Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

Milk Production Estimation Model for Cattle Based on Image Processing using Random Forest, XGBoost, and LightGBM

Za'imatun Niswati[1,*], Sri Nurdiati[1], Agus Buono[1], Cece Sumantri[2]

[1] School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia
[2] Department of Animal Production Science and Animal Husbandry Technology, IPB University, Bogor, Indonesia
Email: [1,*]zaimatunzaimatun@apps.ipb.ac.id, [2]snurdiati@gmail.com, [3]agusbuono@apps.ipb.ac.id, [4]ceces@apps.ipb.ac.id
Correspondence Author Email: zaimatunzaimatun@apps.ipb.ac.id
Submitted: **14/06/2025**; Accepted: **02/09/2025**; Published**: 05/09/2025**

Abstract–Milk is a livestock product consumed by individuals of all ages. Therefore, it is essential to increase milk production in Indonesia to meet domestic demand. The growth of dairy cattle populations and milk production has not been able to keep up with rising consumption, resulting in a reliance on imports for most dairy products and their derivatives, with imports steadily increasing over the years. Therefore, alternative solutions are needed to enhance the milk production. One approach is to develop a milk production estimation model to determine the optimal number of dairy cattle to be cultivated by farmers and livestock companies to meet domestic demand. The objective of this study was to create a dairy milk production estimation model through image analysis using the Random Forest, XGBoost, and LightGBM algorithms. The milk production estimation model used in this study used CLAHE for contrast enhancement and VGG-16 for feature extraction. The results showed that XGBoost provided the best performance, explaining 74% of the data variation in the Y variable with a relatively small estimation error of 0.92. After parameter tuning using Grid Search, an improvement was observed, where XGBoost explained 86% of the data variation in the Y variable, and the estimation error decreased to 0.72. Image processing and machine learning technologies are part of precision agriculture that aims to improve the efficiency, productivity, and sustainability of livestock operations.

Keywords: Milk Production; Random Forest; XGBoost; LightGBM

1. INTRODUCTION

Dairy cow milk production is a key indicator of productivity in dairy farming [1]. Each farm strives to improve both the quality and quantity of milk produced while minimizing the factors that hinder the optimization of milk production [2]. Among the many factors influencing dairy cow maintenance, genetics is highly important. The genetics of cows are closely related to their morphology and physiology [3], [4]. The udder is the structure most associated with a cow's ability to produce milk. Therefore, accurate measurement and assessment of udder morphology and structure can serve as a method to enhance the potential of dairy cows [5].

Manual measurements, often conducted by farmers or technical personnel, frequently rely on their experience [1]. Experience-based measurements can lead to subjective assessments and varied results, which are particularly problematic at a larger scale. With advancements in computer technology and image processing, udder morphology measurements can now be performed automatically, accurately, and efficiently. Digital image acquisition and visual information can be obtained directly from cows using cameras [6].

Milk production estimation is crucial for helping farmers design livestock husbandry and nutrition strategies. Machine learning and image-based morphometric analysis can make milk yield estimation more objective and adaptive. Several estimation methods have been developed, such as EMA-ARIMA with 94.3% accuracy [7], CNN-SAE (84%) and CNN-MLP (77%) by [8] [9], and a deep learning approach with 85.2% accuracy [10]. Meanwhile, demonstrated that algorithms such as Random Forest (91.72%), K-Nearest Neighbors (91.65%), and Support Vector Regression (78.42%) demonstrated high performance in estimating milk production in Ireland [11].

In digital image processing techniques, all udder measurements, including area, length, width, and shape ratio, can be calculated quantitatively. The numerous images generated can be used to correlate visual information with the target variable, that is, milk production per cow. For more accurate estimations, Machine Learning algorithms are based on decision trees, such as Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) [12], [13].

Previous research used a single machine learning algorithm, such as Linear Regression, Support Vector Regression (SVR), or Random Forest, without conducting ensemble learning that combines the advantages of several algorithms at once such as Random Forest, XGBoost and LightGBM. Random Forest is an ensemble algorithm that involves two stages: building multiple decision trees and combining their results [14]. Random Forest is robust to overfitting and can handle data with numerous features. XGBoost and LightGBM are advancements in boosting techniques that can learn complex relationships in data with high efficiency [15], [16].

This study also promotes the application of artificial intelligence (AI) in livestock farming using digital image processing. The automated system analyzes photographs of cows and estimates milk yield to enhance livestock management efficiency and precision in decision-making regarding feed selection. Beyond its technical aspects, the use of this technology aligns with growing trends in smart farming and precision agriculture. The implementation of digital imaging systems and Machine Learning allows both small- and large-scale farmers to maximize output without relying on expensive manual observations and laboratory tests.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

The goal of this study was to develop a model for estimating dairy cow milk production based on udder size image analysis and to evaluate the performance of the Random Forest, XGBoost, and LightGBM algorithms. This study can contribute to the practical and scientific development of smart farming technologies in the future.

2. RESEARCH METHODOLOGY

This study aimed to develop a model for estimating dairy cow milk production based on udder size features using random forest (RF), XGBoost, and LightGBM.

2.1. Data

Data were collected at the Cibugary Livestock Farm and Mandiri Sejahtera Livestock Group. The research data consisted of 202 images of cows taken from the sides and backs. The label for the data was the milk production per cow after the images were captured. Additional information included the age and lactation phase of cows. Two types of data were used in this study.

a. Digital cow images from the side and rear: Photographs were captured using a digital camera with consistent positioning and lighting conditions. Each cow was photographed before milking.
b. Milk production data: Daily milk production per cow was measured as a target variable.

The research data used were side and rear images of dairy cows. The research data are presented in Figure 1.

Figure 1. Research Data

2.2. Research Stages

The stages carried out in this study are shown in Figure 2, which consists of data preprocessing and machine learning model construction.

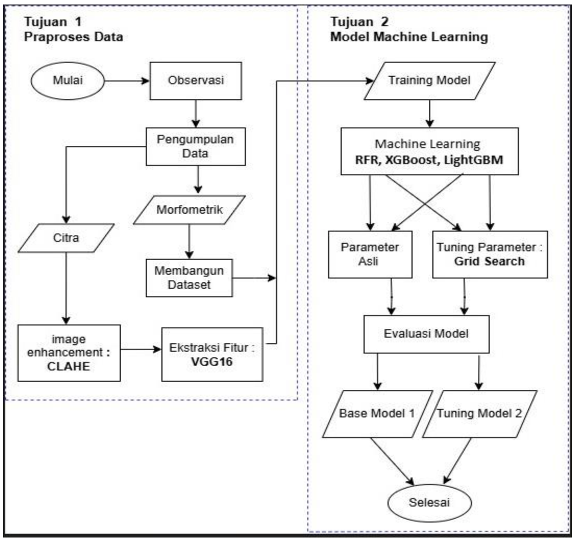

Figure 2. Research Stages

Preprocessing, including noise reduction, size normalization, and contrast enhancement, was performed to enhance the image quality and prepare it for further processing. Enhancement is used to improve the visual aspects of the image so that important information is more visible using the contrast-limited adaptive histogram equalization (CLAHE) technique [17]. CLAHE is effective for images with low contrast (such as nighttime or medical images). From the processed images, the features were then extracted by determining

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

a. Udder surface area
b. Udder length and width
c. Length-to-width ratio = 3:2

Feature extraction was performed using VGG16, which automatically learns feature representations from images. VGG16 was chosen for feature extraction because it is a powerful feature extractor that generates high-dimensional features. The advantages of VGG16 are as follows:

a. Its architecture is very consistent, using only 3 × 3 kernels with stride 1 and 2 × 2 max pooling.
b. It is easy to understand and implement in practice.
c. It can generalize visual features, such as contours, textures, and object shape patterns.
d. It is suitable for feature extraction in new domains, such as cow udder images with limited data [18].
e. It is effective for image classification and basic object detection tasks [14].
f. Transfer learning with VGG16 was highly effective for small datasets.

Feature Extraction Steps Using VGG16:

a. Pre-trained Model: VGG16, which was trained on the ImageNet dataset, was used as the backbone. This indicates that the network weights were well-tuned for general pattern recognition in images.
b. Removal of fully connected layers: In feature extraction, only the convolutional layers are retained, as these layers are responsible for extracting features.
c. Image Input: Images with a resolution that matches the VGG16 specifications (typically 224 × 224 pixels) are provided as input to the network.
d. Extraction from the Last Convolutional Layer: Features are extracted from the output of the last convolutional layer, usually from the layer just before the fully connected layers. These features are tensors that represent various aspects of an image.
e. Flattening or Pooling (optional): After the features are extracted, these tensors can be flattened or subjected to pooling (average or max-pooling) to reduce dimensions and facilitate the use of features in other machine learning models.
f. Use of Features: The extracted features can be used as inputs for simpler machine learning models, such as Random Forests, or even in more complex deep learning models.

Because there is no need to train the model from scratch, feature extraction with VGG16 is very efficient [20]. The VGG16 feature extraction architecture is presented in figure 3.

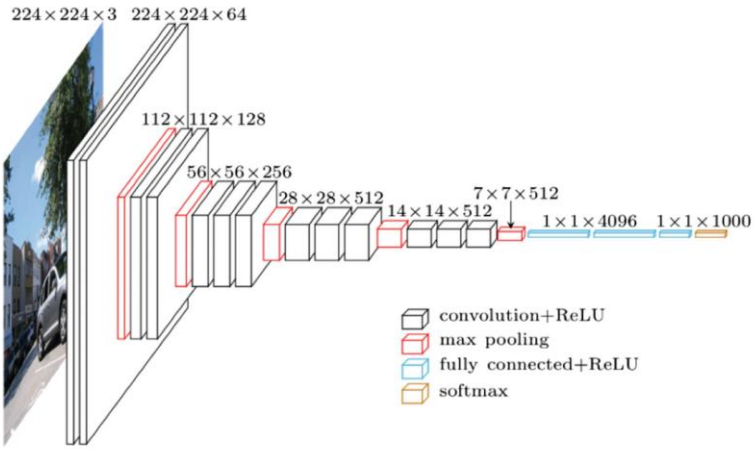

Figure 3. VGG16 Architecture

2.3. Algoritma Machine Learning

Random Forest (RF) is a machine learning algorithm that combines the outputs of multiple decision trees to achieve a single result. A 'forest' is formed from many trees, which are obtained through the process of bagging or bootstrap aggregating.

Each tree in the RF outputs a class prediction. The class prediction with the most votes was the candidate prediction in the model. The more trees there are, the higher the accuracy, which helps prevent overfitting.

The random Forest operates in two phases. The first phase involves combining several N decision trees to create a Random Forest. The second phase involves making predictions for each tree created in the first phase.

The RF model can be used for classification and regression applications. When RF is used for regression, the individual decision tree predictions are averaged to obtain the final estimate. Individual decision trees tend to overfit the training data, but the Random Forest model works by combining the predictions of multiple individual decision trees, which helps reduce the overfitting problem [21].

The Random Forest Regressor is an ensemble algorithm that uses multiple decision trees to improve the accuracy of regression predictions, as shown in Figure 4.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

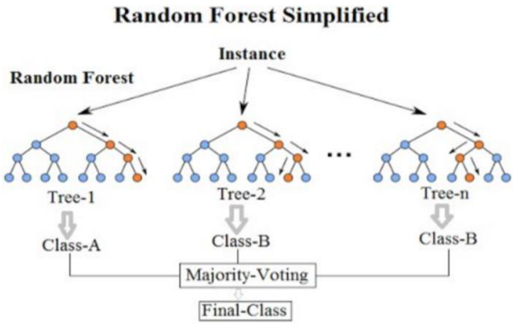

Figure 4. *Random Forest Regressor*

Random Forest is a bagging-based ensemble learning algorithm that combines multiple decision trees to improve accuracy and reduce overfitting.

How it works:

a. Randomly sample the training data (bootstrap sampling).
b. For each tree, a random subset of features is selected at each split.
c. Each tree is trained independently.
d. The predictions of all trees are combined: the average predicted value.

The Random Forest Regressor predicts the target value (dependent variable) y based on the features X in a dataset with N samples and M features. Random Forest computes the average prediction from each decision tree using the equation (1)

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(X) \tag{1}$$

Extreme Gradient Boosting (XGBoost) is a highly efficient and powerful boosting algorithm with many features for handling missing data, parallelization and pruning. It is suitable for complex datasets and generates accurate prediction models.

Gradient boosting is an ensemble technique that builds models sequentially, with each new model correcting the errors of the previous model using the gradient of the loss function [14].

XGBoost is a boosting algorithm that builds a model incrementally, with each new tree attempting to improve on the previous tree's error by minimizing a loss function using gradient descent.

How it works:

a. Initialize the initial model (e.g., the average value for regression).
b. Calculate the residual (error) between the prediction and the target.
c. Train the decision tree to predict the residual.
d. Add new trees to the model with specified weights.
e. Repeat until the number of iterations or stopping criterion is met.

XGBoost also uses regularization techniques to prevent overfitting, a condition in which the model focuses excessively on the training data, leading to poor performance on the test data. The regularization techniques used by XGBoost are L1 and L2 regularization, where each feature is assigned a different weight depending on its significance to the target variable using the equation (2)

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{2}$$

With $\hat{y}_i^{(t)}$ the prediction at iteration t, f_t the new tree built at iteration t and η the learning rate

The Light Gradient Boosting Machine (LightGBM) uses a histogram-based decision tree learning technique that reduces memory usage and improves training speed. LightGBM can effectively capture interactions between udder size features, resulting in accurate models with high generalization [22].

LightGBM can handle large datasets with shorter computation times than other gradient boosting algorithms. It uses histogram compression techniques, allowing for more efficient memory usage and natively supports categorical features without requiring preprocessing or encoding. This algorithm works well on datasets with many features and can handle issues related to high-dimensional data [23], [24]. LightGBM is an extension of gradient boosting that prioritizes speed and memory efficiency.

How it works:

a. Initialize the initial model by predicting the mean value (regression).
b. Calculate the gradient of the loss function against the previous model's prediction (retaining data with large gradients for information gain calculations and combining mutually exclusive features to reduce dimensionality).
c. Train the decision tree to predict with the greatest loss reduction.
d. New trees are added to the model with learning rate weights to control the contribution of improvement.
e. The process continues until the maximum number of iterations is reached.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

Evaluation was used to determine the error rate in milk production estimation. The predicted results were then compared with the actual data to obtain accuracy parameters such as the Coefficient of Determination (R^2 score) and Mean Absolute Error (MAE). R^2 represents the proportion of variance in the dependent variable explained by the independent variables and indicates the goodness-of-fit of the model. MAE is a measure used to evaluate regression models [8].

3. RESULT AND DISCUSSION

Dairy cow milk production is a key indicator of success in dairy farming and is influenced by physiological conditions such as udder size and health [25]. Various studies have shown a significant relationship between udder size and milk production. Udder dimensions, such as length, width, and height, are often used as potential indicators to estimate a cow's milk capacity [26].

Digital image processing can extract visual features from udder images, allowing udder dimensions to be measured with higher precision than manual methods [2]. These features were then used as inputs to build a machine learning-based model for estimating milk production.

Table 1 shows the evaluation results of the milk production estimation model using digital images and machine learning. The Coefficient of Determination (R^2 score) and Mean Absolute Error (MAE) were used to measure the accuracy of the estimation model.

The results of model evaluation using the Random Forest, XGBoost, and LightGBM algorithms are presented in Table 1. The scenarios used were: evaluating the machine learning model with the original parameters to produce a base model, and evaluating the machine learning model with Grid Search parameter tuning to produce a tuned model.

Random Forest is stable and robust to noise, relatively easy to tune, and less sensitive to learning rate. It has a higher bias than boosting and is less efficient when the dataset is very large. XGBoost, an optimized implementation of GDBT, often provides high accuracy but requires tuning. It controls overfitting better and handles sparse/missing values integrally. XGBoost with parameter tuning produces the highest R2 value and the lowest MAE value. LightGBM is very fast and memory-efficient without significant loss of accuracy. Leaf-wise growth risks overfitting on small/large-depth datasets if left unchecked.

Table 1. Evaluation Model MAE and R2 on Algorithm

	Base Model		Tuning Model	
	MAE	R^2	MAE	R^2
Random Forest Regressor	0.96	0.75	0.98	0.74
XGBoost Regressor	0.92	0.74	0.72	0.86
LightGBM Regressor	1.03	0.71	1.02	0.71

In the base model, XGBoost explained 74% of the variation in the data for Y, and the model's estimation error was relatively small at 0.92. After tuning with Grid Search, an improvement was observed, where XGBoost explained 86% of the variation in the data for Y, and the model's estimation error decreased to 0.72. Random Forest and LightGBM did not show significant changes after model tuning.

XGBoost is a highly efficient and powerful boosting algorithm that focuses on correcting previous estimation errors. It can handle missing values and prevent overfitting, thereby improving the model performance.

XGBoost has been validated in many fields and has been shown to perform better than RF [27], [28], [29]. XGBoost is currently the first-choice algorithm for many practitioners and data science competitions [21].

Random Forest is very effective in handling data with complex, nonlinear features and does not require many tuning parameters to achieve good performance. LightGBM is a variation of boosting that is extremely fast and memory-efficient. It is highly efficient for large datasets with diverse features. The model was sufficient to explain most of the variation in milk production from udder images alone (R^2 = 0.86). Scater plot of Model Comparison – R^2 is presented in Figure 5.

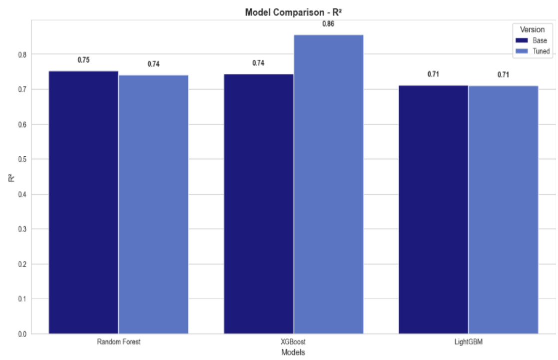

Figure 5. Scater plot of Model Comparison – R^2

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

In this study, observations were also made on the image data used as features influencing the estimation. This image data plays a crucial role in determining the results obtained by the developed estimation model. This process involves analyzing various elements contained in the image to evaluate their impact on estimation accuracy. An illustration of this is presented in Figure 6.

Figure 6. Scatter plot of udder size vs. milk volume

The scatterplot shows that udder width from the rear is the strongest indicator for estimating milk production, followed by udder height from the rear and udder length from the side. However, there is a distribution in the data indicating that other factors influence milk production. However, data dispersion suggests that other factors, including genetics, feed, and age, affect milk production. Image processing technology and machine learning are part of precision agriculture, which aims to improve the efficiency, productivity, and sustainability of livestock farming [13].

Although the current study demonstrates promising results, it has several limitations that need to be addressed in future research. One of the primary limitations of this study is the size and diversity of the dataset used. A broader and more varied dataset would allow for a more generalizable model, as the current dataset may not fully capture the full range of variables that could affect the milk production. Expanding the dataset to include data from different regions, climates, and cow management practices would help to mitigate this limitation. In addition, the current study focused on a specific set of parameters and techniques. Future studies could benefit from exploring additional machine learning algorithms or incorporating ensemble methods to further enhance the model performance. By addressing these limitations and incorporating these suggestions, future studies can significantly improve the robustness, accuracy, and applicability of the model, providing valuable insights for the livestock farming industry.

4. CONCLUSION

Based on the R^2 Score and Mean Absolute Error (MAE) values, XGBoost gets the best results, explaining 74% of the data variation in variable Y, with a relatively small estimation error of 0.92. After performing parameter tuning using Grid Search, an improvement was observed, where XGBoost was able to explain 86% of the data variation in variable Y, and the model's estimation error decreased to 0.72. Image processing technology and machine learning are integral to precision agriculture, aimed at improving the efficiency, productivity, and sustainability of livestock farming. Future research should focus on exploring more complex combinations of parameter tuning and automated optimization techniques, such as Grid Search or Bayesian Optimization. These advanced methods can help refine the model's performance by adjusting the hyperparameters more effectively, leading to better predictions and reduced errors. Additionally, to improve the generalization capability of the model, it is recommended to use larger and more diverse datasets. The inclusion of data from different lactation seasons, cow breeds, and geographical variations would provide a more comprehensive understanding of the factors influencing milk production. Such diversity in the dataset would allow the model to learn more robust patterns and improve its ability to make accurate predictions in different environments.

REFERENCES

[1] N. A. Poulsen and L. Bach Larsen, "Genetic factors affecting the composition and quality of cow's milk," *Burleigh Dodds Ser. Agric. Sci.*, pp. 501–532, 2022, doi: 10.19103/as.2022.0099.15.
[2] L. Liu *et al.*, "Ten Genetic Loci Identified for Milk Yield, Fat, and Protein in Holstein Cattle," *Animals*, vol. 10, no. 11, pp. 1–15, 2020, doi: 10.3390/ani10112048.
[3] J. Moran, *Nutrient requirements of dairy cattle*. The National Academies Press, 2020.
[4] Dairy Training Centre, *Dairy Cattle Feeding and Nutrition Management.*, 2017.
[5] S. Kaskous, "Optimization of Milk Performance and Quality in Dairy Farms by using a Quarter individual Milking System 'MultiLactor,'" *Int. J. Environ. Agric. Biotechnol.*, vol. 5, no. 4, pp. 943–952, 2020, doi: 10.22161/ijeab.54.14.
[6] M. R. Gianegitz *et al.*, "Contributions of genetic improvement programs for dairy livestock farming," *Biol. Agric. Sci. Theory Pract.*, 2024, doi: 10.56238/sevened2024.008-004.
[7] R. da R. Righi, G. Goldschmidt, R. Kunst, C. Deon, and C. A. da Costa, "Towards combining data prediction and internet of

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1302−1308
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.7585

things to manage milk production on dairy cows," *Comput. Electron. Agric.*, vol. 169, no. December 2019, p. 105156, 2020, doi: 10.1016/j.compag.2019.105156.

[8] A. Liseune, M. Salamone, D. Van den Poel, B. Van Ranst, and M. Hostens, "Leveraging latent representations for milk yield prediction and interpolation using deep learning," *Comput. Electron. Agric.*, vol. 175, no. June, p. 105600, 2020, doi: 10.1016/j.compag.2020.105600.

[9] A. Liseune, M. Salamone, D. Van den Poel, B. van Ranst, and M. Hostens, "Predicting the milk yield curve of dairy cows in the subsequent lactation period using deep learning," *Comput. Electron. Agric.*, vol. 180, no. September 2020, p. 105904, 2021, doi: 10.1016/j.compag.2020.105904.

[10] K. S. Themistokleous, N. Sakellariou, and E. Kiossis, "A deep learning algorithm predicts milk yield and production stage of dairy cows utilizing ultrasound echotexture analysis of the mammary gland," *Comput. Electron. Agric.*, vol. 198, no. September 2021, p. 106992, 2022, doi: 10.1016/j.compag.2022.106992.

[11] C. O'Leary and C. Lynch, "An Evaluation of Machine Learning Approaches for Milk Volume Prediction in Ireland," *2022 33rd Irish Signals Syst. Conf. ISSC 2022*, pp. 0–7, 2022, doi: 10.1109/ISSC55427.2022.9826160.

[12] M. Seyyedattar, S. Zendehboudi, A. Ghamartale, and M. Afshar, "Advancing hydrogen storage predictions in metal-organic frameworks: A comparative study of LightGBM and random forest models with data enhancement," *Int. J. Hydrogen Energy*, vol. 69, no. March, pp. 158–172, 2024, doi: 10.1016/j.ijhydene.2024.04.230.

[13] S. Wolfert, L. Ge, C. Verdouw, and M. J. Bogaardt, "Big Data in Smart Farming – A review," *Agric. Syst.*, vol. 153, pp. 69–80, 2017, doi: 10.1016/j.agsy.2017.01.023.

[14] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[16] T. Y. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree.," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[17] M. R. Islam and M. Nahiduzzaman, "Complex features extraction with deep learning model for the detection of COVID19 from CT scan images using ensemble based machine learning approach," *Expert Syst. Appl.*, vol. 195, no. February, p. 116554, 2022, doi: 10.1016/j.eswa.2022.116554.

[18] A. Svyatkovskiy, J. Kates-Harbeck, and W. Tang, "Training distributed deep recurrent neural networks with mixed precision on GPU clusters.," Proceedings of the Machine Learning on HPC Environments , 2017, doi: 10.1145/3146347.3146358.

[19] T. K. G. Taranjit Kaur, "Automated Brain Image Classification Based on VGG-16 and Transfer Learning," International Conference on Information Technology (ICIT), 2019, doi: 10.1109/ICIT48102.2019.00023.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.

[21] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.

[22] Z. Zhou, L. Zhang, and S. Cai, "Application of Ensemble Learning in Agricultural Yield Prediction: A Review.," *Agric. Syst.*, vol. 178, 2020.

[23] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, "Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 2, pp. 228–236, 2022, doi: 10.36456/jstat.vol15.no2.a5548.

[24] A. Rafi, A. Wahyu, E. P. A.R, S. Fadilah, and A. M. Rizki, "Perbandingan Algoritma Lightgbm Dan Ann Untuk Menentukan Kualitas Anggur Merah," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 9, no. 1, pp. 1572–1579, 2025.

[25] P. Jain, S. Singh, and A. Kumar, "Relationship between udder traits and milk production in dairy animals: A review.," *Vet. World*, vol. 12, no. 9, pp. 1452–1457, 2019.

[26] M. R. Machado, S. Karray, and I. T. De Sousa, "LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry," *14th Int. Conf. Comput. Sci. Educ. ICCSE 2019*, no. Nips, pp. 1111–1116, 2019, doi: 10.1109/ICCSE.2019.8845529.

[27] M. S. Oughali and S. A. E.-R. Bahloul, Mariah, "Analysis of NBA players and shot prediction using random forest and XGBoost models," in *International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1–5, doi: • 10.1109/ICCISci.2019.8716412.

[28] A. X. V. I. Simp and S. Remoto, "PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data Mehdi," Ambient Aerosol Measurements in Different Environments , vol. 10, no. 7, 1992, pp. 6425–6432, 2013.

[29] W. Zhang, C. Wu, H. Zhong, Y. Li, and L. Wang, "Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization," *Geosci. Front.*, vol. 12, no. 1, pp. 469–477, 2021, doi: 10.1016/j.gsf.2020.03.007.