Frontiers in Health Informatics
ISSN-Online: 2676-7104
www.healthinformaticsjournal.com
2024; Vol 13: Issue 3
Open Access

Comparative Analysis of Machine Learning Techniques for Predicting Dairy Cow Health Conditions

Devinder Kaur[1a, b], Amandeep Kaur[2]
[1a]Research Scholar, Department of Computer Science, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India.
[1b]Assistant Professor, Department of Computer Science, Mata Gujri College, Fatehgarh Sahib, Punjab, India
devinderkaurcomp@matagujricollege.org
[2]Assistant Professor, Department of Computer Science, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India
amandeep_virk@sggswu.edu.in

Cite this paper as: Devinder Kaur, Amandeep Kaur (2024) Comparative Analysis of Machine Learning Techniques for Predicting Dairy Cow Health Conditions. *Frontiers in Health Informatics, 13* (3), 2109-2116

Abstract:

Cattle are susceptible to several transmissible diseases that could endanger human health due of our reliance on dairy products. It is imperative to safeguard the welfare of dairy cows to stop these illnesses from spreading. This study compares the effectiveness of various machine learning techniques for identifying disease in dairy cows. A specially designed sensor-based Internet of Things (IoT) gadget has been developed to measure the key health indicators of cows. This innovative tool has been used to gather a comprehensive dataset from 150 cows spread over seven districts in Punjab. Principal component analysis, or PCA, is used to reduce the number of dimensions in the dataset. A variety of techniques, such as Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbour (KNN), have been employed to train and assess the models. The models' effectiveness is evaluated by performance indicators that include recall, accuracy, precision, and F1 score. With an accuracy of 99.34%, precision of 97.93%, recall of 97.13%, and F1 score of 97.03%, Random Forest exhibited the best overall results. The results underscore the potential of machine learning, specifically Random Forest, to serve as a reliable tool for early disease detection in dairy cows. This might ultimately enhance animal health and protect public health by limiting the supply of dairy products.

Keywords: *Internet of Things (IoT), Disease Detection, Machine Learning, Principal Component Analysis (PCA).*

Introduction

The primary sources of livelihood in rural India are agriculture and practice of dairy farming. As the population is expanding, food and dairy production is critically needed in the future. Majority of India's historical livestock farming regions are still following traditional practices, with farmers making decisions based only on their own personal experience.

Forecasting the health and illness of cattle in advance is essential, particularly considering the possibility of animal-to-human disease transfer (Chaudhry et al. 2020). However, it can be challenging to ensure a high quality of life for every individual cow because of fragmented approach to various aspects of dairy farming like feeding, milking, reproduction, and health are frequently managed as distinct and separate processes.

Presently, dairy producers are being guided towards the field of precision livestock farming via precision dairy technology (Kaur et al. 2024). Dairy farmers now have a unique opportunity like sensor-based devices, big data, and application of machine learning for health monitoring of cattle (Kaur et al. 2022). These technologies allow for continuous monitoring of vital signs of animal health, such as movement, temperature, pulse rate, and humidity, rather than responding to illnesses only when they manifest. Information and communication technology (ICT), control systems, Data analysis and machine learning (ML) are used in the collection and analysis of data pertaining to cattle health parameters (Banhazi et al. 2012).

Access to current, quantitative data is essential for making well-informed decisions about livestock management. IoT has made a great contribution in this context (Kaur et al. 2024). But in India, only large dairy owners are using these types of devices because they tend to be very expensive. There was a pressing need to develop a cost-effective IoT solution tailored for mid-size to small rural farmers. In this study, an innovative device based on Internet of Things has been created to monitor the health status of cows. This solution will help to reduce production costs, enabling fewer farmers to manage more animals by providing early alerts regarding diseases.

With such advancements, farmers will be able to proactively identify, anticipate, and prevent cattle diseases before they spread to other areas and cause major losses. The continuous collection of data through sensors, combined with the predictive capabilities of machine learning, can make dairy farming more productive.

Material and Methods

In this study, an innovative Internet of Things (IoT)-enabled neck collar device was deployed to continuously monitor the health of dairy cows. This device integrates sensors that measure key physiological metrics such as body temperature, activity levels and pulse rate. Data captured by the device is wirelessly transmitted via Bluetooth to a central database for further analysis.

Figure 1 offers an in-depth review of the methodology employed in predicting the health status of dairy cows by merging machine learning classifiers with Internet of Things (IoT) technologies. The process starts with the deployment of IoT based neck collar on the dairy cows, which continuously monitor and collect various health-related data including temperature, pulse rate, and activity levels.

The collected data is subsequently sent to a central data repository, where it is pre-processed to ensure accuracy and consistency. This pre-processing stage involves cleaning the data and addressing any missing values, and normalizing the input features to prepare it for analysis. Feature extraction techniques were employed, followed by dimensionality reduction using Principal Component Analysis (PCA). Cows were classified as either "sick" or "healthy" based on predefined thresholds for each physiological parameter.

Following pre-processing, the data is fed into machine learning classifiers that have been trained to recognize patterns and anomalies indicative of health issues. Using logistic regression, decision trees, neural networks, and support vector machines, the data is analysed to predict each cow's health state.

The predictions generated by the machine learning models are then evaluated and interpreted to provide actionable insights that allows farmers to adopt proactive strategies, such as adjusting feed, administering medication, or changing management practices to maintain optimal health and productivity among their dairy herd.

Figure 1. Steps for Predicting Health Status of Dairy Cows Using IoT and Machine Learning Classifiers

Data Acquisition

The research study's dataset was acquired using a specially designed IoT-based device equipped with sensors to monitor body temperature, pulse rate, and activity through an accelerometer (depicted in Figure 2). The dataset comprises data from 150 dairy cows housed in registered gaushalas and dairy farms across seven districts of Punjab, including Ludhiana, Patiala, Ropar, Sangrur, Nawan Shahr, Fatehgarh Sahib, and Hoshiarpur. Data collection was conducted over a two-month period for each cow, with daily averages calculated for each parameter.

Figure 2. Smart Neck Collar equipped with sensors

Feature Extraction

Movement intensity was derived from the accelerometer data, determined as the absolute vector sum of the accelerations along the x, y, and z axes, applying the subsequent formula:

Total Acceleration= $\sqrt{x^2 + y^2 + z^2}$

Further calculations of the average and standard deviation of total acceleration were used for the classification of activity levels into mild, moderate, and intense categories. The thresholds for these activity levels were determined using the following criteria:

- Mild Activity: Average Acceleration + 0.5 * Standard Deviation
- Moderate Activity: Average Acceleration + 1 * Standard Deviation
- Intense Activity: Average Acceleration + 1.5 * Standard Deviation

Activity levels were then assigned numerical values of 1, 2, and 3 for mild, moderate, and intense activities, respectively.

The neck collar sent information via Bluetooth at an interval of fifteen minutes each. Every cow had it on for ten hours every day. This yielded approximately 40 data rows per cow per day on average. To generate the dataset for each cow, data was gathered over a two-month period.

Classification

A popular data analysis method, Principal component analysis, is a technique for drawing features and reducing dimensionality. It is extensively employed in numerous fields such as statistics, machine learning, and signal processing (Jolliffe et al, 2016). PCA was applied for decreasing the dimensionality of the collected data to the daily averaged measurements of temperature, pulse rate, and activity levels.

Normal body temperature for cattle typically ranges between 38.5°C and 39.5°C; deviations from this range can signal health issues, such as temperatures below 38.5°C indicating conditions like milk fever, poisoning, or indigestion, and temperatures above 41°C suggesting diseases such as anthrax, influenza, or foot-and-mouth disease (Kumari et al. 2018). Additionally, a healthy cow's resting pulse rate is generally between 40 and 80 beats per minute (BPM) (Chaudhry et al. 2020).

Cows were classified as "sick" if their temperature fell outside the 38.5°C to 39.5°C range, if their pulse rate was outside the 40-80 BPM range, and if their activity level was categorized as mild. Otherwise, cows were classified as "healthy."

Application of Machine Learning Classifiers

After processing, the dataset was utilized as input for several machine learning classifiers to predict the cows' health status. The classifiers used in this study include:

1. Support Vector Machine (SVM): This algorithm works best in high-dimensional spaces, especially when using kernel functions for non-linear classification tasks. Its goal is to find the best hyperplane within the dataset that differentiates between various classes (Schölkopf et al. 2002).
2. Random Forest: An ensemble learning method that, to improve accuracy and reduce overfitting, builds several decision trees during training and combines their predictions (Hastie et al. 2009).
3. Logistic Regression: Logistic regression is a mathematical model used in binary classification issues that provide coefficients indicating the link between predictors and outcomes (Khan et al. 2018). It calculates the likelihood of a binary result based on one or more predictor variables (Peng et al. 2002).
4. K-Nearest Neighbours (K-NN): Suitable for a wide range of classification applications, the K-NN algorithm is a simple and effective method. Data points are categorised based on the majority class of their k-nearest neighbours (Hastie et al. 2009).

To assess and contrast the classifiers' predictive capabilities, the following performance metrics were employed:

1. **Accuracy**: Symbolises the percentage of cases that were successfully classified out of all the instances. According to (Han et al. 2012), accuracy on its own might not be adequate in situations when datasets are imbalanced, even though it offers a general idea of classifier performance.
2. **Precision:** This measure determines the ratio of accurate positive forecasts to all positive forecasts, providing information about accuracy of positive predictions (Lawrence et al. 2007).

3. **Recall**: Often called "sensitivity", recall quantifies the percentage of true positives that are appropriately recognised, indicating the classifier's capacity to find all pertinent occurrences.
4. **F1 Score**: Specifically useful for assessing classifiers on unbalanced datasets. The harmonic mean of recall and precision is known as the F1 Score. It offers a fair measure that considers both false positives and false negatives.

Results and Discussion

The machine learning system's performance classifiers applied to estimate the state of health of dairy cows is summarized in Table 1, with a visual representation provided in Figure 3. The evaluation of predictive models' efficacy in binary classification tasks is contingent upon several essential measures, including accuracy, precision, recall, and F1 score.

The outcomes obtained by applying machine learning classifiers are presented in Table 1, and These outcomes are shown graphically in Figure 3.

Classifier	Accuracy	Precision	Recall	F1 Score
Random Forest	99.34%	97.93%	97.13%	97.03%
Logistic Regression	96.36%	85.04%	88.52%	86.75%
Support Vector Machine	92.77%	75.34%	68.85%	71.95%
K-Nearest Neighbors	97.68%	90.4%	92.62%	91.49%

Table 1. Performance Metrics of Machine Learning Classifiers in Predicting Cow Health Status

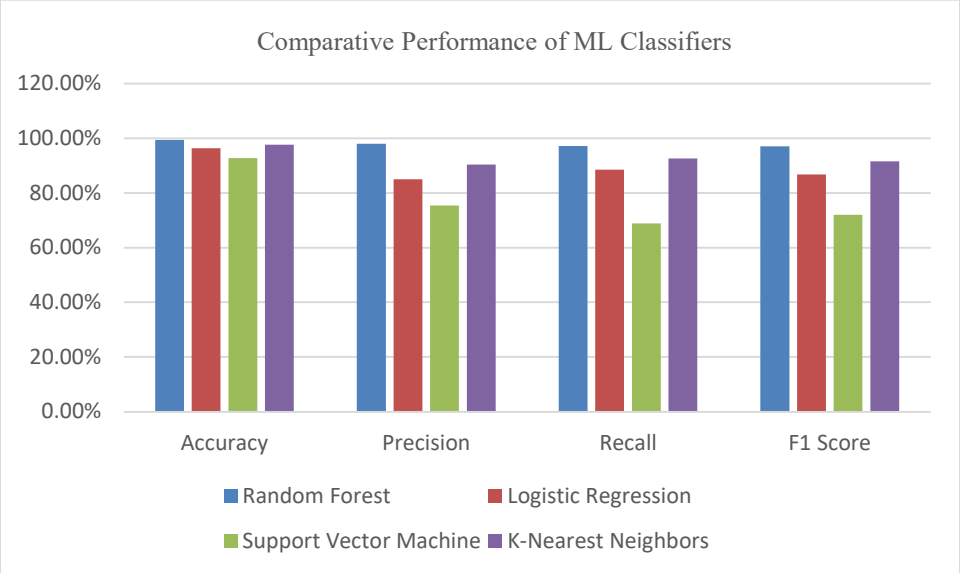

Figure 3. Visual Depiction of comparative performance of ML Classifiers

Availability of large datasets enables utilising machine learning (ML) approaches to develop classification models. These techniques typically involve dividing the dataset into training and testing subsets. In this study, an 80:20 split ratio was applied. However, when a dataset includes repeated measurements for the same animal, randomly dividing records into training and testing sets can lead to data leakage, as the testing set may inadvertently contain information already present in the training set. To ensure that the ML model accurately predicts new, previously unseen data, alternative methods for selecting training and testing data are necessary. This study adopted the approach suggested by (Bobbo et al. 2023), which involves dividing records based on cow ID, ensuring that data from individual animals is only included in one of the two sets. This method enhances accuracy of the model's predictions on the testing set.

Every metric showed that the best performing classifier was the Random Forest by achieving an accuracy of 99.34%, a precision of 97.93%, a recall of 97.13%, and an F1 score of 97.03%. These results imply that Random Forest is highly effective in accurately classifying the health status of cows, with minimal misclassification. Its high precision and recall demonstrate its ability to correctly find genuine positives while reducing false negatives and positives, making it the most reliable model among those tested.

K-Nearest Neighbors (K-NN) also demonstrated a high level of performance, achieving 91.49% F1 score, 92.62% recall, 90.40% precision, and 97.68% accuracy. While slightly less effective than Random Forest, K-NN's metrics still reflect its robustness in classifying cow health, particularly in balancing precision and recall.

Logistic Regression provided a decent but not exceptional performance, demonstrating an accuracy of 96.36%, a precision of 85.04%, a recall of 88.52%, and an F1 score of 86.75%. Although reasonably effective, Logistic Regression in comparison to Random Forest and K-NN, generated more false positives and false negatives, suggesting that it may not be as reliable for this specific classification task.

Out of all the classifiers examined, the Support Vector Machine (SVM) classifier performed worst, with 92.77% accuracy, 75.34% precision, 68.85% recall, and 71.95% F1 score. The lower metrics suggest that SVM struggled with the classification task, potentially due to the complexity of the dataset or suboptimal parameter selection. The significant difference in performance between SVM and the other classifiers indicates that it may not be the best choice for this application.

Based on the examination of various machine learning classifiers, the Random Forest algorithm qualified to be the best model for predicting dairy cow health. The resilience and reliability of the system is demonstrated by its exceptional performance across all metrics such as F1 score, recall, accuracy, and precision. Its high accuracy and good generalisation to new data, with a low number of false negatives and false positives, are probably due to ensemble nature of Random Forest, which combines numerous decision trees. While K-Nearest Neighbors classifier, proved slightly less effective than Random Forest, also performed admirably, particularly in recall, which indicates its strength in identifying actual positive cases. This makes K-NN a viable alternative when high recall is prioritized.

Logistic Regression, although a commonly used classifier, exhibited limitations in this study, particularly in precision and recall. This suggests that while it can be useful for binary classification tasks, in circumstances where the cost of false positives or false negatives is substantial, it might not be the ideal choice, as is often the case in health monitoring applications whereas the relatively poor performance of the Support Vector Machine highlights the importance of selecting appropriate algorithms and tuning parameters for specific datasets. The SVM's lower metrics across the board suggest that it was not well-suited to the characteristics of this dataset, potentially due to its high-dimensional nature or the choice of kernel functions.

This study shows how well machine learning classifiers work to forecast dairy cow health status using information gathered from an Internet of Things (IoT)-based neck collar. The Random Forest method attained the highest F1 score, recall, accuracy, and precision, among the tested classifiers, making it the most dependable. This demonstrates its reliability in correctly recognising cows that are healthy and those that are ill, which makes

it an important tool for real-time health monitoring in dairy production.

The K-Nearest Neighbors classifier also showed strong performance, indicating its potential as an alternative to Random Forest, particularly in scenarios where striking a balance between recall and precision is essential. Logistic Regression, while reasonably effective, was less reliable due to its increased false-positive and false-negative rate. The Support Vector Machine classifier, with the lowest performance, suggests that its application may be limited in this context unless further optimization is undertaken. The superior performance of ensemble methods like Random Forest emphasizes the capacity of these methods to enhance the accuracy and reliability of health diagnostics in livestock.

These findings support the use of ensemble methods in veterinary diagnostics, particularly in scenarios where high accuracy and reliability are critical. Future research should explore the integration of additional physiological and environmental factors to further refine predictive models and improve early disease detection in dairy cows.

References

- Banhazi, T. M., Lehr, H., Black, J. L., Crabtree, H., Schofield, P., Tscharke, M., &
- Berckmans, D. 2012. Precision livestock farming: an international review of scientific and commercial aspects. International Journal of Agricultural and Biological Engineering 5(3): 1-9.
- Bobbo T, Matera R, Pedota G, Manunza A, Cotticelli A, Neglia G, and Biffani S. 2023. Exploiting machine learning methods with monthly routine milk recording data and climatic information to predict subclinical mastitis in Italian Mediterranean buffaloes. Journal of Dairy Science 106: 1942–1952.
- Chaudhry, A. A., Mumtaz, R., Zaidi, S. M. H., Tahir, M. A., & School, S. H. M. 2020. Internet of Things (IoT) and machine learning (ML) enabled livestock monitoring. In Proceedings of 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), pp. 151-155. IEEE.
- Han, J., Kamber, M., & Pei, J. 2012. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
- Hastie, T., Tibshirani, R., & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- Jolliffe I T and Cadima J. 2016. Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2065): 20150202.
- Kaur D and Kaur A. 2022. IoT and machine learning-based systems for predicting cattle health status for precision livestock farming. In: Proceedings of 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), pp. 1-5. IEEE.
- Kaur D and Virk A K. 2024. Leveraging IoT for precision health monitoring in livestock with artificial intelligence. In: Data-Driven Farming, Auerbach Publications. pp. 1-18. ISBN: 978-1-032-61892-0.
- Khan B H, Corbeil J R and Corbeil M E (Eds). 2018. Responsible analytics and data mining in education: Global perspectives on quality, support, and decision making. Routledge
- Kumari, S., & Yadav, S. K. 2018. Development of IoT based smart animal health monitoring system using Raspberry Pi. International Journal of Advanced Studies of Scientific Research 3(8).
- Lawrence K. D., Kudyba S., and Klimberg R. K. (Eds.). 2007. Data Mining Methods and Applications. CRC Press, 295 pages.

- Neethirajan, S. 2020. The role of sensors, big data, and machine learning in modern animal farming. Sensing and Bio-Sensing Research 100367.
- Peng, C. J., Lee, K. L., & Ingersoll, G. M. 2002. An introduction to logistic regression analysis and reporting. The Journal of Educational Research 96(1): 3-14.
- Schölkopf, B., & Smola, A. J. 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press.