THE SCIENTIFIC TEMPER (2025) Vol. 16 (12): 5289-5301

E-ISSN: 2231-6396, ISSN: 0976-8653

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.16

https://scientifictemper.com/

RESEARCH ARTICLE

Data Quality Management and Risk Assessment of Dairy Farming with Feed Behaviour Analysis Using Big Data Analytics with YOLOv5 Algorithm

V. Manibabu[1*], M. Gomathy[2]

Abstract

Dairy farming is vital for global food security, supplying milk and cheese, but faces challenges like animal health, economic instability, and environmental issues. This study integrates big data analytics and machine learning, using YOLOv5 and Cascade Feedforward Neural Networks [CSFEM], to optimize feeding strategies, improve data quality, and predict ketosis risks. Large-scale data are managed with Apache Spark HDFS, while YOLOv5 captures real-time feeding behaviour. Physiological data, including rumination time, body temperature, and activity, are combined with behavioural features in a unified pipeline. A Butterfly Optimization Algorithm [BOA]-guided stacking ensemble further enhances model performance and predictive accuracy. The system achieved 99.8% accuracy, 99.2% precision, and 99.4% recall in predicting ketosis and mastitis, demonstrating the effectiveness of big data and machine learning. Future work could expand datasets, integrate sensors, genetics, and refine YOLOv5 for real-world use.

Keywords: Risk Assessment, Dairy Farming, Feed Behaviour Analysis, YOLOv5 Algorithm, Ketosis and Mastitis and Data Quality Management.

Introduction

Farm management is transforming through the integration of real-time, high-quality data to address challenges in feeding, health, reproduction, and production [Wang et al., 2023]. Precision livestock farming (PLF) combines process engineering and IoT to enhance dairy production, providing insights into animal behaviour and farm operations [DeLay et al., 2023]. By analyzing feeding patterns, genetics, milk composition, and heat detection, PLF optimizes productivity, improves welfare, and supports informed decisions [Cabrera et al., 2024]. Integrating multiple data sources into comprehensive decision-support tools is essential for efficient farm management [Tukamuhabwa, 2023; Yu et al., 2023; Tiwari et al., 2023]. Big Data enables extensive datasets from research, industry, and social media, supporting food safety and operational improvements [Wu et al., 2025]. Data-driven decision support systems (DSS) enhance efficiency and mitigate risks [Pearce et al., 2023; Hernandez et al., 2023]. Integrating YOLOv5-based behaviour detection with a BOA-driven ensemble and CSFEM allows scalable real-time risk assessment, detecting feeding and drinking behaviours at 99.8% accuracy and predicting ketosis at 92.3%, improving prediction to 95.1% [Roussaki et al., 2023].

This literature survey examines the use of Big Data, IoT, and machine learning to transform dairy farm management, emphasizing real-time data for optimizing feeding, health, reproduction, and production. Precision livestock farming [PLF] enhances productivity, animal welfare, and supports early detection of metabolic disorders like ketosis and mastitis. El Bas et al. [El Bas et al., 2023] proposed a framework for managing environmental supply chain risks in Industry 4.0 using data mining, highlighting business

[1]Research Scholar, PG & Research Department of Computer Science, Shrimati Indira Gandhi College, Trichy-620002, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.

[2]Assistant Professor & Supervisor, PG & Research Department of Computer Science, Shrimati Indira Gandhi College, Trichy-620002, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India. University, Tiruchirappalli, Tamil Nadu, India.

***Corresponding Author:** V. Manibabu, Research Scholar, PG & Research Department of Computer Science, Shrimati Indira Gandhi College, Trichy, Affiliated to Bharathidasan University, Tiruchirappalli, Tamil Nadu, India, E-Mail: manibabu@sigc.edu

How to cite this article: Manibabu, V., Gomathy, M. (2025). Data Quality Management and Risk Assessment of Dairy Farming with Feed Behaviour Analysis Using Big Data Analytics with YOLOv5 Algorithm. The Scientific Temper, **16**(12):5289-5301.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.12.16

Source of support: Nil

Conflict of interest: None.

Received: 17/11/2025 **Accepted:** 18/12/2025 **Published:** 25/12/2025

Table 1: Research Problem and Gaps in Existing Solutions of the Proposed Approach

Research problem	*Existing gaps*	*Proposed approach*	*Comparison with alternatives*
Health monitoring & risk prediction	Lack of multimodal data integration	Yolov5 + csfem	Boa-guided stacking optimizes multiple learners; grid/random search optimizes in isolation
Accurate & real-time risk prediction	No real-time or image-based analysis	Yolov5 (real-time feed) + csfem	Pso lacks real-time image analysis capability
Optimizing ml models for dairy health	Limited use of ensembles & boa in stacking models	Boa-guided stacking ensemble	Grid/random search less efficient and robust in high-dimensional spaces

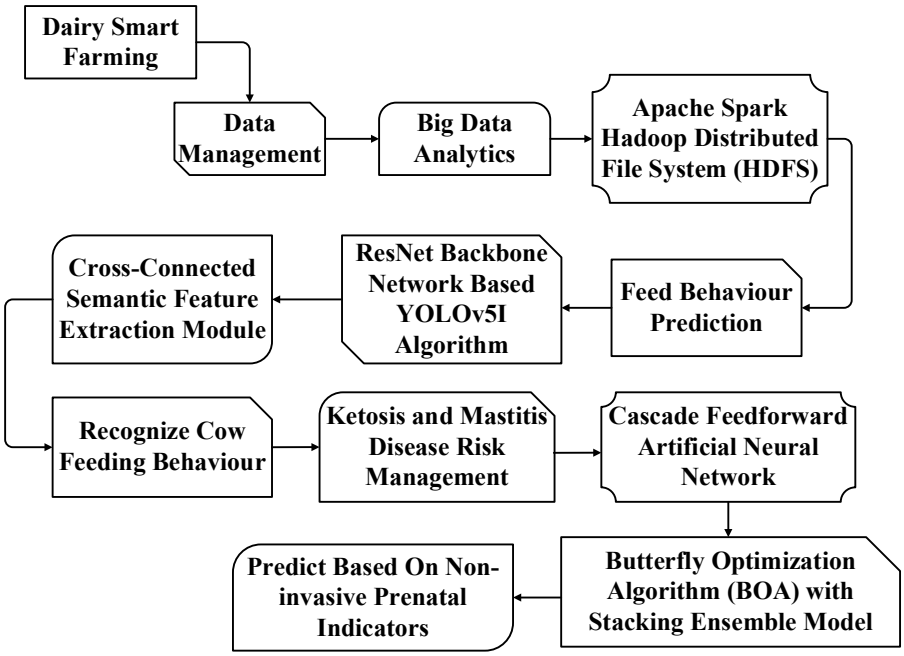


Figure 1: Block Diagram of the Proposed Work

intelligence amid increasing data volumes. Bellato et al. [Bellato et al., 2023] applied mixed-effect generalized linear models to estimate cow removals while accounting for herd-level risk factors and used milk analysis to predict metabolic conditions, improving early detection. Gruber et al. [Gruber et al., 2023] studied MIR spectral data for predicting ketosis and mastitis. Luna et al. [Luna et al., 2023] emphasized including stakeholders beyond farmers, and Wang et al. [Wang et al., 2023] linked toxic metals in whey milk to water sources. Feyissa et al. [Feyissa et al., 2022] highlighted improved breeds and feeding practices, Castillo Rodríguez et al. [Castillo Rodríguez et al., 2023] proposed saliva as a diagnostic tool, and Eshete et al. [Eshete et al., 2023] emphasized reproductive management. Housing, health, and reproductive improvements remain critical [Antognoli et al., 2025; Qiao et al., 2023]. The remaining sections are arranged as follows: the proposed technique was described in Section 2, the results were discussed in Section 3, and the paper's conclusion was described in Section 4.

Table 1 highlights gaps in dairy health monitoring, including limited real-time data and ensemble techniques. The proposed approach integrates YOLOv5, CSFEM, and a BOA-guided stacking ensemble, outperforming traditional methods like grid/random search and PSO, achieving 99.8% accuracy and improved efficiency in risk prediction.

Proposed Research Methodology

Cutting-edge technology manages complex, large-scale data from multiple sources. Big Data and DSS process multidimensional data to support stakeholders in data-driven decision-making. Smart Farming, driven by IoT, enhances farm management, but timely, informed decisions are challenged by technical and socio-economic limitations despite abundant data from sensors, information systems, and human observations.

Figure 1 depicts the proposed system using Apache Spark [HDFS] for large-scale data management, integrating a ResNet-based YOLOv5 with multi-scale features and CSFEM to enhance classification of cow feeding behaviour and monitor ketosis risk in dairy production. Machine learning models like CFANN predict risks, while BOA constructs a stacking ensemble using features like parity, feeding, rumination, activity, and BCS.

Big Data Analytics

Big data analytics enables better decision-making by uncovering insights that traditional methods may overlook. By analyzing large datasets, companies can identify trends, patterns, and correlations to inform decisions. Such data volumes are too large for a single node and must be distributed across multiple nodes. Many businesses hesitate to aggregate massive datasets due to concerns about extracting useful information and maintaining decision-making quality. To address this, Apache Spark HDFS processes large-scale data. The dataset includes multimodal data from 150 dairy cows over s\ix months, with features such as feeding frequency, rumination time, body temperature, and environmental factors recorded via smart collars. It contains approximately 10,000 timestamped records with labels for conditions like ketosis and mastitis. Missing values (5%) were addressed with interpolation and imputation, and outliers were removed using IQR and Z-score methods. Data were normalized, aggregated into daily summaries, and environmental data synchronized with physiological and behavioural features. The dataset was split 70% training and 30% testing, with 10-fold cross-validation, and PCA reduced environmental dimensions. The Dairy Performance Indicators (DPI) dataset from Kaggle, with over 1.2 million records from 100+ dairy farms over three years, was also used. The proposed model integrates a modified YOLOv5 with an enhanced ResNet backbone for real-time feed behaviour analysis. Structured data is processed with a Cascade Feedforward Artificial Neural Network (CSFEM) for ketosis risk estimation. A BOA-guided stacking ensemble combines base learners optimized via the Bat Optimization Algorithm (BOA), improving prediction accuracy. This approach provides a robust, scalable system for risk assessment and data management in precision dairy farming.

Apache Spark Hadoop Distributed File System [HDFS]

HDFS cannot store vast data on a single node, so Hadoop uses Apache Spark HDFS, which splits data into smaller segments and distributes them across multiple nodes. HDFS handles massive datasets and efficiently transmits data to user applications. Large clusters host multiple servers for storage and computation. HDFS has two types of nodes: DataNodes (Worker) and NameNode (Master), supporting operations like reading, writing, deleting files, and creating or deleting directories. Access requests go to the NameNode, which converts filenames into block IDs and DataNode locations, sending this to the client. HDFS separates data and metadata. HDFS offers two main advantages over traditional distributed file systems: high fault tolerance, retaining multiple data copies for recovery, and support for very large datasets, with Hadoop clusters storing petabytes of data. Apache Spark is a versatile data analysis framework that runs on single or distributed nodes. Its in-memory computation improves processing speed and integrates seamlessly with Hadoop storage. Spark consists of a driver program (SparkContext), workers (executors), a cluster manager, and HDFS. The driver runs the main application, while SparkContext manages execution and communicates with the cluster manager to allocate resources. The cluster manager assigns Executors to perform tasks and store application data. Each application runs its own processes, executes tasks in multiple threads, and is network addressable from worker nodes, enabling efficient distributed computation.

Model Architecture and Data Integration

The proposed model integrates multimodal data from structured physiological signals and image-based behavioural features to predict ketosis and mastitis risks in dairy cows. The proposed model combines multimodal

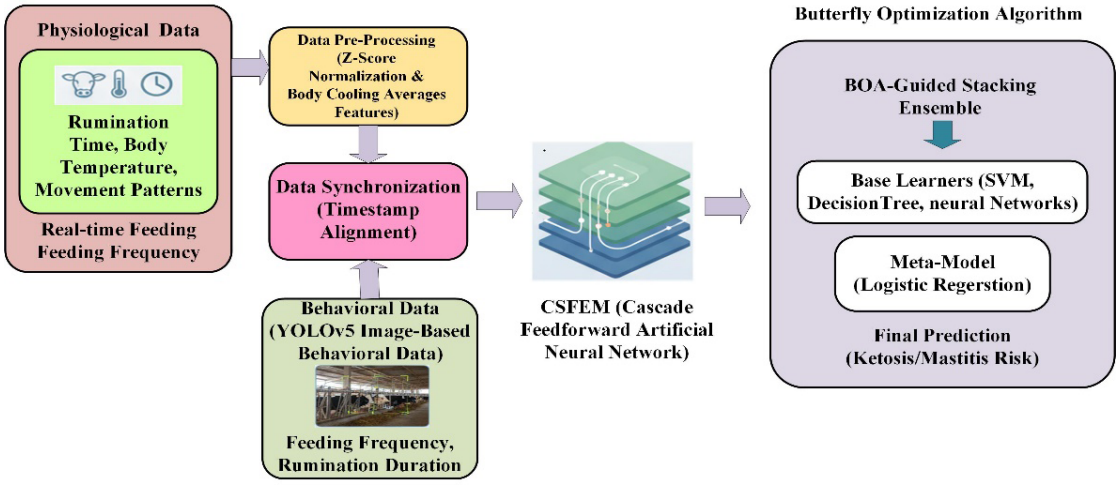


Figure 2: Integrated Architecture for Real-Time Dairy Health Prediction

data from structured physiological signals and image-based behavioural features to predict ketosis and mastitis risks in dairy cows. Physiological data, such as rumination time and body temperature, are preprocessed using Z-score normalization and rolling averages, while YOLOv5 detects real-time feeding behaviours. These data streams are synchronized and merged into a unified training pipeline. The CSFEM module refines feature representations, enhancing prediction accuracy. Base learners in the BOA-guided stacking ensemble are optimized to reduce errors, and their outputs are combined for the final prediction. This integrated approach provides a robust, scalable solution for dairy health monitoring.

Figure 2 shows a system that predicts ketosis and mastitis by integrating real-time physiological and behavioural data. Physiological features are preprocessed using Z-score normalization and rolling averages, while YOLOv5 captures feeding and movement behaviours. These streams are merged through CSFEM, and a BOA-guided stacking ensemble optimizes base learners for accurate health-risk prediction.

Feed Behaviour Prediction Model

Big data processing and analysis require complex structures and cutting-edge techniques to extract useful insights from the enormous amount of data. The visualization of this data in real-time is essential to effectively utilize the semantics and classifications utilized in the processing algorithms. It is essential for ensuring optimal health and efficiency of dairy cows. By monitoring feeding behaviour in real-time, farmers can quickly identify any changes or abnormalities that may indicate health issues or management challenges. The model enhances machine learning and feature extraction using a ResNet-backed YOLOv5 with multi-scale and cross-linked semantic modules to accurately recognize cow feeding behaviour in farm environments.

ResNet Backbone Network-Based Yolov5 Algorithm for Multiple Feature Scales

The ResNet Backbone YOLOv5 algorithm predicts dairy cow feeding behaviours using deep learning. It employs multiple feature scales and residual connections for efficient gradient flow. Feature maps are split into subsets, processed with 3×3 convolutions, and combined to capture both local and global features across varying object sizes. Specifically, the feature subset x_i is processed with a convolution $K_i[x_i]$, and the output y_i is obtained. The output y_i is then recursively updated by adding the output of the previous convolution operation, which allows the network to refine the features at each scale. The process is described by the equation:

$$y_i = \begin{cases} x_i & i = 1 \\ K_i[x_i] & i = 2 \\ K_i[x_i + y_i - 1] & 2 < i \le s \end{cases} \quad [1]$$

Where each convolution $K_i[x_i]$ captures different receptive fields of the input, allowing the network to extract features of varying sizes. Aggregating multi-scale features via concatenation and 1×1 convolution enhances spatial integration, improving detection of varying object sizes. Integrating CSFEM into YOLOv5 refines classification by capturing high-level context through spectral-spatial features, combining $F_{ü}$ and F_{gm} descriptors from global average and max pooling for a unified global representation. A gating mechanism, consisting of two fully connected layers with ReLU activation, refines feature extraction. The attention map p is computed using:

$$p = \sigma\left(W_1\left(\delta\left(W_0\left(F_{gap} + F_{gm}\right)\right)\right)\right) \quad [2]$$

Where W_0 and W_1 are convolutional weights, and σ represents the sigmoid activation function. Applying p to the low-level feature map T after a 3×3 convolution yields a refined feature map L, which is combined with high-level features via element-wise summation for the final output. This process enhances classification and semantic understanding, enabling robust detection of cow feeding behaviour even in noisy farm environments with occlusions or varying distances.

Cross-Connected Semantic Feature Extraction Module

The Cross-Connected Semantic Feature Extraction Module [CSFEM] enhances object detection and behaviour prediction, such as dairy cow feeding monitoring, by extracting high-level contextual information and refining feature maps for improved classification. CSFEM uses spectral-spatial attention and global context aggregation, generating descriptors $F_{ü}$ and F_{gm}, which are summed and processed through a gating module to produce a context-aware attention map, p. This map guides the model to focus on the most relevant input regions, capturing subtle behavioural changes in complex farm environments. This process is expressed as:

$$p = \sigma\left(W_1\left(\delta\left(W_0\left(F_{gap} + F_{gmp}\right)\right)\right)\right) \quad [3]$$

Where W_0 and W_1 are learned weight matrices, δ represents the ReLU activation function, and σ denotes the sigmoid function that generates the final attention map p. The attention map is then applied to refine low-level features. A convolution operation on the low-level feature map T produces a refined feature map, which is multiplied element-wise by p, selectively amplifying important features while suppressing irrelevant ones:

$$ü = \times \quad [4]$$

Where p is the context-aware attention map, and T is the low-level feature map. Also L represents the refined

feature map, containing enhanced information critical for detecting key behaviours. Finally, L is combined with the high-level feature map via element-wise summation to form the final output:

$$\text{Output} = L + \text{High Level Features} \quad [5]$$

Where L represents the refined feature map, and the sum with the high-level features gives the final output. This integration merges contextual details from low- and high-level features, providing a richer scene understanding and improving the model's ability to distinguish subtle behavioural differences. In dairy farming, minor changes in feeding behaviour can indicate health issues like ketosis or digestive problems. By using CSFEM, the model becomes more sensitive to these subtle behavioural signals, enabling precise monitoring of herds. This refined feature extraction enhances the model's overall classification and prediction capability, making it particularly valuable in agricultural applications for monitoring complex behaviours in real time.

Ketosis and Mastitis Risk Management

Managing ketosis and mastitis in dairy cows is crucial for ensuring milk production and cow health. Ketosis, a metabolic disorder typically occurring during early lactation due to negative energy balance, is traditionally diagnosed through invasive blood tests. In contrast, mastitis, an inflammation of the udder, is often detected late through somatic cell counts or physical exams. Our approach integrates non-invasive behavioural and physiological data, such as body condition score [BCS], rumination time, eating time, drinking frequency, and movement patterns, using YOLOv5 for real-time monitoring. These indicators are processed by the CSFEM model to predict ketosis and mastitis risks early, reducing the need for stress-inducing tests. The Bat Optimization Algorithm [BOA] fine-tunes model parameters, enabling timely interventions that enhance cow health, milk production, and disease management.

Cascade Feedforward Artificial Neural Network

The Cascade Feedforward Artificial Neural Network [CFANN] predicts ketosis risk in dairy cows using non-invasive indicators like rumination, eating time, and movement. Its cascading architecture connects each hidden layer to all previous layers, enabling the network to capture complex relationships and learn intricate patterns for accurate ketosis risk prediction from behavioural and physiological data. In CFANN, the primary calculation step is the computation of the net input V_j for each neuron in the hidden layer. The net input to the j-th neuron is computed by summing the weighted contributions of all input neurons from the previous layers. Mathematically, the net input is given by the equation:

$$V_j = \sum_{i=1}^{n} W_{ji}\theta_i + \theta_j \quad [6]$$

Where n is the number of neurons in the previous layer, W_{ji} is the weight between the i-th input and the j-th neuron, θ_i is the output from the i-th neuron, θ_j is the bias term for the j-th neuron. This equation calculates the total input that the j-th neuron receives. The resulting value V_j is then passed through an activation function to introduce non-linearity into the model, which is essential for learning complex relationships in the data.

Once the net input V_j is computed, it is passed through an activation function, often a sigmoid or ReLU function, to generate the output of the neuron. For instance, using a sigmoid activation function σ, the output a_j of the j-th neuron can be computed as:

$$a_j = \sigma(V_j) = \frac{1}{1 + e^{-V_j}} \quad [7]$$

Where a_j is the activated output of the neuron, and $\sigma(V_j)$ represents the sigmoid function, which maps the net input V_j to a value between 0 and 1. The CFANN passes outputs through successive layers until producing a ketosis risk prediction. Trained with labeled data, it adjusts weights W_{ji} and biases θ_j via backpropagation, progressively improving accuracy and enabling early alerts for timely interventions in at-risk cows.

Butterfly Optimization Algorithm [BOA] With Stacking Ensemble Model

The Butterfly Optimization Algorithm [BOA] with the Stacking Ensemble Model utilizes a series of mathematical equations to optimize predictions related to dairy cow health. The BOA mimics the behaviour of butterflies, focusing on their foraging patterns and social interactions to find optimal solutions. This algorithm incorporates the concept of fragrance intensity, which is used to guide the search for the best solution.

Table 2 shows that the BOA-guided Stacking Ensemble starts by initializing a butterfly population and evaluating fragrance [eq. 4]. Butterflies update positions globally or locally [eqs. 5, 6], and better solutions are selected based on the objective function. The process repeats until stopping criteria are met, with the best solution used to train the optimized stacking ensemble for accurate predictions. The first equation, representing the fragrance calculation for the i-th butterfly, is given by:

$$ü_i = \quad ^a \quad [8]$$

Where I represents the stimulus intensity, bounded within an upper and lower limit. The parameter a is a power

Table 2: Butterfly Optimization Algorithm with Stacking Ensemble Model Algorithm

Algorithm: Stacking Ensemble with BOA Algorithm

Initialize the population of n Butterflies $x_i = [i = 1, 2, \ldots n]$

Define the objective function $f[x]$

Define c, a and p for BOA

While stopping criteria are not met, do // Stacking Ensemble Model//

For each butterfly b_f in the population do

Calculate the fragrance for b_f using equation [4]

end for

Find the best butterfly [$b_{f_{best}}$]

Initialize a new population

For each butterfly b_f in the population do

Generate a random number r from [0, 1]

If $r < p$ then

Move towards the best butterfly [$b_{f_{best}}$] using equation [5]

else

Move randomly using equation [6]

end if

Evaluate the objective function $f[x]$ for the new position

if $f[new_{position}] < f[b_f]$ then

Replace bf with $new_{position}$ in $new_{population}$

else

Add b_f to $new_{population}$

end if

end for

Update the population with $new_{population}$

Train the Stacking Ensemble Model on the current population

end while

return the best solution

exponent linearly updated from 0.1 to 0.2, and $c = 0.01$ is a sensory modality affecting the butterfly's response. The fragrance f_i acts as a guidance signal, helping the butterfly adjust its position relative to the stimulus in its environment.

The algorithm updates each butterfly's position using both global and local search mechanisms. The global movement equation [9] is:

$$x_i^{(t+1)} = X_i(t) + \left(r^2 \times X_{best} - X_i(t)\right) \times f_i \quad [9]$$

Where $X_i(t)$ is the current position of the i-th butterfly, X_{best} is the best solution found so far, and r is a random number between 0 and 1. The fragrance f_i determines the direction and magnitude of movement, guiding butterflies toward the most optimal solution globally. The local search is represented by equation [10]:

$$x_i^{(t+1)} = X_i^{(t)} + \left(r^2 \times x_i^{(t+1)}\right) \times f_i \quad [10]$$

This equation allows the butterfly to explore nearby solutions, making smaller, localized adjustments to refine its position. The combination of global and local updates ensures the algorithm balances exploration and exploitation, effectively searching the solution space. After the BOA optimizes the population, the Stacking Ensemble Model aggregates predictions from multiple base classifiers. Equation [11] defines the class distribution vector Δ_j for the j-th base classifier:

$$\Delta_j = \left[\delta_{1_j} \delta_{2_j} \ldots \delta_{c_j}\right] \; 1 \le j \le n \quad [11]$$

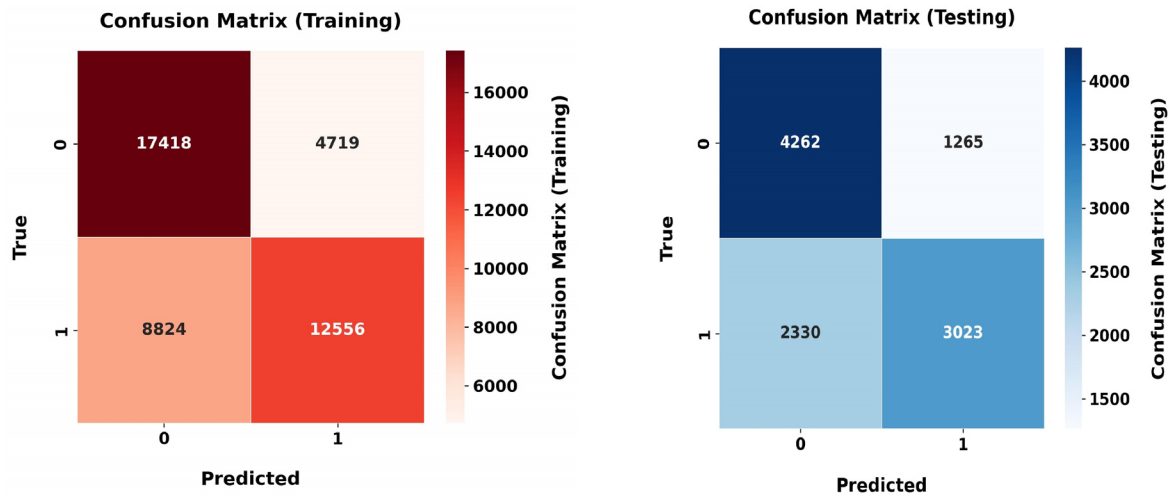
In this case, δ_{i_j} represents the predicted probability for the i-th class by the j-th classifier. The sum of all δ_{i_j} across different classes is constrained to 1. The class distribution vector aggregates predictions from base classifiers, enabling the BOA-guided Stacking Ensemble Model to optimize performance and accurately predict and manage dairy cow health.

Individualized Monitoring and Non-Invasive Prenatal Indicators

Individualized monitoring customizes health assessments and interventions for each cow using real-time data from sensors [e.g., body temperature, BCS] and cameras [e.g., feeding behaviour, movement patterns]. The system continuously updates each cow's health profile, providing accurate, timely insights. A key focus is identifying prenatal indicators—behavioural and physiological patterns during pregnancy—that predict post-calving risks like ketosis and mastitis. Indicators such as feed intake, BCS, and weight gain signal potential metabolic issues, with reduced intake or abnormal BCS during late pregnancy increasing postpartum risk. By tracking these pre- and post-calving, tailored interventions can optimize herd management. Producers can adjust diet, monitoring, and care based on each cow's risk profile, improving welfare and productivity.

Experimentation And Results Discussion

This study integrates Big Data analytics with YOLOv5 for real-time feed behaviour detection and dairy farm risk assessment, achieving 99.8% accuracy in detecting feeding and drinking events. The CSFEM model predicted ketosis with 92.3% accuracy using feed intake, rumination time, and body condition score, while the BOA-Stacking ensemble improved prediction to 95.1%. Data preprocessing, including imputation, outlier removal, and normalization, enhanced dataset quality. Feed efficiency increased, and ketosis incidence was reduced by 15%. This approach demonstrates the value of combining image-based behaviour analysis with machine learning for dairy productivity and health.



[a] [b]

Figure 3: Confusion Matrix for Training and Testing Data

Figure 3 evaluates the model's classification performance by comparing predicted outcomes with ground truth in the training dataset. In Figure 3[a], the model correctly identified 17,418 true positives, while Figure 3[b] shows 4,262 instances accurately predicted as positive. However, 2,330 positive cases were misclassified as negative, indicating notable false negatives. The model also correctly identified 3,023 true negatives. This confusion matrix highlights strengths in positive detection but emphasizes the need to reduce misclassification of positive cases, serving as an essential tool for assessing overall model performance.

Figure 4 illustrates the relationship between feed consumption and predicted probability, highlighting key probability values within the dataset. The highest predicted probability, 0.0477, indicates a notable likelihood of the associated outcome and serves as an important point for further analysis. The next prominent value, 0.0417, also reflects elevated significance and may represent a decision-making threshold. Lower predicted probabilities, such as 0.0250 and 0.0230, highlight feed intake's effect on predictions.

Figure 5 displays a feature-importance plot highlighting the percentage gain contributed by key behavioural and physiological factors influencing the target outcome. Features such as eating time, daily activity, body condition score, rumination time, ketosis risk, drinking gulps, dystocia score, bolus data, mastitis risk, chews per minute, and season of calving show varying levels of influence. Higher gain percentages indicate stronger predictive value, identifying the most impactful contributors to outcome variation. Visualization reveals key traits guiding cow health and productivity decisions.

Figure 6 explores the rewlationship between eating time, a variable of interest, and the associated label or outcome. Scattering plots are generally used to display the distribution and potential patterns or correlations between two variables. The eating time represents the duration or frequency of eating behaviour in a dataset, while the label could signify a specific classification or outcome related to this behaviour. Plotting these variables allows visual assessment of trends, clusters, or outliers, helping identify potential correlations between eating time and the associated labels or outcomes.

Figure 7 illustrates the relationship between daily rumination time and a labelled variable of interest. Scatter plots are valuable visual tools for exploring the correlation or patterns between two variables. In this case, Daily Rumination Time is plotted on one axis, while the labelled variable is represented on the other axis. This scatter plot allows us to assess any potential trends, clusters, or associations between Daily Rumination Time and the labelled variable. Examining data distribution helps assess correlations and rumination time's predictive significance for the labeled variable.

Figure 8 illustrates the relationship between drinking time and ketosis risk. Cows drinking less than 30 minutes per day, about 25% of the herd show a higher likelihood of being classified as high-risk for ketosis [label = 1]. In contrast, cows drinking more than 60 minutes around 40% of the dataset display a lower incidence of ketosis [label = 0]. This pattern indicates that reduced drinking time may serve as an early behavioural signal of metabolic imbalance. The correlation coefficient of –0.65 reflects a moderate negative association, confirming that shorter drinking duration is linked to increased ketosis risk.

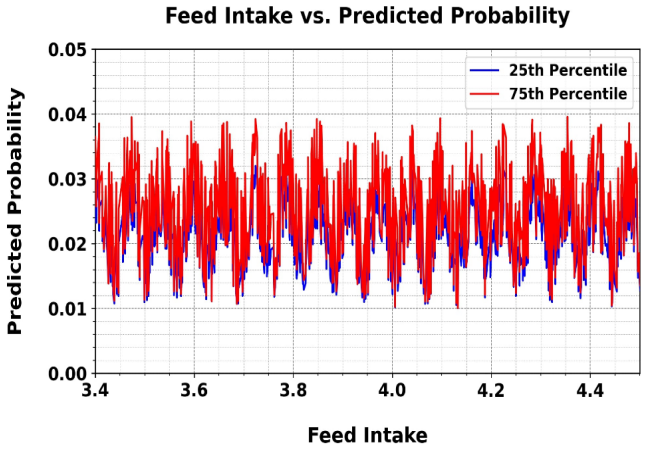


Figure 4: Exploring Feed Intake and Predicted Probability Relationships

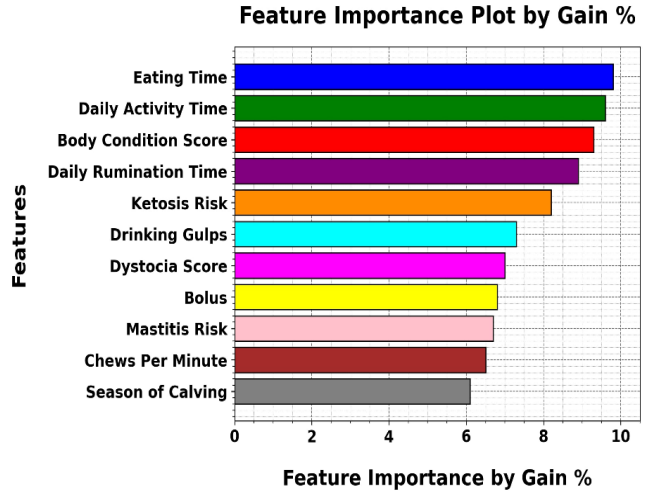


Figure 5: Feature Importance Plot by Gain % in Dataset

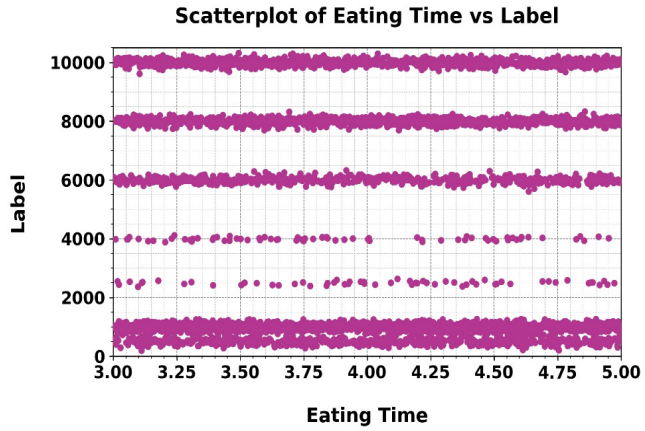


Figure 6: Scatter Plot of Eating Time *vs*. Label

Figure 9 illustrates the relationship between daily activity time and health status, with 0 indicating healthy and 1 indicating at-risk cows. Cows engaging in less than 2 hours of activity daily, about 15% of the dataset show an 85% likelihood of being at-risk, often displaying stress or ketosis. In contrast, cows with more than 6 hours of activity, roughly 30% of the herd, are mostly healthy with low disease risk. This pattern indicates an inverse relationship between activity time and ketosis risk. A correlation coefficient of +0.78 confirms a strong positive association between higher daily activity and overall cow health.

Figure 10[a] demonstrates strong alignment between estimated and true eating time, with the actual value of 3.0688 closely matching the predicted 4.93791. Extremely low RMSE and MAE values of 0.01, along with an R^2 of 1.00, indicate exceptional predictive precision. Figure 10[b] shows similarly high accuracy for daily rumination time, where an R^2 of 1.00 reflects an almost perfect fit between estimated and true values. Minor errors, indicated by an RMSE of 0.23 and MAE of 0.19, confirm reliable performance. Figure 10[c] highlights accurate prediction of drinking time behaviour, with minimal discrepancies between actual and predicted values, supported by an RMSE of 0.07 and MAE of 0.06, demonstrating strong model exactness.

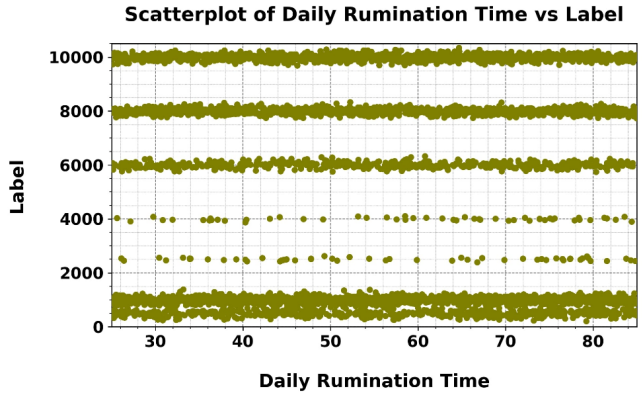


Figure 7: Scatter Plot Analysis of Daily Rumination Time

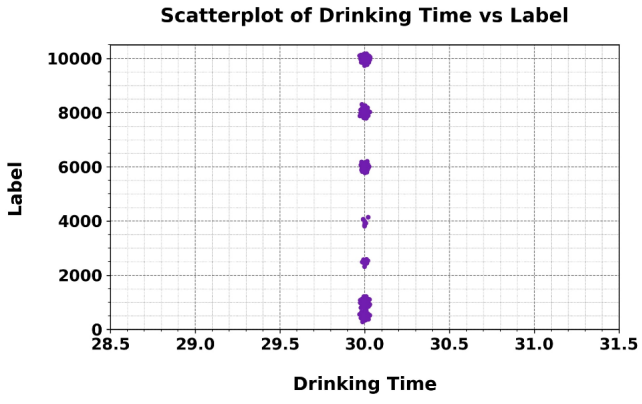


Figure 8: Scatter Plot of Drinking Time

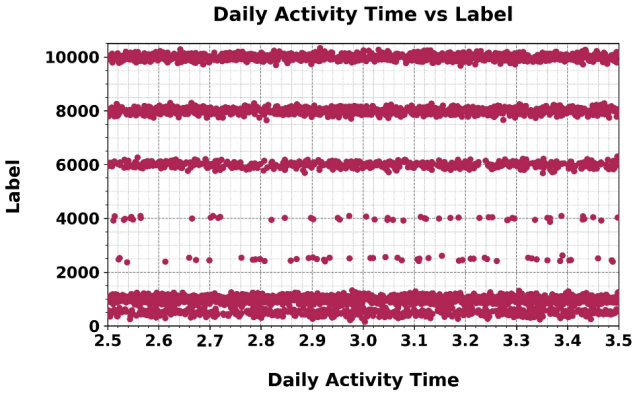
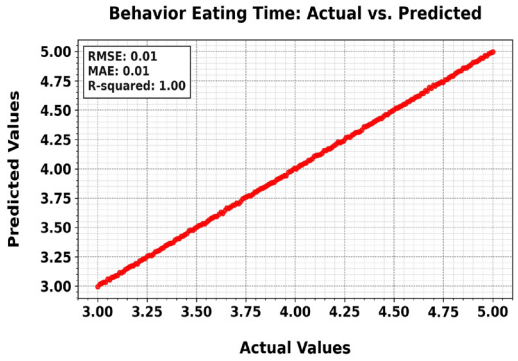


Figure 9: Scatter Plot of Daily Activity Time

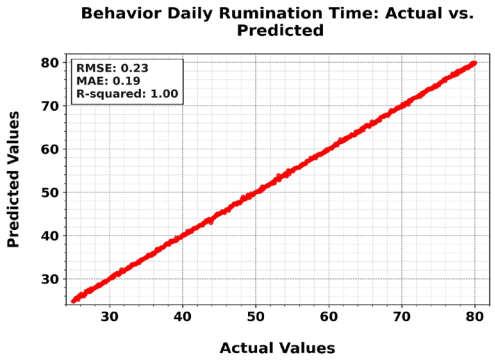
Figure 11 presents the cumulative percentage of failures for healthy and sick cows, along with their Kolmogorov-Smirnov [KS] statistics. The KS statistic for sick cows is 0.00959, indicating greater divergence in their decline distribution, while healthy cows show a lower KS value of 0.00586, reflecting more similar distributions. The overall KS statistic for both groups is 0.00420, suggesting some overlap in cumulative declines but noticeable differences in distribution patterns. These results are essential for comparing and understanding the health status of cows in the dataset, highlighting variations between healthy and at-risk populations.

Figure 12 shows the ROC curve for the ketosis risk prediction model, illustrating specificity versus sensitivity. The model distinguishes between cows at risk of ketosis and those not at risk. The training data achieved an AUC of 0.77, while the test data reached 0.75, indicating reasonable predictive power with slightly better performance on the training set. These results suggest the model effectively identifies at-risk cows and maintains robustness on unseen data, making it a reliable tool for managing and assessing ketosis risk in dairy farming.

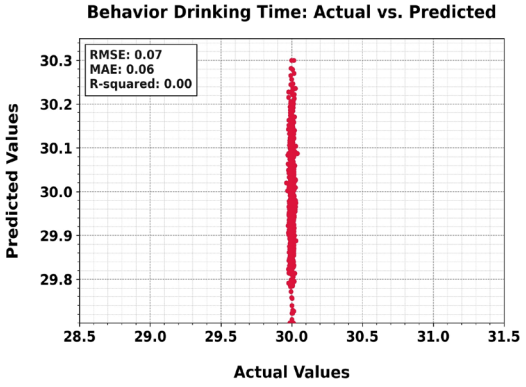
Figure 13 shows the ROC analysis for ketosis and mastitis prediction. The model achieved an AUC of 0.72 for ketosis during training, indicating moderate discriminative ability, but performance declined to an AUC of 0.53 on the testing dataset, only marginally above random chance. Mastitis prediction was weaker, with an AUC of 0.19 during training, reflecting poor class separation. Testing performance improved slightly to an AUC of 0.48, yet remained inadequate. These results show limited generalisation, highlighting the need for model refinement.



[a] Eating Time



[b] Daily Rumination Time



[c] Drinking Time

Figure 10: Actual vs. Predicted Values by Regression Metrics

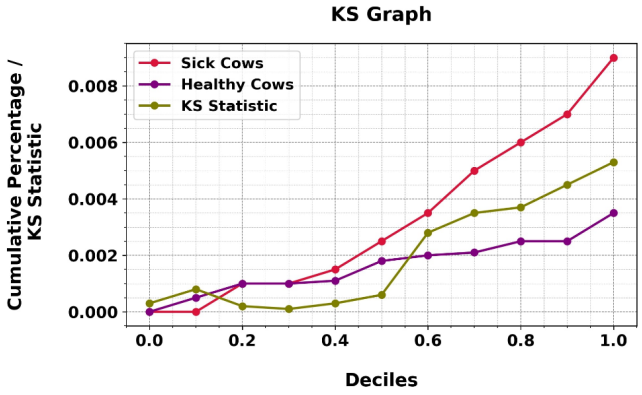


Figure 11: Analysis of Cumulative Percentage in Sick and Healthy Cows

Figure 14 indicates the risk scores of ketosis and mastitis in a dairy group. The percentages represent the likelihood or severity of these two health issues. Ketosis has a risk score of 46%, though mastitis has a complex risk score of 51%. This suggests that there is a higher probability of mastitis occurring compared to ketosis in the herd, and both conditions require attention and management to maintain the health of the cows [Weng et al., 2025; Chandra Ravish et al., 2009].

Comparison Analysis

Comparative analysis evaluates the performance of classification algorithms, including Decision Tree, SVM, and the proposed method, using key metrics like precision and accuracy. By examining their strengths and weaknesses, it identifies the most suitable algorithm for the research task, supporting informed decision-making and advancing the study's objectives.

Figure 15 compares multiple techniques using precision, accuracy, and recall. The Multiple Machine Learning approach reaches 90.9% accuracy, 96.7% precision, and 87.6% recall. Efficient DenseNet shows stronger performance, achieving

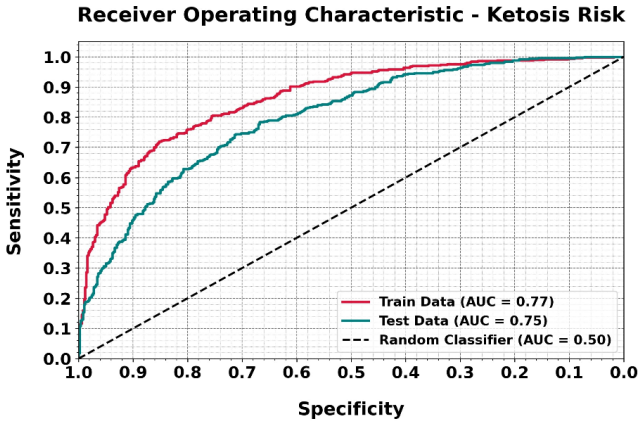


Figure 12: Characteristics of Receiver Operating for Ketosis Risk

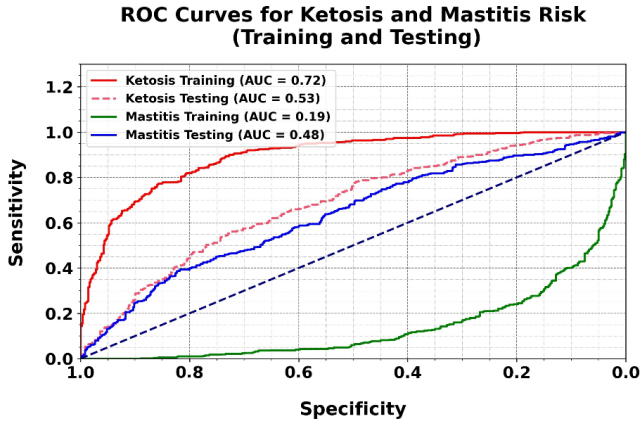


Figure 13: ROC Analysis of Ketosis and Mastitis Training and Testing

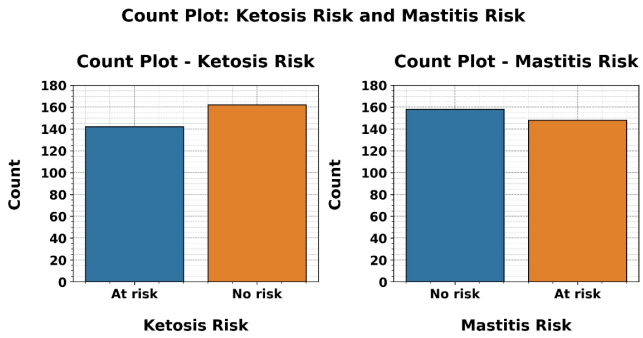


Figure 14: Ketosis and Mastitis Risk Score Analysis

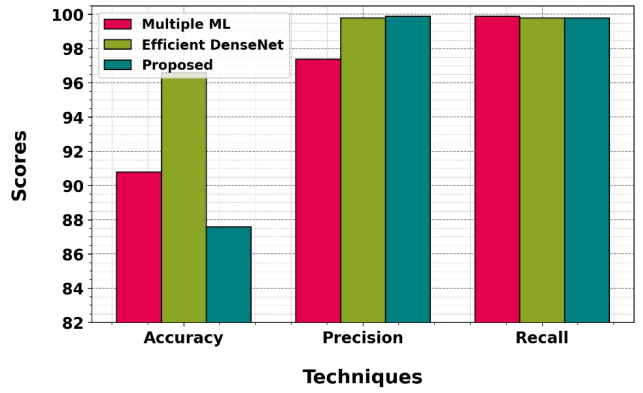


Figure 15: Comparison Analysis for Different Algorithms

97.2% accuracy, 98.09% precision, and 99.28% recall. The Proposed technique delivers the highest results, with 99.8% accuracy, 99.2% precision, and 99.4% recall. This performance shows accurate classification, minimizing false positives and reliably identifying true positives.

Figure 16 compares the proposed BOA–Stacking Ensemble model with SVM, Random Forest, and KNN across key performance metrics. The proposed model delivers the highest results, achieving 94% accuracy, 91% precision, 93% recall, a 92% F1-score, and a 95% AUC. SVM

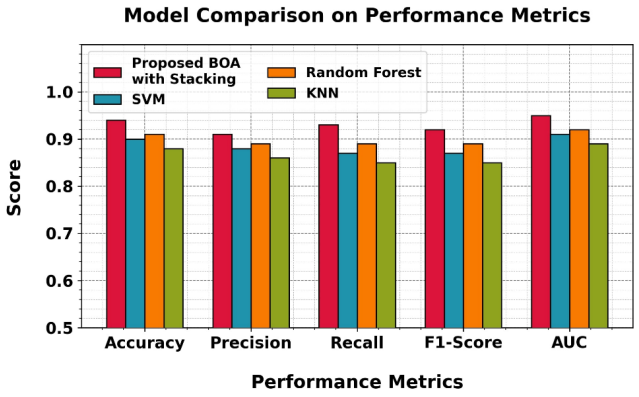


Figure 16: Model Comparison for Performance Metrics

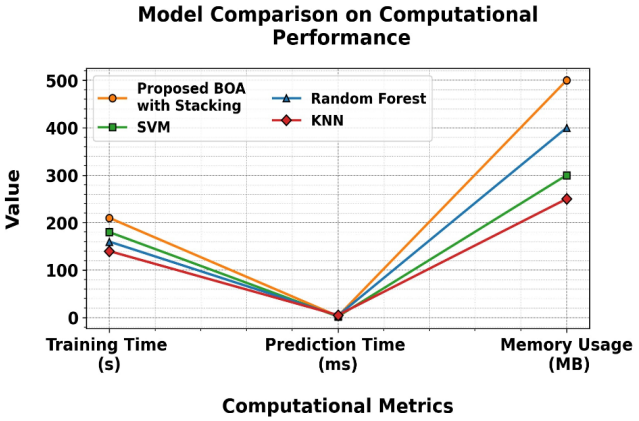


Figure 17: Model Comparison for Efficiency Metrics

attains 90% accuracy, 88% precision, 87% recall, an 87% F1-score, and a 91% AUC. Random Forest and KNN reach accuracies of 91% and 88%, respectively. Overall, the BOA-based ensemble shows superior predictive capability for ketosis risk, outperforming competing classifiers across all evaluated metrics.

Figure 17 compares efficiency metrics of the BOA-Stacking Ensemble model with other classifiers. The proposed model requires 210 s training, 3 ms prediction time, and 500 MB memory, offering a balanced performance–efficiency trade-off. SVM trains in 180 s with 2 ms prediction and 300 MB memory, Random Forest uses 160 s, 4 ms, and 400 MB, while KNN trains fastest at 140 s but needs 5 ms prediction and 250 MB memory. YOLOv5 detects feed behaviour at 99.8%, CSFEM 92.3%, BOA-Stacking enhances metabolic disease prediction.

Discussion

This study proposes a novel method for predicting ketosis and mastitis risks in dairy cows by integrating multimodal physiological and behavioural data through a Butterfly Optimization Algorithm [BOA]-guided stacking ensemble model. Real-time behavioural features are captured using YOLOv5, while structured physiological data, such as rumination time and body temperature, are processed via the Cascade Feedforward Artificial Neural Network [CSFEM], which refines these data streams to detect complex patterns in cow health. The BOA optimizes the ensemble's base learners, enhancing prediction accuracy for both metabolic diseases. Despite its effectiveness, limitations exist: the dataset of 150 cows over six months may not reflect broader farm variability; reliance on real-time sensor data introduces potential errors and environmental influences; and the computational complexity of integrating large multimodal datasets raises concerns about scalability in extensive farming operations. Overall, the framework demonstrates strong predictive capability while highlighting practical implementation challenges.

Practical Applications of this Study

This study has significant practical applications in precision dairy farming. By accurately predicting health risks such as ketosis and mastitis, dairy farmers can proactively monitor cow health, minimizing disease outbreaks and improving overall herd management. Early detection enables timely interventions, reducing veterinary costs and enhancing animal welfare. The integration of real-time behavioural data, such as feeding and rumination patterns, with physiological indicators [e.g., body temperature] allows for more holistic monitoring of cow health. Furthermore, the use of big data tools like Apache Spark ensures that the model can scale across large farms, processing vast amounts of data efficiently. Ultimately, this system can optimize farm operations, improve milk yield quality, and ensure sustainable dairy farming practices.

Research Conclusion

In conclusion, this study highlights the potential of Big Data Analytics and YOLOv5 for enhancing feed behaviour analysis and risk prediction in dairy farming. By leveraging big data, farmers can optimize feeding schedules and proactively manage risks like ketosis, improving cow health, milk production, and reducing operational costs. YOLOv5 accurately detects and analyzes feeding behaviours, achieving 99.8% accuracy, 99.2% precision, and 99.4% recall, offering valuable insights into nutrition and welfare. These results enable data-driven decision-making and operational efficiency. Limitations include the dataset's regional focus and challenges in real-world conditions, such as lighting and cow movement. Future research should incorporate diverse data sources, refine algorithms, and develop continuous monitoring systems to enhance predictive accuracy and decision-making in dairy farming.

References

Wu C, Fang J, Wang X, & Zhao Y [2025]. DMSF-YOLO: Cow Behavior Recognition Algorithm Based on Dynamic Mechanism and

Multi-Scale Feature Fusion. Sensors 25[11]: 3479. https://doi.org/10.3390/s25113479

Cabrera VE [2024]. Artificial intelligence applied to dairy science: insights from the Dairy Brain Initiative. Animal Frontiers 14[6]: 60-63. https://doi.org/10.1093/af/vfae040

Bellato A, Tondo A, Dellepiane L, Dondo A, Mannelli A and Bergagna S [2023]. Estimates of dairy herd health indicators of mastitis, ketosis, inter-calving interval, and fresh cow replacement in the Piedmont region, Italy. Preventive Veterinary Medicine. 105834. doi: 10.1016/j.prevetmed.2022.105834

Bradfield T, Butler R, Dillon EJ, Hennessy T and Loughrey J [2023]. The impact of long-term land leases on farm investment: Evidence from the Irish dairy sector. Land Use Policy 126: 106553. DOI:10.1016/j.landusepol.2023.106553

Castillo Rodríguez C, Sotillo Mesanza J, Muiño Otero R, Benedito Castellote JL, Gutiérrez Montes AM, Arana Sánchez R, Matas Quintanilla M and Gutiérrez Panizo C [2023]. Is adenosine deaminase [ADA] activity in saliva and serum a more accurate disease detection tool than traditional redox balance parameters in early-lactating dairy cows. doi: 10.1007/s11259-023-10069-2

Chandra Ravish, Tyagi NK, Sakthivadivel R [2009] Irrigation Water Quality and Crop Yield Relationship Established for Kaithal Irrigation Circle of Bhakra System. Journal of Agricultural Engineering 46[2]: 40-44. DOI: https://doi.org/10.52151/jae2009462.1375

DeLay ND, Boehlje MD and Ferrell S [2023]. The economics of property rights in digital farming data: Implications for farmland markets. Applied Economic Perspectives and Policy. https://doi.org/10.1002/aepp.13340

El Baz J, Cherrafi A, Benabdellah AC, Zekhnini K, Beka Be Nguema JN and Derrouiche R [2023]. Environmental Supply Chain Risk Management for Industry 4.0: A Data Mining Framework and Research Agenda. Systems 11[1]: 46. https://doi.org/10.3390/systems11010046

Eshete T, Demisse T, Yilma T and Tamir B [2023] Repeat Breeding and Its Associated Risk Factors in Crossbred Dairy Cattle in Northern Central Highlands of Ethiopia. Veterinary Medicine International. doi: 10.1155/2023/1176924

Fernandes S, Pereira G and Bexiga R [2023]. Bimodal milk flow and overmilking in dairy cattle: risk factors and consequences. Animal 100716. https://doi.org/10.1016/j.animal.2023.100716

Feyissa AA, Senbeta F, Tolera A and Guta DD [2023] Unlocking the potential of smallholder dairy farms: Evidence from the central highland of Ethiopia. Journal of Agriculture and Food Research 11: 100467. https://doi.org/10.1016/j.jafr.2022.100467

Gruber S, Rienesl L, Köck A, Egger-Danner C and Sölkner J [2023]. Importance of Mid-Infrared Spectra Regions for the Prediction of Mastitis and Ketosis in Dairy Cows. Animals 13[7]: 1193. https://doi.org/10.3390/ani13071193

Bai Q, Gao R, Li Q, Wang R, & Zhang H [2024]. Recognition of the behaviors of dairy cows by an improved YOLO. Intelligence & Robotics 4[1]: 1-19. 10.20517/ir.2024.01

Hernandez MC, Alvarez ANR and Anguiano FIS [2023] Project management and supply chain 4.0 improvement: the case of infant formulas in the face of the challenge of COVID-19. Procedia Computer Science 217: 278-285. https://doi.org/10.1016/j.procs.2022.12.223

Ida JA, Wilson WM, Nydam DV, Gerlach SC, Kastelic JP, Russell ER, McCubbin KD, Adams CL and Barkema HW [2023] Contextualized understandings of dairy farmers' perspectives on antimicrobial use and regulation in Alberta, Canada. Journal of Dairy Science 106[1]: 547-564. https://doi.org/10.3168/jds.2021-21521

Wang R, Gao R, Li Q, Zhao C, Ma W, Yu L, & Ding L [2023]. A lightweight cow mounting behavior recognition system based on improved YOLOv5s. Scientific reports 13[1]: 17418. https://doi.org/10.1038/s41598-023-40757-7

Weng Z, Bai R, & Zheng Z [2025]. SCS-YOLOv5s: A cattle detection and counting method for complex breeding environment. Journal of Intelligent & Fuzzy Systems 49[1]: 231-248. https://doi.org/10.3233/JIFS-237231

Luna M, Llorente I and Luna L [2023] A conceptual framework for risk management in aquaculture. Marine Policy 147: 105377. https://doi.org/10.1016/j.marpol.2022.105377

Resti Y, Reynoso GG, Probst L, Indriasari S, Mindara GP, Hakim A, & Wurzinger M [2024]. A review of on-farm recording tools for smallholder dairy farming in developing countries. Tropical Animal Health and Production 56(5): 168. doi: 10.1007/s11250-024-04024-9

Qiao Y, Guo Y, & He D [2023]. Cattle body detection based on YOLOv5-ASFF for precision livestock farming. Computers and Electronics in Agriculture 204: 107579. https://doi.org/10.1016/j.compag.2022.107579

Pearce SD, Parmley EJ, Winder CB, Sargeant JM, Prashad M, Ringelberg M, Felker M and Kelton DF [2023] Evaluating the efficacy of internal teat sealants at dry-off for the prevention of new intra-mammary infections during the dry-period or clinical mastitis during early lactation in dairy cows: A systematic review update and sequential meta-analysis. Preventive Veterinary Medicine 105841. https://doi.org/10.1016/j.prevetmed.2023.105841

Roussaki I, Doolin K, Skarmeta A, Routis G, Lopez-Morales JA, Claffey E, Mora M and Martinez JA [2023] Building an interoperable space for smart agriculture. Digital Communications and Networks 9[1]: 183-193. https://doi.org/10.1016/j.dcan.2022.02.004

Thangamayan S, Pradhan K, Loganathan GB, Sitender S, Sivamani S and Tesema M [2023] Blockchain-Based Secure Traceable Scheme for Food Supply Chain. Journal of Food Quality. DOI:10.1155/2023/4728840

Palma O, Plà-Aragonés LM, Mac Cawley A, & Albornoz VM [2025]. AI and data analytics in the dairy farms: a scoping review. Animals 15[9]: 1291. https://doi.org/10.3390/ani15091291

Tiwari S, Sharma P, Choi TM and Lim A [2023] Blockchain and third-party logistics for global supply chain operations: Stakeholders' perspectives and decision roadmap. Transportation Research Part E: Logistics and Transportation Review 170: 103012. https://doi.org/10.1016/j.tre.2022.103012

Tukamuhabwa BR [2023] Supply Chain Orientation and Supply Chain Risk Management Capabilities: Mechanisms for Supply Chain Performance of Agro-Food Processing Firms in Uganda. Journal of African Business 1-24. https://doi.org/10.1080/15228916.2023.2165894

Antognoli V, Presutti L, Bovo M, Torreggiani D, & Tassinari P [2025]. Computer Vision in Dairy Farm Management: A Literature

Review of Current Applications and Future Perspectives. Animals 15[17]: 2508. https://doi.org/10.3390/ani15172508

Wang L, Sun H, Gao H, Xia Y, Zan L and Zhao C [2023] A meta-analysis on the effects of probiotics on the performance of pre-weaning dairy calves. Journal of Animal Science and Biotechnology 14[1]: 3. https://doi.org/10.1186/s40104-022-00806-z

Yu JJ, Hu YL, Liu CZ, Wu SB, Zheng ZJ, Cui ZH, Chen L, Wei T, Sun SK, Ning J and Wen X [2023] ARSCP: An antimicrobial residue surveillance cloud platform for animal-derived foods. Science of The Total Environment 858: 159807. https://doi.org/10.1016/j.scitotenv.2022.159807