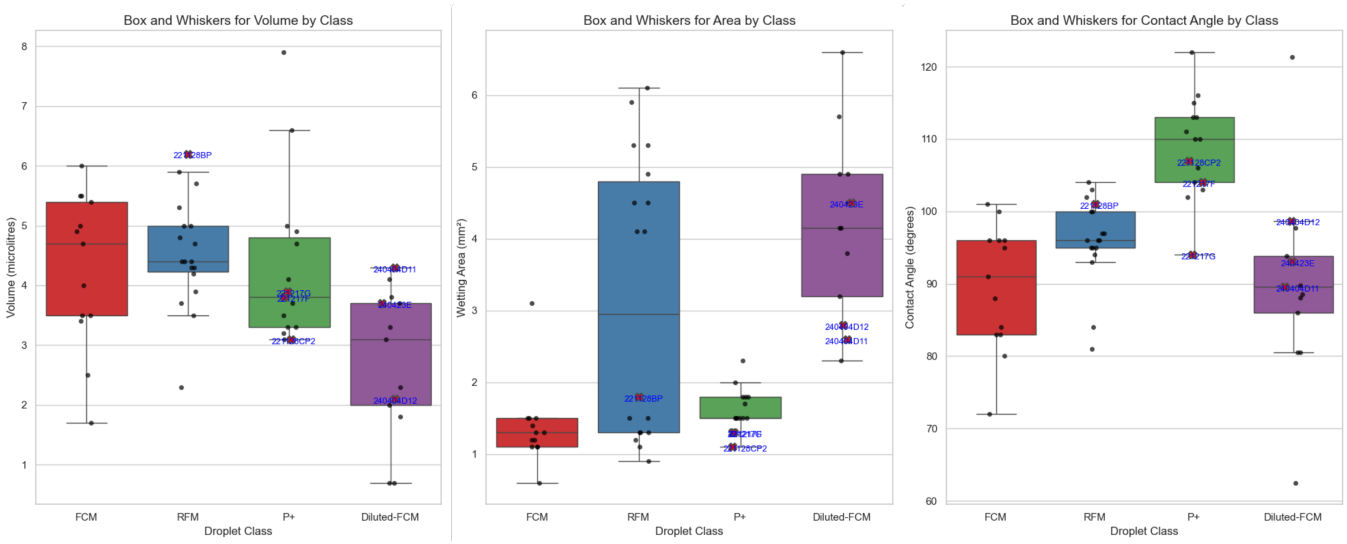

Fig. 11: Three box and whiskers graphs, plotting the distribution of volume, contact angle, and wetting area by class. Data points identified as persistently misclassified samples have are marked as red Xs.

cation rate ranged from 51.8-97.92%, with 86–96% of these misclassifications incorrectly predicting them as FCM. This trend is illustrated in the graph, where these droplets exhibit smaller wetting areas for their class, situating them within the central 50% of the FCM box. Similarly, the two Diluted droplets were misclassified 83.17-100% of the time, with 88-100% of those misclassifications being falsely identified as RFM droplets. Although diluted milk contains reduced fats from added water, aligning its composition more closely with RFM, its wetting areas also resemble the median of the RFM class, which likely contributes to its frequent misclassification.

Overall, these findings indicate that variations in physical characteristics, and consequently evaporation patterns, can cause droplets to exhibit behaviour akin to that of other classes. This similarity impedes the models' ability to precisely distinguish between classes. Therefore, there is a clear need for more refined features in future analyses that consider the physical properties of droplets. Supplementary XY plots of volume against both wetting area and contact angle can be seen in Appendix H and I, respectively.

B. Limitations

While the project fulfils the project requirements it sought out to achieve, it faces two key limitations.

1) Timestep Selection Schemes: The results in the FCM-Diluted classification task demonstrate a notable performance drop under timestep selection schemes, highlighting the challenge of selecting effective timesteps to minimise redundant data, improve model generalisation, and ensure system efficiency. This issue becomes more complex with the addition of classes, each with distinct patterns that need thorough capture. However, relying on schemes based on unadulterated samples risks overlooking the unique evaporation dynamics of adulterated samples, which poses a significant limitation. Developing robust schemes that accurately capture evaporation patterns in both pure and diluted droplets—or implementing a dynamic approach to identify the most informative segments in the time-series—would greatly improve the adaptability, efficiency, and comprehensiveness of the system.

2) Physical Measurements: The investigation into persistent and unexpectedly high misclassification rates uncovered the influence of physical properties on evaporation rates. This exposes a significant limitation in the current system's focus on spatio-temporal characteristics without accounting for these physical factors, thus complicating classification between classes. Integrating these properties into the data or standardising evaporation profiles could lead to significant improvements to the system's comprehensiveness. For instance, scaling evaporation rates by approximating the interaction between volume and contact angle/wetting area would allow droplets with varying physical properties to be represented as how they would exist under standardised conditions.

VI. Conclusion and Future Work

Overall, 'Udderly Advanced' has emerged as a promising initiative in the fight against milk adulteration, demonstrating AI's potential in democratising adulteration detection infrastructure in developing countries. The project has reached significant milestones, including the development of a droplet dataset requiring only a camera and substrate, and substantial upgrades to an existing feature extraction pipeline, which improved accuracy from 50.0% to 85.7%. Additionally, six computationally-efficient ML models were trained, achieving accuracy rates exceeding 98.6% for dilution detection and 89.5% for classifying diluted droplets among broader milk types. These models demonstrate inference times as fast as 54.1 ms on our test set, with disk sizes as small as 1.56 KB. Further analysis of selection schemes for time-series data and frequently misclassified droplets has identified future directions for optimising the dataset. Fundamentally, the project's success lies it its ability to bridge the gap between sophisticated laboratory techniques and practical, field-deployable solutions, embodying core principles crucial for its

adoption: comprehensiveness, accessibility, cost-effectiveness, efficiency, and scalability.

However, the journey of 'Udderly Advanced'—and the challenge of adulteration—is far from over. Adulteration in practice often involves even subtler alterations, with dilution levels ranging from 2% to 20% [2]. This calls for future work with more granular and complex dilution ratios, such as: (1) Pure FCM, (2) 75:25 FCM/Water, (3) 50:50 FCM/Water, and (4) 25:75 FCM/Water. Additionally, post-dilution adulterants, like melamine, pose a particularly insidious threat and present more advanced classification problems. To simulate trickier real-world adulteration scenarios, melamine could be introduced into 50% diluted FCM at a concentration of 250 parts per million (ppm), far exceeding the New Zealand Food Safety's maximum threshold (2.5 ppm) by a factor of 100 [28]; this first ensures the detection of substantial concentrations before tackling lower, more challenging concentrations. A future objective could be to develop a model to classify: (1) Pure FCM, (2) 50:50 FCM/Water, and (3) 250 ppm melamine in 50:50 FCM/Water.

These future classification objectives would not only test the sensitivity of the system but also provide valuable insights into adulteration levels and types, crucial for regulatory bodies and the dairy industry alike.

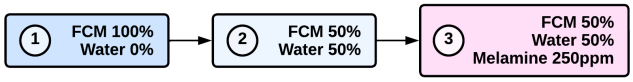

Fig. 12: A potential three-class classification problem for detecting diluted and melamine contaminated droplets.

Lastly, looking ahead, this AI-based approach offers potential far beyond its current application. By prioritising accessibility and efficiency, the project lays the groundwork for mobile integration. A future is conceivable where a smartphone application, utilising models developed in this project, can provide instant milk quality assessments at all points across the supply chain—from farm to consumer. This vision of ubiquitous, real-time food quality assessment could transform global food safety practices, 'udderly advancing' milk quality control in the regions most vulnerable to adulteration.

REFERENCES

[1] T. K. Thorning, A. Raben, T. Tholstrup, S. S. Soedamah-Muthu, I. Givens, and A. Astrup, "Milk and dairy products: Good or bad for human health? an assessment of the totality of scientific evidence," *Food & Nutrition Research*, vol. 60, no. 1, Nov. 2016, Article 32527.

[2] C. Handford, K. Campbell, and C. Elliott, "Impacts of milk fraud on food safety and nutrition with special emphasis on developing countries," *Comprehensive Reviews in Food Science and Food Safety*, vol. 15, Oct. 2015.

[3] A. Yadav, M. Gattupalli, K. Dashora *et al.*, "Key milk adulterants in india and their detection techniques: a review," *Food Analytical Methods*, vol. 16, pp. 499–514, Dec. 2023.

[4] C. M. Gossner, J. Schlundt, P. Ben Embarek, S. Hird, D. Lo-Fo-Wong, J. J. Beltran, K. N. Teoh, and A. Tritscher, "The melamine incident: implications for international food and feed safety," *Environmental Health Perspectives*, vol. 117, no. 12, pp. 1803–1808, Dec. 2009.

[5] K. Jurica, I. Brčić Karačonji, D. Lasić, D. Bursać Kovačević, and P. Putnik, "Unauthorized food manipulation as a criminal offense: Food authenticity, legal frameworks, analytical tools and cases," *Foods*, vol. 10, no. 11, Oct. 2021, Article 2570.

[6] M. Musa and S. Yang, "Common milk adulteration in developing countries cases study in China and Sudan: A review," *Advances in Dairy Research*, vol. 5, Oct. 2017.

[7] M. Guetouache, B. Guessas, and S. Medjekal, "Composition and nutritional value of raw milk," *Issues in Biological Sciences and Pharmaceutical Research*, vol. 2, no. 10, pp. 115–122, Dec. 2014.

[8] V. Kumar and S. Dash, "Evaporation-based low-cost method for the detection of adulterant in milk," *ACS Omega*, vol. 6, no. 41, pp. 27 200–27 207, Oct. 2021.

[9] T. Azad and S. Ahmed, "Common milk adulteration and their detection techniques," *International Journal of Food Contamination*, vol. 3, no. 1, Dec. 2016, Article 22.

[10] K. Goyal, P. Kumar, and K. Verma, "Food adulteration detection using artificial intelligence: A systematic review," *Archives of Computational Methods in Engineering*, vol. 29, no. 1, pp. 397–426, Jan. 2022.

[11] United Nations. Sustainable development goals (sdgs). https://sdgs.un.org/goals. Accessed: May 20, 2024.

[12] A. Haji, K. Desalegn, and H. Hassen, "Selected food items adulteration, their impacts on public health, and detection methods: A review," *Food Sci Nutr*, vol. 11, no. 12, pp. 7534–7545, Oct. 2023.

[13] A. Ayub Khan, A. A. Laghari, and S. Ahmed Awan, "Machine learning in computer vision: A review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8, no. 32, Apr. 2021, Article e4.

[14] C. Acuña, A. Mier y Terán, M. O. Kokornaczyk, S. Baumgartner, and M. Castelán, "Deep learning applied to analyze patterns from evaporated droplets of viscum album extracts," *Scientific Reports*, vol. 12, no. 1, Sep. 2022, Article 15332.

[15] H. A. Neto, W. L. Tavares, D. C. Ribeiro, R. C. Alves, L. M. Fonseca, and S. V. Campos, "On the utilization of deep and ensemble learning to detect milk adulteration," *BioData Mining*, vol. 12, no. 1, Jul. 2019, Article 13.

[16] A. Harindran, S. Hashmi, and M. Vinjanampati, "Pattern formation of dried droplets of milk during different processes and classifying them using artificial neural networks," *Journal of Dispersion Science and Technology*, vol. 43, pp. 1–10, Feb. 2021.

[17] A. M. Pérez-Calabuig, S. Pradana-López, S. Lopez-Ortega, K. d. J. Beleño Sáenz, J. C. Cancilla, and J. S. Torrecilla, "Application of residual neural networks to detect and quantify milk adulterations," *Journal of Food Composition and Analysis*, vol. 122, 2023, Article 105427.

[18] S. Andalib, K. Taira, and H. P. Kavehpour, "Data-driven time-dependent state estimation for interfacial fluid mechanics in evaporating droplets," *Scientific Reports*, vol. 11, no. 1, Jun. 2021, Article 13579.

[19] Y. Suh, J. Lee, P. Simadiris, X. Yan, S. Sett, L. Li, K. Rabbi, N. Miljkovic, and Y. Won, "A deep learning perspective on dropwise condensation," *Advanced Science*, vol. 8, no. 22, Sep. 2021, Article 2101794.

[20] A. Stout, G. Gouws, A. Lensen, and H. Al-Sahaf, "Culturing insights: Lean machine learning transforms high-dimensional dairy data," 2024, In Progress, Final draft in Preparation.

[21] I. Niño-Adan, I. Landa-Torres, E. Portillo, and D. Manjarres, "Influence of statistical feature normalisation methods on k-nearest neighbours and k-means in the context of industry 4.0," *Engineering Applications of Artificial Intelligence*, vol. 111, 2022, Article 104807.

[22] S. Smith, "Chapter 23: Image formation and display," in *Digital Signal Processing: A Practical Guide for Engineers and Scientists*, 1st ed. Elsevier, Oct. 2002, pp. 1–10.

[23] M. S. Nixon and A. S. Aguado, "Chapter Three - Image processing," in *Feature Extraction and Image Processing for Computer Vision (Fourth Edition)*, fourth edition ed., M. S. Nixon and A. S. Aguado, Eds. Academic Press, Nov. 2020, pp. 83–139.

[24] T. Yellman, "Redundancy in designs," *Risk Analysis*, vol. 26, no. 1, pp. 277–286, Jan. 2006.

[25] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, pp. 685–695, Sep. 2021.

[26] R. T. Nakatsu, "Validation of machine learning ridge regression models using monte carlo, bootstrap, and variations in cross-validation," *Journal of Intelligent Systems*, vol. 32, no. 1, Jul. 2023, Article 20220224.

[27] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015.

[28] Y.-C. Tyan, M.-H. Yang, S.-B. Jong, C.-K. Wang, and J. Shiea, "Melamine contamination," *Analytical and Bioanalytical Chemistry*, vol. 396, p. 729–735, Aug. 2009.