



Filter-based Optimized Feature Descriptors for Detecting Clinical Mastitis in Cattle using Random Forest Classifier Model

M. Chengathir Selvi¹, M. Swarnaa¹, S.T. Veena¹, K. Kanagarajadurai²

10.18805/ag.D-6273

ABSTRACT

Background: In the development of small and marginal farming, livestock industries play a significant role. Mastitis, an important crisis that affects the health and economic welfare of dairy farms, is an inflammation in the mammary gland caused from trauma or an infection. Conventional approach relies heavily on identifying mastitis through milk changes, which can be costly and impractical for smaller farms. A mastitis detection algorithm based on machine learning techniques with optimized features through feature selection is proposed.

Methods: To pre-process the dataset, random sampler techniques like SMOTE, Random under sampler, Random over sampler are used. To select highly correlated features, filter-based feature selection techniques such as information gain, Chi-square test, correlation coefficient and mean absolute deviation are applied to a clinical mastitis dataset. The extracted features are trained with machine learning models like support vector machine, Naïve Bayes Classifier, Decision tree and Random Forest classifier.

Result: The experimental outcomes on the clinical mastitis dataset with various machine learning models are evaluated in terms of accuracy, sensitivity, specificity. The findings reveal the use of random forest classifier obtained mean accuracy of 99.2% for the dataset with 99% specificity and sensitivity.

Key words: Infrared images, Machine learning, Mastitis detection, Non-invasive detection.

INTRODUCTION

The livestock industry is a cornerstone of the global economy, contributing approximately 40% to agricultural GDP, according to the Food and Agriculture Organization of the United Nations (Plummer and Plummer, 2012). However, its success is hindered by infectious diseases caused by bacteria, viruses and fungi. Among these, bovine mastitis poses a significant challenge, severely affecting milk quality and production. Waseem *et al.*, (2020) emphasized that this disease is mainly caused by bacterial infections like *Staphylococcus aureus* and *Streptococcus agalactiae*, resulting in inflammation of the udder. These pathogens inhabit the udder and teat skin, eventually colonizing and growing within the teat canal. Early identification and elimination of mastitis during lactation can yield substantial economic benefits by mitigating its adverse effects.

Bovine mastitis is classified into clinical and subclinical forms based on the causative agent. Subclinical mastitis, while not visibly detectable, significantly affects the somatic cell count and the temperature of the udder's skin surface (Vieira *et al.*, 2021) and is higher prevalence compared to clinical mastitis (Seddar-Yagoub *et al.*, 2023). To identify mastitis, various detection methods have been developed, including the California Mastitis Test (CMT) kit (Bouamra *et al.*, 2024; Dingwell *et al.*, 2003) and the Somatic Cell Counter unit (Rychtarova *et al.*, 2021), which enable effective monitoring and diagnosis based on observable symptoms and changes.

Research shows a strong correlation between CMT scores and SCC, establishing the CMT kit as a reliable,

¹Department of Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi-626 005, Tamil Nadu, India.

²Veterinary University Training and Diagnostic Centre, Madurai-625 005, Tamil Nadu, India.

Corresponding Author: M. Chengathir Selvi, Department of Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi-626 005, Tamil Nadu, India. Email: chengathir@mepcoeng.ac.in

How to cite this article: Chengathir Selvi, M., Swarnaa, M., Veena, S.T. and Kanagarajadurai, K. (2025). Filter-based Optimized Feature Descriptors for Detecting Clinical Mastitis in Cattle using Random Forest Classifier Model. Agricultural Science Digest. 1-10. doi: 10.18805/ag.D-6273.

Submitted: 26-12-2024 **Accepted:** 07-05-2025 **Online:** 03-07-2025

cost-effective tool for detecting subclinical mastitis (SCM) (Cai *et al.*, 2018; Ma *et al.*, 2021; Zhou *et al.*, 2022). Machine learning, a branch of artificial intelligence, further enhances early detection capabilities by analyzing large datasets, enabling advancements in disease detection and classification to benefit farmers and scientists.

Mikhail and Keskin (2013) proposed using Support Vector Machine (SVM) for subclinical mastitis detection, achieving 50% training accuracy and 85% testing accuracy and outperforming logistic regression. However, false negatives posed significant risks, particularly with insufficient somatic cell count data. Similarly, (Shaltout *et al.*, 2014) demonstrated the effectiveness of Information Gain (IG) for feature selection, achieving 90% accuracy in influenza classification using a decision tree classifier.

Ryan *et al.* (2021) demonstrated early mastitis detection through continuous cattle monitoring, leveraging changes in milk components such as fat, protein, lactose and somatic cell count (SCC), achieving 85% model accuracy. Bobbo *et al.* (2021) compared machine learning models for mastitis detection using SCC and identified the Random Forest classifier as the most accurate when milk components were utilized for model construction.

Ma *et al.* (2021) proposed a non-invasive method to estimate cattle body temperature using machine learning techniques like Linear Regression and SVM. The approach achieved 63.8% accuracy for detecting common illnesses but struggled to predict outcomes from historical data for individual cattle.

Grodkowski *et al.* (2022) identified key features for designing a mastitis prediction model, demonstrating that logistic regression outperforms artificial neural networks (ANN) when using features like cattle movement, feed intake, resting period and rumination. Rao *et al.* (2023) evaluated various machine learning models for cattle disease prediction using Kaggle datasets and discussed different techniques for disease detection.

Ankhita *et al.*, (2020) compared the performance of KNN and SVM for mastitis detection, recommending SVM for disease detection applications due to its superior performance. (Wang *et al.*, 2022) proposed a deep learning-based supervised learning approach for mastitis detection, achieving 99.9% accuracy, outperforming machine learning models under default parameters. Mohan *et al.* (2019) introduced an expert system for animal disease diagnosis using Convolutional Neural Networks (CNNs), achieving 98.8% accuracy with a diagnostic system consisting of a convolution layer and pooling layer that takes RGB images as input.

Abdul Ghafoor and Sitkowska (2021) proposed a machine learning-based system for detecting clinical mastitis in cattle, improving detection speed based on symptoms exhibited by the cattle. The comparison of machine learning models revealed that the K-nearest neighbors (KNN) model achieved 99.46% accuracy with sensitivity and specificity of 94.7% and 98.9%, respectively. However, since KNN is a lazy training model, it is less efficient in terms of overall model performance. The mastitis detection algorithm aims to accurately predict cattle mastitis status for real-time applications. The subsequent sections cover the proposed methodology, which utilizes optimized feature descriptors and machine learning algorithms, followed by experimental results and a comparison of the algorithm's performance with various models. The conclusion is provided in Section IV.

MATERIALS AND METHODS

The proposed AI driven mastitis detection system is shown in Fig 1 consists of 3 phases namely Feature Selection phase, Mastitis detection Algorithm and Performance

Evaluation phase. The research work was carried out at the host institute, Mepco Schlenk Engineering College, over a research period of approximately six months, from June 2024 to November 2024. The objectives of the mastitis detection system are:

- . To detect Clinical Mastitis in cattle through AI-driven mastitis detection algorithm.
- . To select higher correlation features by filter-based feature selection.
- . To develop mastitis detection algorithm using Machine learning techniques.
- . To validate the performance of mastitis detection system on clinical mastitis dataset.

Feature selection

Abdul Ghafoor and Sitkowska, (2021) emphasized that the feature selection for supervised learning is calculated from the relevance or correlation between features and target variable.

Algorithm: Clinical mastitis detector feature selection.

Input: Clinical mastitis dataset:

$$CMD = (X, y),$$

Where,

Feature set, $X = \{cf_1, cf_2, \dots, cf_n\}$

y = Target variable.

Output: Selected feature subset.

$$X = \{f_k \subseteq X\}, k < n$$

Process initialization

- a) Read the clinical mastitis dataset, CMD and extract the feature set X and target variable, y .
- b) Define k , the desired number of selected features.

For each feature, cf in X :

- i. Compute information gain (IG):

a) Compute entropy, $E(X) = -\sum P(cf) \cdot \log(P(cf))$

b) Compute $E(X|y) = -\sum P(cf|y) \cdot \log(P(cf|y))$

c) $IG = E(cf) - E(X|y)$

- ii. Compute chi-square test score (X^2):

a. Calculate observed frequency, O_i and expected frequency, E_i

$$b. \quad X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

- iii. Compute the Fisher's score (Ω):

a. Calculate the class means, μ_1 and μ_2 ; class variance,

$$\sigma_1^2 \text{ and } \sigma_2^2 \quad \Omega = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

b.

- iv. Compute the correlation coefficient with the target variable,

$$\delta: \quad \delta = |r(cf_i, y)| = \frac{\sum (cf_i - \bar{cf})(y - \bar{y})}{\sqrt{\sum (cf_i - \bar{cf})^2 \sum (y - \bar{y})^2}}$$

- v. Compute the mean absolute deviation (MAD):

a. Calculate mean or median of X_i

$$b. \quad MAD = \frac{1}{n} \sum_{j=1}^n |cf_{ij} - \overline{cf_i}|$$

vi. Feature ranking and selection.

features_{score_i} =

$$\alpha_1 * (IG) + \alpha_2 * X^2 + \alpha_2 * \Omega + \alpha_2 * \delta + \alpha_2 * (MAD)$$

3. Sort the feature in descending order of the features_{score_i} and rank the features.

4. $\hat{X} \leftarrow \{fk \subseteq X : k \leq n\}$

Thus, the feature selection phase identifies the most relevant features that exhibit a strong correlation with the target variable. These selected features are retained for training the machine learning models for the mastitis detection system, while the less relevant features are excluded.

Machine learning models

The Subclinical Mastitis detection model is built using various Machine learning techniques like random forest, ensemble methods and decision tree. The model that

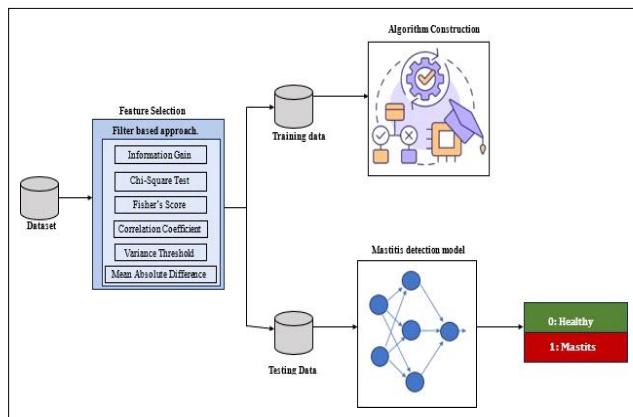


Fig 1: Optimized feature descriptors based mastitis detection system design.

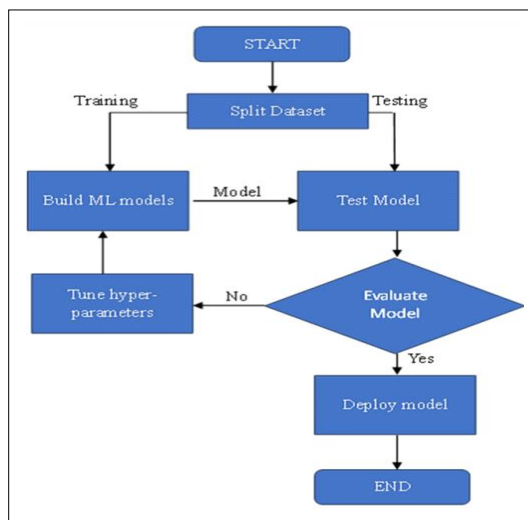


Fig 2: Workflow of machine learning model.

produces more accurate result in terms of time and space complexity is further enhanced with the help of hyper-parameter tuning to produce an optimized model with better performance. The machine learning model in Fig 2, results either subclinical mastitis or healthy cow based on the weight parameters learnt by the model constructed.

RESULTS AND DISCUSSION

Dataset description

The clinical mastitis dataset Ankhita *et al.* (2020) consists of 7922 records stating the health condition of cattle is elaborated with purpose and category of each feature in Table 1. Fig 3 depicts the distribution of data samples in the dataset when the target label is considered.

To reduce the risk of bias and ensure generalization of model, it is important to balance the number of records under each class. Fig 4 shows the number of data entries after the dataset is balanced by using random under-sampler, random over-sampler and SMOTE approach.

- Random over-sampling method aims to address class imbalance by duplicating instances from the minority class to achieve a more balanced distribution.
- Random under-sampling method entails randomly eliminating instances from the majority class to achieve a more equitable class distribution.
- SMOTE (synthetic minority oversampling technique) tackles imbalanced datasets by generating synthetic samples for the minority class.

Once the dataset is randomly sampled to avoid imbalance, for the feature selection filter-based techniques are deployed to identify the highly correlated features.

Feature selection

Fig 5 depicts the contribution of each feature towards achieving the target, where the size of udder parameter shows considerable performance by every feature selection technique. The features of hardness, milk visibility and pain in the udder are most directly correlated to target class.

Information gain, as a purity measure, highlights that parameters like pain, hardness and milk visibility have the highest values due to their strong dependency on the target variable. In contrast, features like day, mastitis after giving birth and previous mastitis status contribute minimally. By MAD, udder parameters (IUFL, EUFL, IUFR, EUFR, IURL, EURL, IURR, EURL) exhibit 80% better correlation compared to others. Across all techniques, temperature, udder size, pain, hardness and milk visibility are the most significant contributors to predicting the target class. Features with a correlation above 30% of the overall range are selected as the best features.

Feature selection highlights hardness, pain and milk visibility as highly significant features, while the udder parameters and temperature exhibit strong correlations with the target variable. Fig 6 reveals inter-feature correlations among IUFL and EUFL, EUFR and IUFR, IURL and EURL, IURR and EURL, as well as hardness, pain

and milk visibility. To optimize the model, one representative feature is retained from each correlated group, reducing the dataset to four key parameters for training the mastitis detection system.

Although pain, hardness and changes in milk color have a higher prevalence in predicting the target, even a small error can pose a risk to cattle health. The associative rule that can be generated using only the features pain, hardness and color visibility in milk is:

```
IF pain = TRUE, THEN class = TRUE
ELSE IF hardness = TRUE, THEN class = TRUE
ELSE IF milk_visibility = TRUE, THEN class = TRUE
ELSE class = FALSE
```

Any incorrect logging or reporting of features can lead to misclassification by the model. Due to these concerns, the features pain, hardness and milk visibility are excluded from model training. where, temperature and eight udder parameters are selected for training the machine learning models.

The mastitis detection system is developed using multiple classification algorithms including random forest classifier, Decision Tree Classifier, K-Nearest Neighbours Classifier and Support Vector Machine Classifier for the selected features. Performance evaluation is essential in machine learning to analyze a model's effectiveness, generalization capability and alignment with task objectives using metrics such as precision, recall, F1-score and accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

$$\text{F1 score} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

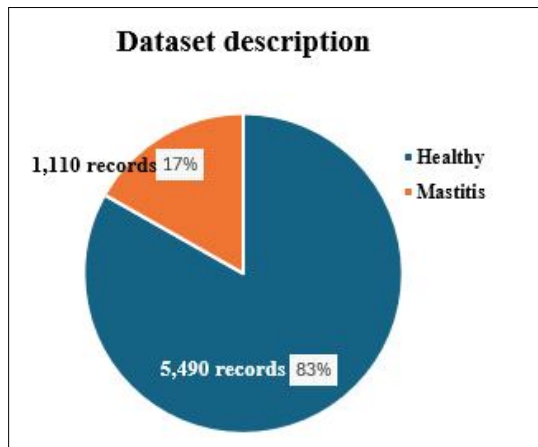


Fig 3: Distribution of records based on target label.

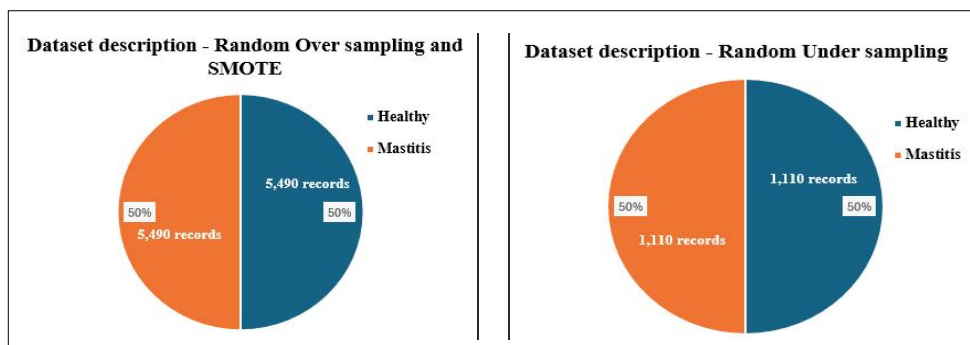


Fig 4: Data distribution based target using (a) Random over-sampler and SMOTE, (b) Random under-sampler.

Table 1: Clinical mastitis dataset.

Feature descriptor	Purpose	Feature type	Value range
Cow ID	To uniquely identify the cattle	Numerical	1-150
Cow breed	Breed to which the cattle belong	Categorical	Jersey, HF
Previous Mastitis Status	Mastitis history for cattle	Categorical	Yes, No
Months after giving birth	To track the lactation period	Numerical	0-7
USST	Udder Skin Surface Temperature	Numerical	
IUFL, EUFL	Udder front left inhale and exhale limit	Numerical	
IUFR, EUFR	Udder front right inhale and exhale limit	Numerical	
IURL, EURL	Udder rear left inhale and exhale limit	Numerical	
IURR, EURL	Udder rear right inhale and exhale limit	Numerical	
Milk visibility	Change in milk	Categorical	Yes, No
Pain	pain in udder	Categorical	Yes, No
Hardness	Change in texture of udder	Categorical	Yes, No

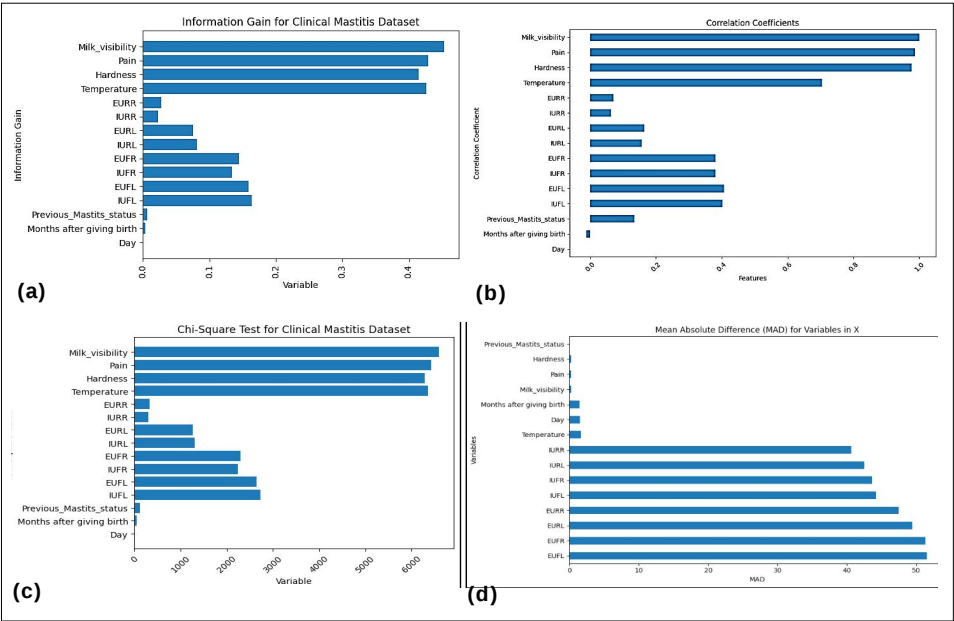


Fig 5: Feature selection based on (a) Information gain; (b) Chi-square test; (c) Correlation coefficient; (d) Mean absolute deviation.

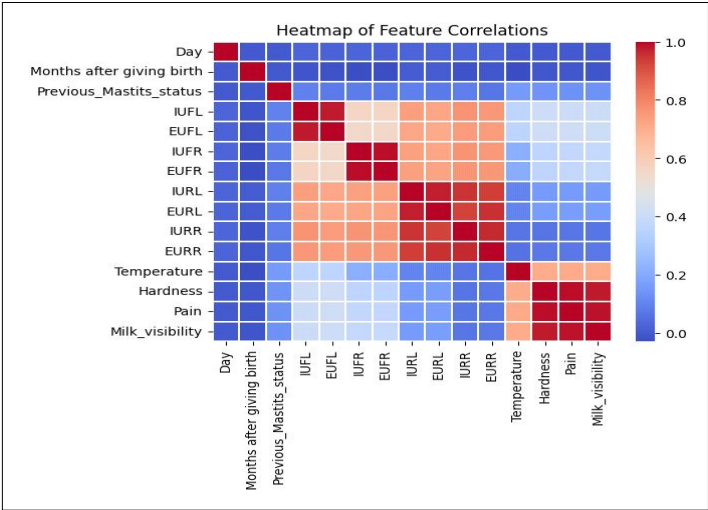


Fig 6: Heatmap of correlated features.

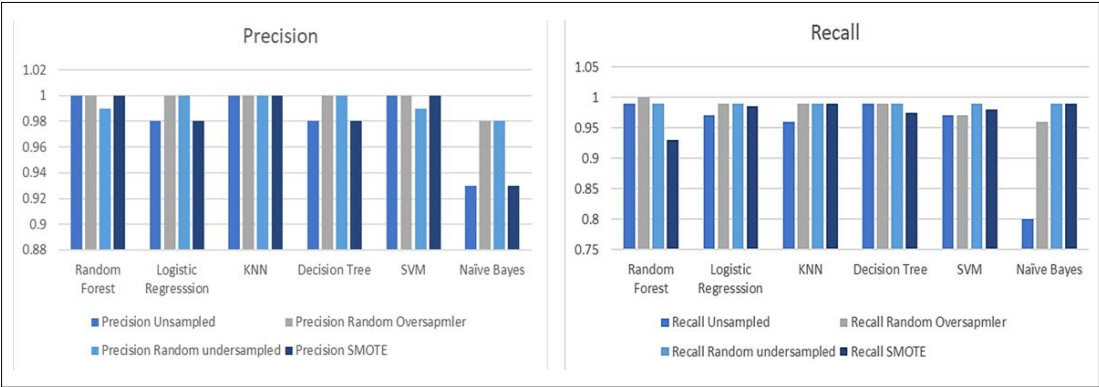


Fig 7: Precision and recall of mastitis detection system.

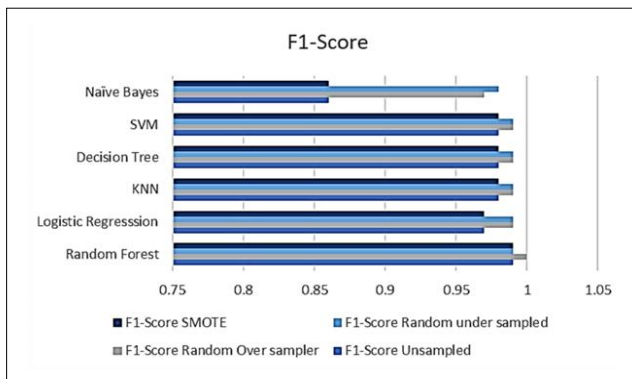


Fig 8: F1 Score of mastitis detection system.

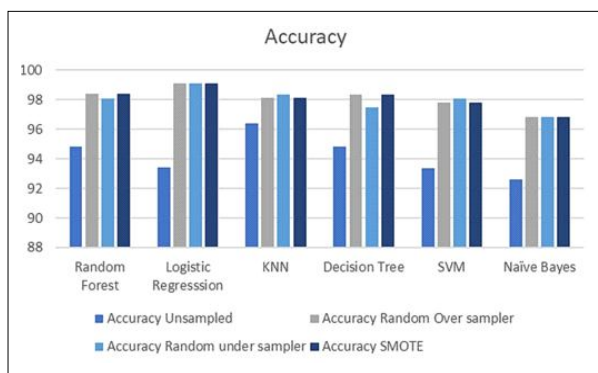


Fig 9: Accuracy of mastitis detection system when dataset is unsampled.

Performance evaluation

Fig 7 to 10 summarize the performance of the mastitis detection system trained with optimized features (IURL, IUFL, IURR and temperature), selected through a two-step feature selection process involving initial screening (Information Gain, Fisher's Score, Chi-Square test) and correlation analysis to avoid redundancy and overfitting.

Fig 7 highlights precision-based performance across Random Forest, KNN, SVM, Logistic Regression, Decision Tree and Naïve Bayes classifiers under unsampled, SMOTE, random oversampling and under-sampling techniques with an 80:20 train-test split. Random oversampling yielded the best precision (1.0) for most models, while Naïve Bayes showed lower precision (0.93) in unsampled and SMOTE datasets but improved to 0.98 with oversampling.

Fig 9 presents accuracy-based analysis, where Logistic Regression excelled across sampling methods, underscoring the importance of sampling to enhance model performance over unsampled datasets.

Tables 2-5 present the performance of various machine learning models for mastitis classification using two feature sets: F1 (IURL, IURR, EUFL, Temperature) and F2 (IUFL, EUFL, IUFR, EUFR, IURL, EUFL, IURR, EUFR, Temperature). The models were trained and evaluated under varying training-testing splits and a 10-fold cross-validation approach (k=10). The results highlight that Random Forest consistently demonstrates high accuracy and robustness, maintaining near-perfect precision, recall

Table 2: Performance evaluation of mastitis detection using machine learning models when dataset is unsampled.

Feature set	F1				F2			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Training: 80%, Testing: 20%								
Random forest	1	0.99	0.99	94.2	1	0.99	0.99	94.21
Logistic regression	0.98	0.97	0.97	93.11	0.98	0.97	0.97	93.01
KNN	1	0.96	0.98	94.02	1	0.96	0.98	94.12
Decision tree	0.98	0.99	0.98	94.1	0.98	0.99	0.98	94.1
SVM	1	0.97	0.98	92.1	1	0.97	0.98	92.2
Naïve bayes	0.93	0.8	0.86	92	0.93	0.8	0.86	91.1
Training: 70% - Testing: 30%								
Random forest	0.99	0.99	0.99	94.0	0.99	0.99	0.99	94.0
Logistic regression	0.98	0.97	0.98	93	0.98	0.97	0.98	93
KNN	1	0.97	0.98	93.98	1	0.97	0.98	93.98
Decision tree	1	0.97	0.98	94.1	1	0.97	0.98	94.1
SVM	0.99	0.97	0.98	91.96	0.99	0.97	0.98	91.96
Naïve bayes	0.93	0.8	0.86	91.8	0.93	0.8	0.86	91.8
Training: 60% - Testing: 40%								
Random forest	0.99	0.99	0.99	94.15	0.99	0.99	0.99	94.15
Logistic regression	0.98	0.97	0.98	92.87	0.98	0.97	0.98	92.87
KNN	1	0.97	0.98	94	1	0.97	0.98	94
Decision tree	1	0.97	0.98	94.1	1	0.97	0.98	94.1
SVM	0.99	0.97	0.98	92.1	0.99	0.97	0.98	92.1
Naïve bayes	0.93	0.8	0.86	91.87	0.93	0.8	0.86	91.87

Table 3: Performance evaluation of mastitis detection using machine learning models when dataset is random under sampled.

Feature set			F1		F2			
Model	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Training: 80%, Testing: 20%								
Random forest	0.99	0.99	0.99	98.06	1	0.99	0.99	99.32
Logistic regresssion	1	0.99	0.99	99	0.99	0.98	0.99	98.19
KNN	1	0.99	0.99	98.3	1	0.99	0.99	99.54
Decision tree	1	0.99	0.99	97.4	1	0.99	0.99	99.32
SVM	0.99	0.99	0.99	98.06	0.99	0.99	0.99	99.09
Naïve bayes	0.98	0.99	0.98	96.8	0.97	0.99	0.98	97.97
Training: 70% - Testing: 30%								
Random forest	0.99	0.99	0.99	98.51	1	0.99	0.99	99.24
Logistic regresssion	1	0.99	0.99	99	0.99	0.98	0.98	98.49
KNN	1	0.98	0.99	98.3	1	0.99	0.99	99.39
Decision tree	1	0.99	0.99	97.4	0.99	0.99	0.99	99.09
SVM	0.99	0.98	0.98	98	0.99	0.99	0.99	98.94
Naïve bayes	0.98	0.98	0.98	96	0.97	0.99	0.98	98.04
Training: 60% - Testing: 40%								
Random forest	1	0.98	0.99	98.46	1	0.98	0.99	98.98
Logistic regresssion	1	0.98	0.99	99.07	0.97	0.97	0.97	97.07
KNN	1	0.97	0.99	98.3	1	0.98	0.99	99.09
Decision tree	1	0.98	0.99	97.4	1	0.98	0.99	98.98
SVM	0.99	0.97	0.98	98.06	0.99	0.98	0.99	98.19
Naïve bayes	0.98	0.98	0.98	96.8	0.98	0.98	0.98	98.08

Table 4: Performance evaluation of mastitis detection using machine learning models when dataset is random over sampled.

Feature set			F1				F2	
Model	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Training: 80%, Testing: 20%								
Random forest	1	1	1	0.984	1	1	1	0.9874
Logistic regresssion	1	0.99	0.99	0.991	0.99	0.97	0.98	0.9725
KNN	1	0.99	0.99	0.981	1	0.99	0.99	0.9897
Decision tree	1	0.99	0.99	0.9833	1	0.99	0.99	0.9808
SVM	1	0.97	0.99	0.978	0.99	0.99	0.99	0.9893
Naïve bayes	0.98	0.96	0.97	0.968	0.98	0.99	0.98	0.9812
Training: 70% - Testing: 30%								
Random forest	1	1	1	98.54	1	1	1	98.54
Logistic regresssion	1	0.99	0.99	0.9911	1	0.99	0.99	0.9911
KNN	1	0.99	0.99	0.981	1	0.99	0.99	0.981
Decision tree	1	0.99	0.99	0.9833	1	0.99	0.99	0.9833
SVM	1	0.97	0.98	0.978	1	0.97	0.98	0.978
Naïve bayes	0.98	0.96	0.97	0.9682	0.98	0.96	0.97	0.9682
Training: 60% - Testing: 40%								
Random forest	1	1	1	98.49	1	1	1	99.74
Logistic regresssion	1	0.99	0.99	0.9911	1	0.98	0.99	99.08
KNN	1	0.99	0.99	0.981	1	0.99	0.99	99.21
Decision tree	1	0.99	0.99	0.9833	1	0.99	0.99	99.39
SVM	0.99	0.97	0.98	0.978	0.99	0.97	0.98	98.21
Naïve bayes	0.98	0.95	0.97	0.9687	0.98	0.95	0.97	96.79

and F1-scores, particularly with larger training datasets (e.g., 80%-20% split). Similarly, KNN performs well, showing minimal accuracy loss as the training size decreases, while Logistic Regression and SVM maintain stable performance with F1-scores close to 0.99, showcasing resilience across feature sets and splits. Decision Tree achieves perfect scores for most metrics but is more consistent at higher training percentages. Naïve Bayes, while slightly less accurate (around 91%), remains

competitive with its F1-scores, making it suitable for specific scenarios.

The results also demonstrate that random sampling techniques such as oversampling, under-sampling and SMOTE improve model performance significantly. KNN excels with SMOTE or unsampled datasets, whereas Logistic Regression performs best with Random Oversampling or Under Sampling. Random Forest and Decision Tree exhibit robust and consistent results across

Table 5: Performance evaluation of mastitis detection using machine learning models when dataset is SMOTE.

Feature set	F1				F2			
Model	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Training: 80%, Testing: 20%								
Random forest	1	0.99	0.99	99.69	1	0.99	0.99	99.69
Logistic regression	0.98	0.97	0.97	99.16	0.98	0.97	0.97	99.16
KNN	1	0.96	0.98	99.31	1	0.96	0.98	99.31
Decision tree	0.98	0.99	0.98	99.47	0.98	0.99	0.98	99.47
SVM	1	0.97	0.98	99.46	1	0.97	0.98	99.46
Naïve bayes	0.93	0.8	0.86	95.75	0.93	0.8	0.86	95.75
Training: 70% - Testing: 30%								
Random forest	0.99	0.99	0.99	99.54	0.99	0.99	0.99	99.54
Logistic regression	0.98	0.97	0.98	99.2	0.98	0.97	0.98	99.2
KNN	1	0.97	0.98	99.39	1	0.97	0.98	99.39
Decision tree	1	0.97	0.98	99.46	1	0.97	0.98	99.46
SVM	0.99	0.97	0.98	99.34	0.99	0.97	0.98	99.34
Naïve bayes	0.93	0.8	0.86	95.68	0.93	0.8	0.86	95.68
Training: 60% - Testing : 40%								
Random forest	0.99	0.99	0.99	99.54	0.99	0.99	0.99	99.54
Logistic regression	0.98	0.97	0.98	99.2	0.98	0.97	0.98	99.2
KNN	1	0.97	0.98	99.39	1	0.97	0.98	99.39
Decision tree	1	0.97	0.98	99.46	1	0.97	0.98	99.46
SVM	0.99	0.97	0.98	99.34	0.99	0.97	0.98	99.34
Naïve bayes	0.93	0.8	0.86	95.68	0.93	0.8	0.86	95.68

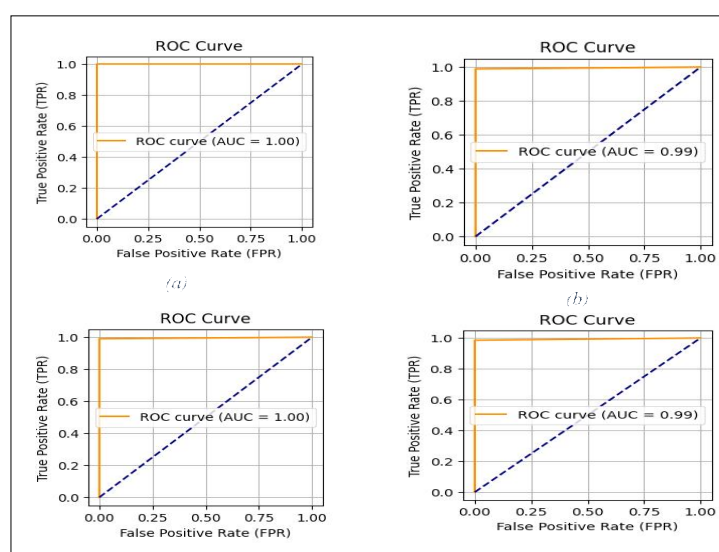


Fig 10: ROC for Random Forest classifier for mastitis detection when dataset is sampled using a) Random over-sampler; b) Random under-sampler; c) SMOTE; d) Unsampled.

sampling methods, with better performance using Feature Set F1 compared to F2. These findings emphasize the importance of sampling techniques and feature optimization for improving the accuracy and reliability of mastitis detection systems.

Overall, Random Forest and Decision Tree emerge as the most reliable classifiers, achieving superior performance across various configurations and the results are further supported by the ROC curve (Fig 10), underscoring the effectiveness of the proposed feature sets and data sampling strategies.

CONCLUSION

A technical investigation on detection of mastitis in dairy cattle is made with use of machine learning models with leveraging features through feature selection. A comparison between mastitis detection model using machine learning model trained using dataset randomly sampled using SMOTE, random over-sampler, random under-sampler proves Random Forest classifier show significant performance with average accuracy (99.25%) and avg. precision (0.99) when the dataset is sampled using random over-sampler, random under-sampler, SMOTE and unsampled data under varying training and testing split and under different feature set for mastitis detection. The use of feature selection approach makes the model generalized. By filter-based feature selection high correlating features to the target are chosen for mastitis detection algorithm construction. The purpose of various random sampling approaches is to balance the dataset alongside defining Random under sampler to be best method for the dataset since the random over-sampler leads to overfitting of model. Our research underscores the promise of machine learning-based methods for precise and effective detection of cow diseases. Infrared thermographic image-based detection of mastitis in cattle during early stages is a promising avenue for future enhancement, stimulating innovative techniques to improve detection and ultimately the health of livestock.

ACKNOWLEDGEMENT

The authors are thankful to DST-TDT-TDP directorate, the Principal and the Management of Mepco Schlenk Engineering College, Sivakasi for their support and facilities provided to carry out this research work. The authors express their gratitude to the anonymous reviewers for their insightful comments and suggestions in improving the work.

Funding

The work presented in this paper was funded by Department of Science and Technology, Technology Development Transfer, Ministry of Science and Technology, New Delhi, India under Grant No. DST/TDT/TDP-33/2022.

Conflict of interest

All authors declared that there is no conflict of interest.

REFERENCES

- Abdul Ghafoor, N., Sitkowska, B. (2021). MasPA: A machine learning application to predict risk of mastitis in cattle from AMS sensor data. *Agri Engineering*. 3: 575-583. <https://doi.org/10.3390/agriengineering3030037>.
- Ankhita K., Manjaiah, D.H., Kartik, M. (2020). Data for: Clinical mastitis in cows based on udder parameter using internet of things (IoT) 1. <https://doi.org/10.17632/kbvcdw5b4m.1>.
- Bobbo, T., Biffani, S., Taccioli, C., Penasa, M., Cassandro, M. (2021). Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Sci. Rep.* 11: 13642. <https://doi.org/10.1038/s41598-021-93056-4>.
- Bouamra, M., Ziane, M., Akkou, M., Bentayeb, L., Titouche, Y. (2024). Effect of subclinical mastitis detected in the first month of lactation on the reproductive performance of dairy cows in western Algeria. *Asian J. Dairy Food Res.* 43(4): 650-656. <https://doi.org/10.18805/ajdrf.DRF-431>.
- Cai, J., Luo, J., Wang, S., Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*. 300: 70-79. <https://doi.org/10.1016/j.neucom.2017.11.077>.
- Dingwell, R.T., Leslie, K.E., Schukken, Y.H., Sargeant, J.M., Timms, L.L. (2003). Evaluation of the California mastitis test to detect an intramammary infection with a major pathogen in early lactation dairy cows. *Can. Vet. J. Rev. Veterinaire Can.* 44: 413-415.
- Grodzowski, G., Swaczkowski, T., Koszela, K., Mueller, W., Tomaszuk, K., Baars, T., Sakowski, T. (2022). Early detection of mastitis in cows using the system based on 3D motions detectors. *Sci. Rep.* 12: 21215. <https://doi.org/10.1038/s41598-022-25275-2>.
- Ma, S., Yao, Q., Masuda, T., Higaki, S., Yoshioka, K., Arai, S., Takamatsu, S., Itoh, T. (2021). Development of noncontact body temperature monitoring and prediction system for livestock cattle. *IEEE Sens. J.* 21: 9367-9376. <https://doi.org/10.1109/JSEN.2021.3056112>.
- Mikail, N., Keskin, İ. (2013). Application of the support vector machine to predict subclinical mastitis in dairy cattle. *Scientific World Journal*. 2013: 603897. <https://doi.org/10.1155/2013/603897>.
- Mohan, A., Raju, R.D., Janarthanan, P. (2019). Animal disease diagnosis expert system using convolutional neural networks. 2019 Int. Conf. Intell. Sustain. Syst. ICISS 441-446. <https://doi.org/10.1109/ISS1.2019.8908108>.
- Plummer, P.J., Plummer, C., 2012. Chapter 15-diseases of the mammary gland, in: Pugh, D.G., Baird, A.N. (Eds.), *Sheep and Goat Medicine (Second Edition)*. W.B. Saunders, Saint Louis. pp: 442-465. <https://doi.org/10.1016/B978-1-4377-2353-3.10015-0>.
- Rao, Mrs.A., H.R., M., B.C.R., Thaseen, S. (2023). Cattle disease prediction using artificial intelligence. *Int. J. Res. Appl. Sci. Eng. Technol.* 11: 2184-2189. <https://doi.org/10.22214/ijraset.2023.50535>.

- Ryan, C., Guéret, C., Berry, D., Corcoran, M., Keane, M.T., Mac Namee, B. (2021). Predicting Illness for a Sustainable Dairy Agriculture: Predicting and Explaining the Onset of Mastitis in Dairy Cows. DOI:10.48550/arXiv.2101.02188.
- Rychtarova, J., Krupova, Z., Brzakova, M., Borkova, M., Elich, O., Dragounova, H., Seydlova, R., Sztankoova, Z., Rychtarova, J., Krupova, Z., Brzakova, M., Borkova, M., Elich, O., Dragounova, H., Seydlova, R., Sztankoova, Z. (2021). Milk quality, somatic cell count and economics of dairy goats farm in the czech republic. IntechOpen. <https://doi.org/10.5772/intechopen.97509>.
- Seddar-Yagoub, F., Dahou, A.A., Meskini, Z., Doukani, K., Homrani, A. (2023). Prevalence and risk factors of bovine mastitis on conventional dairy farms in northwestern Algeria. Asian J. Dairy Food Res. 43(2): 320-326. <https://doi.org/10.18805/ajdfr.DRF-325>.
- Shaltout, N.A., El-Hefnawi, M., Rafea, A., Moustafa, A. (2014). Information gain as a feature selection method for the efficient classification of influenza based on viral hosts. 2014(1): 2078-0958.
- Vieira, R.K.R., Rodrigues, M., Santos, P.K.S., Medeiros, N.B.C., Cândido, E.P., Nunes-Rodrigues, M.D. (2021). The effects of implementing management practices on somatic cell count levels in bovine milk. Animal. 15: 100177. <https://doi.org/10.1016/j.animal.2021.100177>.
- Wang, Y., Kang, X., He, Z., Feng, Y., Liu, G. (2022). Accurate detection of dairy cow mastitis with deep learning technology: A new and comprehensive detection method based on infrared thermal images. Animal. 16: 100646. <https://doi.org/10.1016/j.animal.2022.100646>.
- Waseem, R., Muhee, A., Malik, H.U., Akhoun, Z.A., Munir, K., Nabi, S.U., Taifa, S. (2020). Isolation and identification of major mastitis causing bacteria from clinical cases of bovine mastitis in Kashmir valley. Indian J. Anim. Res. 54(11): 1428-1432. <https://doi.org/10.18805/ijar.B-3848>.
- Zhou, X., Xu, C., Wang, H., Xu, W., Zhao, Z., Chen, M., Jia, B., Huang, B. (2022). The early prediction of common disorders in dairy cows monitored by automatic systems with machine learning algorithms. Anim. Open Access J. MDPI. 12: 1251. <https://doi.org/10.3390/ani12101251>.