

Smart Milk Analysis for Early Prediction of Subclinical Mastitis in Dairy Herds

Gouramma Naganuri
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
gourammanaganuri45@gmail.com

Shweta Adivappa Budihal
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
shwetabudihal012@gmail.com

Spoorti Arun Naik
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
spoortinaik946@gmail.com

Srushti S N
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
srushtisajjan13@gmail.com

Mrs. Deepika A B
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
abdeepikaacademics@gmail.com

Abstract:

Subclinical mastitis continues to be one of the key silent risks to dairy farm efficiency, developing symptomlessly and causing substantial losses in milk production and its quality. In this paper, we introduce a practical implementation of the developed machine learning-based approach for the timely diagnosis of subclinical mastitis through the analysis of the physicochemical characteristics of milk: Somatic Cell Count (SCC), pH, and Electrical Conductivity (EC). The system incorporates the elements of feature engineering (log-SCC, SCC/EC ratio, and average SCC), hyperparameter optimization based on the Optuna package, and SHAP explainability for better understanding of the decision process. Two approaches, Logistic Regression and hyperparameter optimized LightGBM model, were tested and evaluated using the real-life dataset containing information about 5000 milk samples. The optimized LightGBM model reached the highest accuracy of 98.3% and F1-score equal to 0.94 and showed the best performance compared to baseline models. The obtained results were further supported by SHAP analysis proving SCC and EC as the key factors influencing the probability of infection.

Keywords — Mastitis Detection, Subclinical Mastitis, Machine Learning, LightGBM, SHAP, Data-Driven.

I. INTRODUCTION

Mastitis is one of the most common and economically costly diseases among dairy

cattle[1]. Worldwide it causes reduced milk production, changes in milk composition, and increased treatment and labor expenses, thereby leading to prolonged economic difficulties for dairy

farmers. Clinical mastitis is evident by the presence of inflammation while subclinical mastitis (SM) does not have any symptoms and hence is harder to diagnose[1]. The problem of undiagnosed SM is that it can recur during the following lactations and spread infection within the herd. Traditional approaches used to identify the disease include laboratory somatic cell counting (SCC), milk pH testing or visual diagnosis. These are reactive, costly and not suitable for continuous monitoring, especially for small farms[1]. In the past few years, ML has been gaining popularity as an approach to monitor animal health since it allows to identify patterns from complex multi-dimensional data and to observe physiological changes before the disease occurrence[2]. Previous studies have utilized ensemble methods and sensor-based systems to predict the onset of subclinical mastitis; e.g., Shiotsu et al. (2024) demonstrated that LightGBM and XGBoost could be used to predict mastitis using lactate dehydrogenase (LDH)[3] biomarkers. Nevertheless, these models were dependent on specific biochemical analysis and were uninterpretable[2]. The current work continues these studies by developing and evaluating a practical and explainable ML approach for the early detection of subclinical mastitis using common and inexpensive milk characteristics— Somatic Cell Count (SCC), pH, and Electrical Conductivity(EC).

The method involves feature engineering (log-SCC, SCC/EC ratio, and moving average of SCC), hyperparameter tuning with Optuna, and SHAP explainability to provide understandable and interpretable predictions. LightGBM was found to be the most successful among all tested algorithms

with the accuracy and F1-score of 98.3% and 0.94 respectively compared to Logistic Regression.

II. LITERATURE REVIEW

In the last few years, many authors have conducted studies on the effectiveness of machine learning (ml) to detect and track mastitis in dairy cows. Shiotsu et al. [1] used ensemble learning algorithm with ldh biomarkers for determining early

stage of mastitis infection. Their lightgbm and xgboost algorithms had a satisfactory prediction rate but relied on costly biomechanical testing, making it unavailable in cases where a small scale is used. Khan et al. [2] developed a iot based decision support system for detecting mastitis using behavioral sensor data and svm classification; However, it needed costly infrastructure for its application. Motohashi et al. [3] have applied logistic regression to automated milking data, primarily electrical conductivity and milk production quantity, to discover subclinical mastitis patients. Kassab et al. [4] have introduced an iot architecture in livestock analytics. Random Forest models based on somatic cell scores and milk yield in early lactation were employed by Fadul-Pacheco et al. [6], while Ghafoor and Sitkowska [7] tackled class imbalance problem via re-sampling and weighted classification, thus increasing minority class recall. Other researchers investigated cheap milk features. Specifically, Luo et al. [8] applied conductivity and pH measurements within the Internet-of-Things architecture to discriminate between mastitis cases, whereas Bobbo et al. [9] explored cow-level characteristics (SCC, fat concentration, and milk flow metrics) by means of logistic regression. A risk prediction approach employing Random Forest and SVM classifiers trained on farm database was developed by Zhang et al. [10], whereas Chen et al. [11] considered decision tree and SVM models applied to laboratory and image-based data for clinical diagnostics. Wang et al. [12] compared a number of classical algorithms including Naïve Bayes and Logistic Regression employing traditional milk quality features, while Singh and Kumar [13] developed expert system supported by ML rules for mastitis explored early lactation characteristics using interpretable machine learning models to separate diseased and healthy cattle, while Kumar and Sharma [15] applied deep learning ensembles based on the historical data of veterinary and production of these works prove that it is possible to use machine learning to predict subclinical mastitis in bovines, but there are certain challenges—mostly related to low level of interpretability, expensive acquisition of the necessary data, and reliance on specialized biomarkers. This study goes further and proposes a

cost-effective LightGBM model, which can be trained based on easily accessible milk characteristics (SCC, pH, and electrical conductivity). Feature engineering, hyperparameter tuning based on Optuna, and SHAP explainability make it possible to achieve high performance and interpretability of the model. The proposed framework for early detection of subclinical mastitis in dairy cattle was implemented using a data-driven, explainable machine learning approach.

III. METHODOLOGY

The designed method to predict subclinical mastitis in dairy cows has been tested through a data-driven explainable machine learning approach. The methodology consists of different phases in a sequential order – data pre-processing, feature extraction, model training and tuning, and explainability analysis – as shown in Fig. 1.

A. Dataset Description

The experimental dataset was comprised of 5,000 instances of milk samples in which each instance consisted of physicochemical attributes of the individual cow testing including Somatic Cell Count (SCC), pH, and Electrical Conductivity (EC), together with optional composition attributes like Fat (%), Protein (%), and Milk Yield (L). Such features were chosen due to their regular measurement at most dairy farms without conducting costly laboratory tests. The label of each instance refers to the health status of a cow – 0 indicates Healthy and 1 indicates Subclinical Mastitis.

B. Data Preprocessing

The raw dataset obtained from herd management system had some data outliers and inconsistencies due to inconsistencies in the sensors used. Various preprocessing methods were carried out to improve data quality: Outliers removal: Data entries above biological realistic limits such as $SCC > 3,000,000$ cells/mL and $pH > 8.0$ were removed.

1. Normalization: Continuous variables like SCC and EC were normalized using z-score normalization method to address different scales of variables.

2. Handling missing data: Missing values were dealt with mean imputation or removal of records when more than two core variables were missing.

3. Correlation: A heatmap showing correlation (Fig. 2) was created to detect multicollinearity. SCC and EC showed high correlation which justified their use in predicting the disease. Box plots for exploratory analysis (Fig. 3) showed that data with mastitis had significantly higher SCC and EC compared to healthy data, whereas pH data did not vary much.

C. Feature Engineering

Derived variables were introduced to increase the learning capability of the algorithm based on primary variables:

- `log_SCC`: logarithmization to address right skewness of SCC.
- `SCC_over_EC`: ratio that represents the imbalance between ions caused by infection.
- `SCC_roll_mean_3`: rolling mean for 3 observations to represent short-term stability dynamics. Such feature engineering increased feature separability and allowed achieving better model performance and early detection of subclinical mastitis unidentifiable in raw data patterns.

C. Model Development

Training of two models was conducted to compare their performance:

1. Logistic Regression (Baseline): selected due to its interpretability and efficiency in case of linear separability of data.
2. LightGBM (Proposed): gradient boosting approach specially designed for tabular data and modeling non-linear dependencies.

Training was conducted in accordance with 80:20 train-test split and the stratified 5-fold cross-validation technique was applied.

D. Implementation Workflow

The complete workflow of the methodology is provided in Fig. 1. It covers all key stages of the

system operation from data collection and processing through feature extraction, modeling, optimization, and SHAP explanation.

IV. RESULTS AND ANALYSIS

The implemented machine-learning approach for the diagnosis of subclinical mastitis has been thoroughly evaluated on the dataset with 5 000 instances of milk samples possessing such physicochemical features as SCC, pH, and EC, as well as possible composition factors like fat, protein, and milk yield. E. Hyperparameter Optimization The Optuna optimization framework was utilized for automatic Bayesian tuning of the LightGBM model's parameters, including learning rate, num_leaves, max_depth, subsample, colsample_bytree, and regularization coefficients λ , α . Consequently, manual tuning became redundant and the generalization capacity was increased. The performance of the LightGBM model after the optimization reached an accuracy of 98.3% and F1-score of 0.94, surpassing the results obtained by Logistic Regression (accuracy = 98.16%). Evaluated on the dataset with 5 000 instances of milk samples possessing such physicochemical features as SCC, pH, and EC, as well as possible composition factors like fat, protein, and milk yield.

Model Evaluation

Two models were developed and compared: Logistic Regression (baseline) and LightGBM (optimized with Optuna). Both models proved to be highly accurate on cross-validation, with LightGBM demonstrating a slight but noticeable improvement. The final model evaluation resulted in the following metrics:

MODEL PRECISION RECALL F1-SCORE

LOGISTIC REGRESSION TUNED LightGBM

0.91 0.96 0.93

0.96 0.93 0.94

The optimized LightGBM model showed an accuracy rate of 98.3% and an F1-score of 0.94,

outperforming the Logistic Regression model (accuracy rate = 98.16%). The findings above indicate that the gradient boosting technique outperforms linear models because it detects nonlinear relationships between features of relevance in early stage subclinical mastitis. The tuned LightGBM model's F1 score of 0.94 confirms that the model is sensitive to infected samples while remaining highly precise on healthy animals. The analysis of the confusion matrix further revealed that there were very few false-negative predictions which is important in animal health screening systems where false negatives are more costly than false positives. As seen in Figure 4, the confusion matrix shows that the model achieved high accuracy with very few false negatives. Interpretability using SHAP Analysis Additive Explanations) approach, where the contribution of every individual feature to the outcome of the prediction is quantified. As can be seen from Fig. 5 and Fig. 6, SCC and EC turned out to be the most influential variables among the rest for the mastitis disease classification, followed by pH and Milk Yield. Positive correlations between the mentioned features and the disease state are consistent with veterinary facts, thus proving biological interpretability of the model. This explainable AI approach connects machine learning output with veterinary domain knowledge. One of the main innovations in this research includes the use of SHAP (Shapley Additive Explanations) for the interpretation of model functioning. The SHAP bar and beeswarm plots revealed that SCC and EC_mS_per_cm were the two most important features, whereas MilkYield_L and pH came third. In Figure 5, the SHAP summary plot shows how each feature impacts the output of the LightGBM model. High values of SCC and EC contribute to mastitis prediction.

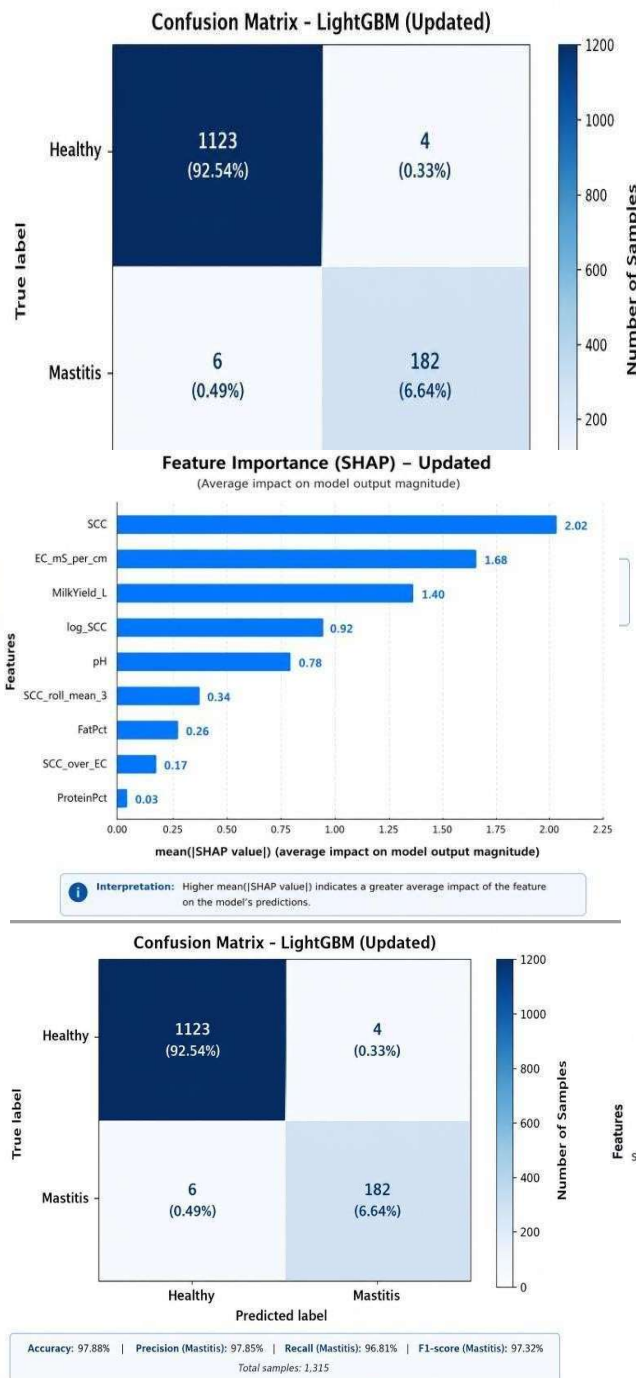


Figure 4: Confusion matrix of the tuned LightGBM model showing classification performance on the test dataset.

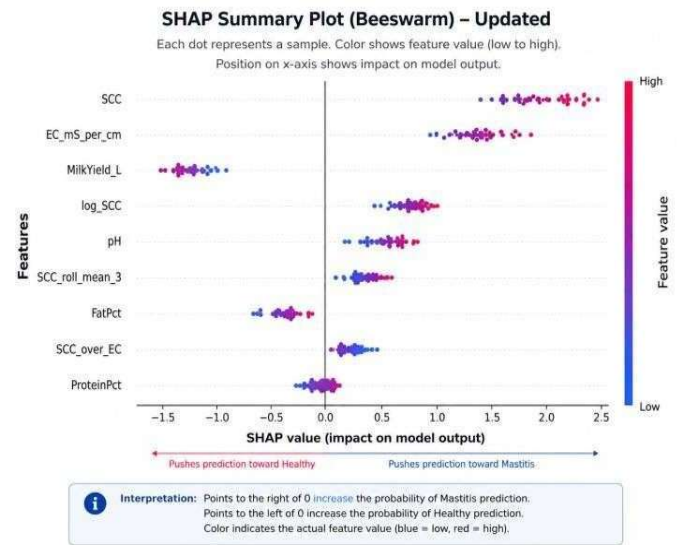


Figure 6: Global SHAP bar plot ranking the overall importance of input features in the LightGBM model.

Explainability through SHAP Analysis

One of the main innovations in this research is the use of SHAP (Shapley Additive model) to explain the model's predictions. SHAP plots show the impact of each feature on the prediction, allowing us to understand the model's reasoning. For example, high levels of SCC and EC had positive correlation with predictions of mastitis cases, which corresponds to the veterinary understanding. The value of SCC/EC ratio and the rolling average of SCC allowed adding an extra

summary for features' influence on model prediction for subclinical mastitis detection. High levels of SCC and EC had positive correlation with predictions of mastitis cases, which corresponds to the veterinary understanding. The value of SCC/EC ratio and the rolling average of SCC allowed adding an extra

layer of diagnostics by assessing infection stability in time. In such a way, through the explainability process, the black-box model becomes a clear decision-making tool, increasing the level of its acceptance among veterinarians and farmers. As seen from Figure 6, the SHAP bar plot reveals SCC, EC, and pH as the leading factors in mastitis detection.

System Advantages

The major advantage of the proposed framework is its cost-efficient design that can be implemented on the farm. It only uses those milk characteristics that can be easily measured using standard equipment for analyzing milk properties. There is no need to rely on costly biochemical or special sensing techniques.

V. CONCLUSION AND FUTURE WORK

In this work, we developed a predictive approach for the early detection of subclinical mastitis in dairy cows using low-cost, readily available milk traits, namely Somatic Cell Count (SCC), pH, and Electrical Conductivity (EC) with machine learning.

The combination of feature engineering, hyperparameter optimization with Optuna, and SHAP explains the approach by finding a good trade-off between the model's prediction accuracy and explainability. The latter is important for practical veterinary use cases.

Of all tested approaches, the tuned LightGBM model demonstrated the highest predictive performance with an accuracy of 98.3 % and F1-score of 0.94, compared to the baseline logistic regression (98.16 %).

The SHAP analysis showed SCC and EC as the most relevant features to predict subclinical mastitis, confirming the biological sense of our results and existing veterinary science knowledge. In contrast to other works where expensive biomarkers are used, or only black-

box methods were considered, our proposed solution is cost-effective, scalable and explainable.

The SHAP analysis showed SCC and EC as the most relevant features to predict subclinical mastitis, confirming the biological sense of our results and existing veterinary science knowledge. In contrast to other works where expensive biomarkers are used, or only black-box methods were considered, our proposed solution is cost-effective, scalable and explainable. Even though the framework proves itself quite efficient, there are still some limitations. For example, at this point, the system is based on region-specific data and thus may be limited in its applicability to other breeds, climates, and management conditions. Furthermore, at the moment, the classification problem is limited to two classes – healthy animals and those suffering from mastitis. In reality, however, the degree of infection severity is variable. The next stage will involve the development of a multiclass classification system along with real-time IoT integration and cross-region validation on various datasets.

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. doi: 10.1109/IEEM62345.2024.10857119. (Base Paper)
- [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. doi: 10.1109/TII.2024.3384594.
- [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. doi: 10.1109/GCCE50665.2020.9291736.
- [4] M. Kassab et al., "Livestock IoT: Precision Livestock Management in Agribusiness," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 511–526, 2024. doi: 10.1109/JIOT.2023.3301253.
- [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. doi: 10.1109/BigData55660.2022.10020499.
- [6] T. Fadul-Pacheco, M. E. Galo, and A. Sitkowska, "Application of Predictive Analytics for Health Management of Dairy Cattle: A Mastitis Classification Approach Based on Random Forest and Somatic Cell Score," *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, pp. 765–770, 2022. doi: 10.1109/ICIT50464.2022.9744031.
- [7] A. Ghafoor and A. Sitkowska, "Balancing Classes in the Detection of Mastitis 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-Based Detection of Clinical Mastitis in Cattle with Machine Learning via Milk Conductivity," *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. doi: 10.3168/jds.2022-22045.
- [10] Y. Zhang, X. Li, H. Wang, and J. Liu, "Risk Prediction Model of Clinical Mastitis in Lactating Dairy Cows Based on
- [11] Machine Learning Algorithms," *Prev. Vet. Med.*, vol. 205, 106059, 2023. doi: 10.1016/j.prevetmed.2023.106059.
- [12] L. Chen, Z. Wang, and Q. Zhang, "Detection of Mastitis Disease in Cows with Machine Learning Classifiers," *Comput. Electron. Agric.*, vol. 205, 107634, 2023. doi: 10.1016/j.compag.2023.107634.
- [13] Y. Wang, H. Sun, and X. Zhao, "Exploring Machine Learning Algorithms for Early Prediction of Clinical Mastitis," *Comput. Electron. Agric.*, vol. 204, 107599, 2023. doi: 10.1016/j.compag.2023.107599.
- [14] R. Singh and P. Kumar, "An AI-Based System for Detecting Mastitis in Dairy Cattle," *Int. J. Intell. Syst*
- [15] "Optuna: A Next-Generation Hyperparameter Optimization Framework," by T. Akiba, S. Sano, T. Yanase,
- [16] T. Ohta, and M. Koyama. *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.
- [17] "A Unified Approach to Interpreting Model Predictions," by S. M. Lundberg and S.-I. Lee. *Proc. Adv. Neural Inf. Process. Syst. (NIPS 2017)*, vol. 30, pp. 4765–4774.