TIDS: A Thermal Imaging Dataset for Subclinical Mastitis in Dairy Sheep

Georgios Botsoglou[1 [0009−0005−1812−281X]], Marios Lysitsas[2 [0009−0007−0648−6004]], Dimitris Dimitriadis(✉)[1 [0000−0002−9404−0331]], Constantina Tsokana[1 [0000−0003−4180−6310]], George Valiakos[2 [0000−0002−3869−5026]], and Grigorios Tsoumakas[1 [0000−0002−7879−669X]]

[1] Aristotle University of Thessaloniki, Greece
{gbotso,dndimitri,greg}@csd.auth.gr, ctsokana@vet.auth.gr
[2] University of Thessaly, Greece
{mlysitsas,georgevaliakos}@uth.gr

Abstract. Subclinical mastitis (SCM) in dairy sheep is a significant challenge in the agricultural and veterinary sectors, leading to substantial financial losses for farmers and negatively impacting overall dairy sheep productivity. It often goes unnoticed due to the absence of clinical signs, making early diagnosis particularly difficult. However, early identification of SCM is critical, as it allows for timely intervention that can prevent disease progression, reduce economic losses, and minimize the need for costly treatments. This work focuses on detecting SCM in dairy sheep using a non-invasive thermal imaging approach. A major limitation in this field is the lack of available datasets, as acquiring such data presents several challenges. Capturing clear thermal images of dairy sheep udders is hindered by factors such as animal movement, environmental conditions, and variability in breed, health, and udder size. Furthermore, large, diverse sample sizes are required, making data acquisition resource-intensive. Ethical concerns regarding animal welfare and the high cost of thermal imaging equipment add to the complexity. These challenges hinder the use of data-driven techniques, such as deep learning models, which require large datasets. In this paper, our contributions are two-fold: first, we introduce a novel dataset, TIDS, along with an explanatory analysis supported by domain expertise. Second, we apply deep learning models to detect SCM in dairy sheep and provide a comprehensive methodology, marking a novel approach in this area.

Keywords: Subclinical Mastitis, Thermal Images, Convolutional Neural Networks, Deep Learning

1 Introduction

Mastitis is defined as inflammation of varying degrees of severity of the mammary gland, and constitutes a crucial pathological condition for dairy ruminants, including lactating ewes [1,2,3]. The two primary types of mastitis encountered are clinical and subclinical. Clinical mastitis has symptoms that can be perceived

macroscopically during clinical examination of the animal and/or is associated with alteration, visible with the naked eye, in the characteristics of the milk. It is severe and requires treatment. However, the majority of cases encountered in sheep belong to the subclinical type of the disease, where there are no visible symptoms or observable changes in the macroscopic characteristics of the obtained milk [1,4]. However, inflammation and histological lesions are present [5]. Therefore, in animals with SCM, a decrease in the milk production, milk yield, and cheese-making properties is documented.

Moreover, the prevalence of the disease in dairy flocks is estimated at 5-30%, or even more [4,6]. Through this considerable reduction in productivity and quality, significant economic losses are caused in sheep farms worldwide [7,8,9], while animal welfare concerns are also considerable [3,10]. Finally, public health and food safety concerns regarding SCM are not negligible, since pathogens with zoonotic potential are regularly implicated in SCM cases and the production of enterotoxins in milk is possible [4,11].

Identification of SCM is challenging, due to the total absence of any detectable clinical changes and requires specific tests in milk. As a result, it is regularly underdiagnosed. The tool currently utilized most for screening is the estimation of somatic cell count in milk using the California Mastitis Test (CMT). However, the gold standard for identification of the etiologic agents of SCM, which are usually bacteria, is aerobic culture. This approach has certain limitations and requirements, mostly regarding cost and time. Therefore, evaluation and establishment of new fast and non-invasive diagnostic tools is considered beneficial [1,2].

Infrared thermography (IRT) is a technology with numerous potential applications in both human and veterinary medicine, since it provides real-time and non-invasive measurements of body temperature through converting infrared radiation emitted by the heat source into respective pixel intensity [12,13]. Promising implementations have been described in literature and adopted in various medical fields, that mainly concern its evaluation as a diagnostic tool through the detection of increased temperature values, as a result of pyrexia or localized inflammation in variable pathological conditions [14]. Similarly, in veterinary medicine, infrared thermography has been investigated for the detection of body temperature variations of the examined animals, with promising results in various fields, like bovine mastitis [12]. Besides, in mastitis, inflammatory processes occurring in the mammary gland increase the udder's inner and surface temperatures [13,15]. This allows for the employment of IRT for the detection of these variations, even in cases of SCM [15,16,17]. In that regard, it has also been examined in ovine mastitis with promising results [18,19,20,21], but current data are rather limited yet.

Deep learning has been applied to predictive tasks using IRT [22]. Specifically, computer vision-based learning models analyze patterns in thermal images to classify them accordingly. However, in our study of thermal imaging for SCM detection in dairy sheep, several challenges arise. To the best of our knowledge, no publicly available datasets exist for training and testing deep learning models in

this domain. Consequently, there is also a lack of related research leveraging deep learning for SCM prediction in dairy sheep. This leads to our central research question: "Can we develop a high-quality thermal image dataset with input from domain experts to enable the creation of effective deep learning models?".

To address this question, we collected thermal images from 16 dairy sheep farms and carried out extensive manual work to classify the images as healthy or affected by SCM, using somatic cell count (SCC) measurements conducted in laboratories. The dataset was carefully curated, and various preprocessing steps were performed to ensure compatibility with our deep learning models. Since this dataset is newly introduced, we also conducted an exploratory analysis with domain experts to extract relevant features. This analysis highlighted the challenges of addressing the classification problem using statistical or traditional machine learning approaches. Finally, several deep learning models were applied as baselines for evaluation. Our results are promising, while the dataset also presents several challenges that the research community can further explore.

Figure 1 provides an overview of the workflow presented in this paper. Specifically, we first collected thermal images and milk samples from sheep farms to construct the TIDS dataset. This dataset was then used both for in-depth analysis and for training and evaluating classification models using machine learning and deep learning techniques.

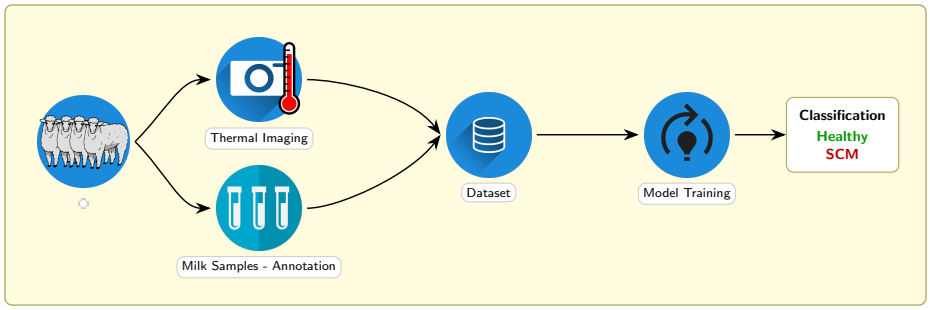

Fig. 1. Schematic overview of the creation of TIDS, combining thermal images and annotated milk samples to support the classification of healthy and SCM cases.

The remainder of this paper is organized as follows. Section 2 provides a detailed description of the dataset, including the acquisition process conducted by domain experts, the preprocessing steps required for data preparation, and the exploratory analysis. Section 3 outlines the baseline models used and the learning process and presents the results along with a discussion of key findings. Section 4 reviews related work, Section 5 discusses the results in detail, and Section 6 concludes the paper with suggestions for future research.

2 The Dataset

In this section, we introduce the TIDS (Thermal Imaging Dataset for Subclinical Mastitis in Dairy Sheep)[3]. We first outline the dataset acquisition process, detailing the camera used, its settings, the selected farms, and the criteria for image collection. Next, we describe the preprocessing steps applied to prepare the images for classification. Given that this dataset is newly introduced, we also conducted an exploratory analysis to examine key statistical insights related to SCM. With guidance from domain experts, we defined relevant features and trained various machine learning classifiers on the features produced.

2.1 Acquisition

Farms and animals included in this study A total of 16 dairy sheep farms located in central Greece were selected for this study according to the following criteria: breeding Lacaune sheep (crossbred or purebred), having an automatic milking system, documentation and availability of detailed history of every previous pathological condition and treatment of the selected animals, and vaccination for contagious agalactia. All ewes were examined prior to the sampling by the same experienced personnel, to ensure that they are phenotypically healthy, free of clinical mastitis and other systemic diseases. Moreover, to achieve uniformity regarding the lactation stage and reduce its effect in the obtained results, each selected animal was between the 2nd and the 4th month of its lactation period at the moment of the sampling. The number of animals included from each farm ranged from 25 to 66.

Collection of milk samples A single visit was carried out to each farm and milk samples were collected during routine morning milking. All the relevant procedures were carried out according to the guidelines provided by the National Mastitis Council. In particular, pledgets and 70% ethyl alcohol solution were used to thoroughly scrub both teats of each udder. Initially, CMT was performed. The first three streams were discarded, and then the necessary quantity of milk was added to the paddle disc for the test. Subsequently, two sterile vials with preservative (0.1 g sodium azide, Merck KGaA, Darmstadt, Germany) were aseptically filled with approximately 40 mL of milk for the SCC test with Lactoscan SCC counter. Each vial was labeled with the date, animal code, and side of the udder half and milk samples were transported to the laboratory within two hours. All the procedures mentioned above were performed by the same experienced personnel to reduce the subjectivity of CMT assessment and ensure similar sampling conditions.

Thermal Imaging Thermal images were received from each animal, just before obtaining milk samples, using a FLIR E96 24° camera (Teledyne FLIR LLC.,

[3] The dataset can be accessed at `https://doi.org/10.5281/zenodo.15619247`

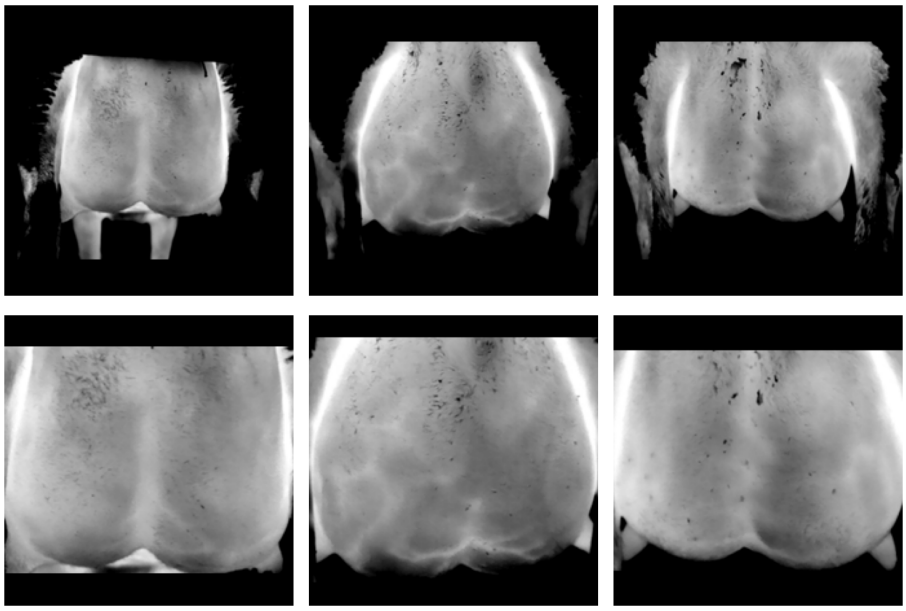

Fig. 2. Thermal imaging examples showing paired full and cropped ROI images. Samples 1-2 (columns 1-2) are healthy, while Sample 3 (column 3) is affected by SCM. Top row shows full thermal images, bottom row shows corresponding cropped regions of interest. Thermal intensity is displayed with white indicating higher temperatures and black indicating lower temperatures.

Wilsonville, OR, USA). Initially, environmental conditions (temperature and humidity) were recorded using a COMET U3120 Datalogger (COMET SYSTEM s.r.o., Roznov pod Radhostem, Czech Republic) and the obtained measurements were evaluated in the corresponding camera settings fields. Emissivity was set to 0.98. Then, images were taken at a distance of 70 cm from the posterior surface of the udder. All procedures were carried out before milking and in the milking parlor to avoid the effects of climatic conditions (wind, rain, etc.). Examples of thermal images are shown in Figure 2.

Somatic Cell Count [4] A direct fluorescent image low-magnification microscopic recognition method was utilized for the SCC, using a LACTOSCAN SCC counter and the compatible kit, according to the guidelines provided by the manufacturer (Milkotronic Ltd., Nova Zagora, Bulgaria). This technique enables the detection and quantification of somatic cells in milk by fluorescent staining of cellular DNA, followed by low-magnification microscopic recognition, providing a rapid and reliable assessment of udder inflammation, while it has been previously employed in dairy sheep [23], demonstrating reliability for SCM detection.

[4] the number of somatic cells found in a millilitre of milk

Before the test, milk samples were heated to 40°C using a water bath, cooled to 20°C and vortexed for 15-20 sec to distribute fat uniformly. Then each milk sample was diluted with water at a ratio of 1:1 ratio immediately before testing, as recommended by the manufacturer, since ovine milk typically contains >5% fat. A quantity of 100μl of the diluted samples was inoculated in microtubes containing SOFIA GREEN lyophilized dye and they were vortexed for approximately 10 sec. Subsequently, 8μl were received from the microtubes and added in a predefined chamber of a four-chamber disposable chip that was inserted in the analyzer. All udder halves were classified in five categories according to the number of somatic cells per ml of the respective milk sample (1: <250,000; 2: 250,000–500,000; 3: 500,000–1,500,000; 4: 1,500,000–5,000,000; 5: >5,000,000), based on the results obtained from the Lactoscan SCC counter. Thresholds suggested in literature for discrimination of subclinical mastitis from healthy udder halves[1,4,24] and the empirical SCC ranges corresponding to positive CMT results were used in this categorization, to further classify SCM cases according to the severity of the inflammation process. All samples of the categories 3, 4 and 5 were defined as SCM-positive, while samples of the categories 1 and 2, as healthy. The classification was performed at the udder half level; however, for the purposes of model development and evaluation, the health status was assigned at the animal level. Specifically, animals with at least one udder half classified as category 3, 4, or 5 were labeled as SCM, while animals with both udder halves in categories 1 or 2 were considered healthy. All animals included in the analysis had high-quality thermal images with clear visibility of the udder surface, ensuring consistent input quality across cases.

2.2 Preprocessing

The dataset consists of 418 thermal images, each corresponding to a different dairy sheep - 207 images from sheep affected by SCM and 211 from healthy sheep. The dataset was further preprocessed, as this is a crucial step in training deep learning models to enhance their generalization ability and robustness. A number of preprocessing steps taken are explained, alongside various transformations meant to help the model given the limited amount of samples.

The udder region was manually cropped from each thermal image to remove irrelevant areas outside the region of interest. The cropped images were then resized to 224 x 224 pixels to standardize input dimensions for further processing. Resizing was performed by adjusting the largest dimension to 224 pixels while maintaining the original aspect ratio. In experiments involving deep learning, padding was then applied as necessary to achieve the final square dimensions. Figure 2 shows some examples before and after cropping.

2.3 Exploratory analysis

Thermal imaging analysis for SCM detection has received limited attention in the literature. Exploratory data analysis was performed to examine the patterns

and characteristics in the thermal data, which can support the later development of predictive models.

Feature Extraction Key features for thermal image classification were identified through a comprehensive review of relevant literature. A set of features was extracted, encompassing statistical measures such as average (mean) temperature, maximum (hottest) temperature, and temperature variation (standard deviation) within the thermal image. To better understand how temperatures are spread out in the thermal image, percentiles (25th, 50th, and 75th) were used, along with the interquartile range (IQR), which shows the range where most temperatures fall. The difference between the highest and lowest temperatures (maximum and minimum intensity) was also measured to show the full range of temperature variation in the area. Shape characteristics were described using skewness and kurtosis, which provide insight into the asymmetry and peakedness of the temperature distribution, respectively. Information entropy was calculated to assess the complexity and variability of the pixel intensity patterns, with higher values indicating more diverse temperature distributions. Morphological features were derived by applying image erosion, and the mean pixel value of the eroded images was used to capture the structural characteristics and spatial distribution of high-temperature regions.

Although it might appear self-evident that feature distributions differ between healthy and affected udders, Mann-Whitney U tests were employed to rigorously quantify these differences. Significant differences ($p < 0.05$) were observed in mean and percentile intensities, IQR, skewness, kurtosis, and eroded mean, indicating that these features capture meaningful physiological variations and serve as effective discriminators. In contrast, standard deviation, temperature range difference, and entropy did not show significant differences, suggesting limited discriminatory power when considered individually.

Figure 3[5] illustrates the distributions of several image-derived features across healthy and mastitis-affected groups. The mean intensity for healthy samples shows a bimodal distribution, suggesting possible subgroups within the class, while the mastitis group is more unimodal and shifted toward lower values. The percentile features (25th, 50th, and 75th) consistently exhibit leftward shifts in the mastitis group, reflecting reduced thermal intensity. All percentile histograms show sharply peaked central bins, likely due to value rounding or discrete measurement effects. The IQR distribution is narrower and more concentrated in healthy samples, whereas mastitis samples show higher variability and a long tail. In terms of shape, skewness is more negative in healthy samples, while kurtosis is slightly higher, indicating heavier tails. Both features again show pro-

[5] The histograms generated using Seaborn's [25] histplot function with the default binning strategy (bins='auto'). This method automatically determines the number and width of bins based on the data distribution, optimizing the balance between resolution and smoothness. This adaptive binning allows each feature's histogram to best represent the underlying data characteristics without manual bin width specification.

nounced central peaks, indicating non-Gaussian behavior. The eroded mean is higher in healthy cases and displays more symmetry, while mastitis samples show a slight leftward skew. Temperature range difference is heavily right-skewed with a dominating spike near 1.0 for both groups, suggesting a ceiling effect. Entropy is tightly distributed around 6.5–7.0 with little variation, but mastitis samples trend slightly lower.

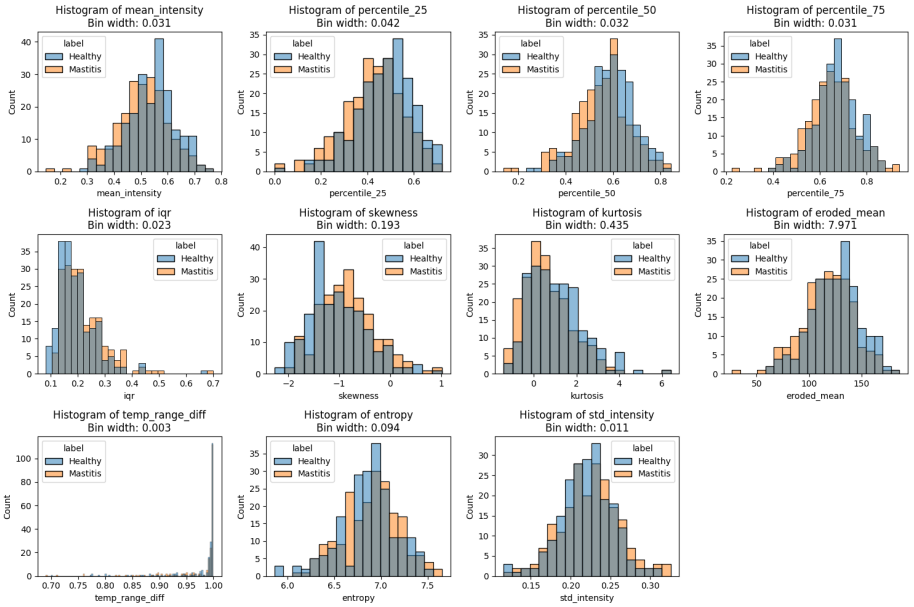

Fig. 3. Distribution of the Features for Healthy and SCM-Affected Udders. Histograms show the count distribution for each feature, with separate distributions for healthy and mastitis cases

Finally, the standard deviation of intensity is among the most normally distributed features, with healthy samples slightly more peaked and centered. These patterns highlight not only differences between the two classes but also the non-normal and often discretized nature of several features, which has implications for statistical modeling and classifier choice.

Classification Various machine learning models were trained and optimized. The dataset (418 thermal images) is split into training and testing sets using an 80-20 split. We excluded temperature range difference feature from the set of features, since we observed extremely low variance across both healthy and mastitis samples. Feature scaling is applied using z-score normalization to stan-

Table 1. Best Parameters for Different Models

Model	**Best Parameters**
SVM	C: 1, Gamma: 0.01, Kernel: RBF
XGBoost	Learning Rate: 0.1, Max Depth: 4, Estimators: 300
Logistic Regression	C: 0.01, Penalty: L2
Random Forest	Max Depth: 5, Min Split: 5, Estimators: 200
KNN	Neighbors: 9, Weights: Uniform
Gradient Boosting	Learning Rate: 0.2, Max Depth: 5, Estimators: 300
AdaBoost	Learning Rate: 0.01, Estimators: 200

dardize the features, which is critical for models sensitive to input scale, such as SVM and K-Nearest Neighbors (KNN).

Hyperparameter tuning is performed for each model with a grid search in a 5-fold cross-validation manner in the training set. The models evaluated include SVM, XGBoost, Logistic Regression, Random Forest, KNN, Gradient Boosting, AdaBoost, Naive Bayes and Linear Discriminant Analysis (LDA). Most of these models were implemented using the scikit-learn library[6], a widely used Python toolkit for machine learning that provides efficient and user-friendly tools for model training, evaluation, and validation. The XGBoost model was integrated separately using its dedicated Python package due to its specialized implementation for gradient boosting[7]. For each model, a predefined grid of hyperparameters is explored to identify the optimal configuration. The best hyperparameters are used to train the final models, and predictions are made on the test set.

For the models where parameter tuning was conducted the best performing parameters found are presented in Table 1.

The resulting classification performance metrics for each model are presented in Table 2. Accuracy is defined as

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

The F1 score is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where precision and recall are given by

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}.$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. All metrics were calculated using the default classification threshold of 0.5 on predicted probabilities. Feature importance analysis revealed that entropy, skewness, 25th percentile and Kurtosis were the most significant features for the model's predictive performance.

[6] https://scikit-learn.org/
[7] https://pypi.org/project/xgboost/

Table 2. Performance metrics for different models. The highest F1-score is shown in bold.

Model	F1 Score	Precision	Recall	Accuracy
SVM	0.5823	0.6571	0.5227	0.6071
XGBoost	**0.7010**	0.6415	0.7727	0.6548
Logistic Regression	0.5526	0.6562	0.4773	0.5952
Random Forest	0.6667	0.6744	0.6591	0.6548
KNN	0.5600	0.6774	0.4773	0.6071
Gradient Boosting	0.6875	0.6346	0.7500	0.6429
AdaBoost	0.6990	0.6102	0.8182	0.6310
Naive Bayes	0.5789	0.6875	0.5000	0.6190
LDA	0.5385	0.6176	0.4773	0.5714

The feature importance of models which support their identification are shown in Table 3. The feature importance analysis reveals that skewness and entropy are consistently among the top three most important features across all models, highlighting their strong predictive power for SCM detection. Notably, AdaBoost places an exceptionally high importance on skewness (0.5129), making it the most dominant feature in any model. Feature importance analysis indicates that statistical properties such as skewness and entropy tend to have higher importance scores compared to other features, suggesting a relatively greater influence on the models' predictions.

Table 3. Top 3 Most Important Features for Each Model with Importance Scores

Model	Feature	Importance Score
XGBoost	Skewness	0.1111
	Entropy	0.1072
	25th Percentile	0.1040
Random Forest	Skewness	0.1324
	Entropy	0.1205
	Kurtosis	0.1060
Gradient Boosting	Entropy	0.2168
	Skewness	0.1509
	Kurtosis	0.1272
AdaBoost	Skewness	0.5129
	Entropy	0.2965
	25th Percentile	0.1907

3 Experimental Design and Results

3.1 Baseline Models

As baselines to this dataset we used the ResNet-18 [26], DenseNet-121 [27] and EfficientNet B0 [28]. These models are widely used and have demonstrated strong generalization across a variety of vision tasks [29,30,31]. They present an excellent balance between computational efficiency, accuracy, and implementation simplicity. They have extensive pretrained weights and stable implementations available in libraries like PyTorch [32] and TensorFlow [33], which is especially useful in small-data regimes. The models, pretrained on ImageNet, were used with their original classification heads replaced by a binary classification head. Due to the small sample size only the final layers were finetuned. The dataset was split into training, validation, and test sets in a 60-20-20 ratio. The training set was used for model optimization, the validation set for hyperparameter tuning and model selection, and the test set for final performance evaluation.

To augment the dataset and improve the model's generalization ability, a series of transformations were applied. First, a random horizontal flip was used with a probability of 50%, enabling the model to become invariant to horizontal orientations. A small random rotation, within a range of ± 10 degrees, was applied to help the model handle minor variations in object orientation without excessive distortion. Additionally, a random affine transformation was performed, translating the image by up to 10% of its width and height in both directions. This introduced slight shifts in object position, enhancing the model's robustness to positional changes. A random resized crop was also employed, where the image was cropped to between 80% and 100% of its original size, ensuring that the model could adapt to different object scales. Finally, a mild Gaussian blur with a kernel size of 5 and a random sigma value between 0.1 and 3.0 was applied to simulate varying focus levels, while maintaining sufficient image clarity. These transformations collectively increased the diversity of the dataset, aiding in the development of a more robust model capable of handling variations in orientation, position, scale, and focus.

The hyperparameter search space for the optimization process was tuned using Optuna [34]. Each model was trained with the Tree-structured Parzen Estimator (TPE) search algorithm executed for 50 trials. The optimization target set is the maximization of validation F1-score. To accommodate the specific challenges of training on a small dataset, we carefully defined the search space to include both core training parameters and regularization strategies. The learning rate, a parameter controlling update magnitude during training, was logarithmically sampled between 1e-5 and 1e-2. This wide range was chosen to allow exploration from cautious to aggressive updates, particularly important when fine-tuning on small datasets where large updates can easily lead to overfitting or instability. For batch size, we explored discrete values (16, 32, 64) to balance between training stability (smaller batches) and computational efficiency (larger batches). We tested both Adam and SGD optimizers. Adam offers adaptive learning rates and typically converges faster, while SGD provides a useful

Table 4. Optimal training parameters for each model

Model	Optimizer	Learning Rate	Weight Decay	Batch Size	Dropout
ResNet-18	SGD	2.60e-4 (FC), 7.79e-4 (Conv.)	1.30e-3	32	0.06
DenseNet-121	Adam	8.5e-4	2.9e-4	16	0.13
EfficientNet B0	SGD	8.98e-4 (FC), 2.15e-5 (Conv.)	4.68e-3	16	0.40

contrast in update dynamics and is known for strong generalization in some contexts. Weight decay—a form of L2 regularization—was sampled from a log scale between 1e-5 and 1e-2. It helps prevent overfitting by penalizing large weights, which is useful in settings with limited data. We also varied the number of unfrozen layers (1 or 3) to control the degree of fine-tuning. Unfreezing more layers allows the model to adapt more to the new data, but increases overfitting risk—especially in data-scarce scenarios. Finally, dropout rates between 0.0 and 0.5 were explored as a regularization method to improve generalization. For each run, binary cross-entropy loss was the optimization target and F1-score was monitored per epoch for train and validation sets. Experiments were set to run for 100 epochs with early stopping set, having a patience of 20 epochs.

The best-performing configuration for each architecture involved different fine-tuning strategies. For ResNet-18, unfreezing the final convolutional layer along with the fully connected layer resulted in the highest validation performance. For DenseNet-121, optimal results were achieved by training only the fully connected layer. In the case of EfficientNet B0, unfreezing just the last convolutional layer led to the best generalization. Table 4 presents the optimal parameters for each model.

3.2 Results

The hyperparameter optimization (Table 4) revealed architecture-specific preferences: DenseNet-121 favored the adaptive learning of Adam, while ResNet-18 and EfficientNet-B0 benefited from SGD's fine-grained control. Notably, EfficientNet-B0 required a markedly lower learning rate for its convolutional layers (2.15e-5) compared to its FC layer (8.98e-4), highlighting the importance of preserving pretrained feature extractors while adapting task-specific heads.

The experimental results offer a baseline for the novel dataset provided. DenseNet-121 achieved the highest overall performance, with an F1-score of 0.6575 and accuracy of 63.77%. Its precision was 63.16%, and recall was 68.57%. EfficientNet-B0 exhibited the highest recall 74.29%, indicating superior sensitivity to positive cases, but suffered from lower precision 56.52%, reflecting a higher rate of false positives. ResNet-18 delivered intermediate results, with moderate precision 56.82% and the lowest accuracy 57.97% among the models.

The moderate performance ceiling (best F1 < 0.66) likely reflects fundamental challenges in SCM detection.

Table 5. Comparison of Model Performance Metrics

Model	F1 Score	Precision	Recall	Accuracy
ResNet-18	0.6329	0.5682	0.7143	0.5797
DenseNet-121	0.6575	0.6316	0.6857	0.6377
EfficientNet B0	0.6420	0.5652	0.7429	0.5797

4 Related Work

Subclinical mastitis detection is crucial in livestock management. IRT offers a non-invasive method by detecting udder temperature changes indicative of inflammation [35,36].

Early studies showed strong correlations between udder skin surface temperature (SST) and CMT scores in cows [16]. Subsequent research validated the use of IRT for mastitis diagnosis in sheep [18], Girolando and Jersey cows [37], and Holstein Friesian cows [38], demonstrating its diagnostic capabilities comparable to traditional methods like CMT and SCC. Studies also explored IRT's use in detecting E.coli-induced mastitis [39] and in various environmental conditions [40,41].

Various algorithms have been designed and implemented for the automated analysis of thermal images [42,43,44], and IRT's utility has been extended to dairy goats [45] and buffaloes [46]. These studies consistently highlight IRT's potential for early subclinical mastitis detection across species, showing strong correlations with CMT and SCC.

A similar approach utilizing SVMs and stochastic neighbor embedding was recently published [19]. While the results are promising, achieving 84% accuracy, the dataset is not publicly available, and the test set is highly imbalanced.

This paper contributes a novel sheep udder thermal image dataset and a perspective of using deep learning models for detecting SCM, establishing a baseline for future research.

5 Discussion

The feature analysis revealed distinct thermal patterns between healthy and SCM-affected udders. Statistically significant differences were observed in features describing overall temperature (mean intensity) and distribution shape (skewness, kurtosis). Features related to central tendency, such as mean intensity and percentiles (25th, 50th, 75th), exhibited statistically significant differences. Standard deviation, and entropy, showed limited univariate discriminative power though entropy's inclusion in machine learning models training hinted at complementary roles in combination with other features.

Among traditional machine learning models, ensemble methods such as AdaBoost, XGBoost, and Random Forest demonstrated a relatively balanced performance. XGboost achieved the highest F1-score (0.701) while the AdaBoost

achieved the highest recall (0.8182) indicating stronger sensitivity to SCM cases, though at the cost of precision (0.6102). Random Forest exhibited comparable F1-scores (0.6667), with feature importance rankings highlighting skewness, entropy, and percentile values as key contributors. Skewness, in particular, emerged as a prominent feature across models, potentially reflecting its ability to encode asymmetric thermal patterns linked to inflammation. Linear models, such as logistic regression, lacked in performance, likely due to their limited capacity to model non-linear relationships in the data.

Experiments with neural networks, involving DenseNet-121, ResNet-18 and EfficientNet-B0, resulted in lower performance compared to traditional models. DenseNet-121 achieved the highest F1-score (0.6575) and accuracy (63.77%) among the evaluated architectures, though its performance remained slightly lower than XGBoost. Notably, traditional machine learning models achieved moderately higher performance compared to neural networks in this study. This observation may reflect the advantages of handcrafted features, which incorporate domain-specific knowledge and reduce reliance on large-scale data. Neural networks, faced limitations due to the small number of samples, which was prohibitive for the effective learning of discriminative feature extraction from the images.

The overall performance ceiling observed across all models highlights the inherent challenges involved in detecting SCM through thermal imaging. One of the key difficulties stems from the complex nature of the dataset itself, which presents several factors that complicate the classification of udders as either healthy or affected by mastitis.

Firstly, the thermal images contain various sources of noise, such as the presence of sheep hair, which can interfere with accurate temperature measurements. These irregularities can introduce distortions in the image data, making it harder to reliably differentiate between healthy and SCM-affected udders. Additionally, thermal images often suffer from occlusions, where parts of the udder may be obscured by the animal's body or other factors, reducing the amount of visible data available for analysis.

Furthermore, environmental conditions significantly influence the temperature of the udders, adding another layer of complexity to the task. Factors such as ambient temperature, humidity, and sunlight can cause variations in the thermal readings, making it difficult to distinguish between healthy udders and those affected by SCM. This environmental variability introduces a degree of uncertainty, as the temperature difference between a healthy and an SCM-affected udder may be subtle and heavily influenced by external conditions.

6 Conclusions and Future Work

Detecting SCM using thermal imaging and artificial intelligence is a challenging task due to the absence of visible clinical symptoms and the need for high sensitivity and specificity in detection algorithms. Nevertheless, overcoming these challenges is crucial, as early and accurate detection of SCM can significantly im-

prove animal welfare, reduce economic losses in the dairy industry, and enhance milk quality and herd health management.

This paper tackles these challenges by introducing a new, expert-validated dataset to the research community, which, to the best of our knowledge, is the first of its kind for dairy sheep. This makes it a significant contribution to the field. Additionally, the paper presents initial baseline results on this dataset. To our knowledge, no other researchers have applied deep learning for SCM detection using thermal imaging in dairy sheep, emphasizing the novelty and importance of this work.

In future work, we should focus on expanding the dataset by collecting more thermal images from diverse environments and improving annotation methods based on confirmed inflammation. Closer-range imaging could enhance resolution and detection accuracy. Exploring advanced models like Vision Transformers and leveraging transfer learning could further improve performance, if a larger dataset is available. Enhancing model interpretability through better visualization techniques is also crucial. Additionally, long-term studies tracking individual sheep could help develop predictive models, improving early detection and disease management in real-world farm settings.

Acknowledgments. This research was co-funded by Greece and the European Union in the framework: Sub-Measure 16.1–16.2–Establishment and operation of Operational Team (O.T.) of the European Innovation Partnership (EIP) for agricultural productivity and sustainability–Establishment and operation of Operational Team (O.T.) of the European Innovation Partnership (EIP) for agricultural productivity and sustainability under grant number M16SYN2-00202 THERMASHEEP-Application of Infrared Thermography (IRT) Technology as a Subclinical Mastitis Diagnostic Tool: Improving Welfare and Productivity Indicators in Sheep and Goat Farms.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Fragkou, I., Boscos, C. & Fthenakis, G. Diagnosis of clinical or subclinical mastitis in ewes. *Small Ruminant Research.* **118**, 86-92 (2014)
2. Libera, K., Konieczny, K., Grabska, J., Smulski, S., Szczerbal, I., Szumacher-Strabel, M. & Pomorska-Mól, M. Potential novel biomarkers for mastitis diagnosis in sheep. *Animals.* **11**, 2783 (2021)
3. Gelasakis, A., Mavrogianni, V., Petridis, I., Vasileiou, N. & Fthenakis, G. Mastitis in sheep–The last 10 years and the future of research. *Veterinary Microbiology.* **181**, 136-146 (2015)
4. Contreras, A., Sierra, D., Sánchez, A., Corrales, J., Marco, J., Paape, M. & Gonzalo, C. Mastitis in small ruminants.. *Small Ruminant Research.* **68**, 145–153 (2007)
5. Arteche-Villasol, N., Fernández, M., Gutiérrez-Expósito, D. & Pérez, V. Pathology of the mammary gland in sheep and goats. *Journal Of Comparative Pathology.* **193** pp. 37-49 (2022)

6. Giadinis, N., Arsenos, G., Tsakos, P., Psychas, V., Dovas, C., Papadopoulos, E., Karatzias, H. & Fthenakis, G. "Milk-drop syndrome of ewes": Investigation of the causes in dairy sheep in Greece. *Small Ruminant Research.* **106**, 33-35 (2012)
7. Leitner, G., Chaffer, M., Shamay, A., Shapiro, F., Merin, U., Ezra, E., Saran, A. & Silanikove, N. Changes in milk composition as affected by subclinical mastitis in sheep. *Journal Of Dairy Science.* **87**, 46-52 (2004)
8. De Olives, A., Díaz, J., Molina, M. & Peris, C. Quantification of milk yield and composition changes as affected by subclinical mastitis during the current lactation in sheep. *Journal of Dairy Science.* **96**, 7698–7708 (2013)
9. Martí-De Olives, A., Peris, C. & Molina, M. Effect of subclinical mastitis on the yield and cheese-making properties of ewe's milk. *Small Ruminant Research.* **184**, 106044 (2020)
10. Gougoulis, D., Kyriazakis, I., Papaioannou, N., Papadopoulos, E., Taitzoglou, I. & Fthenakis, G. Subclinical mastitis changes the patterns of maternal–offspring behaviour in dairy sheep. *The Veterinary Journal.* **176**, 378-384 (2008)
11. Benkerroum, N. Staphylococcal enterotoxins and enterotoxin-like toxins with special reference to dairy products: An overview. *Critical Reviews In Food Science And Nutrition.* **58**, 1943-1970 (2018)
12. Rekant, S., Lyons, M., Pacheco, J., Arzt, J. & Rodriguez, L. Veterinary applications of infrared thermography. *American Journal Of Veterinary Research.* **77**, 98-107 (2016)
13. Machado, N., Da Costa, L., Barbosa-Filho, J., De Oliveira, K., De Sampaio, L., Peixoto, M. & Damasceno, F. Using infrared thermography to detect subclinical mastitis in dairy cows in compost barn systems. *Journal Of Thermal Biology.* **97** pp. 102881 (2021)
14. Lahiri, B., Bagavathiappan, S., Jayakumar, T. & Philip, J. Medical applications of infrared thermography: a review. *Infrared Physics & Technology.* **55**, 221-235 (2012)
15. Tommasoni, C., Fiore, E., Lisuzzo, A. & Gianesella, M. Mastitis in dairy cattle: On-farm diagnostics and future perspectives. *Animals.* **13**, 2538 (2023)
16. Çolak, A., Polat, B., Okumus, Z., Kaya, M., Yanmaz, L. & Hayirli, A. Early detection of mastitis using infrared thermography in dairy cows. *Journal Of Dairy Science.* **91**, 4244-4248 (2008)
17. Sinha, R., Bhakat, M., Mohanty, T., Ranjan, A., Kumar, R., Lone, S., Rahim, A., Paray, A., Khosla, K. & Danish, Z. Infrared thermography as non-invasive technique for early detection of mastitis in dairy animals-A review. *Asian Journal Of Dairy And Food Research.* **37**, 1-6 (2018)
18. Martins, R., Prado Paim, T., Abreu Cardoso, C., Dallago, B., Melo, C., Louvandini, H. & McManus, C. Mastitis detection in sheep by infrared thermography. *Research In Veterinary Science.* **94**, 722-724 (2013)
19. Tselios, C., Alexandropoulos, D., Pantopoulos, C. & Athanasiou, G. Thermal Imaging and Dimensionality Reduction Techniques for Subclinical Mastitis Detection in Dairy Sheep. *Animals.* **14**, 1797 (2024)
20. Lysitsas, M., Spyrou, V., Billinis, C. & Valiakos, G. Coagulase-negative staphylococci as an etiologic agent of ovine mastitis, with a focus on subclinical forms. *Antibiotics.* **12**, 1661 (2023)
21. Korelidou, V., Simitzis, P., Massouras, T. & Gelasakis, A. Infrared Thermography as a Diagnostic Tool for the Assessment of Mastitis in Dairy Ruminants. *Animals.* **14**, 2691 (2024)
22. Ramesh, V. A review on application of deep learning in thermography. *Int. J. Eng. Manag. Res.* **7** pp. 489-493 (2017)

23. Michael, C., Lianou, D., Vasileiou, N., Mavrogianni, V., Petinaki, E. & Fthenakis, G. Longitudinal study of subclinical mastitis in sheep in Greece: An investigation into incidence risk, associations with milk quality and risk factors of the infection. *Animals*. **13**, 3295 (2023)
24. Vasileiou, N., Chatzopoulos, D., Sarrou, S., Fragkou, I., Katsafadou, A., Mavrogianni, V., Petinaki, E. & Fthenakis, G. Role of staphylococci in mastitis in sheep. *Journal Of Dairy Research*. **86**, 254-266 (2019)
25. Waskom, M. Seaborn: statistical data visualization. *Journal Of Open Source Software*. **6**, 3021 (2021)
26. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770-778 (2016)
27. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Densely connected convolutional networks. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 4700-4708 (2017)
28. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference On Machine Learning*. pp. 6105-6114 (2019)
29. Chauhan, T., Palivela, H. & Tiwari, S. Optimization and fine-tuning of DenseNet model for classification of COVID-19 cases in medical imaging. *International Journal Of Information Management Data Insights*. **1**, 100020 (2021)
30. Majumder, R. Efficient Classification of Pulmonary Pneumonia and Tuberculosis Alongside Normal and Non-X-ray Images with Minimal Resources and Maximum Accuracy. *MedRxiv*. pp. 2024-12 (2025)
31. UÇan, M., Kaya, B. & Kaya, M. Multi-class gastrointestinal images classification using EfficientNet-B0 CNN model. *2022 International Conference On Data Analytics For Business And Industry (ICDABI)*. pp. 1-5 (2022)
32. Ketkar, N., Moolayil, J., Ketkar, N. & Moolayil, J. Introduction to pytorch. *Deep Learning With Python: Learn Best Practices Of Deep Learning Models With PyTorch*. pp. 27-91 (2021)
33. Singh, P., Manure, A., Singh, P. & Manure, A. Introduction to tensorflow 2.0. *Learn TensorFlow 2.0: Implement Machine Learning And Deep Learning Models With Python*. pp. 1-24 (2020)
34. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings Of The 25th ACM SIGKDD International Conference On Knowledge Discovery & Data Mining*. pp. 2623-2631 (2019)
35. Sharun, K., Dhama, K., Tiwari, R., Gugjoo, M., Iqbal Yatoo, M., Patel, S., Pathak, M., Karthik, K., Khurana, S., Singh, R. & Others Advances in therapeutic and managemental approaches of bovine mastitis: a comprehensive review. *Veterinary Quarterly*. **41**, 107-136 (2021)
36. Sathiyabarathi, M., Jeyakumar, S., Manimaran, A., Jayaprakash, G., Pushpadass, H., Sivaram, M., Ramesha, K., Das, D., Kataktalware, M., Prakash, M. & Others Infrared thermography: A potential noninvasive tool to monitor udder health status in dairy cows. *Veterinary World*. **9**, 1075 (2016)
37. Ribeiro, I., Gonçalves, P., Rodrigues, M., Nascimento, G., Baptista, R., Calil Filho, J., Wolf, A. & Wolf, S. Infrared thermography for detection of clinical and subclinical mastitis in dairy cattle: comparison between Girolando and Jersey breeds. *Ciência Animal Brasileira*. **24** pp. e-76726 (2023)
38. Sathiyabarathi, M., Jeyakumar, S., Manimaran, A., Pushpadass, H., Sivaram, M., Ramesha, K., Das, D., Kataktalware, M., Jayaprakash, G. & Patbandha, T. Investigation of body and udder skin surface temperature differentials as an early

indicator of mastitis in Holstein Friesian crossbred cows using digital infrared thermography technique. *Veterinary World*. **9**, 1386 (2016)
39. Metzner, M., Sauter-Louis, C., Seemueller, A., Petzl, W. & Zerbe, H. Infrared thermography of the udder after experimentally induced Escherichia coli mastitis in cows. *The Veterinary Journal*. **204**, 360-362 (2015)
40. Pamparienė, I., Veikutis, V., Oberauskas, V., Žymantienė, J., Želvytė, R., Stankevičius, A., Marčiulionytė, D. & Palevičius, P. Thermography based inflammation monitoring of udder state in dairy cows: sensitivity and diagnostic priorities comparing with routine California mastitis test. *Journal Of Vibroengineering*. **18**, 511-521 (2016)
41. Velasco-Bolaños, J., Ceballes-Serrano, C., Velásquez-Mejía, D., Riaño-Rojas, J., Giraldo, C., Carmona, J. & Ceballos-Márquez, A. Application of udder surface temperature by infrared thermography for diagnosis of subclinical mastitis in Holstein cows located in tropical highlands. *Journal of Dairy Science*. **104**, 10310–10323 (2021)
42. Lima, M. & Pandorfi, H. Thermal Image Thresholding for Automatic Detection of Bovine Mastitis. *International Journal Of Computer Applications*. **975** pp. 8887
43. Bradski, G., Kaehler, A. & Others OpenCV. *Dr. Dobb's Journal Of Software Tools*. **3** (2000)
44. Khakimov, A., Pavkin, D., Yurochka, S., Astashev, M. & Dovlatov, I. Development of an algorithm for rapid herd evaluation and predicting milk yield of mastitis cows based on infrared thermography. *Applied Sciences*. **12**, 6621 (2022)
45. FA, P., BP, P., RG, S. & Others Application of Infrared Thermography as a Determinant of Sub-Clinical Mastitis in Sapera Dairy Goats.. *Indonesian Journal Of Animal & Veterinary Sciences/Jurnal Ilmu Ternak Dan Veteriner*. **27** (2022)
46. Kittur, P., Satheesan, L., Madhusoodan, A., Sriranga, K., Kumar, D., Kamboj, A. & Dang, A. Correlation of udder thermogram and somatic cell counts as a tool for detection of subclinical mastitis in buffaloes. *Veterinary Research Communications*. pp. 1-9 (2024)