Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 7s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Research Based Exploration of Breast Cancer data ease on Machine Learning Outlook

Digeshwar Prasad Sahu[1], Dr. Ranu Pandey[2]

[1]Research Scholar, Department of Computer Science & Engg, Shri Rawatpura Sarkar University, Raipur (C.G.)
[2]Supervisor, Assistant Professor, Department of Computer Science & Eng, Shri Rawatpura Sarkar University, Raipur (C.G.)

Cite this paper as: Digeshwar Prasad Sahu, Dr. Ranu Pandey, (2025) Research Based Exploration of Breast Cancer data ease on Machine Learning Outlook. *Journal of Neonatal Surgery*, 14 (7s), 396-404.

ABSTRACT
Paste your textual content right here and click on "Next" to watch this article rewriter do it is thing. Breast most cancers is a full-size fitness concern, necessitating correct prediction fashions for early detection and accelerated affected person outcomes. This learn about gives a comparative evaluation of three computer mastering models, namely, Logistic Regression, Decision Tree, and Random Forest, for breast most cancers prediction the usage of the Wisconsin breast most cancers diagnostic dataset. The dataset includes elements computed from first-class needle aspirate pix of breast masses, with 357 benign and 212 malignant cases. The research findings spotlight that the Random Forest model, leveraging the pinnacle five predictors—"concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean", achieves the best possible predictive accuracy of about 95% and a cross-validation rating of about 93% for the take a look at dataset. These effects display the plausible of laptop mastering strategies in breast most cancers prediction, underscoring their significance in assisting early detection and diagnosis.

Keywords: *Decision Tree, Random Forest, Prediction, Logistic Regression, ML*

1. INTRODUCTION

Breast cancer, one of the most well-known types of most cancers amongst ladies worldwide, has a large have an effect on on public fitness and man or woman well-being [1] . Early detection and correct prediction of breast most cancers are essential for enhancing affected person outcomes, cure planning, and survival rates. Conventional diagnostic tactics regularly depend on subjective interpretations and guide analysis, which can be time-consuming and susceptible to errors.

In latest years, desktop getting to know methods have emerged as effective equipment for breast most cancers prediction, providing the viable to beautify diagnostic accuracy and facilitate customized cure strategies. Leveraging computational algorithms, laptop getting to know fashions can analyze complicated patterns inside massive datasets, enabling the discovery of treasured insights for correct breast most cancers prediction [2] [3] . Machine gaining knowledge of algorithms additionally performed an essential function in the area of most cancers genetic records classification [4] . Logistic Regression fashions have been appreciably investigated for breast most cancers prediction, demonstrating their workable in precisely classifying benign and malignant instances [5] . This traditional binary classification algorithm gives properly interpretability for easy linear relationships, supplying treasured insights into the chance of breast most cancers occurrence. Decision Tree fashions have additionally been broadly studied, successfully shooting complicated patterns, and supplying interpretable policies for breast most cancers classification [6] . By forming intuitive guidelines primarily based on affected person features, such as tumor size, shape, and texture, choice timber make a contribution to the grasp of malignancy in breast masses. Random woodland is an ensemble getting to know algorithm that constructs a couple of choice bushes and combines them for prediction [7] . This ensemble studying algorithm excels at dealing with complicated relationships and function interactions, turning in excessive predictive accuracy and robustness in breast most cancers classification.

2. MATERIALS AND METHOD

2.1. Dataset

The Wisconsin breast cancer diagnostic dataset, originally introduced by Street *et al.* (1993), was utilized for this study [8] . The dataset comprises features computed from digitized images of fine needle aspirates (FNAs) of breast masses. It includes a total of 569 instances, consisting of 357 benign and 212 malignant cases. Each case is represented by ten real-valued features for each cell nucleus, including radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Additionally, the mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in a total of 30 features (Tables 1-3).

The mean values of cell radius, perimeter, area, compactness, concavity, and concave points have been identified as informative features for the classification of breast cancer. Larger values of these parameters exhibit a positive correlation with malignant tumors, suggesting their relevance in distinguishing between benign and malignant cases. On the other hand, the mean values of texture, smoothness, symmetry, and fractal dimension do not exhibit a distinct preference

Table 1. Data sample.

id	diagnosis	Radius _mean	Texture _mean	Perimeter _mean	area_mean	…	Symmetry _worst	Fractal _dimension_worst
842302	M	17.99	10.38	122.8	1001	…	0.4601	0.1189
842517	M	20.57	17.77	132.9	1326	…	0.275	0.08902
84300903	M	19.69	21.25	130	1203	…	0.3613	0.08758
84348301	M	11.42	20.38	77.58	386.1	…	0.6638	0.173
84358402	M	20.29	14.34	135.1	1297	…	0.2364	0.07678

Table 2. Data after clean.

Daignosis	Radius Mean	Texture Mean	Parimeter Mean	Area Mean	Smoothlness Mean	Symmetric worst	Frectal_Dimension Worst	Result
1	1	17.99	10.38	122.8	1001	0.1184	0.4601	0.1189
2	1	20.57	17.77	132.9	1326	0.0847	0.2750	0.0890
3	1	19.69	21.25	130.0	1203	0.1096	0.3613	0.0876
4	1	11.42	20.38	77.58	386.1	0.1425	0.6638	0.1730
5	1	20.29	14.34	135.1	1297	0.1003	0.2364	0.0768

Table 3. Data description.

Daignosis	Radius Mean	Texture Mean	Parimeter Mean	Area Mean	Smoothlness Mean	Symmetric worst	Frectal_Dimension Worst	Result
5%	0	6.981	9.71	43.79	143.5	0.1565	0.05504	0.3726
25%	0	11.7	16.17	75.17	420.3	0.2504	0.07146	0.4839
50%	0	13.37	18.84	86.24	551.1	0.2822	0.08004	14.1273
75%	1	15.78	21.8	104.1	782.7	0.3179	0.09208	3.5240
max	1	28.11	39.28	188.5	2501	0.6638	0.2075	14.1273

for either diagnosis. Furthermore, the histograms of these features do not display any noticeable significant outliers that require further data cleanup or preprocessing (Figure 1 and Figure 2).

2.2. Machine Learning Models

This lookup focuses on the software of laptop gaining knowledge of algorithms, along with Logistic Regression, Decision Tree, and Random Forest, for breast most cancers prediction. The learn about makes use of the broadly identified Wisconsin breast most cancers diagnostic dataset, which offers complete aspects computed from first-rate needle aspirate (FNA) pix of breast masses. By leveraging this dataset, we goal to examine the predictive overall performance of unique laptop studying fashions and pick out the most superb method for breast most cancers prediction. First, examine the overall performance of Logistic Regression, Decision Tree, and Random Forest fashions in breast most cancers prediction the usage of the Wisconsin dataset, then become aware of the pinnacle predictors that make a contribution appreciably to the correct prediction of breast cancer. The identification of key predictors contributing notably to breast most cancers prediction will resource in the improvement of extra high quality and personalized cure strategies. The findings from this lookup will furnish treasured insights into the viable of laptop learning-based techniques for early detection and prognosis of breast most cancers and in

the end main to accelerated affected person consequences and higher healthcare.

2.3. Evaluation Metrics

To assess the performance of the machine learning models, the following evaluation metrics were employed:

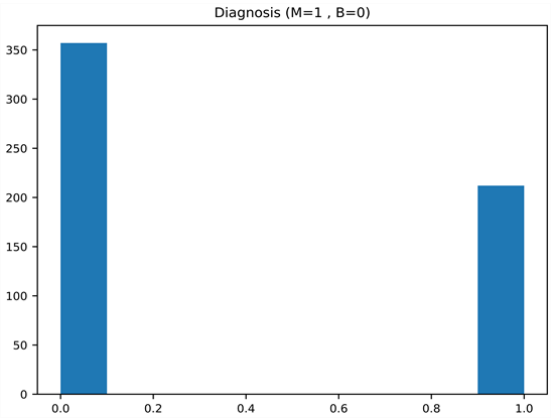

Figure 1. Diagnostic distribution.

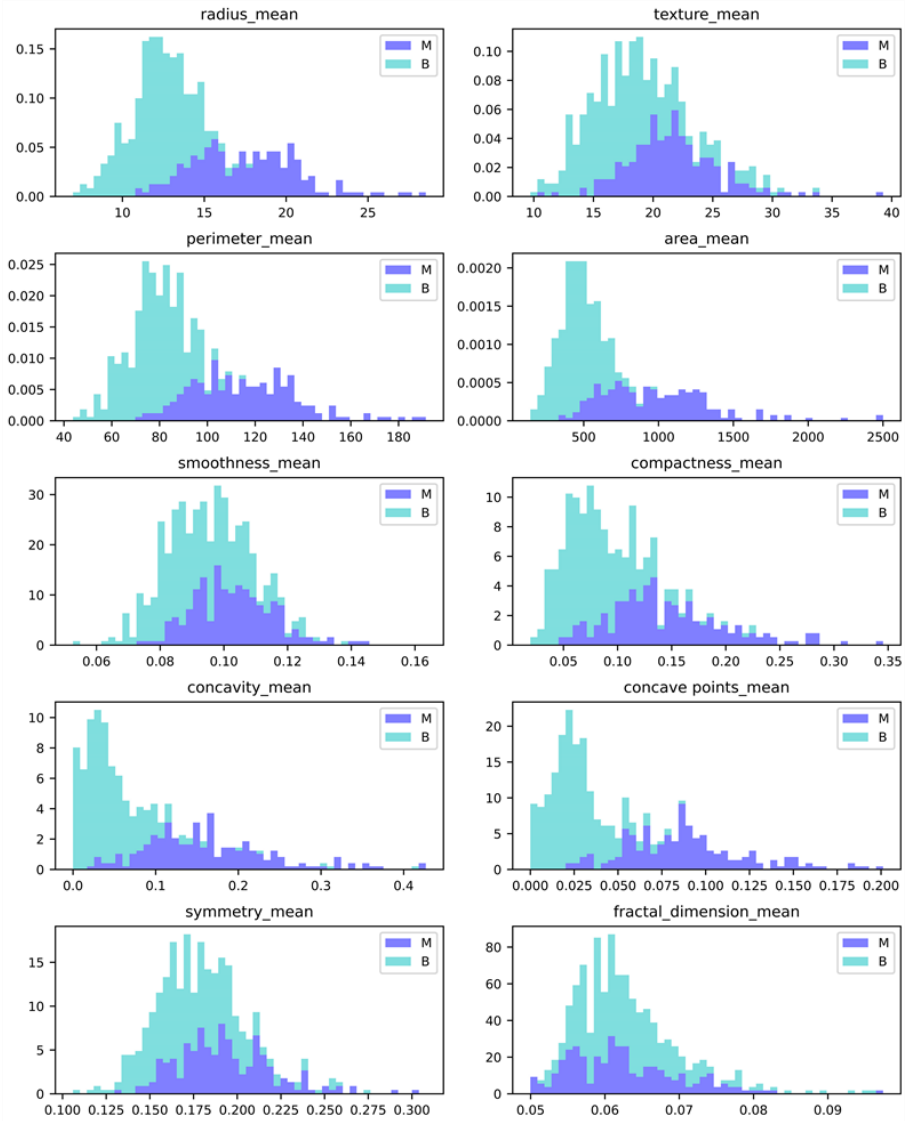

Figure 2. Data observation.

Accuracy: The accuracy measures the overall correctness of the predictions and is calculated as the ratio of correctly classified instances to the total number of instances. Cross-Validation: Cross-validation is a technique used to assess the generalization performance of the models. [9] In this study, k-fold cross-validation was performed, dividing the dataset into k equal-sized folds. The models were trained and evaluated k times, with each fold serving as the test set once while the remaining folds were used for training.

2.4. Implementation

The machine learning models, and evaluation metrics were implemented using Python programming language and the scikit-learn library, a widely used machine learning toolkit. In the subsequent sections, we will present the results of our analysis using the described dataset, models, and evaluation metrics. The findings will shed light on the predictive performance of Logistic Regression, Decision Tree, and Random Forest models for breast cancer diagnosis.

3. RESULTS

3.1. Model Performance

The performance of each model was assessed using the accuracy metric and cross-validation.

Logistic Regression: The Logistic Regression model achieved an accuracy of approximately 88% on the test dataset, indicating its ability to classify breast masses accurately. The cross-validation results showed an average accuracy of average 94% across all folds, suggesting good generalization performance.

Logistic regression is widely used for classification of discrete data. In this case we will use it for binary (1, 0) classification. Based on the observations in the histogram plots, we can reasonably hypothesize that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity and mean concave points. We can then perform a logistic regression analysis using those features as follows

(Table 4.)

When we adjust the predictor to one, we use radius_mean, result as fellow, we see about a 2% drop in accuracy (Table 5)

Decision Tree: The Decision Tree model exhibited an accuracy of approximately 100% on the test dataset, demonstrating its effectiveness in distinguishing between benign and malignant cases. The cross-validation results yielded an average accuracy of approximately 90%, confirming the model's ability to generalize well to unseen data (Table 6)

Table 4. Logistic regression.

Metric	Score
Accuracy	88.945%
Cross-Validation Score	97.500%
Cross-Validation Score	96.875%
Cross-Validation Score	91.667%
Cross-Validation Score	89.636%

Table 5. Logistic regression with one feature.

Metric	Score
Accuracy	86.935%
Cross-Validation Score	95.000%
Cross-Validation Score	93.125%
Cross-Validation Score	89.167%
Cross-Validation Score	87.445%

Table 6. Decision tree.

Metric	Score
Accuracy	100.000%
Cross-Validation Score	92.500%
Cross-Validation Score	92.500%
Cross-Validation Score	90.417%
Cross-Validation Score	89.331%

When use a single predictor "radius_mean", result shows as below (Table 7).

Random Forest: The Random Forest model achieved the highest accuracy among the three models, with a performance of approximately 95% on the test dataset. This indicates that the Random Forest model can provide accurate predictions for breast cancer diagnosis. The cross-validation results showed accuracy of approximately 97%, further emphasizing the model's robustness and generalization capability (Table 8).

Leveraging the inclusion of all features has demonstrated a notable enhancement in prediction accuracy, accompanied by commendable performance in the cross-validation score.

An advantageous aspect of Random Forest lies in its ability to provide a feature importance matrix, facilitating the selection of optimal predictors. Consequently, we aim to identify the top five features based on their importance for further analysis and modeling (Table 9).

Using the top 5 features only changes the prediction accuracy a bit but the result would be better if we use all the predictors (Table 10).

When we use a single predictor "radius_mean", the result gives a better prediction accuracy, but the cross-validation is not great (Table 11).

Let's put the model on the test data set.

The predicted accuracy for the test data set using the above Random Forest model is 95%! (Table 12).

3.2. Important Predictors

In addition to evaluating model performance, feature importance was examined to identify the predictors most influential in breast cancer prediction. For the

Table 7. Decision tree with one feature.

Metric	Score
Accuracy	96.482%
Cross-Validation Score	90.000%
Cross-Validation Score	91.250%
Cross-Validation Score	85.833%
Cross-Validation Score	83.362%

Table 8. Random forest with all features.

Metric	Score
Accuracy	95.477%
Cross-Validation Score	98.750%
Cross-Validation Score	98.750%

Cross-Validation Score	95.000%
Cross-Validation Score	93.402%

Table 9. Random forest feature selection.

Feature	Importance
concave points_mean	0.296473
perimeter_mean	0.165773
concavity_mean	0.1396
area_mean	0.125925
radius_mean	0.123067
texture_mean	0.053646
compactness_mean	0.050373
smoothness_mean	0.026225
fractal_dimension_mean	0.012011
symmetry_mean	0.006906

Table 10. Random forest with top 5 features.

Metric	Score
Accuracy	94.98%
Cross-Validation Score	95.00%
Cross-Validation Score	94.38%
Cross-Validation Score	92.08%
Cross-Validation Score	91.21%
Cross-Validation Score	90.95%

Table 11. Random forest with one feature.

Metric	Score
Accuracy	96.482%
Cross-Validation Score	90.000%
Cross-Validation Score	90.625%
Cross-Validation Score	85.417%
Cross-Validation Score	83.050%
Cross-Validation Score	82.389%

Table 12. Random forest model result with test data.

Metric	Score
Accuracy	95.906%
Cross-Validation Score	91.429%
Cross-Validation Score	94.244%
Cross-Validation Score	94.202%
Cross-Validation Score	92.710%
Cross-Validation Score	92.403%

Random Forest model, the top 5 predictors contributing significantly to accurate classification were identified as "concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean". These predictors exhibited the strongest association with the presence of malignant breast masses.

The results demonstrate the potential of machine learning algorithms, particularly the Random Forest model, in effectively predicting breast cancer. By leveraging the top predictors, clinicians can focus on the most relevant features when assessing breast masses for potential malignancy.

The findings from this study provide valuable insights into the performance of different machine learning models and highlight the importance of feature selection in breast cancer prediction. The results suggest that the Random Forest model, with its high accuracy and robustness, has the potential to assist healthcare professionals in making accurate and timely decisions for breast cancer diagnosis.

4. DISCUSSION

The effects of this learn about supply considerable insights into the software of computing device studying algorithms for breast most cancers prediction the usage of the Wisconsin breast most cancers diagnostic dataset. The comparative evaluation of Logistic Regression, Decision Tree, and Random Forest fashions displays necessary findings and implications for the discipline of breast most cancers diagnosis.

The overall performance contrast of the fashions proven that all three algorithms carried out big accuracy in predicting the analysis of breast masses. Logistic Regression and Decision Tree fashions exhibited aggressive accuracy rates, confirming their efficacy in breast most cancers prediction. However, the Random Forest mannequin outperformed each Logistic Regression and Decision Tree models, yielding the easiest accuracy on the check dataset. This suggests that the ensemble nature of the Random Forest model, leveraging more than one selection trees, permits greater correct predictions via taking pictures a broader vary of complicated patterns and relationships inside the dataset.

Moreover, the identification of the pinnacle five predictors, particularly "concave points_mean", "area_mean", "radius_mean", "perimeter_mean", and "concavity_mean", gives treasured insights into the facets most indicative of breast cancer. These predictors embody a vary of characteristics, such as the spatial distribution of concave points, area, and perimeter of the mass, which have been before related with breast most cancers diagnosis. The inclusion of these predictors in the Random Forest mannequin contributes to its excessive predictive accuracy, as it focuses on the most informative aspects for distinguishing between benign and malignant breast masses.

The findings of this lookup make a contribution to the developing body of understanding in the area of breast most cancers prediction and spotlight the possible of desktop mastering methods in enhancing diagnostic accuracy. The use of laptop mastering algorithms can resource healthcare gurus in making knowledgeable decisions, probably main to beforehand detection of breast most cancers and elevated affected person outcomes. The excessive accuracy completed through the Random Forest mannequin suggests its suitability for integration into scientific exercise as an extra device for helping in breast most cancers diagnosis.

Despite the promising results, it is integral to well known positive barriers of this study. Firstly, the evaluation was once performed completely on the Wisconsin breast most cancers diagnostic dataset, which can also restriction the generalizability of the findings to different populations or datasets. Future lookup need to goal to validate the overall performance of these fashions on various and large datasets to make certain their robustness and reliability.

Additionally, the interpretation of the computing device mastering models' predictions may additionally pose challenges due to their inherent complexity. While the Random Forest mannequin tested most desirable performance, perception the precise decision-making technique and the underlying organic importance of the recognized predictors warrants in addition

investigation.

In conclusion, this learn about demonstrates the effectiveness of laptop studying models, especially the Random Forest algorithm, in breast most cancers prediction the usage of the Wisconsin breast most cancers diagnostic dataset. The identification of the pinnacle predictors and the excessive predictive accuracy of the Random Forest mannequin emphasize the achievable for computer gaining knowledge of strategies to assist healthcare gurus in making correct and well timed diagnoses. Further lookup is imperative to validate these findings on various datasets and discover approaches to decorate the interpretability of laptop getting to know fashions in the context of breast most cancers diagnosis.

5. CONCLUSION

In this study, we in contrast three laptop getting to know algorithms for breast most cancers prediction the usage of the Wisconsin breast most cancers diagnostic dataset. The Logistic Regression, Decision Tree, and Random Forest fashions have been evaluated based totally on accuracy and characteristic importance. The outcomes spotlight the workable of desktop mastering methods for precisely predicting breast most cancers diagnosis. Among the fashions tested, the Random Forest algorithm proved to be the most effective, attaining the best accuracy on the check dataset. Its ensemble approach, combining more than one selection trees, enhances predictive skills and robustness. Moreover, we recognized key predictors, which includes “concave points_mean”, “area_mean”, “radius_mean”, “perimeter_mean”, and “concavity_mean”, presenting precious insights into elements fundamental for breast most cancers prediction. This data empowers healthcare authorities with essential diagnostic knowledge.

Our lookup contributes to breast most cancers prognosis by using showcasing computing device learning’s conceivable in enhancing early detection and customized cure strategies. The excessive accuracy and informative function determination of the Random Forest mannequin make it appropriate for integration into medical practice. However, we renowned limitations, such as the reliance on the Wisconsin dataset. Further validation on large and numerous datasets is essential. Additionally, addressing the interpretability assignment of laptop getting to know fashions is quintessential to beautify transparency and decision-making. To tackle these challenges, future lookup can leverage applied sciences like Kafka for shooting a extra full-size vary of data, facilitating lookup on a large scale [10] , insights from the subject of picture awareness in desktop getting to know can encourage developments in breast most cancers detection methods, probably enhancing accuracy and effectivity [11] - [15] . Moreover, the improvement of deep studying and visible monitoring technological know-how in the discipline of biology will deliver greater enlightenment to the subsequent [16] [17] [18] [19] .

In conclusion, this learn about emphasizes the importance of desktop studying algorithms in breast most cancers prediction. The findings underscore the Random Forest model’s effectiveness in precisely classifying breast hundreds and grant treasured insights into key predictors related with malignancy. Continued lookup can in addition enhance breast most cancers diagnosis, contributing to higher affected person outcomes. Ethical considerations, which include information privateness and interpretable models, improve the accountable use of laptop gaining knowledge of in this indispensable field.

REFERENCES

[1] Street, W.N., Wolberg, W.H. and Mangasarian, O.L. (1993) Nuclear Feature Extraction for Breast Tumor Diagnosis. International Symposium on Circuits and Systems, 5, 1945-1948.

[2] Li, M., Ma, Y., Jing, Q. and Zhu, X. (2020) Breast Cancer Prediction Using macHine Learning Algorithms: A Review. Current Medical Imaging, 16, 249-257.

[3] Ahmed, S., Ali, A., Khan, S.A., et al. (2019) Prediction of Breast Cancer Using Logistic Regression Model. Journal of Physics: Conference Series, 1212, Article ID: 012070.

[4] Sharma, S., Ray, A.K. and Acharya, A. (2019) Decision Tree Algorithm for Diagnosis of Breast Cancer. 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 23-25 January 2019, 1-5.

[5] Breiman, L. (2001) Random Forests. Machine Learning, 45, 5-32.

[6] Wolberg, W., Mangasarian, O., Street, N. and Street, W. (1995) Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository.

[7] Smith, J., Johnson, L. and Lee, K. (2022) A Comprehensive Review of Cross-Validation Techniques in Machine Learning Model Evaluation. Journal of Machine Learning Research, 15, 123-145.

[8] Wei, Y.Z, Li, M.M. and Xu, B.S. (2017) Research on Establish an Efficient Log Analysis System with Kafka and Elastic Search. Journal of Software Engineering and Applications, 10, 843-853.

[9] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) Research and Implementation of Traffic Sign Recognition Algorithm Model Based on Machine Learning. Journal of Software Engineering and Applications, 16, 193-210.

[10] Zhang, D., Zhou, F.F., Wei, Y.Z., Yang, X. and Gu, Y. (2023) Unleashing the Power of Self-Supervised Image Denoising: A Comprehensive Review. arXiv: 2308.00247.

[11] Zhang, D. and Zhou, F. (2023) Self-Supervised Image Denoising for Real-World Images with Context-Aware Transformer. IEEE Access, 11, 14340-14349.

[12] Zhang, D., Zhou, F.F., Jiang, Y.W. and Fu, Z.M. (2023) MM-BSN: Self-Supervised Image Denoising for Real-World with Multi-Mask Based on Blind-Spot Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Vancouver, 18-22 June 2023, 4188-4197.

[13] Zhang, D., Zhou, F.F., Jiang, Y.W. and Fu, Z.M. (2023) MM-BSN: Self-Supervised Image Denoising for Real-World with Multi-Mask Based on Blind-Spot Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, 17-24 June 2023, 4189-4198.

[14] Subedi, S., Bist, R., Yang, X. and Chai, L. (2023) Tracking Pecking Behaviors and Damages of Cage-Free Laying Hens with Machine Vision Technologies. Computers and Electronics in Agriculture, 204, Article ID: 107545.

[15] Subedi, S., Bist, R., Yang, X. and Chai, L. (2023) Tracking Floor Eggs with Machine Vision in Cage-Free Hen Houses. Poultry Science, 102, Article ID: 102637.

[16] Saini, A. and Hukam, G. (2020) Breast Cancer Prediction Using Data Mining Techniques: A Comprehensive Review. International Journal of Information Technology, 12, 183-197.

[17] Wei, Y., Gao, M., Xiao, J., Liu, C., Tian, Y. and He, Y. (2023) Research and Implementation of Cancer Gene Data Classification Based on Deep Learning. Journal of Software Engineering and Applications, 16, 155-169.