

Machine Learning Methods and Their Applications in Biomedical Sensors

Jingyun Yang

{2413530329@st.usst.edu.cn}

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, China

Abstract. In recent years, biomedical sensors have achieved significant breakthroughs. Machine learning technology has facilitated advancements in sensor applications by improving signal quality and automatically extracting features. This article systematically reviews relevant research, focusing on three machine learning methods, including Random Forest Regression (RFR), Partial Least Squares (PLS), and Polynomial Regression (PR), and provides a comprehensive comparison in terms of performance, advantages, disadvantages, and application scenarios. Research indicates that RFR possesses strong anti-overfitting capabilities and it's suitable for malaria detection, gait disorder identification, and other scenarios; PLS efficiently processes high-dimensional collinear data and performs well in hemoglobin testing and cfDNA analysis; PR adapts to nonlinear relationships and offers fast computation, meeting the needs of tuberculosis and dopamine detection. This study offers insights for researchers. In the future, with developments of flexible integration, multi-modal fusion, and non-invasive continuous monitoring technology, machine learning methods will be further refined to accommodate higher-quality sensing data, leading to broader application of biomedical sensors.

Keywords: Biomedical science, Random Forest Regression, Partial Least Squares, Polynomial Regression.

1 Introduction

In recent years, biomedical sensors have achieved significant breakthroughs. The incorporation of nanomaterials has markedly enhanced the sensitivity and stability of sensors. For instance, graphene, with its high conductivity and flexibility, has been employed in the development of single-molecule-level biosensors [1]. Artificial intelligence algorithms are deeply integrated throughout the entire sensor signal processing workflow. For instance, convolutional neural networks (CNN) are employed for feature extraction from medical imaging sensor data [2]. Implantable sensors are evolving from "short-term monitoring" to "long-term intervention". For instance, the self-healing and biodegradable bladder pressure sensor developed by the Tianjin University team enables real-time monitoring of abnormal bladder activity [3].

Machine learning achieves breakthroughs in biomedical sensor applications by increasing raw signal quality and automatically extracting features. Random Forest Regression conducts multi-parameter physiological monitoring and noise-reduced signal prediction; Partial Least Squares processes multi-wavelength signals such as near-infrared and Raman spectroscopy and subsequently establishes regression models linked to pathological indicators; Polynomial Regression can compensate for non-linear impacts of environmental factors such as temperature and pH on sensor signals, thereby improving the stability of detection. Due to the scarcity of review literature integrating biomedical sensors with machine learning methods published in recent years, this research's main contribution lies in supplementing such reviews by synthesising relevant literature from the past two years, offering insights for researchers.

This research summarises relevant literature and compares three machine learning approaches (Polynomial Regression, Partial Least Squares regression, and Random Forest Regression) in biomedical sensors. The comparison focuses on three specific aspects: performance, advantages, and disadvantages. It further outlines the characteristics of each method and highlights their application in biomedical sensors through representative achievements.

2 Introduction to typical methods

2.1 Random Forest Regression(RFR)

Random Forest Regression refers to constructing multiple independent decision trees and employing a "voting/averaging" way to reduce the overfitting risk of a tree, increasing prediction stability. This method possesses robust resistance to overfitting and noise, requires no data preprocessing, and handles non-linear relationships and multivariate interactions. The fundamental formulas are as follows:

$$\bar{y}(x) = \frac{\sum_{b=1}^B f_b(x)}{B} \quad (1)$$

$\bar{y}(x)$ is the model's final prediction value, used to fuse outputs from multiple base learners to improve stability; $f_b(x)$ is the forecast result of the b th decision tree; B represents the number of decision trees.

$$P = 1 - \left(1 - \frac{1}{N}\right)^n \quad (2)$$

P is the probability that an event occurs at least once; N is the total number of trials; n is the number of times the event occurs.

$$G(m) = \sum_{k=1}^K PMK(1 - PMK) \quad (3)$$

$G(m)$ is a diversity metric for decision trees; PMK denotes the prediction value of a single decision tree for a sample, reflecting the predictive outcome of an individual tree.

$$M = \frac{\sum_{i=1}^N (R_p - R_0)}{N} \quad (4)$$

M denotes the average error, reflecting the overall direction and magnitude of deviation between the random forest's predicted values and the true values; N represents the sample size; R_p denotes the predicted value; R_0 denotes the true value.

$$ME = \frac{\sum |R_p - R_0|^2}{N} \quad (5)$$

ME denotes the mean square error.

$$RE = \sqrt{\frac{\sum (R_p - R_0)^2}{N}} \quad (6)$$

RE is the root mean square error.

$$R^2 = \frac{\sum_{i=1}^N (\text{Predicted Target Value}_i - \text{Actual Target Value}_i)^2}{\sum_{i=1}^N (\text{Actual Target Value}_i - \text{Average Target Value})^2} \quad (7)$$

R^2 denotes the coefficient of determination; *Predicted Target Value_i* represents the model's predicted value for the *i*th sample; *Actual Target Value_i* denotes the actual value for the *i*th sample; *Average Target Value* signifies the mean of the actual values.

2.2 Partial Least Squares(PLS)

Partial Least Squares achieves regression modelling for multiple dependent variables and multiple independent variables by extracting latent variables whilst simultaneously maximising both the covariance between independent and dependent variables and the variance within independent variables. This method effectively handles strong inter-variable correlations whilst reducing data dimensionality and preserving key information.

2.3 Polynomial Regression(PR)

Polynomial Regression refers to approximating non-linear data trends using polynomial functions. This method offers a straightforward model with rapid computational speed, enabling intuitive capture of non-linear relationships between variables. The fundamental formulas go as follows:

$$Cov(X, Y) = \frac{\sum_{i=1}^N [(x_i - \bar{X})(y_i - \bar{Y})]}{N} \quad (8)$$

$Cov(X, Y)$ denotes the covariance, measuring the direction and strength of the linear relationship between two variables X and Y ; N represents the sample size; x_i and y_i denote the *i*th sample values of variables X and Y respectively; $\bar{X}$ and $\bar{Y}$ denote the sample means.

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad (9)$$

$\rho(X, Y)$ is the standardised form of the covariance; σ_X and σ_Y denote the standard deviations of variables X and Y respectively.

$$\rho(R(X), R(Y)) = \frac{Cov(R(X), R(Y))}{\sigma_{R(X)} \sigma_{R(Y)}} \quad (10)$$

After applying transformations such as standardisation and feature extraction to variables X and Y in equation (9), the transformed variables $R(X)$ and $R(Y)$ are obtained. The coefficient of correlation is then calculated to analyse the linear correlation between these transformed variables.

$$R^2 = \frac{\sum_{i=1}^N (\text{Predicted Target Value}_i - \text{Actual Target Value}_i)^2}{\sum_{i=1}^N (\text{Actual Target Value}_i - \text{Average Target Value})^2} \quad (7)$$

3 Comparison of the application of typical methods

3.1 Random Forest Regression(RFR)

Jacob Wekalao and others employed RFR for detecting early-stage malaria. At incident angles ranging from 0° to 80° , the scatter plot exhibited R^2 ranging from 99% to 100%, while the heatmap showed that R^2 was 93% to 100%. When GCP values ranged between 0.1 and 0.9 eV, the scatter plot showed that R^2 was 43% to 100%, with thermograms achieving R^2 between 95% and 100% [4]. Jacob Wekalao employed RFR for hemoglobin detection, yielding R^2 ranging from 100% for both scatter plots and thermograms at GCP ranging from 0.1 to 0.9 eV; similarly, square ring resonators achieved R^2 ranging from 100% for scatter plots and thermograms at wavelengths between 1 and $3\ \mu\text{m}$ [5]. The more closely R^2 approaches 100%, the higher the fit between the machine learning model's predicted values and actual measurements is, indicating stronger predictive capability. Manting Chen and others achieved early detection and precise identification of gait disorders, with the decision tree model demonstrating optimal performance. The gait disorder recognition accuracy, sensitivity, and AUC all reached 100% [6], signifying the model's ability to distinguish target categories without error. These biomedical sensors can identify gait disturbances, providing an efficient solution for gait assessment in elderly populations and disease monitoring in Parkinson's patients; they can detect malaria, addressing the insufficient sensitivity of traditional microscopy; and they can measure hemoglobin levels, aiding rapid diagnosis of conditions such as anaemia and heart disease.

RFR possesses the following advantages.

Initially, it is blessed with high robustness and resistance to overfitting. When detecting early-stage malaria, the R^2 value remained stable above 93% even though significant variations occurred in GCP or incident angle; during hemoglobin detection, the R^2 consistently achieved 100% [4]. Overfitting fluctuations weren't observed, demonstrating the model's high robustness.

Besides, it possesses a strong ability for multi-parameter collaborative analysis. Sensor performance is influenced by multi-parameter coupling; a comprehensive basis for structural optimisation is provided because the impact of multiple inputs on the output is simultaneously analysed. When detecting early-stage malaria, the inputs included the angle of incidence, GCP, and resonator dimensions [4].

However, RFR has the following disadvantages:

RFR has higher computational cost. When detecting early-stage malaria, RFR training time was 12.5 hours, whereas traditional methods, Support Vector Regression (SVR) and Multiple Linear Regression (MLR), required 1.8 hours and 1.5 hours respectively under identical hardware conditions [4]. The computation time for RFR is 7 to 8 times that of traditional models.

3.2 Partial Least Squares(PLS)

Renjie Fang and others detected hemoglobin. PLS is combined with wavelet packet-fuzzy shrinkage denoising (WPT-FS) and whale optimisation algorithm (WOA), modelling near-infrared spectroscopy (NIRS) data. A root mean square error (RMSEP) was 2.0409 and a coefficient of determination (R_p^2) was 0.9746 in the prediction set. Without preprocessing, RMSEP was 3.2593 and R_p^2 was 0.9644. Traditional preprocessing methods airPLS yielded the results that RMSEP was 2.4511 and R_p^2 was 0.9723. The smaller RMSEP was and the closer to

1 R^2 was, the higher the predictive accuracy was. Therefore, the model combined with PLS, WPT-FS, and WOA demonstrated the strongest performance [7]. Manting Chen and others conducted cfDNA detection by combining PLS with the Meanshift algorithm to process dark-field microscopy (DFM) image data, resulting in quantitative analysis of the BRCA-1 gene (a cfDNA biomarker) in breast cancer. The method demonstrated excellent linearity across a concentration range of 0.1 fM to 500 pM, with a detection time of only 40 minutes [8]. These biomedical sensors are used for non-invasive quantitative analysis of whole blood hemoglobin, addressing problems such as water molecule absorption interference and noise affecting accuracy in traditional NIRS detection, which makes them suitable for rapid diagnosis of conditions such as anemia and cardiovascular diseases. They also achieve highly sensitive detection of the BRCA-1 gene associated with breast cancer without requiring nucleic acid amplification, making them applicable for early-stage liquid biopsy in tumour detection.

PLS offers the following advantages.

Initially, it handles high-dimensional, highly collinear data efficiently. Spectral data typically contains hundreds to thousands of variables with strong collinearity. PLS reduces dimensionality by extracting principal components while preserving information. When detecting hemoglobin, the raw near-infrared spectral data contained 3,632 wavenumber points, with co-linear interference between water molecule absorption peaks and hemoglobin signals. PLS reduced dimensions by extracting latent variables (LVs). After screening effective wavenumber nodes with WOA, the model achieved an R^2 of 0.9746, outperforming the R^2 (0.9644) of the unprocessed data [7]. This demonstrates PLS's ability to resolve high-dimensional collinearity issues, thereby avoiding the curse of dimensionality.

Besides, it embraces low computational complexity. The linear regression nature of PLS results in little computational demand. When detecting cfDNA, the entire workflow, combined with Meanshift image analysis, could be completed within 40 minutes, with PLS modelling time negligible [8].

However, PLS has the following disadvantages.

Initially, it possesses inadequate adaptation to nonlinear relationships. When sample components possess nonlinear correlations, the fitting accuracy of PLS decreases significantly. During cfDNA detection, when cfDNA concentrations overtook 500 pM, the “gold nanoparticle monomer ratio-cfDNA concentration” showed a nonlinear saturation trend, with R^2 dropping below 0.8 [8], preventing accurate quantification under nonlinear conditions.

Besides, it possesses strong dependence on data preprocessing. Noise in raw spectra severely impacts PLS performance. For hemoglobin detection, the original R^2 was 0.9644, increasing to 0.9746 after processing [7]. If preprocessing was skipped, water absorption noise and redundant wavenumbers in the spectrum drastically reduced PLS modelling accuracy, demonstrating PLS's strong reliance on preprocessing.

3.3 Polynomial Regression (PR)

Amuthakkannan Rajakannu and others detected tuberculosis with PR, achieving R^2 between 98% and 99% at incident angles between 0° and 80° . R^2 was 97% to 100% when GCP ranged from 0.1 to 0.9 eV. When silver-based circular ring resonators measured 0.5 to 2.5 μm and black phosphorus square ring resonators measured 16 to 18 μm in inner side and 17 to 19 μm in outer side, R^2 was 98% to 100%. When the tuberculosis sample refractive index ranged from 1.343 to 1.351 RIU, R^2 was 97% to 99% [9]. Jacob Wekalao and others detected dopamine with PR,

showing that R^2 was 100% at incident angles ranging from 0° to 80° ; R^2 was 96% to 100% when GCP ranged from 0.1 to 0.9 eV; R^2 was 100% under the circumstance of resonator sizes ranging from 2 to 4.5 μm ; when dopamine refractive index ranged from 1.256 to 1.309 RIU, R^2 was 98% to 99% [10]. These metrics prove that the PR model maintains high predictive accuracy despite variations in sample refractive index and resonator-related parameters in general. These biomedical sensors enable non-invasive quantitative analysis of hemoglobin in whole blood, addressing issues such as water molecule absorption interference and noise affecting accuracy in traditional NIRS detection. Consequently, they are suitable for rapid diagnosis of conditions such as anemia and cardiovascular diseases. They also detect the BRCA-1 gene in breast cancer sensitively without requiring nucleic acid amplification, used for early-stage liquid biopsy in tumors. These biomedical sensors fulfill highly sensitive detection of tuberculosis markers and meet the precise quantification needs for dopamine concentration.

PR offers the following advantages:

Initially, it efficiently handles nonlinear relationships. The electromagnetic response of terahertz sensors to input variables is predominantly nonlinear. When detecting dopamine, R^2 was 99% in fitting the dopamine refractive index, whereas linear regression showed R^2 was 82% [10].

Besides, it has low computational complexity. The model complexity of PR is solely related to the polynomial degree without the need for GPU acceleration. Training and prediction times are shorter than many other machine learning models, suitable for real-time optimisation at the sensor site. When detecting dopamine, the parameter dimension was low ($d=3$), the sample size was small ($N<100$), and the polynomial order was moderate ($K\leq 10$) [10], which indirectly demonstrated PR's minimal computational load.

However, PR has the following disadvantages:

Initially, higher-order polynomials are prone to overfitting. When the polynomial degree is excessively high, the model tends to fit noise rather than true patterns, particularly when small sample sizes or highly fluctuating input variables exist. In tuberculosis detection, at GCP was 0.1 eV, a 6th-degree polynomial yielded R^2 was 0.98, while a 10th-degree polynomial yielded R^2 was 0.90 [9].

Besides, it has insufficient generalisation capability in small-sample scenarios. During tuberculosis detection, R^2 was 0.92 and RMSE was 0.003 under a small sample (5 gradients). After BEADS baseline correction and 3σ outlier removal, R^2 was improved to 0.97 and RMSE decreased to 0.001, effectively suppressing error diffusion [9].

3.4 Comparison of three methods

Table 1. Performance of Three Methods Under Different Indicators

Methods		Random Regression (RFR)	Forest	Partial Squares	Least (PLS)	Polynomial Regression (PR)
Overfitting resistance		strong		/		weak
Non-linear Adaptability	Relationship	/		weak		strong

Computational complexity	high	low	low
--------------------------	------	-----	-----

The three methods are compared based on several indicators, as shown in Table 1. RFR exhibits strong resistance to overfitting but has high computational complexity; PLS has low computational complexity but weak adaptability to nonlinear relationships; PR demonstrates strong adaptability to nonlinear relationships and low computational complexity but weak resistance to overfitting.

4 Challenges and prospects

4.1 Challenges

Even though Random Forest Regression, Partial Least Squares, and Polynomial Regression can achieve breakthroughs in biomedical sensor applications, some technical challenges still exist. Initially, the quality of data is difficult to adapt to. Biomedical sensor signals are susceptible to noise interference. For example, human movement and environmental temperature fluctuations can distort data. All three methods possess limited robustness, so they need more complex preprocessing. In addition, significant variability exists between samples and individuals. Physiological parameters such as heart rate and blood glucose concentration vary considerably across individuals. When sample sizes are insufficient or distributions are uneven, model generalisation capabilities decline substantially, making it difficult to cover diverse clinical scenarios. A lack of clinical interpretability exists in biomedical sensor applications. All three methods feature "black-box", particularly for Random Forest Regression and Polynomial Regression, making it difficult to explain the physiological significance of predictions. This fails to meet clinical diagnosis needs for causal logic.

4.2 The future development of biomedical sensors

The future development of biomedical sensors will present three core trends. The first is flexible integration, with medical devices such as flexible electronic skin patches and flexible neural probes currently under development and soon to be deployed. The second is multimodal fusion. Multimodal fusion systems follow a three-tier architecture of perception-processing-decision-making. Technologies across these tiers collaborate to achieve precise medical monitoring, with projected adoption rates reaching 60% in ICUs of top-tier hospitals and diagnostic accuracy improving by 30%. The third trend is non-invasive continuous monitoring. Within the next three years, non-invasive blood glucose monitoring technology will achieve 95% accuracy while reducing costs to one-third of traditional CGM systems. Biolinq's FDA-certified needle-free blood glucose sensor will achieve large-scale commercialisation.

5 Conclusions

This paper systematically summarises findings on three machine learning methods: Random Forest Regression (RFR), Partial Least Squares (PLS), and Polynomial Regression (PR) in biomedical sensors. The comparative analysis is conducted based on performance, advantages,

disadvantages, and application scenarios. Findings reveal that RFR possesses strong resistance to overfitting and significant advantages in multi-parameter analysis, making it suitable for malaria detection and gait disorder recognition, though it incurs higher computational cost. PLS efficiently tackles high-dimensional collinear data with low computational complexity, excelling in hemoglobin detection and cfDNA analysis, but it struggles with nonlinear relationship fitting. PR effectively captures nonlinear relationships with rapid computation, meeting demands for tuberculosis and dopamine detection, but it suffers from high-order overfitting and insufficient small-sample generalisation. This study supplements the literature on integrating biomedical sensors with machine learning methods, providing critical guidance for researchers to select suitable algorithms and optimise sensor performance. It also prompts the development of high-precision biomedical detection technologies. In the future, with the breakthroughs in flexible integration, multimodal fusion, and non-invasive continuous monitoring, machine learning methods will further optimise to adapt to higher-quality sensor data. Biomedical sensors are poised to deliver stronger technical support for precision medicine.

References

- [1] Wu, S., Ma, H., Chen, X., Zhong, W., Gu, Y., Miao, Y., An, Y.: Collard-like Bi₂S₃@Au nanocomposites-based label free electrochemical immunosensor for quantitative detection of CA19-9. *Talanta* 285, 127299 (2024)
- [2] Pathak, D.M., Tewari, T.K., Gupta, S., Kalra, M.K., Saba, L., Laine, A.F., Suri, J.S.: Optimal feature selection for medical image fusion using deep learning with transformer. *Biomedical Signal Processing and Control* 111, 108377 (2025)
- [3] Yang, K., Du, L., Yu, Y., Zhang, T., Zhao, W., Liu, X., Zhang, L., Zhang, H., Lan, X., Tian, S., Sun, B., Zhang, L., Yang, J.: Self-healing, degradable, and implantable tissue-like sensor for monitoring of abnormal bladder activity. *Biosensors and Bioelectronics* 288, 117767 (2025)
- [4] Wekalao, J., Njoroge, S.M., Elamri, O.: Enhanced malaria detection using a hybrid borophene-based terahertz biosensor with random forest regression analysis. *Brazilian Journal of Physics* 55(3) (2025)
- [5] Wekalao, J.: High-sensitivity graphene-MoS₂ hybrid metasurface biosensor with machine learning optimization for hemoglobin detection. *Plasmonics* (2025)
- [6] Chen, M., Yu, L., Zhao, Q., Wang, X., Yeung, E.H.K., Tsui, K.L., Wang, H., Zhao, Y.: A data-knowledge driven approach with wearable sensors for intelligent gait disorders identification. *Biomedical Signal Processing and Control* 111, 108289 (2025)
- [7] Fang, R., Wang, J., Han, X., Li, X., Tong, J., Gao, M., Huang, X., Wang, H.: Hemoglobin feature extraction based on wavelet packet-fuzzy shrinkage denoising model for improving the accuracy of whole blood NIRS quantitative analysis. *Biomedical Signal Processing and Control* 112, 108550 (2025)
- [8] Li, X., Yang, X., Pan, Z., Zhuo, S., Lin, Z., Chen, J.: Nucleic acid amplification-free biosensor for sensitive and specific cfDNA detection based on CRISPR-Cas12a and single nanoparticle dark-field microscopy (DFM) imaging. *Sensors and Actuators B: Chemical* 406, 135363 (2024)
- [9] Rajakannu, A., Wekalao, J., Elsayed, H.A., El-Sherbeeney, A.M., Abukhadra, M.R., Hajjiah, A., Alarifi, N.S., Mehaney, A.: A high sensitivity terahertz biosensor with hybrid metasurfaces for tuberculosis detection leveraging the integration of machine learning and multi-material resonators. *Brazilian Journal of Physics* 55(4) (2025)
- [10] Wekalao, J., Elsayed, H., Bin-Jumah, M., Alqhtani, H., Abukhadra, M.R., Bellucci, S., Rajakannu, A., Mehaney, A.: Advanced terahertz-range dopamine detection using a 2D material-based metasurface biosensor. *Applied Optics* 64(16), 4625 (2025).