Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 2, 2026

A Hybrid Data-Driven Machine Learning Approach for Anomaly Detection

Peiyu Chen [1, *, #], Tairan Zhang [1, #], Yuxuan Duan [2, #]

[1] School of Accountancy, Central University of Finance and Economics, Beijing 100081, China
[2] School of Materials Science and Engineering, Hebei University of Technology, Tianjin 300401, China

***Corresponding author:** cpy2023310607@163.com

#These authors contributed equally.

Abstract: Anomaly detection is a vital application field of machine learning. Given the widespread problems of extreme class imbalance and complex nonlinearity in high-dimensional data, traditional anomaly detection methods can hardly achieve both high-precision identification and model interpretability. Taking non-invasive prenatal testing data as the research background, this paper proposes a two-stage expert model and constructs an anomaly detection strategy integrating statistical analysis and machine learning. After appropriate data oversampling, the random forest algorithm is used to preliminarily verify the reliability of the results, and the recall rate of each expert model on the test set exceeds 0.9. SHAP is further introduced to analyze the main effects and interaction effects, indicating that the Z-score of each chromosome plays a core role in its corresponding model. Combined with index quality control in NIPT theory, this paper innovatively applies SHAP single-feature dependence analysis to explore the stability of quality control indicators, constructs a standard quality control score, and divides samples into high, medium, and low risk regions, with an abnormality rate of only 7.6% in the low-risk region. On this basis, the results of multi-index collaborative quality control are analyzed and predicted.

Keywords: Random Forest; SHAP; Statistical Modeling; Anomaly Detection; NIPT.

1. Introduction

Anomaly detection is an important field in machine learning, with broad application prospects in fields such as healthcare, economics, and computer networks [1-2]. For example, financial fraud detection, satellite data anomaly identification, medical testing, and industrial fault diagnosis [1]. Non-invasive prenatal testing (NIPT) is an important method for identifying chromosomal variations in the medical field.

Lo first discovered cell-free fetal DNA in 1997, providing the theoretical basis for the emergence of NIPT technology [3]. Compared with invasive prenatal diagnosis, it has attracted considerable attention due to its non-invasive nature and many other advantages. In a sample study conducted in Wuhan, China, the medical field adopts whole-genome sequencing of DNA as the data foundation, uses the Z-score determination method to identify positive results, and analyzes the impact of other characteristic variables on the results [4]. In addition, there are some relatively mature representatives in the application of machine learning and statistical modeling to anomaly detection. Petropoulos A et al. utilized machine learning algorithms to construct an early warning system for sovereign debt default [5]. Khan M M et al. achieved efficient anomaly detection of IoT attacks by means of data balancing and dimensionality reduction combined with supervised machine learning [6].

Some studies only use a single indicator for reporting and analysis without considering the comprehensiveness of the results. Moreover, traditional econometric methods have certain limitations in explaining nonlinearity, interaction effects, and other aspects. Therefore, against the background of NIPT, this paper uses the official dataset provided by the China Undergraduate Mathematical Contest in Model (CUMCM), proposes a hybrid strategy of supervised learning and multi-indicator collaborative quality control based on a two-stage expert model, and performs anomaly detection through the hybrid analysis of machine learning and statistical modeling.

2. Data-Driven and Multi-Indicator Collaborative RF Model

2.1. Data-Driven Random Forest Model

2.1.1. SMOTE Algorithm

Data imbalance is a common issue in machine learning. Chromosomally abnormal samples account for only a small minority, leading to low recognition accuracy and failure to capture the data patterns of minority classes during model training [7]. Data augmentation can address this problem. As a mainstream oversampling method, SMOTE generates new synthetic minority samples via linear interpolation. The generated data follow a nonlinear distribution, while excessive oversampling may cause over-generalization. Its principle is shown in Figure 1.

The algorithm flow of SMOTE is basically as follows:

(1) For each minority class sample x, calculate its Euclidean distances to all samples in the minority class set to obtain its k-nearest neighbors.

(2) Set a sampling ratio according to the sample imbalance ratio to determine the sampling multiplier N; for each minority class sample X, randomly select several samples from its k-nearest neighbors, denoted as X_n

(3) For each randomly selected neighbor X_n generate new samples with the original sample X in accordance with the following formula.

$$X_{new} = X + rand(0,1) * X_n, \quad (1)$$

The SMOTE algorithm was adopted to balance the data of four categories (T13, T18, T21, and multi-chromosomal variations) at a 1:3 ratio, and generate synthetic positive

samples, enabling the random forest to better learn the inherent patterns in the data.

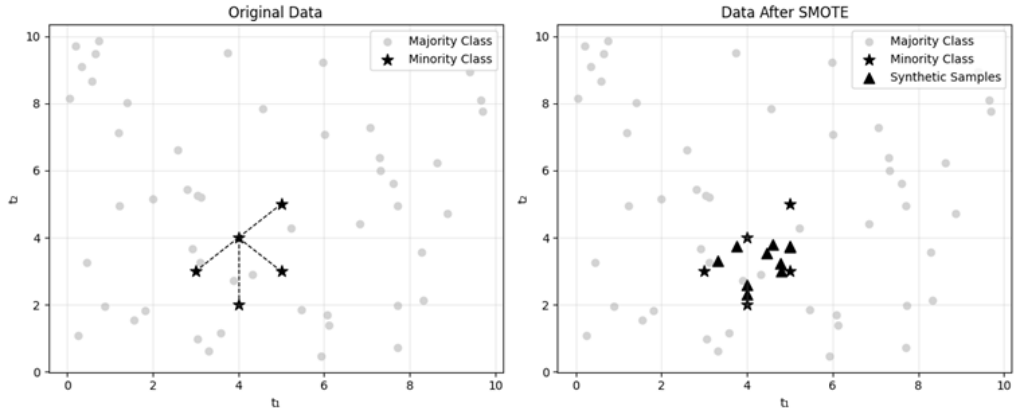


Figure 1. SMOTE structure

2.1.2. Random Forest Algorithm

Random Forest is an ensemble learning algorithm for classification and regression. Based on decision trees, it trains multiple trees simultaneously via Bootstrap resampling, which improves computational efficiency and enhances model diversity. When constructing each tree, a random feature subset is selected at each node, and the optimal feature within the subset is used for splitting. This makes different trees focus on different features and further boosts diversity and generalization. Owing to threshold splitting, random feature selection, and ensemble learning, Random Forest is insensitive to feature scaling and multicollinearity, and reduces overfitting [8]. Its principle and procedure are as follows:

(1) Bootstrap random sampling with replacement is employed to extract n samples from the dataset $\boldsymbol{D}$ as the training subset $\boldsymbol{D_j}$ for a single tree.

(2) Randomly select explanatory variables as the splitting features for the classification subset $\boldsymbol{D_j}$

(3) The features are used to classify the training subset $\boldsymbol{D_j}$ and construct a classification tree $\boldsymbol{h_j(X_j)}$

(4) Repeat steps 1 to 3 k times to construct k classification trees. Based on the classification outputs of each sample in dataset D from the k trees, the final class of each sample is determined by majority voting, thereby yielding the classification result of the random forest model.

A schematic diagram illustrating the working principle of the random forest is shown in Figure 2.

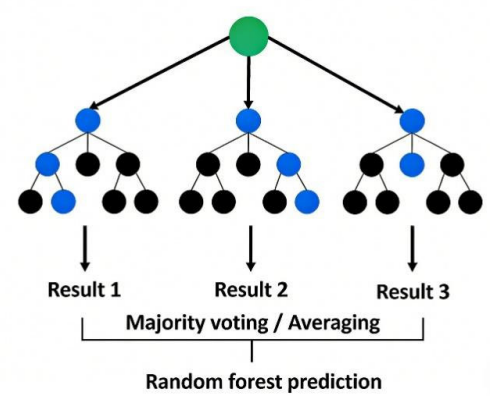


Figure 2. Random Forest structure

In this paper, Random Forest is trained on the augmented dataset. After performance analysis, the classifier is used to explore classification results of balanced data, and the regressor is adopted for risk regression in SHAP analysis.

2.2. SHAP-Based Multi-Indicator Collaborative Quality Control Model

2.2.1. SHAP Algorithm

With the development of machine learning, models have gradually shifted from traditional interpretable approaches to black-box models. Relying only on prediction accuracy is insufficient, while traditional interpretability tools have limitations in explaining the contribution of input features. SHAP is derived from the Shapley value in cooperative game theory [9-10]. It essentially decomposes the model prediction for a single sample into the sum of a base value and the SHAP values of individual features. The detailed process is as follows.

Let $\phi_0 = E[f(X)]$ be the mean of predictions over all samples. For a prediction task with p

features, the predicted value can be expressed as

$$f(x) = \phi_0 + \sum_{i=1}^{p} \phi_i(x), \quad (2)$$

where $\boldsymbol{\phi_i(x)}$ denotes the SHAP value of feature i, calculated based on the concept of Shapley values.

$$\phi_i(x) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)] \quad (3)$$

In the formula, S denotes the feature subset containing feature i, and $f_x(S)$ is the model prediction using only subset S. This value can be decomposed into main effects and interaction effects, which reflect their roles in prediction when the correlation between variables is weak.

2.2.2. Measurement of Quality Control Score

A potential cause of reduced fetal cell-free DNA fraction (FF) is the dilution effect, referring to the relative decrease in cell-free fetal DNA (cffDNA) content caused by elevated maternal cell-free DNA (cfDNA) levels. Any maternal factors that increase cfDNA concentrations—including maternal BMI, age, and conception method—can lower cffDNA levels in maternal peripheral blood, thereby reducing FF.

Previous studies have indicated that one possible cause for the decrease in fetal cell-free DNA fraction is the dilution effect. Fragments with moderate GC content (~40-60%) have the highest amplification and sequencing efficiency, leading to more reads by the sequencer. An excessively high proportion of duplicate reads reduces effective sequencing depth and exacerbates GC bias.

Therefore, this study defines variable quality control based on maternal factors excluding the X-chromosome Z-value, combined with SHAP single-feature dependence results and existing NIPT theory. This study adopts the method from the

mathematical modeling review meeting. For variables with monotonic trends in numerical quality, we truncate samples at the upper or lower 10% quantile and calculate robust standardized Z-scores.

$$Z_{st} = \frac{X - X_{tr}}{\sigma_{tr}}, \quad (4)$$

where X_{tr} denotes the robustly processed mean. For intermediate quality control indicators, the robust standardized Z-score is calculated after computing the distance from the robust mean.

The final quality control score is obtained by summing the three categories according to their respective implications.

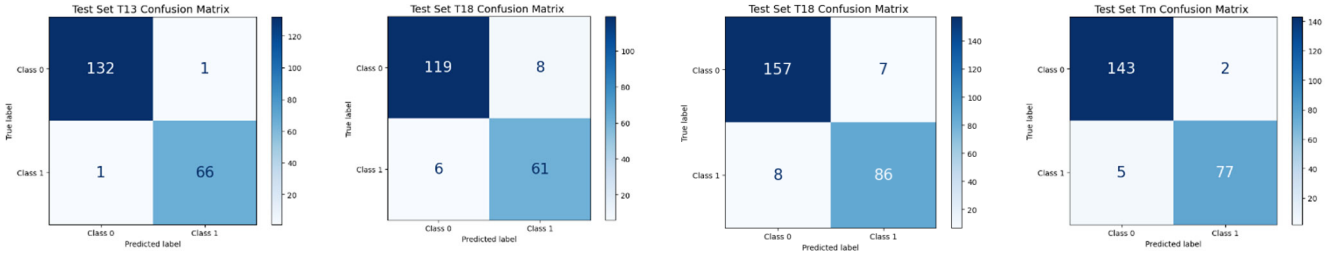
$$Zquality = \sum z(P_i) - \sum z(N_j) + \sum z(M_k), \quad (5)$$

where P_i denotes a positive indicator, N_j a negative indicator, and M_k an intermediate indicator.

3. Results

3.1. Analysis of the Expert Model Results in the First Stage

Firstly, this study conducts preliminary data cleaning. To prevent model overfitting and data leakage during the data balancing process, oversampling is performed on the training set at ratios of 1:2 and 1:3, respectively. The resulting datasets are then used to train random forest classifiers, with parameters set based on empirical rules. The results of the relevant evaluation metrics are presented in Figure 3.



Test set recall:0.985
T13 Expert Model
(a)

Test set recall:0.910
T18 Expert Model
(b)

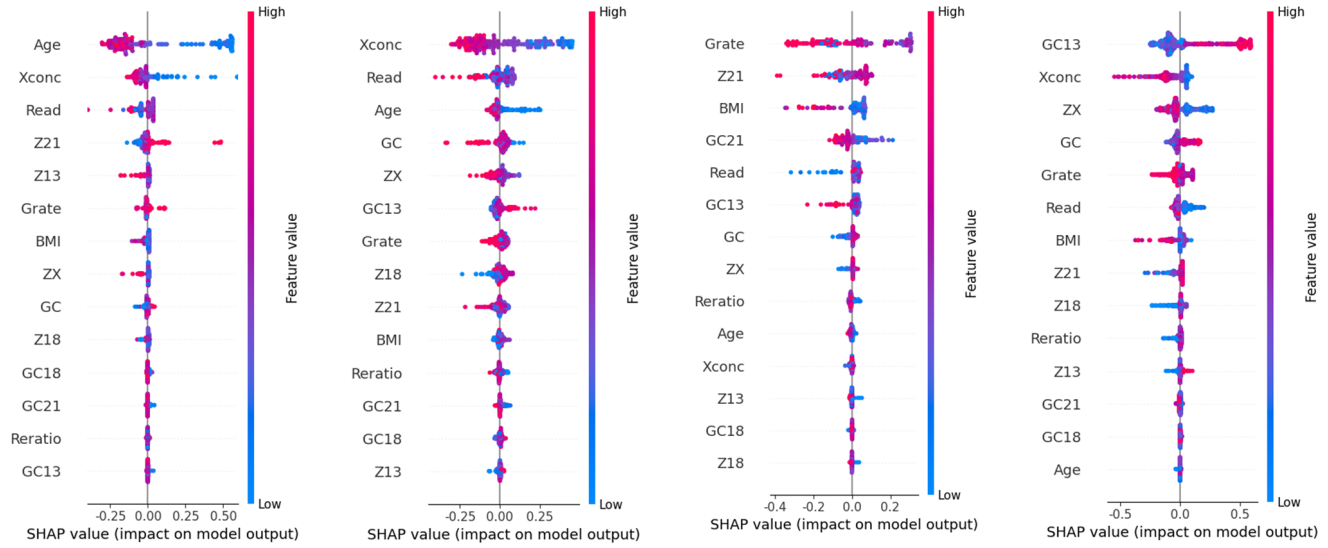
Test set recall:0.915
T21 Expert Model
(c)

Test set recall:0.912
Tm Expert Model
(d)

Figure 3. Identification Evaluation Results of the Expert Model

According to the results shown in the table above, the recall rate on the test set of each expert model is above 0.9, indicating a good model fitting effect. The features can effectively identify aberrant events. The purpose of the first-stage model in this study is to explore variables rather than provide direct interpretation. We will further investigate the mechanism of each variable in the following work.

The objective of this phase is to identify the main effect factors driving nonlinear outcomes. Random forest hyperparameters are set via empirical rules, with fixed training set splits and model training seeds. The best-performing results are selected for the SHAP main effect analysis, as shown in Figure 4.



T13 Expert Model
(a)

T18 Expert Model
(b)

T21 Expert Model
(c)

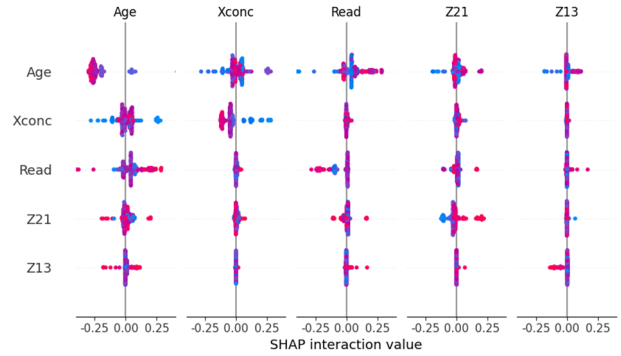
Tm Expert Model
(d)

Figure 4. Main Effect SHAP Results of the Expert Model

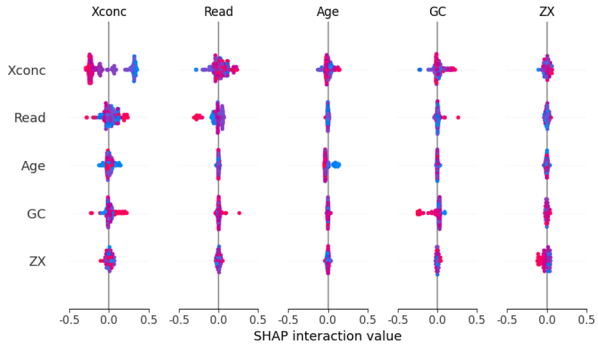
Xconc denotes X-chromosome concentration, Read denotes the number of raw reads, and Reratio denotes the number of remapped reads. From the SHAP main effect analysis of expert models for single-chromosome aneuploidy identification and multi-chromosome aneuploidy identification, the corresponding chromosome Z-score contributes significantly to each type of chromosomal polyploidy variation, ranking first in feature importance among all indicators except maternal variables. For example, in the T13 expert model, the Z-scores of chromosome 21 and chromosome 13 exhibit positive and negative correlations, respectively. Other variables show inconsistent patterns across different expert identification models.

For clarity, Figure 5 only presents the SHAP interaction

effects of the T13 and T18 models.



T13 Interaction Effect
(a)



T18 Interaction Effect
(b)

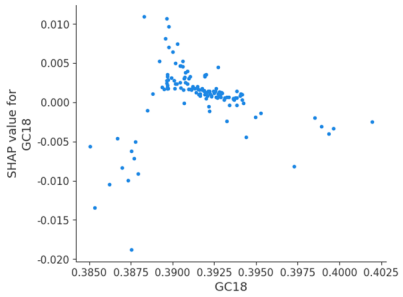
Figure 5. SHAP Interaction Effect Analysis of the Expert Model

The interaction effect analysis based on RF-SHAP shows that key quality control factors in the model, including age, X-chromosome concentration, and DNA fragment ratio, have certain interaction effects. However, the interaction between chromosomal Z-scores is found to be insignificant.

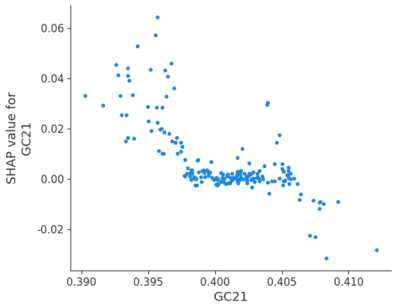
Given the development of NIPT, we perform an exploratory analysis of detection quality by examining the stability of indicators throughout the testing process.

3.2. Results of Multi-Indicator Collaborative Quality Control Analysis

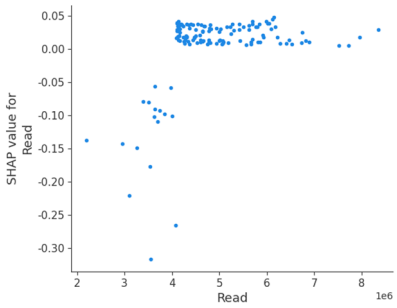
For clarity, this paper presents the single-feature dependence plots of three quality control indicators in the expert model, and the results are shown in Figure 6.



T13 Model GC18 content
(a)



T18 Model GC21 content
(b)



T21 Model DNA raw reads
(c)

Figure 6. Feature Dependence Plot of Quality Control Variables

As can be seen from the figure above, the SHAP value for the single-feature contribution of quality control features is relatively stable within a certain range, while abnormal fluctuations appear beyond the normal range. This confirms the medical reality cited earlier in the paper.

For the quality control of detection performance, this study focuses on the stability of variables and adopts robust standardized Z-score processing based on relevant theories and the identified variables. The final quality control score is calculated by summing the positive and negative items. The quality scores are sorted and truncated at the 30th/70th percentiles and 30th/80th percentiles, respectively, followed by statistical analysis. The results are as shown in Table.1.

Table 1. Collaborative Quality Control Risk Stratification Results

Region	Number	Ratio	Number	Ratio
Low Risk	15	8.2%	9	7.6%
Medium Risk	22	9.5%	28	9.5%
High Risk	29	16.0%	29	16.0%
Truncation	30% 70%		30% 80%	
$Z_{30\%}$=-5.85 $Z_{70\%}$=-2.19 $Z_{80\%}$=-1.31				

The results show that after dividing the samples by the Z-score quality scores, significant differences exist among different groups. Using the 30th and 70th percentiles as the cutoff criteria, the abnormality rate is only 8.2% in the low-risk zone and 9.5% in the medium-risk zone. Samples in the high-risk zone deserve close attention, as they are much more likely to be abnormal.

Based on the proportion of quality control scores, test samples can be classified into high-risk, medium-risk, and low-risk zones. In the high-risk zone, abnormal indicators caused by chromosomal variations may lead to a decline in quality scores, which may in turn negatively affect detection performance and give rise to false-positive results. Such classification contributes to abnormal identification under multivariate adjustment.

4. Conclusions

This study proposes a two-stage framework integrating expert models and exploratory analysis. First, after sample classification and balancing, we employ a random forest for nonlinear fitting. We then conduct variable analysis using the SHAP approach combined with NIPT-related theories, and further perform data stratification via statistical modeling, achieving favorable performance. The innovation of this work lies in incorporating SHAP local feature plots to identify the

stability of feature contributions, thereby integrating the concept of multi-indicator collaborative quality control into anomaly detection. By adopting stepwise hierarchical analysis and modeling based on existing theories instead of relying solely on standalone machine learning, this research provides valuable references for applications involving anomaly detection and identification.

The feasibility and future application potential of the model developed in this study are as follows. Different anomaly detection fields usually have their own theoretical foundations or practical experiences, with varying economic costs across domains. Meanwhile, high-dimensional and complex variables present nonlinear interactive relationships. Integrating theoretical analysis, statistical modeling, and interpretable machine learning analysis can improve the efficiency and interpretability of anomaly detection and identification.

This study has certain limitations, including insufficient data comprehensiveness and potential overfitting in machine learning models. The data domain involved in this paper is highly specialized, and more suitable domain-specific rules may exist in real-world scenarios instead of big data methods. Therefore, the modeling proposed herein only provides a reference for machine learning modeling ideas in anomaly detection, rather than a direct application system.

Acknowledgement

All the authors contributed equally.

References

[1] Nassif, A. B., Talib, M. A., Nasir, Q., & Albadarneh, M. (2021). Machine learning for anomaly detection: A systematic review. IEEE Access, 9, 78658-78700.

https://doi.org/10.1109/ACCESS.2021.3083060

[2] Kang, M. (2018). Machine learning: Anomaly detection. In Prognostics and health management of electronics: Fundamentals, machine learning, and the internet of things (pp. 131-162). Wiley.

[3] Benn, P., Cuckle, H., & Pergament, E. (2013). Non-invasive prenatal testing for aneuploidy: Current status and future prospects. Ultrasound in Obstetrics & Gynecology, 42(1), 15-33. https://doi.org/10.1002/uog.12513

[4] Huang, Q., Xu, Q., Chen, M., Zou, Y., Li, S., & Liang, B. (2025). Application of non-invasive prenatal testing for fetal chromosomal disorders in low-risk pregnancies: A follow-up study in central China. Frontiers in Genetics, 16, 1574775. https://doi.org/10.3389/fgene.2025.1574775

[5] Petropoulos, A., Siakoulis, V., & Stavroulakis, E. (2022). Towards an early warning system for sovereign defaults leveraging on machine learning methodologies. Intelligent Systems in Accounting, Finance and Management, 29(2), 118-129. https://doi.org/10.1002/isaf.1516

[6] Khan, M. M., & Alkhathami, M. (2024). Anomaly detection in IoT-based healthcare: Machine learning for enhanced security. Scientific Reports, 14(1), 5872.

https://doi.org/10.1038/s41598-024-56126-x

[7] Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. Applied Soft Computing, 143, 110415. https://doi.org/10.1016/j.asoc.2023.110415

[8] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. Babylonian Journal of Machine Learning, 2024, 69-79.

https://doi.org/10.58496/BJML/2024/007

[9] Lundberg, S. M., & Lee, S. I. (2017). Consistent feature attribution for tree ensembles. arXiv, arXiv:1706.06060. https://doi.org/10.48550/arXiv.1706.06060

[10] Bernal, L., Rastelli, G., & Pinzi, L. (2025). Improving machine learning classification predictions through SHAP and features analysis interpretation. Journal of Chemical Information and Modeling, 65(21), 11716-11732.

https://doi.org/10.1021/acs.jcim.5c00894