Machine Learning Approaches for Detecting Abnormalities in Female Fetuses

Yuxuan Cai

JiangSu Normal University KeWen College, JiangSu, China

Received: 02 Oct 2025
Revised: 05 Oct 2025
Accepted: 09 Oct 2025
Published: 11 Oct 2025

Abstract: Non-invasive prenatal testing (NIPT) plays a vital role in the early detection of female fetal abnormalities, which is essential for birth defect prevention. In this study, clinical data containing Z-scores of chromosomes 21, 18, and 13, GC content, X chromosome concentration, read count ratio, and maternal BMI were analyzed. To address the class imbalance caused by the limited number of abnormal cases, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, and stratified sampling was used to divide the dataset into training, validation, and testing sets (7:2:1). Multiple machine learning models, including XGBoost, Decision Tree, CNN, MLP, SVM, and Random Forest, were developed and evaluated with accuracy, precision, recall, F1-score, and AUC-ROC metrics. Results demonstrated that Random Forest outperformed other models, achieving an AUC of 0.997 with strong stability and generalization. These findings highlight the effectiveness of machine learning combined with proper data preprocessing in enhancing female fetal abnormality detection.

Keywords: Female fetal abnormality; NIPT; Machine learning; Random Forest; SMOTE

1 INTRODUCTION

With the development of NIPT technology, chromosome abnormality screening based on high-throughput sequencing has gradually become an important method for prenatal detection [1]. However, the detection of female fetal abnormalities still faces challenges such as data sample scarcity, complex features, and insufficient accuracy. The "Healthy China 2030 Plan" calls for the improvement of a comprehensive birth defect prevention and control system. Against this backdrop, how to integrate big data and intelligent algorithms to build an efficient and reliable model for female fetal abnormality detection has become a pressing scientific and practical issue. This paper will systematically compare various machine learning models to explore their application potential in female fetal abnormality detection, providing theoretical foundations and practical references for clinical screening and decision-making.

2 RELATED WORK AND HYPOTHESES

This paper focuses on the detection of female fetal abnormalities. The main tasks include data augmentation using the SMOTE method, reasonable data partitioning through stratified sampling, constructing and comparing multiple machine learning models, and evaluating their performance based on multi-dimensional metrics.

Hypothesis 1: Data imbalance is a significant cause of decreased detection accuracy, and the model's performance will improve significantly after SMOTE augmentation.

Hypothesis 2: Different features contribute differently to the classification results, with the chromosome Z-scores and GC content playing a more significant role in detecting female fetal abnormalities.

Hypothesis 3: There is an interactive effect between GC content and the read count ratio on model stability, and abnormal feature distribution may lead to instability in certain models.

3 DATASET PROCESSING

3.1 Data augmentation based on SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to address the problem of class imbalance by generating synthetic data for the minority class to balance the dataset [2]. In the context of predicting abnormal female fetuses, SMOTE can help generate more cases of abnormal fetuses, thereby improving the training effectiveness of the model, especially when the cases of abnormal fetuses are relatively rare.

The core idea of SMOTE is to synthesize new minority class samples by interpolation between existing minority class samples. Specifically, SMOTE creates new samples by interpolating between a minority class sample and its neighboring samples. If we have a minority class sample $\mathrm{x_i}$, SMOTE will generate new synthetic samples using the following formula:

$$\mathrm{x_{new}} = \mathrm{x_i} + \lambda \cdot (\mathrm{x_j} - \mathrm{x_i}) \tag{1}$$

In this context, x_i is the minority class sample, and x_j is a randomly selected nearest neighbor of $\mathrm{x_i}.\lambda$ is a random number between [0,1] that determines the interpolation ratio between x_i and x_j for generating the new sample.

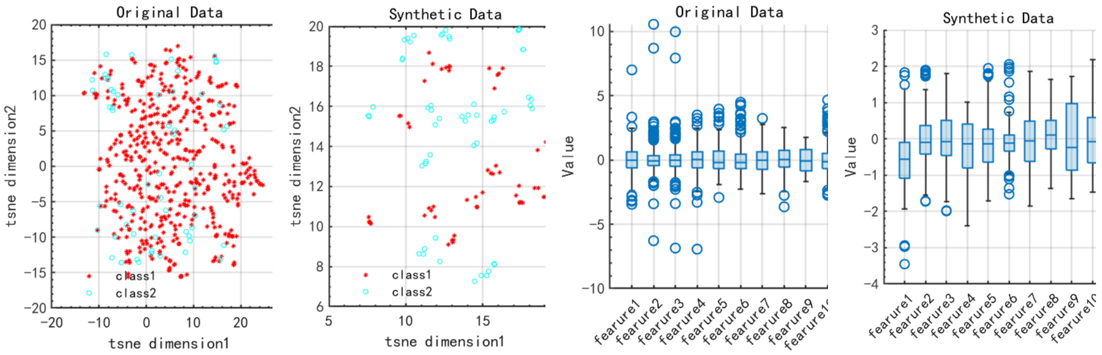

Fig. 1: Illustration of SMOTE Processing Results.

3.2 Dataset split based on stratified sampling

In the task of predicting abnormal female fetuses, stratified sampling can effectively ensure a balanced distribution of factors such as chromosome Z values, GC content, BMI, etc., across the training, validation, and test datasets. For the prediction of abnormal female fetuses, we follow the steps below for stratified sampling and dataset splitting:

Divide the data into strata: Based on the aneuploidy of chromosomes 21, 18, and 13, and incorporating factors such as the X chromosome Z values, GC content, and BMI, the data is divided into different strata. The data is grouped according to chromosome abnormality type, BMI range, and other factors.

Sampling within each stratum: Within each stratum, samples are drawn according to the desired proportion, ensuring that different BMIs and chromosome abnormality conditions are reasonably represented in the training, validation, and test sets.

Dataset splitting: Based on the results of stratified sampling, the sample dataset is divided into training, validation, and test sets with a ratio of 7:2:1. Specifically:

Training set (70%): Used to train the model, allowing the model to learn the relationship between features and labels.

Validation set (20%): Used for model tuning and validation, adjusting hyperparameters and preventing overfitting.

Test set (10%): Used for the final evaluation of the model, testing its performance on unseen data.

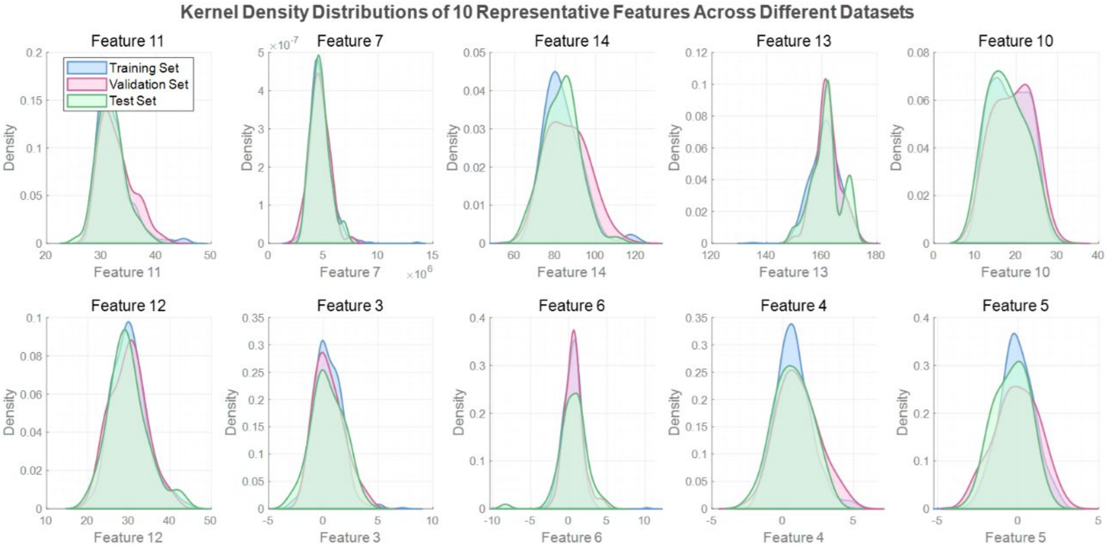

Fig. 2: Kernel Density Distribution Plot of Some Metrics.

4 MODEL ESTABLISHMENT AND SOLUTION

4.1 Establishment of Six Machine Learning Models

(1) XGBoost

XGBoost is an efficient gradient boosting algorithm that performs ensemble learning by building multiple decision trees and uses regularization to reduce overfitting. After preprocessing the data, the dataset is split into training and test sets. The column "Fetal health status" is used as the target variable, and other features required in Question 4 are used as inputs. The XGBoost algorithm is then used for training, and the model's performance is evaluated on the test set, providing accuracy, classification report, and confusion matrix.

(2) Decision Tree

Decision Tree is an intuitive supervised learning algorithm that performs classification by constructing a tree-like structure. In the abnormal fetal judgment task, the decision tree will make a series of hierarchical decisions based on features such as "Whether the Z-value of chromosome 21 is greater than the threshold" and "Whether the X chromosome concentration is normal." Each internal node represents a feature test, the branches are the test results, and the leaf node represents the final classification of "normal" or "abnormal."

(3) CNN

Convolutional Neural Networks (CNNs) are powerful models that automatically learn hierarchical features from data. In the abnormal fetal judgment task, the pregnant woman's BMI, gestational age, and the fetal's Z-values, GC content, and read counts are transformed into a one-dimensional input sequence. The CNN captures local patterns and relationships through filters in the convolutional layers, while pooling layers reduce dimensionality and abstract features. Finally, these high-level features are fed into fully connected layers to output the probability of fetal abnormality.

(4) MLP

Multilayer Perceptron (MLP) is a feedforward neural network model consisting of multiple fully connected layers. Each neuron in each layer is connected to all neurons in the previous layer, which is why it is called a "fully connected network." MLP maps the input data to a high-dimensional feature space through nonlinear activation functions, enabling the completion of regression or classification tasks.

(5) SVM

Support Vector Machine (SVM) is a supervised learning algorithm commonly used for classification and regression tasks [3]. The core idea is to find a hyperplane that separates the samples of different classes and maximizes the minimum distance from the hyperplane to the samples of each class. This method uses kernel tricks to map the data into a high-dimensional space and finds the optimal separating hyperplane in that space. After preprocessing the data, the RBF kernel function is used to construct the SVM model, and the model is fitted using the training data.

(6) Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates the prediction results of each tree by voting or averaging to obtain the final prediction [4]. Each decision tree is trained on randomly selected samples and features, and this randomness gives Random Forest strong resistance to overfitting.

The core idea of Random Forest is to improve the accuracy and stability of predictions by combining multiple decision trees [5]. The specific steps are as follows:

Step 1: Constructing the Dataset

The training dataset is represented as $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where x_i are the input features and y_i is the target variable. In this paper, the input features x_i include X color intensity, the values of 21,18, and 13 color components of the GC content, the ratio of being within a segment, and BMI. The target variable is: Whether the female fat is abnormal, labeled as 0 (Normal) or 1 (Abnormal).

Using the bootstrap method, N samples are randomly selected with replacement from the dataset DDD to construct each decision tree. The dataset used for each tree is obtained by randomly sampling from the original dataset, with the same size as the original dataset.

Step 2: Decision Tree Training

For each decision tree T_j, it randomly selects m features from the total number of features p (Where $m < p$) [6]. In this paper, the input features x are already determined, such as: X color intensity, the values of 21,18, and 13 color components of the GC content. The random forest will generate the corresponding decision tree based on these features. At each node, the feature selected is the one that maximizes the information gain or minimizes the Gini impurity, which can be expressed as:

$$\text{Split feature at node} = \arg \max_{f\in\{f_1,f_2,\ldots,f_m\}} \text{Gain}(f) \tag{2}$$

Where $f_1, f_2, \ldots, f_m$ are the m selected features, and Gain (f) is the increase in the selected feature's gain.

Step 3: Prediction for Each Tree

For a classification problem, the prediction value for each decision tree T_j is obtained by averaging the values through the tree's path from root to leaf node:

Let $y_j(x)$ be the predicted output for input sample x from tree T_j, then the final predicted value is the average of the predictions from all trees:

$$\hat{y}(x)=\frac{1}{N_{\text{trees}}}\sum_{j=1}^{N_{\text{trees}}}\hat{y}_j(x) \tag{3}$$

For a classification problem, the random forest determines the final classification by performing a majority vote based on each tree's prediction:

$$\hat{y}(x)=\arg\boxed{f0}\max_{c\in\{1,2,\ldots,C\}}\sum_{j=1}^{N_{\text{trees}}} 1\left(\hat{y}_j(x)=c\right) \tag{4}$$

Where $1(\hat{y}(x) = c)$ is the indicator function. If tree T_j 's prediction for x is class c, it contributes 1, otherwise, it contributes 0.

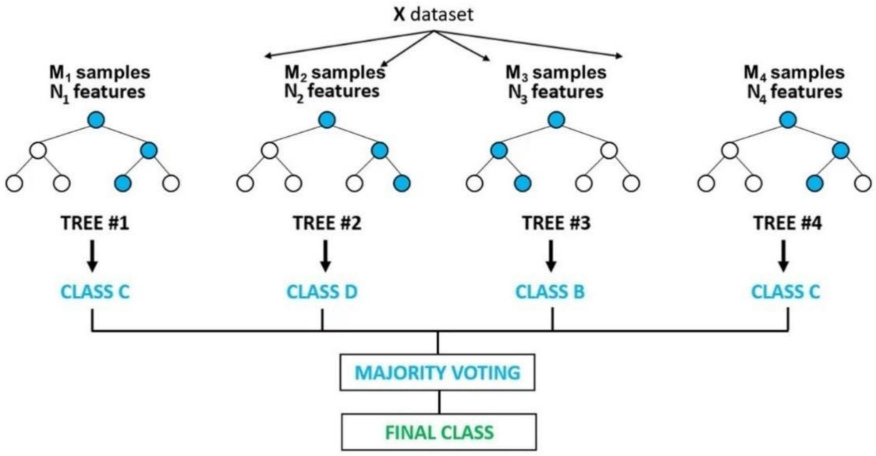

Fig. 3: Diagram of the Random Forest Model.

4.2 Establishment and Calculation of Performance Metrics

(1) Establishment and Calculation of the Confusion Matrix

The Confusion Matrix is a classification model used to evaluate the performance of a model.

By displaying the model's predictions across different categories, it helps analyze the model's performance in each category. It is mainly used in machine learning and statistics for binary or multi-class classification problems.

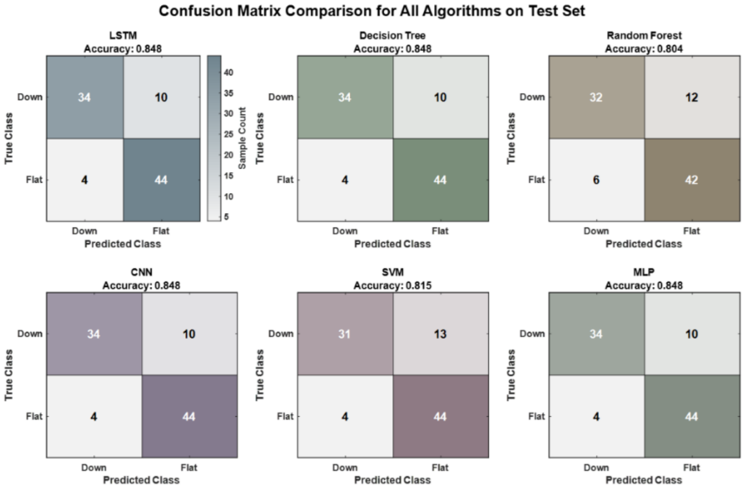

Fig. 4: Confusion Matrix Image.

(2) Accuracy

Accuracy is a fundamental metric for evaluating machine learning models, representing the proportion of correctly predicted samples out of all samples. It is an overall performance metric that reflects the model's effectiveness in classification tasks. The formula for calculating accuracy is as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

In this, TP is the number of true positives, which refers to the number of instances correctly predicted as positive (Abnormal fetus) by the model. TN is the number of true negatives, which refers to the number of instances correctly predicted as negative (Normal fetus) by the model [7]. FP is the number of false positives, which refers to the number of instances incorrectly predicted as positive (Normal fetus misclassified as abnormal fetus) by the model. FN is the number of false negatives, which refers to the number of instances incorrectly predicted as negative (Abnormal fetus misclassified as normal fetus) by the model [8], [9], [10].

(3) Precision

Precision is the proportion of actual positive samples among all the samples predicted as positive by the model. The higher the precision, the greater the proportion of true positives among the predicted positives. The formula for calculating precision is:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

(4) Recall

Recall measures the proportion of actual positive samples that are correctly predicted as positive by the model. A higher recall indicates that the model performs well in identifying positive samples. The formula for calculating recall is:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

(5) F1 score

The F1 score is the harmonic means of Precision and Recall, considering the model’s performance when predicting positive samples. Especially in situations with class imbalance, the F1 score can effectively avoid bias toward a single class and provide a more comprehensive evaluation. The formula for F1 is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Where Precision refers to the proportion of true positives among the predicted positives, and Recall refers to the proportion of actual positives that are correctly predicted by the model.

Table. 1: Training Set Evaluation Metrics Results.

Metric Type	LSTM	Decision Tree	Random Forest	CNN	SVM	MLP
Accuracy	0.8365	0.9128	0.9768	0.7289	0.8828	0.8460
Precision	0.8357	0.9141	0.9770	0.7423	0.8823	0.8451
Recall	0.8315	0.9089	0.9760	0.7085	0.8795	0.8416
F1	0.8332	0.9110	0.9765	0.7101	0.8807	0.8431

Table. 2: Test Set Evaluation Metrics Results.

Metric Type	LSTM	Decision Tree	Random Forest	CNN	SVM	MLP
Accuracy	0.8913	0.8587	0.8478	0.7500	0.8804	0.9130
Precision	0.9068	0.8586	0.8472	0.7715	0.8919	0.9229
Recall	0.8851	0.8574	0.8472	0.7397	0.8749	0.9084
F1	0.8332	0.8579	0.8472	0.7389	0.8780	0.9115

Table. 3: Validation Set Evaluation Metrics Results.

Metric Type	LSTM	Decision Tree	Random Forest	CNN	SVM	MLP
Accuracy	0.8152	0.9128	0.8261	0.7065	0.8152	0.8261
Precision	0.8288	0.9141	0.8324	0.7299	0.8288	0.8324
Recall	0.8106	0.9089	0.8229	0.6989	0.8106	0.8229
F1	0.8115	0.9110	0.8240	0.6934	0.8115	0.8240

(6) AUC-ROC curve

The AUC-ROC curve is used to evaluate the performance of a classification model. AUC (Area Under the Curve) represents the area under the ROC curve. The closer the AUC value is to 1, the better the model's performance. The ROC curve shows the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different threshold values.

In the detection of pregnant women and female fetuses, determining fetal abnormalities is critical because neither pregnant women nor female fetuses carry the Y chromosome. By combining factors such as the aneuploidy of chromosomes 21, 18, and 13, X chromosome, Z-values, GC content, read counts, related ratios, and BMI, an abnormality detection model can be developed. This model is evaluated using the ROC curve and AUC value to optimize the accuracy and recall of detection results. The decision thresholds are adjusted based on different thresholds to improve the effectiveness of abnormality detection.

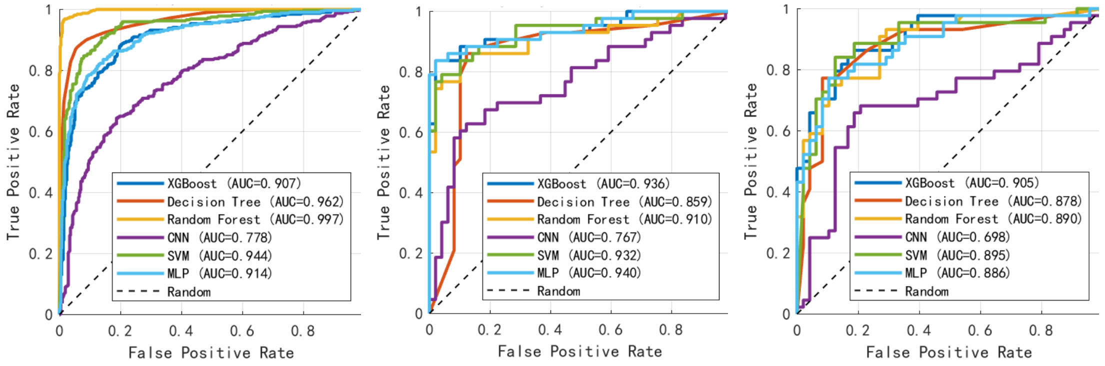

Fig. 5: The ROC Curve Comparison of Different Machine Learning Models.

The ROC curve comparison shows the AUC (Area Under Curve) values for each model, with three separate plots corresponding to the training set, validation set, and test set results. In the training set, Random Forest performs the best with an AUC value of 0.997, higher than other models such as Decision Tree (AUC = 0.962) and XGBoost (AUC = 0.907). In the validation set, Random Forest (AUC = 0.910) still maintains its advantage, approaching XGBoost (AUC = 0.936) and MLP (AUC = 0.940), and is significantly better than CNN (AUC = 0.767). In the test set, Random Forest still shows strong predictive ability (AUC = 0.890). The ROC curve shows that models such as Random Forest and MLP exhibit high classification performance, with AUC values close to or exceeding 0.90, demonstrating strong discriminative power. In contrast, CNN performs relatively weakly, with AUC values not exceeding 0.80. Therefore, this paper selects Random Forest as the abnormality detection model.

5 CONCLUSION

This paper systematically studies the detection of abnormal female fetuses using various machine learning methods. By employing data augmentation and stratified sampling, the issues of class imbalance and uneven sample distribution are effectively addressed. In model comparison, Random Forest performs the best, offering both high accuracy and strong stability. After optimization through grid search and cross-validation, its performance is further enhanced, and it demonstrates good robustness in sensitivity testing. The research results show that a comprehensive strategy involving data preprocessing, model optimization, and performance evaluation can significantly improve the accuracy and application value of abnormal female fetus detection. This study not only provides a reliable technical reference for clinical prenatal screening but also offers methodological support and practical insights for the application of artificial intelligence in medical detection.

REFERENCES

[1] Tímea Csákvári, Diána Elmer, Krisztina Palkovics, Luca Fanni Sántics Kajos, Bettina Kovács, Kálmán Kovács... & Imre Boncz. (2025). Trends and Projections of the Prevalence of Diabetes Mellitus in Pregnancy and Fetal–Neonatal Metabolic Disorders, 2010–2035: A N

ationwide Population-Based Study from Hungary. Journal of Clinical Medicine,14(16),5740-5740. DOI: https://doi.org/10.3390/jcm14165740

[2] Ji Eun Hong, Yeon Eun Kim, Yun Soo Kang, Dong Hyeok Choi, So Hyun Ahn & Jeon gshin An. (2025). SMOTE-augmented machine learning model predicts recurrent and meta static breast cancer from microbiome analysis. Scientific Reports,15(1),33096-33096. DOI: https://doi.org/10.1038/s41598-025-16790-z

[3] Akash Chauhan & Indrajeet Kumar. (2025). Deep feature extraction and optimized VGG1 6-SVM architecture for breast cancer characterization. Discover Computing,28(1),208-208. DOI: https://doi.org/10.1007/s10791-025-09736-6

[4] Yesim Yekta Yuruk. (2025). Uncover This Tech Term: Random Forest. Korean journal of radiology,26(10),998-1001. DOI: https://doi.org/10.3348/kjr.2025.0800

[5] Molly Asher, Yannick Oswald & Nick Malleson. (2025). Understanding pedestrian dynami cs using machine learning with real-time urban sensors. Environment and Planning B: Urb an Analytics and City Science,52(8),1994-2017. DOI: https://doi.org/10.1177/23998083251319058

[6] Zhao, T., Chen, G., Suraphee, S., Phoophiwfa, T., & Busababodhin, P. (2025). A hybrid TCN-XGBoost model for agricultural product market price forecasting. *PLoS One*, *20*(5), e 0322496. DOI: https://doi.org/10.1371/journal.pone.0322496

[7] Aliasghar Bazrafkan, Hannah Worral, Nonoy Bandillo & Paulo Flores. (2025). Multispectr al data and random forest model outperform deep learning in predicting lentil maturity usi ng UAS imagery. Journal of Agriculture and Food Research,23,102202-102202. DOI: https://doi.org/10.1016/j.jafr.2025.102202

[8] Luigi Lavazza, Sandro Morasca & Gabriele Rotoloni. (2025). Software Defect Prediction e valuation: New metrics based on the ROC curve. Information and Software Technology,18 7,107865-107865 DOI: https://doi.org/10.1016/j.infsof.2025.107865

[9] Bruno X Ferreira, Alline V B de Oliveira, João Cajaiba, Vinicius Kartnaller & Brunno F Santos. (2025). Machine learning models for measurement of pH using a low-cost image analysis strategy. Measurement Science and Technology,36(9),096013-096013. DOI: https://doi.org/10.1088/1361-6501/adffa0

[10] Chenglong Yao, Yinglan A, Guoqiang Wang, Baolin Xue, Jin Wu & Xianglong Dai. (202 5). Evaluation of grassland biomass and driving factors in the Hailar river basin based on random forest model. Journal of Cleaner Production,526,146590-146590. DOI: https://doi.org/10.1016/j.jclepro.2025.146590