

Machine Learning-Based Predictive Modeling for Early Detection of Subclinical Mastitis in Dairy Herds

Chethan M.S¹, Anand Inamdar², Chinmay P Gowda³, Samarth Biradar⁴,

Prof. Dinesh Babu K B⁵

Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India¹⁻⁴

Assistant Professor, Dept. of CSE, Don Bosco Institute of Technology, Bengaluru, India⁵

Abstract: Subclinical mastitis (SCM) is a symptomless yet economically damaging udder infection that silently reduces milk yield and quality in dairy cattle, making early identification difficult through conventional visual inspection. This paper presents a machine-learning based predictive framework for early SCM detection using three routinely measurable and low-cost milk parameters: Somatic Cell Count (SCC), pH, and Electrical Conductivity (EC). A dataset of 5,000 milk samples (14.6% mastitis-positive) was enhanced through feature engineering, including logarithmic transformation and SCC-to-EC ratio, and rolling statistical features that capture infection trends over time. Logistic Regression and LightGBM classifiers were trained and compared, and the tuning of LightGBM using Optuna-based automated hyperparameter tuning achieved a test accuracy of approximately 98.8%, with precision of 0.93, recall of 0.96, and an F1-score of 0.94, outperforming Logistic baseline. The SHAP (Shapley Additive Explanations) analysis confirmed SCC and EC as the dominant predictive features, consistent with established veterinary knowledge, and validated the biological relevance of the engineered features. The resulting framework provides an accurate, interpretable, and cost-effective decision-support tool suitable for real-world dairy farm deployment.

Keywords: subclinical mastitis, machine learning, LightGBM, somatic cell count, electrical conductivity, explainable AI, SHAP, precision dairy farming

I. INTRODUCTION

Mastitis is among the most common economical disease affecting dairy cattle worldwide, causing reduced milk yield, degraded milk quality, and increased veterinary costs. Subclinical mastitis (SCM) is particularly concerning as it produces no visible symptoms, allowing infected cows to appear healthy yet silently contributing to long-term productivity decline and disease transmission within the herd.

Conventional detection involves lab-based Somatic Cell Count (SCC) analysis, milk pH measurement, and periodic physical examination. While effective, these methods are reactive, time-consuming, and often expensive for continuous monitoring on small and resource-constrained farms, resulting in infections being identified only after considerable damage has occurred.

Recent advances in machine learning (ML) enable automated analysis of routinely collected milk data to discover subtle infection patterns that precede clinical signs. This paper proposes an explainable machine-learning-based predictive framework that uses three low-cost, easily-measurable milk parameters — SCC, pH, and Electrical Conductivity (EC) — to detect subclinical mastitis at an early stage, by combining feature engineering, optimized ensemble learning, and explainable AI, to provide accurate and interpretable decision support for dairy health management.

The objectives of this work are to: (i) analyze routinely collected milk parameters for mastitis detection; (ii) engineer features such as log-transforms, ratios, and rolling statistics to enhance predictive capability; (iii) implement and compare Logistic Regression and LightGBM classifiers; (iv) optimize model performance using automated hyperparameter tuning; (v) apply SHAP-based explainability to verify biological relevance; and (vi) evaluate performance using classification metrics.

II. RELATED WORK

A growing body of research has employed machine learning and IoT-based sensing for mastitis detection. Shiotsu et al. [1] examined the applicability of ML methods to improve mastitis prediction in livestock, while Khan et al. [2] proposed an IoT-enabled intelligent decision support system to identify mastitis cases in dairy farms. Motohashi et al. [3] used automated milking system (AMS) data for early subclinical mastitis detection, and Kassab et al. [4] reviewed IoT applications for precision livestock health monitoring more broadly.

Sensor-driven ensemble approaches have also been explored: Jensen et al. [5] combined sensor data with ensemble models for mastitis prediction, Fadul-Pacheco et al. [6] applied Random Forest classifiers with somatic cell score, and Ghafoor et al. [7] addressed class imbalance through up-sampling techniques. Luo et al. [8] used milk conductivity data for IoT-driven detection of clinical mastitis, while Bobbo et al. [9] identified subclinical cases using cow-level features. Zhang et al. [10], Chen et al. [11], and Wang et al. [12] further compared machine-learning classifiers for clinical mastitis risk prediction, and Singh et al. [13], Lee et al. [14], and Kumar et al. [15] evaluated various ML algorithms for early and lactation-stage mastitis diagnosis.

Two supporting techniques underpin the present work: Optuna, an automated hyperparameter optimization framework proposed by Akiba et al. [16], and SHAP (Shapley Additive Explanations), introduced by Lundberg and Lee [17] for interpreting complex model predictions. Table I summarizes representative studies and their limitations, most of which involve limited datasets or high deployment complexity or lack of interpretability — which the proposed framework aims to address using a low-cost, explainable, and deployable approach.

TABLE I
Summary of Related Work

Ref.	Author / Year	Focus	Limitation
[1]	Shiotsu et al., 2024	ML-based mastitis prediction accuracy improvement	Limited dataset; no real-time farm deployment
[2]	Khan et al., 2024	IoT-enabled decision support system for mastitis	High system complexity and deployment cost
[3]	Motohashi et al., 2020	Subclinical mastitis detection via AMS data	Restricted to AMS-equipped farms
[5]	Jensen et al., 2022	Sensor-based ensemble ML for mastitis prediction	Dependent on sensor quality
[6]	Fadul-Pacheco et al., 2022	Random Forest classification using SCC	Environmental factors not included
[9]	Bobbo et al., 2023	Cow-level feature-based early detection	Requires detailed farm records
[17]	Lundberg & Lee, 2017	SHAP for explainable model predictions	Computationally expensive at scale

III. DATASET AND FEATURE ENGINEERING

The study uses a dataset of 5,000 milk samples collected across multiple cows, each labelled as healthy (0) or mastitis-affected (1). The dataset is imbalanced with 4,270 healthy samples (85.4%) and 730 mastitis-positive samples (14.6%). Each record includes cow identifier, sample date, SCC (cells/mL), pH, Electrical Conductivity (EC, mS/cm), milk yield (L), fat percentage, and protein percentage.

Raw values were cleaned of missing entries and outliers, and numerical features were scaled before model training. Feature engineering was then applied to expose biologically meaningful infection signatures: a log-transform of SCC (`scc_log`), an SCC-to-EC ratio (`scc_cond_ratio`), rolling mean and rolling standard deviation of SCC over a 3-sample window per cow, the deviation of the current SCC from its rolling mean, a per-cow SCC trend (first difference), an $\text{SCC} \times \text{pH}$ interaction term, and the number of days since the previous sample per cow. Boxplot comparisons (Fig. 1)

confirmed that SCC and EC distributions shift noticeably higher in mastitis-positive samples, validating the discriminative value of these features before model training.

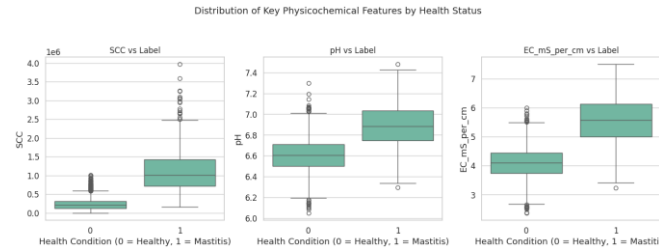


Fig. 1. Boxplot comparison of SCC, pH, and EC distributions for healthy vs. mastitis-affected samples.

IV. PROPOSED SYSTEM ARCHITECTURE

The proposed system follows a six-module sequential pipeline, taking raw milk sensor/laboratory data through preprocessing, feature engineering, model training, optimization, evaluation, and explainability stages, to produce an interpretable prediction.

A. Data Preprocessing Module

Removes missing values and outliers that could distort learning, and normalization (min-max scaling or standardization) so that no single feature dominates model training due to different numeric ranges.

B. Feature Engineering Module

Computes the ratio- and time-based features described in Section III, to translate raw physicochemical readings to indicators that better reflect the physiological progression of infection.

C. Model Selection Module

Logistic Regression is used as an interpretable linear baseline, while LightGBM — a gradient-boosted tree ensemble — is selected as the primary model for its ability to capture nonlinear feature interactions and to handle imbalanced, tabular data efficiently.

D. Hyperparameter Tuning Module (Optuna)

Optuna [16] automatically searches the LightGBM hyperparameter space (learning rate, tree depth, number of estimators, and related parameters) using intelligent sampling to maximize validation performance while reducing overfitting and manual tuning effort.

E. Model Evaluation Module

Cross-validation is used to obtain unbiased performance estimates. Accuracy, precision, recall, and F1-score are calculated on a held-out test split, with particular attention to minimizing false negatives, given the cost of missed infections in veterinary screening.

F. Explainability Module (SHAP)

SHAP [17] values quantify each feature's contribution to individual and global predictions, exposing which physicochemical indicators drive the model's decisions and allowing veterinary experts to validate the model's reasoning against known biology.

V. IMPLEMENTATION

The system was implemented in Python using Pandas and NumPy for data handling and feature engineering, Scikit-learn and LightGBM for model construction and evaluation, Optuna for hyperparameter optimization, SHAP for explainability, and Matplotlib/Seaborn for visualization. No specialized hardware was required; the pipeline runs on a standard laptop-class processor, and SCC, pH, and EC readings are assumed to originate from milk testing devices already in use on dairy farms or in laboratories, making the approach practically deployable for resource-constrained deployment.

VI. RESULTS AND DISCUSSION

The dataset (5,000 samples) was split into training and held-out test sets. Logistic Regression achieved strong baseline performance due to clear linear relationships between SCC/EC and infection status; however, the Optuna-tuned LightGBM model outperformed it, achieving a test accuracy of approximately 98.8%. Table II reports precision, recall, and F1-score for both models.

TABLE II
Model Evaluation on Held-Out Test Set

Model	Precision	Recall	F1-Score
Logistic Regression	0.91	0.96	0.93
Tuned LightGBM	0.93	0.96	0.94

The tuned LightGBM model's F1-score of 0.94 indicates balanced sensitivity to infected samples and precision on healthy cases. On the test set of 1,250 samples, the confusion matrix (Fig. 2) recorded 1,061 true negatives, 7 false positives, 8 false negatives, and 174 true positives — an overall accuracy of 98.8% with very few missed infections, which is critical in disease-screening applications where false negatives are costlier than false alarms.

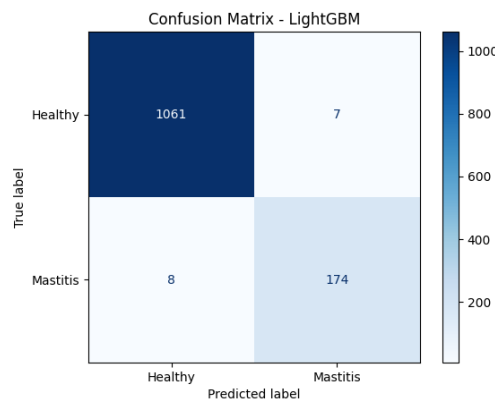


Fig. 2. Confusion matrix of the tuned LightGBM model on the test dataset.

SHAP-based explainability analysis (Fig. 3, Fig. 4) identified SCC and EC (mS/cm) as the two most influential predictors, followed by milk yield, the log-transformed SCC, and pH. High SCC and EC values were consistently associated with mastitis-positive predictions, aligning with established veterinary knowledge that infection raises somatic cell counts and ionic concentration in milk. Engineered features — the SCC-to-EC ratio and the rolling mean of SCC — provided additional diagnostic value by capturing the stability and trend of infection markers over time versus relying on single-timepoint readings.

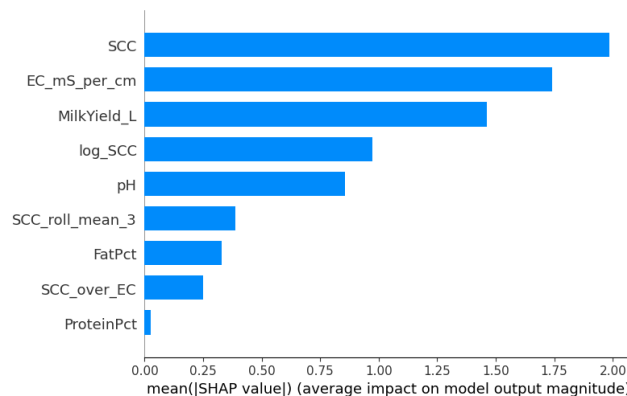


Fig. 3. Global SHAP feature-importance ranking for the LightGBM model.

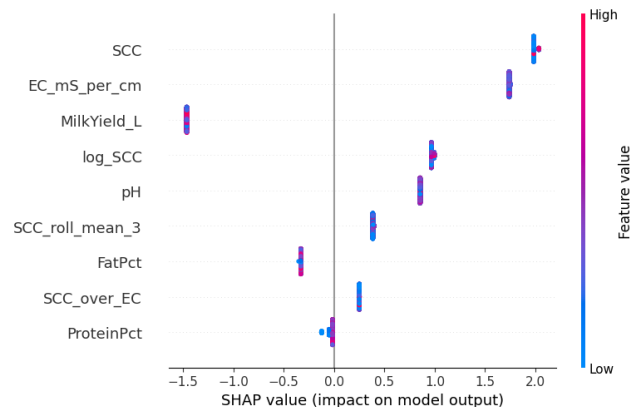


Fig. 4. SHAP summary (beeswarm) plot showing the direction and magnitude of each feature's effect on individual predictions.

These findings confirm that gradient-boosted ensembles capture nonlinear feature interactions relevant to early-stage subclinical mastitis more effectively than a linear baseline, and the SHAP analysis transforms the model from an opaque predictor to an interpretable decision-support tool, an important property for building trust among veterinarians and farmers.

VII. CONCLUSION AND FUTURE SCOPE

This paper presented a machine-learning-based predictive framework for early detection of subclinical mastitis using low-cost, routinely measurable milk parameters — SCC, pH, and EC. Feature engineering enhanced the model's ability to capture subtle infection patterns, and Optuna-based hyperparameter tuning and SHAP-based explainability produced a classifier that is both highly accurate (98.8% test accuracy, F1-score of 0.94 for the tuned LightGBM model) and interpretable. The framework offers a practical, cost-effective decision-support tool for veterinarians and farmers, enabling timely intervention, improved milk quality, and sustainable herd health management.

Future work will extend the binary classification to multiclass mastitis-severity staging, integrate the framework with IoT-enabled milk sensors and automated milking systems for real-time monitoring, validate the model across diverse geographic regions and cattle breeds, explore edge/cloud deployment for large-scale adoption, and develop farmer-friendly visual dashboards to translate the model's outputs into actionable on-farm decisions.

REFERENCES

- [1]. H. Shiotsu, N. P. Martono, and H. Ohwada, "Applicability of Machine Learning to Improve Mastitis Prediction in Livestock," *Proc. IEEE Int. Conf. Ind. Eng. Eng. Manag. (IEEM)*, pp. 1149–1153, 2024.
- [2]. M. F. Khan, V. M. Thorup, and Z. Luo, "Delineating Mastitis Cases in Dairy Cows: Development of an IoT-Enabled Intelligent Decision Support System for Dairy Farms," *IEEE Trans. Ind. Informat.*, vol. 20, no. 7, pp. 9508–9513, 2024.
- [3]. R. Motohashi, K. Tanaka, H. Morikawa, and T. Hayashi, "Early Detection Method for Subclinical Mastitis Using Data from Automated Milking Systems," *Proc. IEEE Glob. Conf. Consum. Electron. (GCCE)*, pp. 199–200, 2020.
- [4]. M. Kassab et al., "Livestock IoT: Precision Livestock Management in Agribusiness," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 511–526, 2024.
- [5]. T. Jensen, P. Larsen, and M. Hansen, "Prediction of Mastitis in Dairy Cattle Using Sensor Data and Ensemble Models," *Proc. IEEE Int. Conf. Big Data*, pp. 1231–1240, 2022.
- [6]. T. Fadul-Pacheco, M. E. Galo, and A. Sitkowska, "Predictive Analytics in Dairy Cattle Health: Mastitis Classification Using Random Forest and Somatic Cell Score," *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, pp. 765–770, 2022.
- [7]. A. Ghafoor and A. Sitkowska, "Addressing Class Imbalance in Mastitis Detection Using Up-Sampling and Random Forest," *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, pp. 88–93, 2023.
- [8]. Z. Luo, Y. Wang, and B. Wang, "IoT-Driven Detection of Clinical Mastitis in Cattle Using Milk Conductivity and Machine Learning," *IEEE Access*, vol. 11, pp. 6201–6214, 2023.
- [9]. T. Bobbo, E. Fiore, M. Ganesella, M. Morgante, and A. Cecchinato, "Early Detection of Subclinical Mastitis in Lactating Dairy Cows Using Cow-Level Features," *J. Dairy Sci.*, vol. 106, no. 5, pp. 4321–4332, 2023.

- [10]. Y. Zhang, X. Li, H. Wang, and J. Liu, "Risk Prediction Model of Clinical Mastitis in Lactating Dairy Cows Based on Machine Learning Algorithms," *Prev. Vet. Med.*, vol. 205, 106059, 2023.
- [11]. L. Chen, Z. Wang, and Q. Zhang, "Detection of Mastitis Disease in Cows with Machine Learning Classifiers," *Comput. Electron. Agric.*, vol. 205, 107634, 2023.
- [12]. Y. Wang, H. Sun, and X. Zhao, "Exploring Machine Learning Algorithms for Early Prediction of Clinical Mastitis," *Comput. Electron. Agric.*, vol. 204, 107599, 2023.
- [13]. R. Singh and P. Kumar, "An AI-Based System for Detecting Mastitis in Dairy Cattle," *Int. J. Intell. Syst. Appl.*, vol. 15, no. 3, pp. 45–52, 2023.
- [14]. H. Lee, J. Park, and S. Kim, "Machine Learning Approach for Early Lactation Mastitis Diagnosis in Dairy Cows," *Animals*, vol. 13, no. 2, p. 274, 2023.
- [15]. S. Kumar and R. Sharma, "Mastitis Diagnosis with Machine Learning Algorithms," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 12345–12358, 2023.
- [16]. T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, pp. 2623–2631, 2019.
- [17]. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Proc. Adv. Neural Inf. Process. Syst. (NIPS 2017)*, vol. 30, pp. 4765–4774, 2017.