Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 14, No. 1, 2025

Study on the Efficacy of a Novel Sedative Medication Based on Wilcoxon Rank-Sum Test and Multiple Machine Learning Models

Yuxiao Chen*

School of Mechatronics and Mold Engineering, Taizhou Vocational College of Science & Technology, Taizhou, China
*Corresponding author

Abstract: This study investigates the efficacy of a novel sedative medication compared to an existing drug using the Wilcoxon Rank-Sum test and multiple machine learning models. The Wilcoxon Rank-Sum test revealed statistically significant differences in petco200, petco2005, IPI005, and moaas005 indicators, primarily within the first 1 to 3 minutes post-induction. Basic information between the novel and existing drug groups was comparable, suggesting effective variable control and attributing differences to the novel sedative. Subsequently, various evaluation metrics including MSE, RMSE, MAE, MAPE, and R^2 were employed. Exploratory predictions using the Random Forest (RF) model yielded suboptimal results. After comparing the performance of RF, XGBoost, CatBoost, LightGBM, and SVR, a combination of the RF model and logistic regression was selected for regression predictions based on data type. Visualization of results demonstrated good predictive performance and mitigated overfitting.

Keywords: Wilcoxon Rank-Sum test, RF, XGBoost, LightGBM.

1. Introduction

Sedation is a crucial aspect of medical procedures, ensuring patient comfort and cooperation while allowing medical staff to perform tasks efficiently. The development of novel sedative medications aims to improve efficacy, reduce side effects, and enhance patient outcomes. In this study, we focus on evaluating a newly developed sedative drug against an established medication.

Phong B. Dao, in his work[1], applied the Wilcoxon rank-sum test, a non-parametric statistical test from the field of statistics, to monitor the operational status and automate fault detection in wind turbines. He introduced a five-step computational method grounded in statistical hypothesis testing, with the null hypothesis stipulating normal, fault-free turbine operation. Rejection of the null hypothesis in favor of the alternative suggests turbine malfunction, indicated by an abrupt change from 0 to 1 in the test decision. M.R. Simi et al. [2] utilized the Wilcoxon rank-sum test and non-parametric statistical methods to adjust the parameter rates of the DRASTIC model, an enhanced version incorporating anthropogenic factors (land use and/or land cover). Alexander P. Nocera et al. [3] found a positive correlation between h-index, m-index, and academic ranking among authors. Among the 2,253 urological academics evaluated, department chairs/directors and professors exhibited the highest median h- and m-indices (26 for chairs/directors with an m-index of 1.046, and 30 for professors with an m-index of 1.094). Yiwei Jia et al. [4] extracted clinical data from 1,230 inflammatory breast cancer (IBC) patients between 2010 and 2020 from the Surveillance, Epidemiology, and End Results (SEER) database. Cox analysis was employed to identify clinicopathological features associated with overall survival (OS) in IBC patients. A Random Survival Forest (RSF) algorithm was used to develop an accurate prognostic prediction model for IBC patients, with survival analysis conducted using Kaplan-Meier methods. Benjamin D. Simon et al. [5] employed a deep learning-based AI workflow for automatic Extraprostatic Extension (EPE) grading of prostate T2W MRI, ADC map, and high B-value DWI. Results indicated a lesion detection probability threshold of 0.45 and a balanced accuracy score for distance features of 0.390±0.078. When applied to the test set, the ROC AUC values for AI-assigned EPE grades 0-3 were 0.70, 0.65, 0.68, and 0.55, respectively.

The primary objective is to identify any statistically significant differences in vital signs and other relevant indicators between patients receiving the novel sedative and those receiving the existing drug. To achieve this, we utilize the Wilcoxon Rank-Sum test, a non-parametric statistical hypothesis test suitable for comparing two groups of samples when the underlying distributions are unknown or non-normal.

Furthermore, we aim to employ machine learning models to predict the effectiveness of both sedatives based on various physiological and demographic factors. By selecting appropriate evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2), we can quantify the predictive performance of the models. This study explores multiple machine learning algorithms, including Random Forest (RF), XGBoost, CatBoost, LightGBM, and Support Vector Regression (SVR), to identify the most suitable model for predicting sedative efficacy.

2. Test Modeling and Solving

2.1. Normality test

In order to make a choice of the test, the above data were subjected to descriptive statistics, and the normality test was determined. The data were plotted on a QQ matrix to roughly visualize their distribution, as shown in Fig.1.

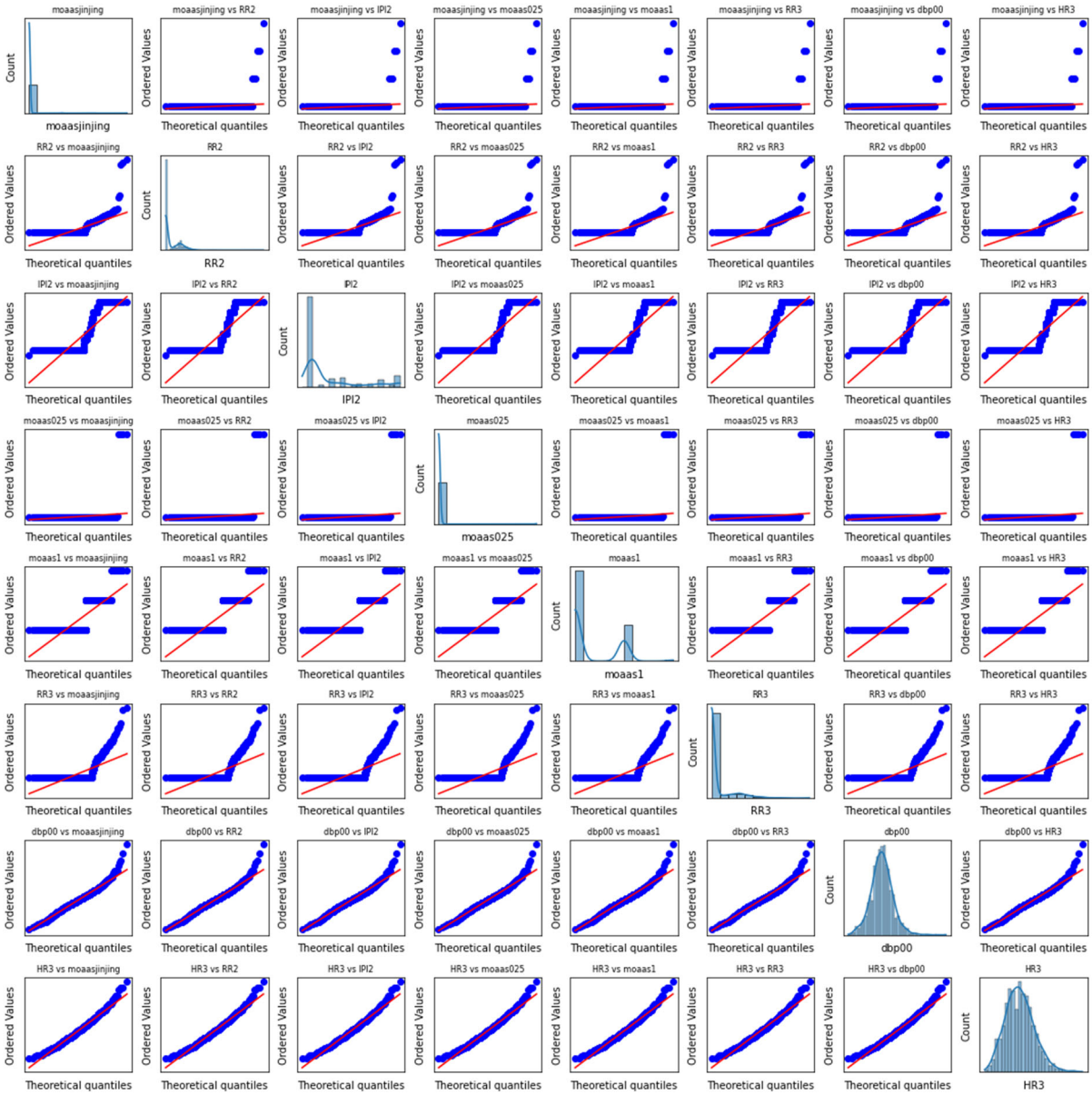

Figure 1. Randomized 9-individual test for normality of signs

As seen in Figure 1, some data may have a normal relationship with each other and some data may be normally distributed, in order to further analyze the data distribution.

2.2. Normality Test: Shapiro-Wilk Test

The Shapiro-Wilk test is a statistical method used to test whether a sample of data comes from a normal distribution. This test is particularly effective for small samples and is one of the most common ways of detecting normality The Shapiro-Wilk test assumes that the sample data is from a normal distribution and the alternative assumption is that the sample data is not from a normal distribution.

The statistic W of the Shapiro-Wilk test is calculated as:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (1)$$

Where:

$x_{(i)}$ is the i-th smallest value in the sample data.

$\bar{x}$ is the mean of the sample data.

a_i is the coefficient associated with the theoretical rank under normal distribution.

All p-values are very small, less than 0.05, presenting significance and therefore all data do not satisfy normal distribution.

2.3. Test of Variance: Wilcoxon Rank Sum Test

The Wilcoxon rank sum test, also known as the Mann-Whitney U Test, is a nonparametric statistical test used to compare the distributions of two independent samples for significant differences. This test is particularly useful in situations when the data do not meet the assumption of normal distribution. It can be used to compare the difference in medians between two groups of samples without assuming a specific distributional pattern for the data.

Steps in Wilcoxon's rank sum test:

Step1: Sorting and Rank Assignment: the two sets of samples are combined and all samples are sorted in order from smallest to largest. Assign a rank value to each sample.

Step2: Calculate the rank sum for each group of samples.

Step3: Calculate the test statistic. Calculate the test statistic:

$$U_A = R_A - \frac{n_A(n_A+1)}{2} \quad (2)$$

$$U_B = R_B - \frac{n_B(n_B+1)}{2} \quad (3)$$

$$U_{min} = min(U_A, U_B) \quad (4)$$

Where UA, UB are the U-statistics of sample A and sample B respectively. RA, RB are the rank sums of sample A and sample B respectively. nA,nB are the sample sizes of sample A and sample B respectively.

In order to use the paired-sample Wilcoxon signed rank test for both sets of data, the data were first screened for the same number of subjects to ensure that the sample sizes were the same, even though the sample sizes for both drug B and drug R data were 475.

From the extracted data, it can be found that most of the vital signs that would show significant differences are mainly concentrated in 1min to 3min after induction.

The basic information of the subjects was screened from the basic information of the subjects with each of the 475 subjects sampled above, and the basic information of the subjects was subjected to descriptive statistics and plotted as shown in Fig.2.

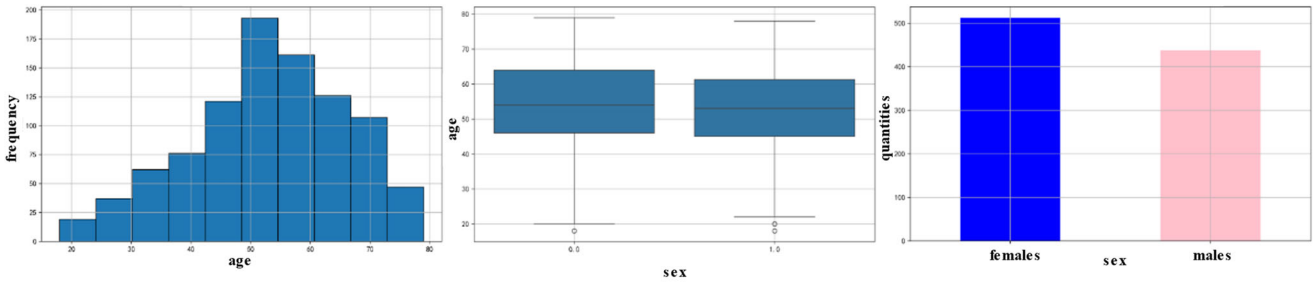

Figure 2. Basic information on all subjects

From the above basic information of all subjects, it can be seen that the distribution of information of all subjects is balanced and basically meets the normal distribution, the total frequency of age and gender does not differ much, and the relationship between gender age and weight is more uniform in line with the reality.

Further comparison of the distribution of data between the two groups, compare the data between the two groups whether there are other factors that may lead to significant differences in vital signs found that the age distribution of the two groups of subjects is more or less the same, and the two groups of data subjects appear to have a history of the situation is more or less the same.

Therefore, it was concluded that both groups of subjects had the same basic profile. All the variables were considered to be controlled in this experiment and the reason for the difference in vital signs indicators was the type of sedative drug.

3. Prediction Based on Multiple Machine Learning Models

3.1. Selection of evaluation indicators

Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R2) are introduced as performance and evaluation indicators.

The Mean Square Error (MSE) is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

The Mean Absolute Error (MAE) is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

The Mean Absolute Percentage Error (MAPE) is calculated in the format:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (8)$$

In the above 4 equations, y_i is the ith actual value and $\hat{y}_i$ is the ith predicted value.

The coefficient of determination (R2) is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

where, $\bar{y}$ is the mean of the actual values.

MSE is used to measure the mean squared difference between the predicted and actual values. rMSE is the square root of MSE, which indicates the standard deviation of the prediction error. maE denotes the mean of the absolute difference between the predicted and actual values. mape is the mean of the absolute percentage of the prediction error, which is used as a measure of the relative magnitude of the error. r2 denotes the ability of the model to explain the data, which reflects the effectiveness of the model's fit.

3.2. RF Exploratory Prediction

The random forest model was used to predict the IPI005 indicator data, and the prediction results were obtained as shown in Table 1.

Table 1. RF model performance evaluation metrics

	MSE	RMSE	MAE	MAPE(%)	R2
training set	1.682	1.297	0.98	12.215	0.59
test set	4.619	2.149	1.468	17.039	-0.053

From the content of Table 1, it can be seen that the test set data of MSE, RMSE, MAE, and MAPE were higher than the training set, indicating that the model overfitted the training set data; the R^2 value was negative, indicating that the model was unable to explain the variation of the data on the test set.

Observation of the data characteristics revealed that gender, history of surgery, smoking, alcoholism, and history of motion sickness were discrete categorical data the rest of the data such as age, height, and weight were continuous data, so the indicator columns of the two data categories were split and predicted the IPI data within 3 minutes of administration of the medication, respectively.

The extensive use of the category mapping method with solo thermal coding for the indicators of gender, history of surgery, smoking, alcohol use, and history of motion sickness led to the problem of dimensional catastrophe, i.e., increased computational complexity, overfitting of the model, increased sparsity of the data, and decreased feature relevance. This ultimately resulted in the inability to accurately predict IPI data within 3 minutes of medication administration.

The use of alternative coding methods for the prediction of multiple IPI data would increase the computational and processing costs significantly.

Taking the above factors into account, the data indicators were split, continuous data and discrete data were predicted separately, and the two predictions were finally summed.

3.3. Multi-model comparison

Random forest regression is in the process of generating many decision trees, is through the modeling dataset of sample observations and feature variables are randomly sampled, each sampling result is a tree, and each tree will generate rules and judgment values that match its own attributes, and the forest finally integrates the rules and judgment values of all the decision trees, to achieve the regression of the random forest algorithm.

XGBoost is an efficient implementation of GBDT. Unlike GBDT, xgboost adds a regularization term to the loss function; and since some loss functions are difficult to compute derivatives, xgboost uses a second-order Taylor expansion of the loss function as the loss function fit. In regression problems, the goal of XGBoost is to minimize a loss function in order to predict continuous variables as accurately as possible.

CatBoost is a GBDT framework based on the symmetric decision tree algorithm, which mainly addresses the pain point of efficiently and rationally handling category-based features and dealing with gradient bias and prediction bias to improve the accuracy and generalization ability of the algorithm.

Support vector machine regression (SVR) maps the data into the high-dimensional data feature space with a nonlinear mapping, which makes the independent variable and the dependent variable in the high-dimensional data feature space have good linear regression characteristics, and the fitting is performed in this feature space and then return to the original space.

LightGBM is an efficient implementation of XGBoost, the idea is to discretize continuous floating-point features into k discrete values and construct a histogram of width k. Then traverse the training data and calculate the cumulative statistic of each discrete value in the histogram. In feature selection, we only need to traverse to find the optimal segmentation points based on the discrete values of the histogram; and we use the grow-by-leaf strategy with depth restriction, which saves a lot of time and space overhead.

For the continuous data metrics, IPI005 is chosen as the test metric, and the metrics are predicted using RandomForest, XGBoost, CatBoost, SVR, and LightGBM, respectively.

The results show that the RandomForest model has the optimal prediction results for this data, so the RandomForest model is used for the continuous data to evaluate the continuity indicators, and the logistic regression model is used to predict the values of the discontinuous dichotomous data for the two different types of data, respectively.

The regression results of the above two types of data were combined, and the weights of the two predictions in the actual prediction results were determined by comparing the two types of data with the actual errors.

4. Conclusion

Our findings indicate that the novel sedative medication exhibits statistically significant differences in specific vital signs compared to the existing drug, primarily within the first 1 to 3 minutes post-induction. These differences were observed in indicators such as petco200, petco2005, IPI005, and moaas005, suggesting early and notable effects of the novel sedative. The comparability of basic information between the two groups suggests that these differences are attributable to the novel sedative itself, rather than confounding variables.

Regarding machine learning predictions, while the initial exploratory Random Forest (RF) model yielded suboptimal results, a more comprehensive analysis involving multiple algorithms led to the selection of a combination of RF and logistic regression models. These models provided good predictive performance, as evidenced by the evaluation metrics, and effectively avoided overfitting. The visualization of results further supported the reliability and accuracy of the selected models.

In conclusion, this study contributes to the understanding of the efficacy of novel sedative medications and demonstrates the potential of machine learning in predicting sedative effects based on patient data. Future research could explore additional physiological indicators, larger sample sizes, and longer observation periods to further validate and refine these findings.

References

[1] P. B. Dao, “On Wilcoxon rank sum test for condition monitoring and fault detection of wind turbines,” Applied Energy, vol. 318, p. 119209, Jul. 2022.

[2] M. R. Simi, B. K. Bindhu, A. Varghese, and M. R. Rani, “Optimization of DRASTICA vulnerability assessment model by Wilcoxon rank sum non parametrical statistical test,” Materials Today: Proceedings, vol. 58, pp. 121–127, Jan. 2022.

[3] A. P. Nocera, H. Boudreau, C. J. Boyd, A. Tamhane, K. D. Martin, and S. Rais-Bahrami, “Correlation Between H-Index, M-Index, and Academic Rank in Urology,” Urology, vol. 189, pp. 150–155, Jul. 2024.

[4] Y. Jia et al., “Prognostic prediction for inflammatory breast cancer patients using random survival forest modeling,” Translational Oncology, vol. 52, p. 102246, Feb. 2025.

[5] B. D. Simon et al., “Automated Detection and Grading of Extraprostatic Extension of Prostate Cancer at MRI via Cascaded Deep Learning and Random Forest Classification,” Academic Radiology, vol. 31, no. 10, pp. 4096–4106, Oct. 2024.