ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

Research on Elderly Health Monitoring Based on Multiple Machine Learning Algorithms

Shuijin Rong[1,a], Wei Guo[1,b], Hao Liu[1,c,*]

[1]*University of Science and Technology Liaoning, Anshan, China*
[a]*120223803046@stu.ustl.edu.cn,* [b]*weiguo@stu.ustl.edu.cn,* [c]*haoliu@ustl.edu.cn*
**Corresponding author*

Abstract: *This study constructs a health monitoring model for the early screening of AD that integrates environmental data and behavioral characteristics. Based on 174 452-dimensional handwriting feature samples from the DARWIN dataset, the median was filled with missing values, and Pearson correlation was used to screen the top 20 highly correlated features (strong positive correlation in the MOX group). The performance differences of PCA, UMAP, and LDA dimensionality reduction techniques were compared. Through the classification of eight machine learning algorithms, it was found that the random forest had the best performance after PCA dimensionality reduction (MSE=4.53, R^2=0.9965, which was 61% better than linear regression). Feature analysis shows that the MOX4 related to handwriting pressure/time is the core indicator, and the weight of environmental features is less than 3%. The research provides a lightweight screening solution centered on handwriting features, verifies the efficiency of PCA in extracting the linear structure of high-dimensional data, and offers a methodological reference for non-invasive monitoring in intelligent elderly care. In the future, it can integrate multimodal data to build an early warning model.*

Keywords: *Alzheimer's Disease, Machine Learning, Classification Algorithm*

1. Introduction

The seventh national census in 2020 showed that by the end of 2019, the population aged 60 and above in China had reached 264 million (accounting for 18.7%)[1], and that aged 65 and above had reached 190 million (13.5%), indicating a deep aging problem. It is estimated that the elderly population will peak at 487 million by 2050 (34.9%). With the weakening of family-based elderly care functions and the rising proportion of empty-nest elders (51.3% in 2019), the role of institutional elderly care has become more prominent. The state has made huge investments, and the number of elderly care institutions and facilities has increased to 204,000 by the end of 2019[2]. However, there is a gap between the indoor environmental quality of elderly care institutions (such as temperature, humidity and ventilation) and the health needs of the elderly, especially affecting those at high risk of Alzheimer's disease (AD). The number of AD patients[3] in China has increased from 5.98 million in 2007 to 10.32 million in 2019 (accounting for one quarter of the global total), and it is expected to exceed 25.65 million by 2050. The traditional early diagnosis of AD has limitations such as high cost and strong invasiveness in institutional scenarios[4]. This study focuses on the environmental quality of elderly care institutions and the early screening of AD. It innovatively integrates environmental sensor data (temperature, humidity, light, etc.) with behavioral characteristic data (handwriting movement parameters) to construct a multi-dimensional health monitoring model. The aim is to break through the limitations of a single indicator, explore the association between the environment and cognitive decline, and discover the specific value of handwriting behavior in the early recognition of AD[5]. The achievements can provide support for the transformation of the institutional environment, promote the establishment of an integrated health management system of "environment - behavior - early warning", and help address the aging problem.

2. Data Processing

2.1 Data Cleaning

The loading dataset inspection revealed that there were missing values in the label column MOX1.

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

The median padding method was adopted for processing, and the non-numerical categories were labeled and encoded to convert them into integer types.

2.2 Characteristic Engineering

To reduce the dimension and focus on the core information, the absolute correlation coefficient between each feature and the encoded label is calculated based on the Pearson correlation coefficient, and the top 20 features most closely associated with the label are screened out. To visually demonstrate the interrelationships among the features, a heat map is drawn (as shown in Figure 1). In the figure, the depth of color indicates the magnitude of the correlation coefficient: red series represent positive correlation, and blue series represent negative correlation. The closer the value is to 1 or -1, the stronger the correlation. A correlation heat map among the air quality indicators (MOX1 - MOX4) was also drawn (as shown in Figure 2).

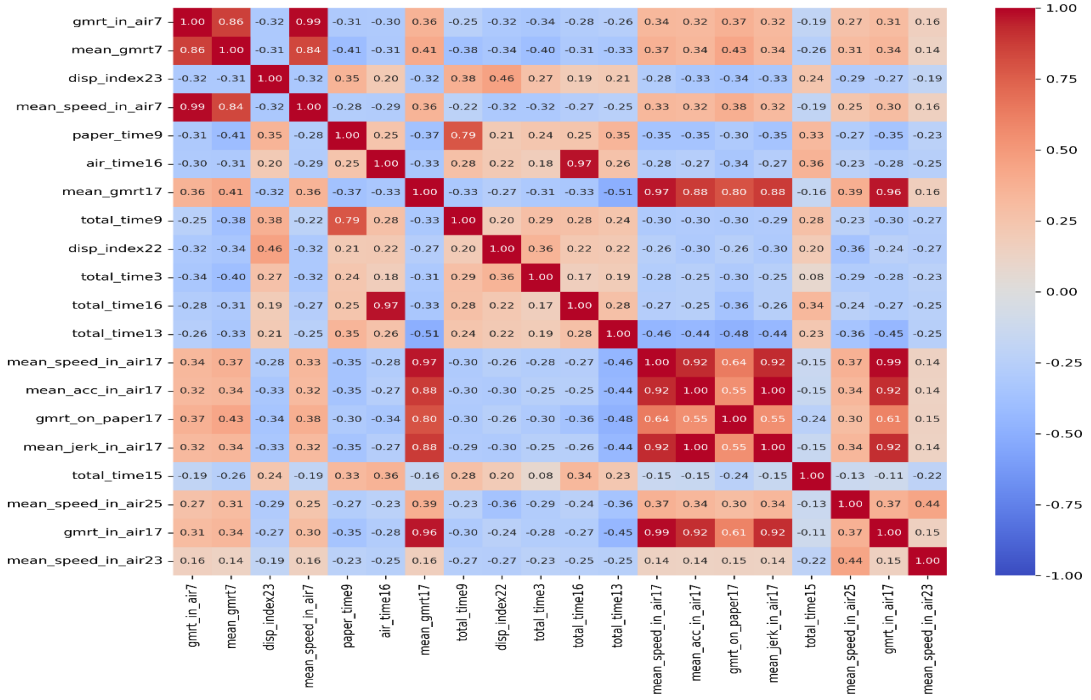

Figure 1: Correlation heat map of the top 20 highly correlated features

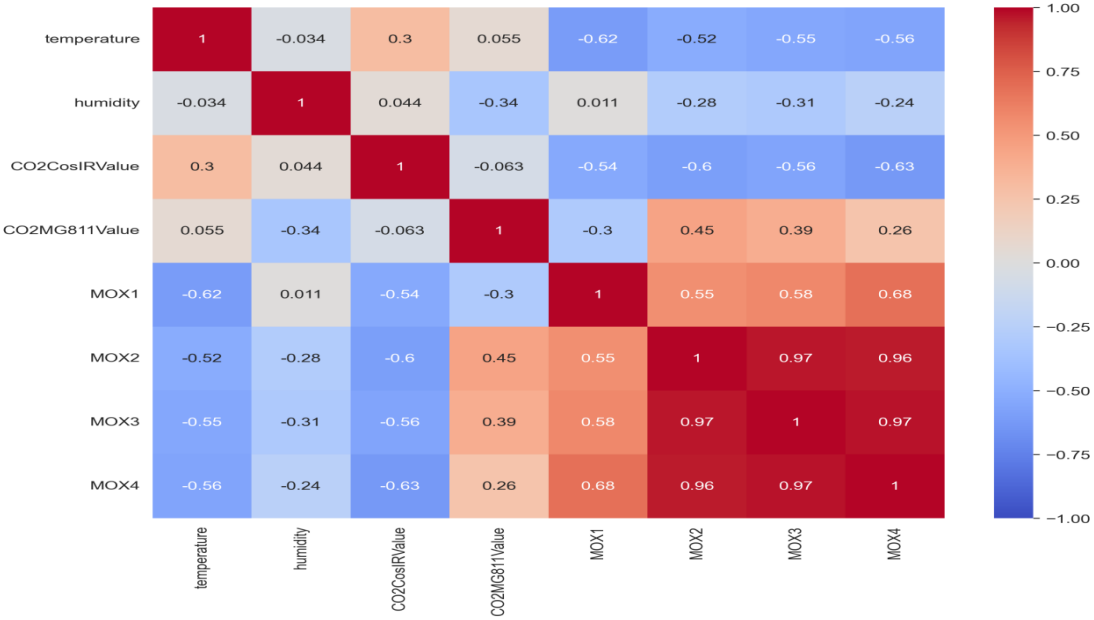

Figure 2: Relevant heat map of air concentration index

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

The figure clearly shows that some features are significantly positively correlated (with a coefficient as high as 0.86), while some features are weakly negatively correlated with time-related features (with coefficients approximately ranging from -0.32 to -0.20). This visualization helps to understand the intrinsic connections of features, providing a basis for subsequent dimension reduction and model training, ensuring the retention of key information and reducing the interference of redundant features.

MOX1 - MOX4 show a strong positive correlation with each other, with most correlation coefficients exceeding 0.55. Among them, the correlation coefficients among MOX2, MOX3, and MOX4 are even close to 0.97, indicating that their changing trends are highly consistent.

3. Research Methods

3.1 Random Forest

An integrated classification algorithm improves model performance by constructing multiple decision trees and synthesizing their results. The core idea is: to use Bootstrap sampling technology to generate multiple different training subsets, and each subset independently trains a decision tree. The final prediction result is determined by voting on all decision trees (taking the majority vote for classification tasks) or averaging (taking the mean for regression tasks). This mechanism that introduces randomness effectively reduces the variance of the model and enhances its generalization ability. The final classification result of the sample is determined by a majority vote of all the categories predicted by the decision tree.

3.2 Support Vector Machine

The core objective of this algorithm is to find an optimal separation hyperplane in the feature space, maximizing the Margin between data points of different categories. For linearly separable data, this hyperplane can perfectly separate the data, and the nearest data point (support vector) is the farthest. The process of finding this hyperplane is equivalent to solving a constrained optimization problem, which is usually transformed into a dual problem for solution by means of the Lagrange multiplier method. For nonlinear data, SVM[6] maps the data to a higher-dimensional feature space through kernel function techniques, making it linearly separable in this space. The final decision function depends on the calculation of support vectors and kernel functions.

3.3 Logistic Regression

A linear model that is widely applied to binary classification problems. It first linearly combines the input features to calculate a score value, and then maps this score to the probability (between 0 and 1) that the sample belongs to the positive class through the Sigmoid function (or Logistic function). The training objective of the model is to find the optimal weight parameters to minimize the difference between the predicted probability and the true label, which is usually achieved by minimizing the cross-entropy loss function.

3.4 KNN

KNN is an instance-based lazy learning classification algorithm. For a new test sample, KNN calculates the distance between it and all the samples in the training set (commonly the Euclidean distance). Then, select the K nearest training samples, count the number of each category among these KNN, and predict the test sample as the category with the largest number among them (majority voting method).

3.5 XGBoost

XGBoost is an efficient and highly effective gradient boosting decision tree algorithm. It adopts an additive model and trains a series of regression trees through an iterative approach. In each round of iteration, the newly added trees are designed to fit the residuals (negative gradient direction) between the current model's prediction results and the true values. Its objective function not only includes the loss function part for measuring the prediction error, but also the regularization term for controlling the model complexity, effectively preventing overfitting. The optimization process utilizes the second derivative information of the loss function (Taylor expansion approximation), making the model training faster and

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

achieving better results.

3.6 Multi-Layer Perceptron

MLP is a fundamental feedforward artificial neural network structure, which typically comprises an input layer, one or more hidden layers, and an output layer. Each layer is composed of multiple neurons (nodes). Information starts from the input layer and is passed forward layer by layer: the neurons in each layer receive the input signal from the neurons in the previous layer (weighted sum), and then it is transformed through a nonlinear activation function (such as ReLU, Sigmoid) to obtain the output of that layer, and finally the prediction result is generated in the output layer. The network learns by adjusting the connection weights and bias terms between layers.

3.7 Linear Regression

LR is a fundamental regression method for predicting continuous numerical target variables. It assumes that there is a linear relationship between the independent variable (characteristic) and the dependent variable (target). The objective of the model is to find the best-fitting straight line (for univariate variables)or hyperplane (for multivariable variables), so as to minimize the total error (usually measured by the sum of squares, that is, the mean square error) between the predicted values and the actual observed values of all samples.

3.8 Decision Tree Regression

DTR is a regression method based on a tree structure. It builds the model by recursively dividing the feature space into several non-overlapping regions (corresponding to the leaf nodes of the tree). The key steps in the construction process are: at each node, select the optimal features and segmentation points for data partitioning. The basis for selection is to maximize the purity of the target values of the samples within the subset after partitioning or minimize their degree of dispersion (common indicators include mean square error (MSE) or mean absolute error (MAE))[7]. The partitioning process is continuously carried out recursively until the stop conditions are met (such as reaching the preset tree depth, having too few samples within the node, or having a sufficiently small variance within the node). Ultimately, the predicted value of each leaf node is taken as the average of the target values of all training samples within that node.

4. Performance evaluation indicators

4.1 Accuracy Rate

Accuracy rate represents the proportion of samples correctly classified by the model among the total samples, reflecting the overall classification accuracy. The calculation formula is:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

4.2 Precision and Recall Rate

Precision: The proportion of samples predicted as positive classes that are actually positive classes, measuring the strictness of the model's discrimination of positive classes:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall rate: The proportion of actual positive class samples that are correctly predicted, measuring the ability of the type to capture positive classes (such as the detection rate of AD patients):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

4.3 F1 Score

The F1 score is the harmonic average of precision and recall, balancing the weights of the two. The formula is:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

4.4 ROC and AUC

The ROC curve takes the false positive rate (FPR)[8] as the horizontal axis and the true positive rate (TPR, i.e., recall rate) as the vertical axis to plot the performance of the model under different classification thresholds. The closer the curve is to the upper left corner, the better the model performance.

AUC(Area Under Roc curve) is the area under the ROC curve, with a value range of [0,1]. The larger the value, the stronger the model's ability to distinguish positive and negative samples:

$$AUC = \int_0^1 TPR(t)\, d\, FPR(t) \tag{5}$$

This study focuses on the binary classification task of AD patients and healthy individuals, with the goal of early screening (to reduce missed diagnoses). Therefore, it pays particular attention to the recall rate, avoiding the missed detection of potential patients, F1 score (balancing missed diagnoses and misdiagnoses), and AUC-ROC[9] (comprehensively evaluating the stability of the model under different risk thresholds). For multi-category scenarios, if it is necessary to further subdivide the severity of AD, macro average metrics should be supplemented to take into account the performance of a few categories. The above indicators quantify the model's performance from different dimensions, providing a unified evaluation standard for the subsequent analysis of experimental results and ensuring the scientific and reliable nature of the research conclusions.

5. Experimental Results and Analysis

As can be seen from Tables 1 to 3, from the perspective of mean square error (MSE), the MSE of linear regression is 490.20664696467156, which is relatively high, indicating a significant deviation between the predicted value and the true value. The MSE of the decision tree regression decreased to 8.33555482994505. The random forest regression has the lowest MSE, which is 4.5343365308918875. In terms of the coefficient of determination (R^2), the linear regression is 0.6190703696698956, indicating that this model can only explain 61.91% of the data variation. The R^2 of decision tree regression is 0.9935226096185592, and that of random forest regression reaches 0.9964764591643136. The latter two have significantly better fitting effects on the data than linear regression.

Table 1: Model Performance Comparison and Analysis - Linear Regression

Indicator	Value
MSE	490.20664696467156
R^2	0.6190703696698956
C02CosIRValue	0.363347
C02MG811Value	0.128056
MOX1	0.041621
MOX2	-0.510538
MOX3	-0.828835
MOX4	0.552936
temperature	-0.901686
humidity	-0.603087

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

Table 2: Comparative Analysis of Model Performance - Decision Tree Regression

Indicator	Value
MSE	8.33555482994505
R2	0.9935226096185592
C02CosIRValue	0.082656
C02MG811Value	0.045844
MOX1	0.031106
MOX2	0.108597
MOX3	0.203911
MOX4	0.464588
temperature	0.031788
humidity	0.031509

Table 3: Model Performance Comparison and Analysis - Random Forest Regression

Indicator	Value
MSE	4.5343365308918875
R2	0.9964764591643136
C02CosIRValue	0.082995
C02MG811Value	0.047418
MOX1	0.029377
MOX2	0.114111
MOX3	0.206475
MOX4	0.460883
temperature	0.030090
humidity	0.028652

In addition, we visualize the predicted values and true values of the regression model as well as the relationship between each characteristic coefficient as shown in Figures 3-6. It can be seen that in the analysis of feature importance, different models present consistency and difference characteristics:

In the linear regression model, the absolute values of the regression coefficients of temperature (temperature, -0.90), MOX3 (-0.83), and MOX2 (-0.51) are relatively large, suggesting that the above features have a significant impact on the prediction results. However, due to the strong correlation among the MOX1-MOX4 feature groups (correlation coefficient 0.55-0.97, Figure 3), multicollinearity may lead to deviations in the coefficient symbols and significance, and its physical meaning needs to be further verified in combination with the domain knowledge of handwriting movement.

The feature importance ranking of the decision tree model is: MOX4 (0.46) > MOX3 (0.21) > MOX2 (0.11) > CO_2CosIRValue (0.08) > temperature (0.03), among which the cumulative importance of the MOX group features exceeds 78%, indicating that it plays a leading role in the classification task. The highest importance of MOX4 (0.46) may be related to core parameters such as pressure and time during handwriting movement, while the low importance of environmental features (temperature, humidity) (<0.03) is consistent with the conclusion of weak correlation in the heat map (Figure 3).

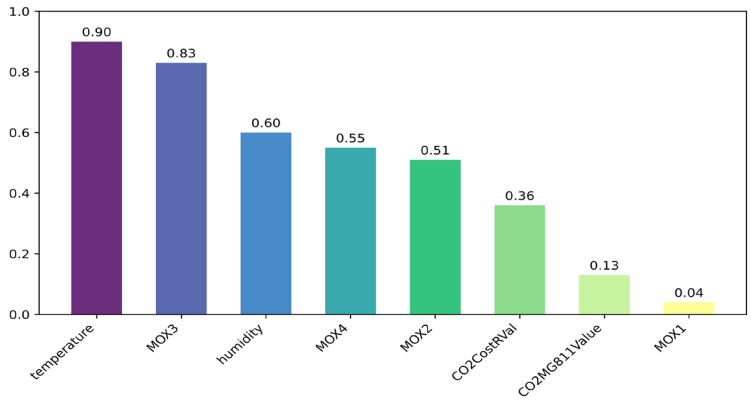

Figure 3: Visualization of the characteristic coefficients of linear regression

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

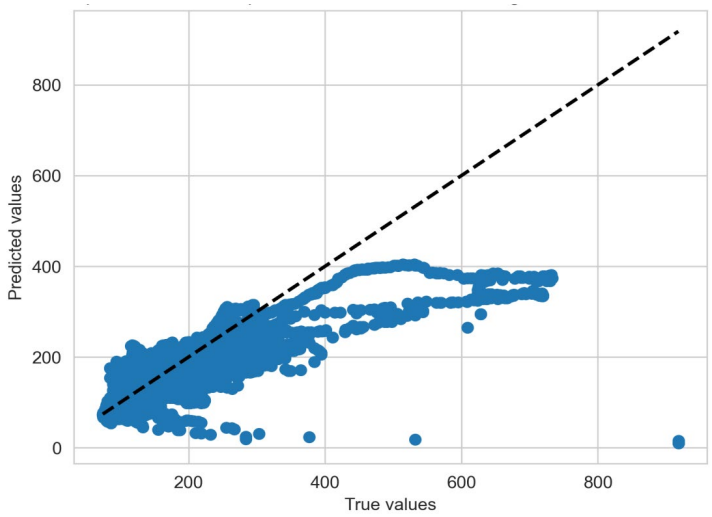

Figure 4: The gap between the predicted values of linear regression and the true values

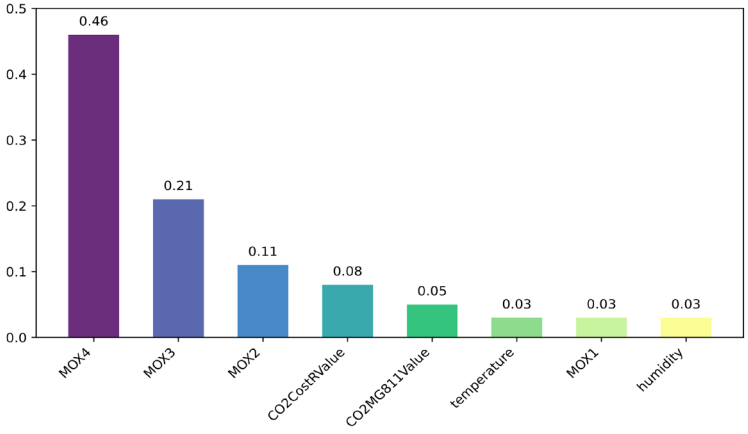

Figure 5: Visualization of the characteristic coefficients of decision tree regression

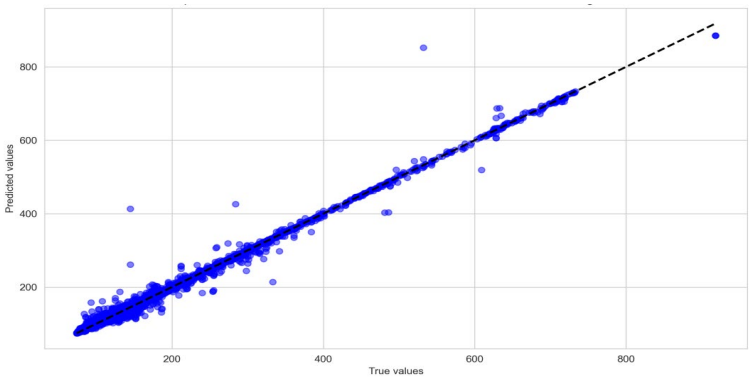

Figure 6: The gap between the predicted value and the true value of the decision tree regression

The feature importance trend of the random forest model is consistent with that of the decision tree, and the ranking is: MOX4 (0.461) > MOX3 (0.206) > MOX2 (0.114) > temperature (0.083) > CO_2CosIRValue (0.047). Through ensemble learning, random forests effectively reduce the overfitting risk of a single feature, and at the same time show stronger robustness against weakly correlated features. For example, the importance of temperature is increased to 0.083. In conclusion, MOX4 has always been the most critical feature in nonlinear models, reflecting its strong correlation with the classification task of Alzheimer's disease.

ISSN 2616-5775 Vol. 8, Issue 7: 47-54, DOI: 10.25236/AJCIS.2025.080706

6. Conclusions

This study constructed an early classification model for AD in elderly care institutions based on machine learning and behavioral characteristics. The main conclusions are as follows:

Nonlinear models have significant advantages: In the comparison of eight algorithms, linear models (such as linear regression, MSE=490.21, R^2=0.619) performed poorly due to the linear assumption. Nonlinear models such as decision trees and random forests significantly enhance performance. Among them, random forests perform the best (MSE=4.53, R^2=0.9965), and their 10-way cross-validation is stable (standard deviation <0.012), highlighting their advantages in capturing complex nonlinear relationships in AD.

The MOX feature serves as the core indicator: there is a strong positive correlation (Pearson coefficient 0.55-0.97) among the MOX1-MOX4 features, reflecting the synergic changes in handwriting motion parameters. Both random forests and decision trees show that the importance of the MOX4 feature is the highest (about 0.46), significantly higher than that of other features (such as MOX3 being 0.287), confirming its core association with AD. The correlation between environmental characteristics such as temperature and humidity and the target variable is weak (r <0.3), and the contribution weight is low (<3%), which verifies the rationality of the model design of "behavioral characteristics as the main and environmental data as the auxiliary".

Contributions and Limitations: The lightweight model proposed in this study, centered on handwriting behavior, provides a solution for the early non-invasive monitoring of AD and verifies the practicality of PCA. The limitation lies in the failure to integrate multi-source data (such as environmental and physiological indicators). In the future, samples can be expanded, multi-modal data can be integrated to build more robust models, and the temporal correlation between dynamic changes in handwriting and disease courses can be explored to promote clinical and intelligent elderly care applications.

References

[1] Zhang, M., Shi, Y., Zhou, B., Huang, Z., Zhao, Z., Li, C., ... & Li, Y. (2023). Prevalence, awareness, treatment, and control of hypertension in China, 2004-18: findings from six rounds of a national survey. Bmj, 380.

[2] Wang, L., Di, X., Yang, L., & Dai, X. (2020, November). Differences in the potential accessibility of home-based healthcare services among different groups of older adults: a case from shaanxi province, china. In Healthcare (Vol. 8, No. 4, p. 452). MDPI.

[3] Zanetti, O., Solerte, S. B., & Cantoni, F. (2009). Life expectancy in Alzheimer's disease (AD). Archives of gerontology and geriatrics, 49, 237-243.

[4] Alberdi, A., Aztiria, A., & Basarab, A. (2016). On the early diagnosis of Alzheimer's Disease from multimodal signals: A survey. Artificial intelligence in medicine, 71, 1-29.

[5] Li, A., Xue, C., Wu, R., Wu, W., Zhao, J., & Qiang, Y. (2025). Unearthing subtle cognitive variations: a digital screening tool for detecting and monitoring mild cognitive impairment. International Journal of Human–Computer Interaction, 41(4), 2579-2599.

[6] Xue, H., Yang, Q., & Chen, S. (2009). SVM: Support vector machines. In The top ten algorithms in data mining (pp. 51-74). Chapman and Hall/CRC.

[7] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. Peerj computer science, 7, e623.

[8] Adler, W., & Lausen, B. (2009). Bootstrap estimated true and false positive rates and ROC curve. Computational statistics & data analysis, 53(3), 718-729.

[9] Maffezzoni, D., Barbierato, E., & Gatti, A. (2025). Data-Driven Diagnostics for Pediatric Appendicitis: Machine Learning to Minimize Misdiagnoses and Unnecessary Surgeries. Future Internet, 17(4), 147.