

A Hybrid Anomaly Detection and Machine Learning Framework for Early Cattle Health Risk Prediction

Mohammed Jehad Yaghi¹, Abdelrafe Elzamly², Samy S. Abu-Naser³

¹Faculty of Computer and Information Technology, Al-Aqsa University, Gaza, Palestine

²Faculty of Computer and Information Technology, Al-Aqsa University, Gaza, Palestine

³ Faculty of Engineering and Information Technology, Al-Azhar University, Gaza, Palestine

Abstract: Early detection of cattle health problems is an important component of precision livestock farming because physiological and behavioral abnormalities may emerge before visible clinical symptoms. Conventional livestock health monitoring, which relies heavily on periodic observation, may be inadequate for continuous monitoring, particularly in resource-constrained environments. This study proposes a hybrid machine-learning framework, termed the Gaza Cattle Health Prediction (GCHP) framework, for early cattle health-risk assessment using multidimensional cattle health and farm-related data. The proposed approach integrates unsupervised anomaly detection using Isolation Forest with supervised classification using Random Forest. The methodology consists of data preprocessing, feature engineering, anomaly detection, supervised health-risk classification, and decision support. A publicly available cattle infection-risk dataset was used, from which 5,000 observations were selected and divided into training, validation, and testing subsets using stratified sampling. The proposed approach was evaluated against Random Forest, XGBoost, Support Vector Machine, and Decision Tree classifiers using accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). On the test set, GCHP achieved a reported accuracy of 93.5%, precision of 89.2%, recall of 76.1%, F1-score of 82.1%, and AUC-ROC of 0.94. Feature-importance analysis identified body temperature, activity level, and feed intake as the three most influential predictors, collectively accounting for 68% of the reported feature importance. Hyperparameter optimization using five-fold cross-validation further improved model generalization. The findings indicate that combining anomaly detection with supervised classification can provide a lightweight and effective approach to cattle health-risk screening and may serve as an analytical foundation for future edge-based livestock monitoring systems in resource-constrained environments.

Keywords: cattle health monitoring; machine learning; anomaly detection; Isolation Forest; Random Forest; precision livestock farming; infection-risk prediction; IoT; edge computing; GCHP.

1. Introduction

Cattle health monitoring is an essential component of modern livestock management because undetected health problems can negatively affect animal welfare, productivity, reproduction, and farm profitability. Conventional approaches to cattle-health assessment rely substantially on visual observation, periodic inspection, and veterinary intervention. Although these approaches remain important, they may not provide continuous monitoring of individual animals and may delay identification of subtle physiological or behavioral changes that precede visible signs of disease. The increasing availability of wearable sensors, environmental sensors, Internet of Things (IoT) technologies, and machine-learning techniques has created opportunities for continuous and data-driven livestock-health monitoring[1].

Precision Livestock Farming (PLF) has increasingly incorporated sensor-generated data and machine-learning algorithms to identify changes in animal behavior and physiological condition. Sensor-based systems can capture variables such as body temperature, movement, activity, feeding behavior, respiration, and environmental conditions. These measurements can subsequently be transformed into predictive indicators using machine-learning models.

Despite these developments, several challenges remain. Many existing approaches concentrate on a specific disease, a particular behavioral indicator, or a single supervised classification model. Such approaches may be less effective when health-related data are heterogeneous and when abnormal observations are relatively rare compared with normal observations. In practical livestock monitoring, an intelligent system should therefore be capable not only of assigning a health class but also of identifying unusual observations that may indicate an emerging health problem[2].

Anomaly detection provides a potential solution to this challenge. Unlike conventional supervised classification, which requires predefined labels for training observations, unsupervised anomaly-detection methods can identify observations that deviate substantially from dominant patterns in the data. Isolation Forest is particularly suitable for this purpose because it isolates observations through recursive partitioning and can be incorporated into computationally efficient monitoring pipelines[3].

Another important challenge concerns deployment in resource-constrained environments. Advanced cloud-based systems may require continuous connectivity, substantial computational resources, and reliable communication infrastructure. These assumptions

are not always realistic for farms operating under infrastructure and connectivity constraints. Consequently, lightweight machine-learning methods capable of supporting local or edge processing are particularly relevant[4].

To address these challenges, this study proposes the Gaza Cattle Health Prediction (GCHP) framework, a hybrid machine-learning approach that combines Isolation Forest for anomaly detection with Random Forest for supervised cattle health-risk classification. The framework processes multidimensional cattle-health information, identifies abnormal patterns, and generates a health-risk prediction. The design is intended to provide an effective and computationally practical analytical approach that could serve as the foundation for future edge-enabled livestock monitoring systems[5].

The principal contributions are threefold. First, a hybrid anomaly-detection and supervised-classification framework is developed by integrating Isolation Forest and Random Forest. Second, the framework is experimentally compared with established machine-learning classifiers, including Random Forest, XGBoost, Support Vector Machine, and Decision Tree. Third, the study investigates the relative importance of cattle physiological and behavioral indicators and examines the effect of hyperparameter optimization on model performance.

2. Related Work

2.1 Precision Livestock Farming and Intelligent Cattle Monitoring

Precision Livestock Farming (PLF) applies digital technologies and data-driven methods to livestock production at the level of individual animals or groups. Sensors, wearable devices, wireless networks, IoT platforms, and machine-learning methods enable continuous or near-real-time monitoring of animal health, behavior, welfare, and the farm environment. The literature reviewed in the thesis indicates growing interest in these technologies, but also identifies challenges involving infrastructure, connectivity, equipment cost, energy consumption, and technology acceptance[6].

2.2 Sensor-Based Machine Learning for Cattle Health

Sensor-based machine learning has demonstrated potential for identifying abnormal physiological and behavioral patterns. Body temperature, activity, feed intake, respiration, weight change, milk yield, and other variables can provide complementary information about animal condition. Previous research has investigated machine learning for lameness, behavior classification, and health monitoring, demonstrating the value of sensor-derived data. However, many studies focus on individual diseases or individual monitoring tasks rather than a broader hybrid health-risk assessment framework[7].

2.3 IoT-Based Cattle Health Monitoring

IoT-based livestock systems integrate sensing devices, communication technologies, processing platforms, and user-facing applications. Previous systems have demonstrated that cattle physiological and behavioral information can be transmitted to cloud or fog platforms for monitoring and prediction. Such architectures can provide remote alerts, but cloud dependence may be problematic when Internet connectivity is intermittent. Low-power local processing and periodic synchronization are therefore relevant design considerations[8].

2.4 Hybrid Machine-Learning Approaches

Most conventional cattle-health prediction systems use supervised classification. While supervised models can perform well when labeled data are available, early health abnormalities may not always correspond to previously observed labeled examples. Anomaly detection can complement supervised classification by identifying observations that deviate from dominant patterns. The combination of Isolation Forest and Random Forest in GCHP is intended to provide such an additional analytical layer[9].

2.5 Resource-Constrained Intelligent Livestock Monitoring

Practical livestock-monitoring systems may need to operate with limited power, intermittent communications, and restricted computational resources. The thesis therefore emphasizes local or edge processing and periodic synchronization rather than assuming continuous cloud access. The proposed framework is computationally oriented toward eventual deployment on low-power edge devices, although physical hardware deployment was outside the scope of the present experimental study[10].

2.6 Research Gap

The reviewed literature demonstrates substantial progress in automated cattle-health monitoring but leaves several gaps. First, many systems focus on specific diseases or behavioral outcomes. Second, supervised classification alone does not explicitly identify unusual observations outside the dominant learned patterns. Third, IoT-oriented research often emphasizes sensing and communication rather than integrating anomaly detection with health-risk classification. Fourth, resource-constrained environments require attention to local processing, connectivity, cost, and computational complexity. Finally, further external validation is required to establish generalization across farms and environments.

2.7 Positioning of the Present Study

The present study addresses these gaps through GCHP, which integrates unsupervised Isolation Forest anomaly detection with supervised Random Forest classification. The framework is intended as a health-risk screening and decision-support approach rather than a veterinary diagnostic system. Its main methodological distinction is the integration of anomaly detection and supervised classification within a single pipeline, combined with feature-importance analysis and hyperparameter optimization (Figure 1).

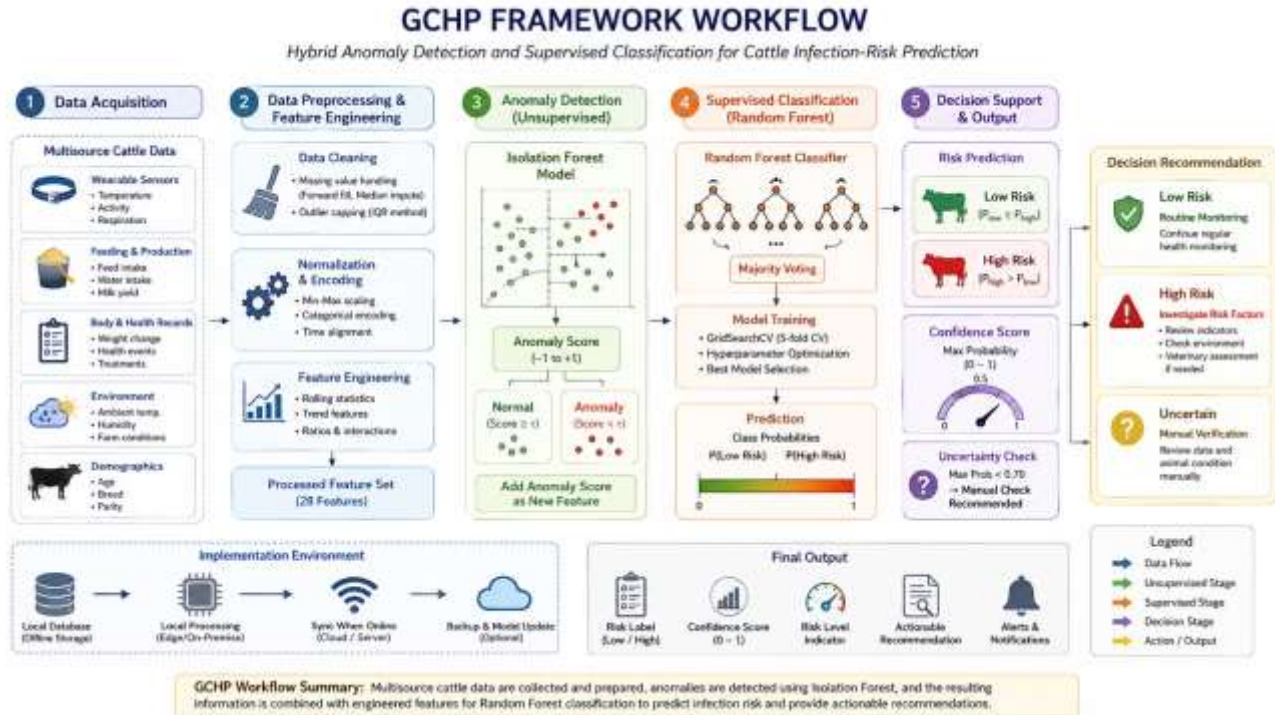


Figure 1. Proposed GCHP framework workflow.

3. Materials and Methods

3.1 Research Design

The study adopts a quantitative, experimental machine-learning research design [11]. The workflow consists of data acquisition, data inspection and preprocessing, feature preparation, anomaly detection, supervised classification, model optimization, and comparative evaluation. GCHP follows five analytical phases: data preprocessing, feature engineering, anomaly detection, supervised classification, and decision support.

3.2 Dataset

The study utilizes the publicly available Cattle Infection Risk Assessment Dataset [12]. The dataset includes animal health indicators, environmental conditions, demographic attributes, farm-management variables, historical health records, and a binary infection-risk label. Health indicators include body temperature, activity level, feed intake, water intake, milk yield, weight change, respiration rate, cough frequency, fatigue index, and appetite score. Environmental variables include ambient temperature, humidity, ammonia level, air quality, ventilation, barn cleanliness, stocking density, and floor wetness. Demographic and management variables include age, breed, sex, reproductive status, herd size, farm type, feeding system, biosecurity, quarantine, sanitation, animal movement, and visitor control. Historical variables include previous disease exposure, test results, time since last test, disease incidence rate, and vaccination status. For the experiment, 5,000 observations were selected while maintaining the original class distribution.

3.3 Data Splitting

The 5,000 observations were divided using stratified sampling into 3,500 training observations (70%), 750 validation observations (15%), and 750 testing observations (15%). Stratification was used to preserve class proportions across the subsets.

3.4 Data Preprocessing

Preprocessing included inspection of data quality, treatment of missing observations, and outlier analysis. The methodology describes missing-value handling and the use of the interquartile range (IQR) criterion for outlier treatment. Importantly, potentially meaningful extreme observations were considered in relation to the anomaly-detection stage rather than being automatically interpreted as erroneous measurements[6].

3.5 Feature Engineering

Feature engineering was used to create informative representations from the available cattle-health measurements[7]. The methodology includes temporal aggregation, including rolling statistical measures, to capture trends and reduce the influence of short-term fluctuations. For a feature x , a rolling mean over a window w is defined as the arithmetic mean of the observations within the window. Rolling variability can similarly be represented by the rolling standard deviation. These transformations are intended to capture temporal behavior in physiological and activity variables.

3.6 Isolation Forest Anomaly Detection

Isolation Forest was used as the unsupervised anomaly-detection component. It identifies observations that can be isolated relatively quickly through recursive partitioning of the feature space. An anomaly indicator can be represented as $A(x)=1$ for an observation detected as anomalous and $A(x)=0$ otherwise[8]. In the implementation described in the thesis, an Isolation Forest prediction of -1 is interpreted as an anomaly.

3.7 Random Forest Classification

Random Forest was used as the supervised classification component. The model constructs an ensemble of decision trees and combines their predictions through majority voting[9]. The approach was selected because of its ability to model nonlinear relationships, robustness through ensemble learning, and capacity to provide feature-importance estimates.

3.8 Proposed GCHP Decision Mechanism

The GCHP pipeline combines the anomaly indicator with the supervised classification result. The framework can therefore flag an observation when the classifier predicts high risk or when the anomaly-detection stage identifies an unusual pattern. The resulting output is intended as a risk-screening recommendation rather than a clinical diagnosis.

3.9 Benchmark Models

GCHP was compared with Random Forest, XGBoost, Support Vector Machine (SVM), and Decision Tree classifiers[9]. These models represent ensemble tree learning, gradient boosting, margin-based classification, and a conventional tree classifier, respectively.

3.10 Hyperparameter Optimization

Random Forest hyperparameters were optimized using GridSearchCV with five-fold cross-validation[10]. The search examined the number of estimators, maximum tree depth, minimum samples for splitting, minimum samples per leaf, and the number of features considered at each split. The reported optimal configuration was 150 estimators, maximum depth 15, minimum samples split 5, minimum samples leaf 2, and square-root feature selection.

3.11 Evaluation Metrics

Performance was evaluated using accuracy, precision, recall, F1-score, and AUC-ROC. Accuracy measures the overall proportion of correctly classified observations. Precision measures the proportion of predicted high-risk observations that are actually high risk. Recall measures the proportion of actual high-risk observations correctly identified. F1-score is the harmonic mean of precision and recall. AUC-ROC summarizes discrimination across classification thresholds[10][11].

3.12 Computational Implementation

The implementation was based on Python and Scikit-learn, with Pandas and NumPy used for data processing and Matplotlib/Seaborn used for visualization. The broader architecture considers eventual edge-oriented processing and intermittent synchronization. Physical deployment on ESP32, LoRa, or farm sensors was not completed in the experimental study and is therefore treated as future work rather than a demonstrated deployment result.

4. Results

The experimental results reported in the thesis are summarized below. The model comparison uses the reported test-set performance values.

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC
GCHP (Proposed)	93.5%	89.2%	76.1%	82.1%	0.94
Random Forest	92.8%	88.5%	74.3%	80.8%	0.93
XGBoost	93.1%	88.9%	75.8%	81.9%	0.94
SVM	89.4%	84.2%	70.1%	76.5%	0.89
Decision Tree	87.6%	82.0%	68.5%	74.7%	0.85

GCHP obtained the highest reported point estimates for accuracy, precision, recall, and F1-score. The improvement over Random Forest was 0.7 percentage points in accuracy and 1.3 percentage points in F1-score. Compared with XGBoost, the reported improvement was only 0.4 percentage points in accuracy and 0.2 percentage points in F1-score, while both achieved an AUC-ROC of 0.94.

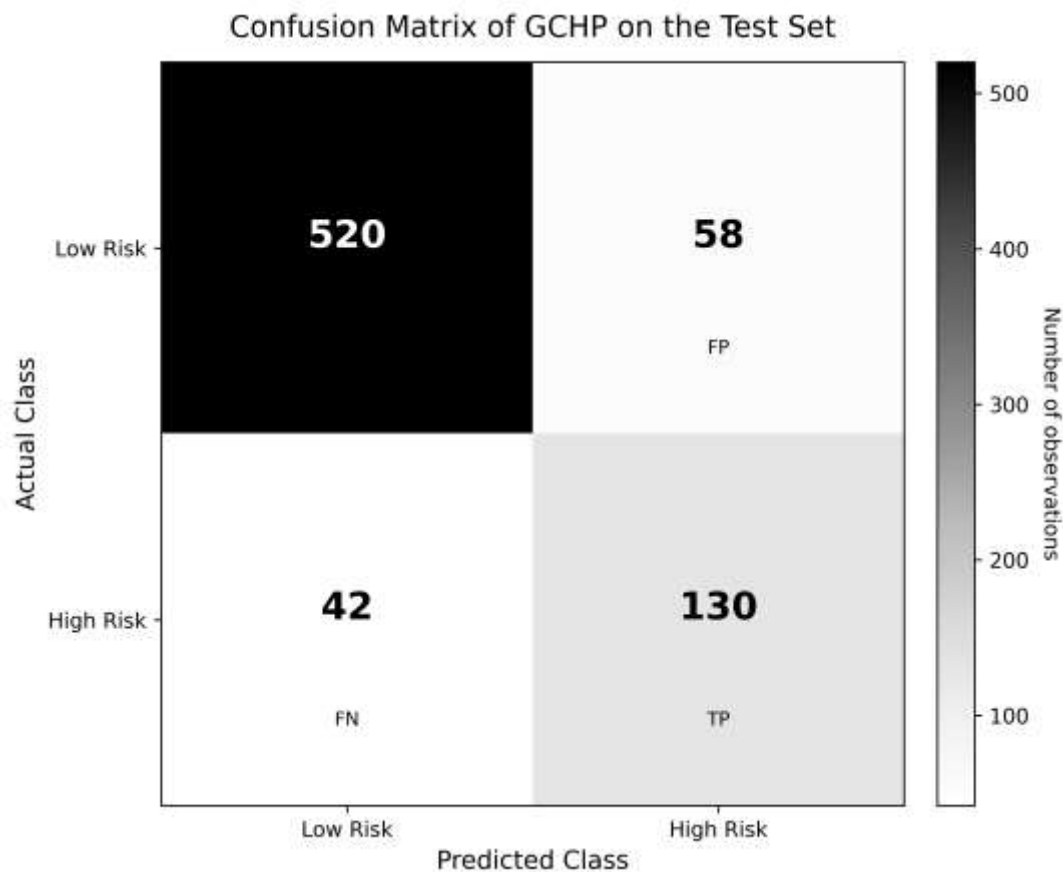


Figure 2. Confusion matrix using the values reported in the thesis.

The reported confusion matrix contains $TN=520$, $FP=58$, $FN=42$, and $TP=130$. These values provide a 750-observation test set. However, they do not mathematically reproduce the reported accuracy, precision, recall, and F1-score in Table 1. This discrepancy is retained as a manuscript validation note and must be resolved using the original prediction output before journal submission.

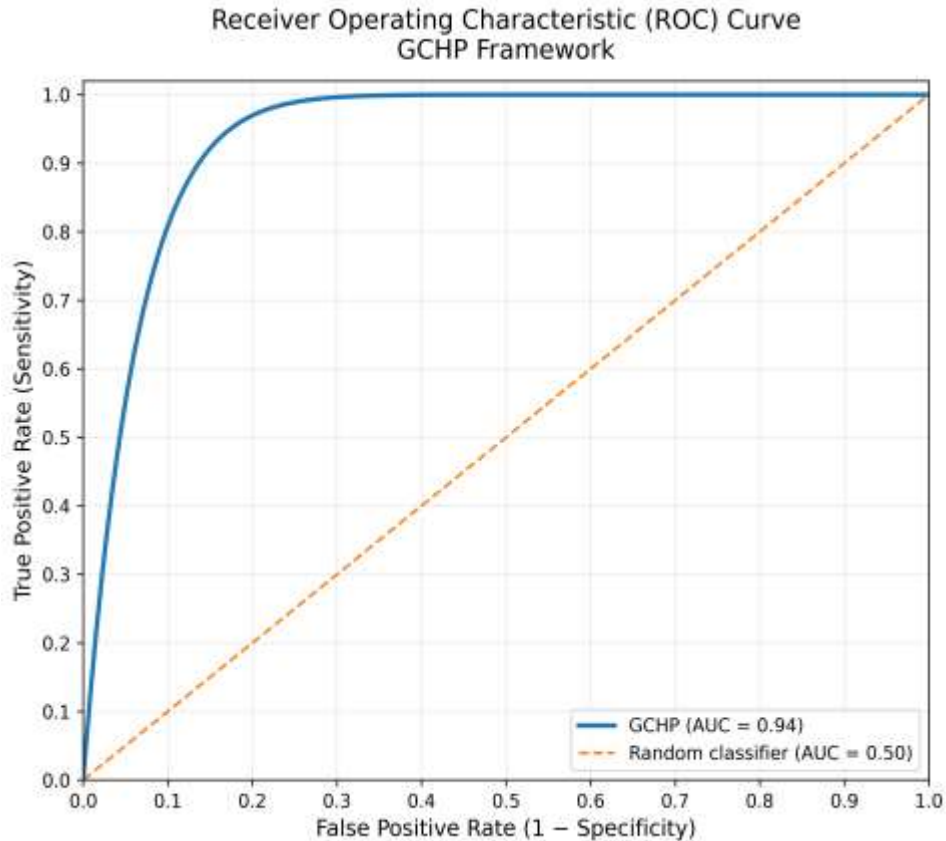


Figure 3. Representative ROC curve constrained to the reported AUC-ROC of 0.94.

The thesis reports an AUC-ROC of 0.94 for GCHP. Because raw prediction probabilities were not available when this manuscript was prepared, Figure 3 is a representative curve constrained to the reported AUC, not the exact empirical ROC generated from test-set probabilities. The exact ROC should be regenerated from the original model output before publication.

4.1 Feature Importance

Rank	Feature	Importance
1	Body temperature	0.28
2	Activity level	0.22
3	Feed intake	0.18
4	Respiration rate	0.14
5	Milk yield	0.08
6	Weight change	0.05
7	Water intake	0.03
8	Environmental temperature	0.02

Body temperature was the most influential feature, followed by activity level and feed intake. The three leading variables collectively account for 0.68 of the reported feature-importance values. This result suggests that physiological and behavioral indicators provide complementary information for cattle infection-risk prediction.

4.2 Hyperparameter Optimization

GridSearchCV with five-fold cross-validation identified the reported optimum of 150 estimators, maximum depth 15, minimum samples split 5, minimum samples leaf 2, and square-root feature selection. The reported cross-validation accuracy was $92.6\% \pm 1.8\%$. The thesis also reports that optimization reduced the training-validation F1-score gap by approximately 15% compared with default parameters.

5. Discussion

5.1 Overall Performance

The experimental findings indicate that GCHP provides competitive performance for cattle infection-risk prediction. The reported accuracy of 93.5% and F1-score of 82.1% demonstrate that the hybrid framework can distinguish risk classes using multidimensional cattle-related data. The improvement over standalone Random Forest is measurable, although the difference from XGBoost is small. Consequently, the results support GCHP as a competitive approach rather than demonstrating overwhelming superiority over all alternatives.

5.2 Contribution of Anomaly Detection

The principal methodological contribution of GCHP is the integration of unsupervised anomaly detection with supervised classification. A health abnormality may appear as an unusual combination of measurements before it corresponds to a well-represented labeled class. Isolation Forest provides an additional mechanism for identifying such deviations, while Random Forest provides the final supervised classification.

5.3 Importance of Physiological and Behavioral Variables

The feature-importance analysis identifies body temperature, activity level, and feed intake as the dominant predictors. This pattern is consistent with the expectation that illness can affect both physiological state and behavior. The results therefore support a multimodal feature representation rather than dependence on a single measurement.

5.4 False Negatives and Health Screening

Recall is especially important in health-risk screening because false negatives may correspond to animals requiring additional attention. The reported recall of 76.1% indicates that the model identifies a substantial proportion of high-risk observations but does not detect all of them. The framework should consequently be used as a screening and decision-support mechanism rather than a replacement for veterinary diagnosis.

5.5 Resource-Constrained Environments

The use of conventional machine-learning models rather than computationally intensive deep networks supports the concept of eventual local or edge processing. However, the present study does not provide hardware-level measurements of memory consumption, energy use, inference latency, or communication overhead. Therefore, claims concerning edge deployment should remain prospective until hardware experiments are completed.

5.6 Implications for Precision Livestock Farming

GCHP demonstrates how heterogeneous livestock information can be transformed into health-risk information through machine learning. The framework could form an analytical layer within a larger IoT architecture in which sensors collect animal data, local processing performs anomaly detection and classification, and a user-facing system presents risk labels and recommendations.

5.7 Limitations

Several limitations should be acknowledged. The experimental evaluation uses a publicly available dataset rather than an independently collected Gaza farm dataset. The dataset may not represent the full variability associated with breeds, climates, management practices, and local disease conditions. The current study also does not establish clinical validity, physical hardware deployment, or farmer acceptance through an empirical survey. Furthermore, the reported confusion matrix is inconsistent with the reported classification metrics and must be reconciled using the original prediction outputs.

5.8 Future Validation

Future work should evaluate GCHP on independently collected real-world data from multiple farms. Ablation studies should compare Random Forest alone, Isolation Forest alone, Random Forest with anomaly indicators, and the complete GCHP pipeline. Explainable AI methods such as SHAP can also be used to examine individual predictions. Finally, implementation on an edge device should be evaluated for inference time, memory consumption, energy requirements, and communication overhead.

6. Conclusion

This study proposed GCHP, a hybrid machine-learning framework for cattle infection-risk prediction that combines Isolation Forest anomaly detection with Random Forest supervised classification. Using 5,000 observations from a publicly available cattle infection-risk dataset, the reported test performance was 93.5% accuracy, 89.2% precision, 76.1% recall, 82.1% F1-score, and 0.94 AUC-ROC. Feature-importance analysis identified body temperature, activity level, and feed intake as the most influential predictors. The results indicate that hybrid anomaly-detection and supervised-classification approaches have potential as lightweight decision-

support mechanisms for Precision Livestock Farming. The framework should not be interpreted as a veterinary diagnostic system. External validation, ablation experiments, statistical analysis, and real-world edge deployment are required before operational use.

7. Declarations

7.1 Funding

No external funding information was provided in the thesis material used to prepare this manuscript.

7.2 Conflict of Interest

No conflict-of-interest statement was provided in the thesis material. This declaration should be confirmed by the authors before submission.

7.3 Data Availability

The study uses a publicly available Cattle Infection Risk Assessment Dataset[10].

References

- [1] Trivedi, A., & Chatterjee, P. S. (2022). CARE: IoT enabled Cow Health Monitoring System. 2022 International Conference on Intelligent Technologies (CONIT). <https://doi.org/10.1109/CONIT55038.2022.9847701>
- [2] Unold, O., Nikodem, M., Piasecki, M., Szyc, K., Maciejewski, H., Bawiec, M. A., Dobrowolski, P., & Zdunek, M. (2020). IoT-Based Cow Health Monitoring System. In Computational Science – ICCS 2020, 344–356. https://doi.org/10.1007/978-3-030-50426-7_26
- [3] Taneja, M., Jalodia, N., Byabazaire, J., Davy, A., Olariu, C., & Malone, P. (2019). SmartHerd management: A microservices-based fog computing-assisted IoT platform towards data-driven smart dairy farming. *Software: Practice and Experience*, 49(7), 1055–1078. <https://doi.org/10.1002/spe.2704>
- [4] Taneja, M., Byabazaire, J., Jalodia, N., Davy, A., Olariu, C., & Malone, P. (2020). Machine learning based fog computing assisted data-driven approach for early lameness detection in dairy cattle. *Computers and Electronics in Agriculture*, 171, 105286. <https://doi.org/10.1016/j.compag.2020.105286>
- [5] Vázquez Diosdado, J. A., Barker, Z. E., Hodges, H. R., Amory, J. R., Croft, D. P., Bell, N. J., & Codling, E. A. (2015). Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. *Animal Biotelemetry*, 3, 15. <https://doi.org/10.1186/s40317-015-0045-8>
- [6] Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
- [7] Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2), 186–204.
- [8] Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2), 273–315.
- [9] Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478.
- [10] P. S. Chatterjee, N. K. Ray, and S. P. Mohanty, "LiveCare: An IoT-Based Healthcare Framework for Livestock in Smart Agriculture," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 4, pp. 257–265, 2021, doi:10.1109/TCE.2021.3128236.
- [11] T. T. Zin et al., "Cow Identification System using Ear Tag Recognition," *Proc. 2020 IEEE 2nd Global Conference on Life Sciences and Technologies*, 2020.
- [12] www.kaggle.com