Feature Selection and Model Evaluation Using Machine Learning in Traffic Flow Prediction

Shengzhe Huang

School of Transportation Engineering, Chang'an University, Shaanxi,710064, China

2022902661@chd.edu.cn

Abstract. Urban traffic congestion is a growing challenge, making accurate prediction of traffic flow at intersections essential for optimizing signal control and reducing congestion. This study leverages machine learning techniques to construct a prediction model based on historical traffic flow data. The Pearson coefficient is used to analyze the correlations between numerical features in the dataset, which are then visualized through a heatmap. The Extreme Gradient Boosting (XGBoost) model is employed to train a classifier and evaluate the importance of these features. Ablation experiments are conducted across 14 models—comprising Boosting, Bagging, Linear, and Traditional Lightweight models—using various numerical features. Results show that only three key features, "BusCount," "TruckCount," and "Total," must be retained in the majority of models to maintain stable classification accuracy. Removing these features causes a significant decline in prediction accuracy, with the exception of the Adaptive Boost (AdaBoost) model. This highlights the critical role these features play in effective traffic flow prediction and model stability.

Keywords: Machine Learning, Traffic Flow Prediction, Pearson coefficient, XGBoost model, Ablation experiments.

1. Introduction

The rapid growth of motor and non-motor vehicles in recent years has outpaced infrastructure development, leading to severe traffic congestion. This issue not only disrupts daily life but also limits urban growth [1]. Consequently, reducing congestion and improving traffic management is a key focus for both academia and traffic authorities. Traffic flow prediction plays a crucial role in intelligent transportation systems, guiding subsequent optimization measures, such as traffic signal control and road infrastructure improvements [2]. To enhance system efficiency, it is essential to develop and refine traffic flow prediction techniques.

Traditional traffic flow prediction systems often struggle to capture the complex, real-time changes in traffic patterns, relying heavily on prior knowledge and data from decision makers, which may lead to errors. In recent years, the powerful data processing capabilities of machine learning have gained significant attention, leading to advancements in traffic flow prediction. Liu et al. proposed a feature selection technique based on the community-based dandelion algorithm, which significantly enhances prediction accuracy [3]. To further improve prediction results, a dynamic framework utilizing long short-term memory (LSTM) networks was introduced, optimizing inputs through feature organization [4]. In another study, deep learning models incorporating attention mechanisms were used to enhance short-term traffic flow predictions [5]. Xiu et al. demonstrated that feature selection via correlation analysis, coupled with a parallel spatiotemporal neural network model, improves subway passenger flow forecasting [6]. Similarly, Zhang et al. improved model accuracy and stability by introducing an attention mechanism for optimizing multidimensional feature inputs [7]. Liu et al. combined Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), LSTM, and Extreme Gradient Boosting (XGBoost) models in a prediction framework, employing principal component analysis (PCA) to reduce feature redundancy [8, 9]. For long-term predictions, a deep learning approach based on Spatiotemporal Graph Convolutional Networks (ST-GCN) was proposed, integrating external parameters such as weather [10, 11]. These studies underscore the growing importance of feature selection and optimization in improving prediction precision.

This study utilizes historical traffic flow data and applies Pearson correlation coefficient heatmaps to analyze relationships between numerical features. The quantitative importance of these features is assessed using the XGBoost classifier. To further evaluate the generalizability of the findings, experiments are conducted across 14 models, including XGBoost. Results indicate that three key features in the dataset—BusCount, TruckCount, and Total—significantly impact the performance of all models except the Adaptive Boosting (AdaBoost) and Dummy models, with accuracy dropping substantially after their removal. The dataset includes 9 distinct dimensional parameters. Through a dual verification approach involving both statistical data analysis and code testing, this study highlights the critical role of core features in traffic flow prediction. These insights not only enhance the efficiency and accuracy of traffic flow forecasting but also contribute to the development of intelligent transportation systems, supporting the intelligent transformation and sustainable development of urban traffic networks.

2. Methodology

2.1. Dataset Description

The abbreviated dataset used in this study are shown in Table 1, the 14 machine learning models that followed in this work were trained and their accuracy verified using a two-month historical traffic flow dataset [12]. The first three columns in the dataset are the time characteristics, including the time data with days, months, and weeks as cycles, the columns 4-9 are the traffic flow characteristics, and the tenth column is the traffic simulation of the corresponding time segment. Daily data contained 24 hours per hour sampled at 15 minutes for a fragment, totaling 5952 data. This dataset was used directly as research material in this study without any preprocessing.

Table 1. Dataset Abbreviations view

	Time	Date	Day of the week	Car	Time	Date	Day of the week	Car	Time
0	12:00:00 PM	10	Tuesday	13	2	2	24	41	normal
1	12:15:00 PM	10	Tuesday	14	1	1	36	52	normal
2	12:30:00 PM	10	Tuesday	10	2	2	32	46	normal
…	…	…	…	…	…	…	…	…	…
5949	11:15:00 PM	9	Thursday	15	4	1	25	45	normal
5950	11:30:00 PM	9	Thursday	16	5	0	27	48	normal
5951	11:45:00 PM	9	Thursday	14	3	1	15	33	low

2.2. Proposed Approach

This study aims to identify key numerical features for predicting traffic flow using machine learning models. The XGBoost model, known for its efficiency and performance across various tasks, is used to evaluate feature importance through correlation analysis of historical traffic data [13]. The dataset's features are categorized into three groups based on their correlation and importance: low correlation and low importance, high correlation and low importance, and high correlation and high importance. Feature ablation experiments are then conducted using several machine learning models, including XGBoost, Random Forest, and Logistic Regression. The most important features are identified by eliminating them and validating the impact on model performance. The common

features across models are selected for high-precision traffic flow prediction. The process flow is illustrated in Fig. 1.

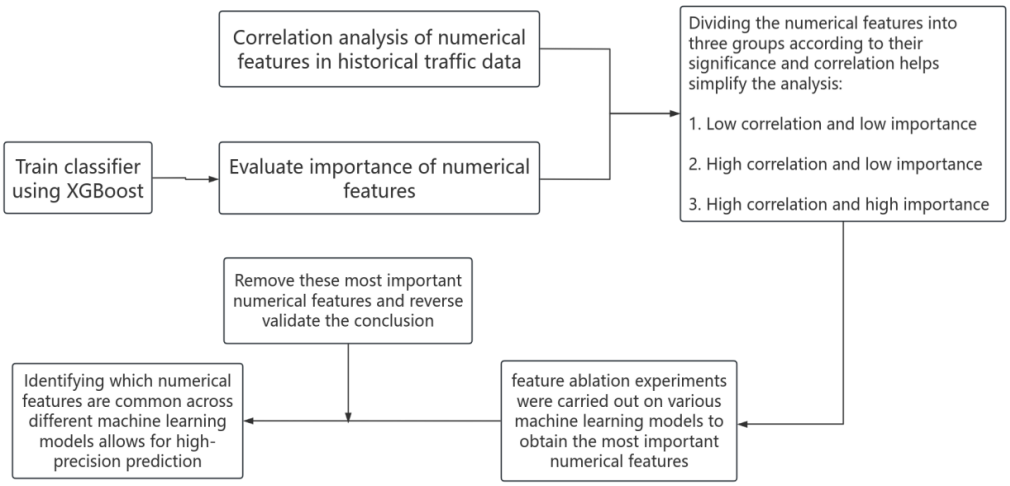

Figure 1. Flowchart of the whole process (Picture credit: Original)

2.3. Machine Learning Models

This set of models all belong to the Boosting method. Boosting is an ensemble learning strategy that strings together multiple weak classifiers (usually decision stumps), each adjusting for the classification errors of the previous round to gradually improve the accuracy of the overall model. This kind of method can usually deal with complex nonlinear problems, and has excellent performance on structured data with high accuracy. It is a model type widely used in practical applications. At the same time, both of them can focus on samples that are difficult to classify by continuously optimizing sample weights, which makes the model accuracy continuously optimized [14]. The boosting models used in this study are AdaBoost, XGBoost (XGB), Gradient Boosting (GB), Histogram-based Gradient Boosting (HGB).

This model adopts Bagging (Bootstrap Aggregation) strategy, which is a parallel integrated learning method. By training multiple independent models to improve the overall performance, each sub-model is trained on the samples obtained from the original data set. This strategy makes the data used by each model different, so there are differences between models, and finally the results of each model are integrated by majority voting or average. Bagging model can reduce the variance of the model, help to prevent overfitting, improve the overall robustness and generalization ability, and is suitable for noisy data or situations with many numerical features [15]. The Bagging models used in this study are Random Forest (RF), Extremely Randomized Trees (ET), Bagging model (BM).

Linear model assumes a linear relationship between features and objectives, and the structure of the model, as:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \tag{1}$$

Where w is the weight and b is the bias term.

This type of model is based on the linear assumption that the relationship between features and targets is linear or nearly linear. They are fast to train, highly interpretable, suitable for high-dimensional sparse data, such as text or simplified feature scenes, and perform very well when the features are good [16]. The linear models used in this study are Stochastic Gradient Descent (SGD), Logistic Regression (LG), Ridge model (RM).

Traditional lightweight model is suitable for rapid modeling, basic reference, or classification tasks under specific conditions. They are simple, easy to implement, and computationally efficient, often serving as baselines, teaching tools, or auxiliary analysis for small datasets [17]. The traditional lightweight models in this study include K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), Support Vector (SV), and the Dummy model (DM). KNN is an instance-based, non-parametric

method that classifies based on the distance between test and training samples, using majority voting. GNB assumes feature independence and Gaussian distribution to calculate posterior probabilities via Bayes' theorem. SV, a discriminative model, constructs an optimal hyperplane to separate classes, using kernel functions for nonlinear problems. The Dummy model makes predictions using preset strategies, providing a baseline for evaluating other models' performance by offering a minimum reference point for accuracy

3. Results and Discussion

3.1. Analysis Results on Correlation and Importance of Numerical Characteristics

The correlation matrix used in this study to examine the relationships between numerical features. Specifically, each pair of characteristics has its Pearson coefficient calculated. The properties are more connected while the correlation coefficient's absolute value is closer to 1. On the other hand, the closer the absolute value of the correlation coefficient is to 0, the less correlated the features are. Additionally, there is a negative correlation between the two features if the correlation coefficient is less than 0.

Table 2. Correlation matrix of numerical features

	Time	Date	Day of the week	CarCount	BikeCount	BusCount	TruckCount	Total	Traffic Situation
Time	1.00	0.00	0.00	0.49	0.29	0.35	-0.33	0.47	0.21
Date	0.00	1.00	-0.02	-0.01	0.00	0.00	0.02	0.00	0.00
Day of the week	0.00	-0.02	1.00	0.00	0.04	-0.04	0.00	0.00	-0.01
CarCount	0.49	-0.01	0.00	1.00	0.71	0.66	-0.62	0.97	0.69
BikeCount	0.29	0.00	0.04	0.71	1.00	0.58	-0.61	0.78	0.57
BusCount	0.35	0.00	-0.04	0.66	0.58	1.00	-0.56	0.76	0.66
TruckCount	-0.33	0.02	0.00	-0.62	-0.61	-0.56	1.00	-0.55	-0.27
Total	0.47	0.00	0.00	0.97	0.78	0.76	-0.55	1.00	0.76
Traffic Situation	0.21	0.00	-0.01	0.69	0.57	0.66	-0.27	0.76	1.00

The traffic flow status is positively correlated with features like "time," "CarCount," "BikeCount," "BusCount," and "Total," as shown in Table 2. This means that an increase in the values of these features will result in an increase in traffic flow. Additionally, the traffic flow status is negatively correlated with "TruckCount," indicating that when there are a great number of trucks, traffic flow is typically low. In the end, the correlation coefficients with other features for the features "Date" and "Day of the week" tend to be zero, indicating that these two features have a weak correlation with the traffic flow status.

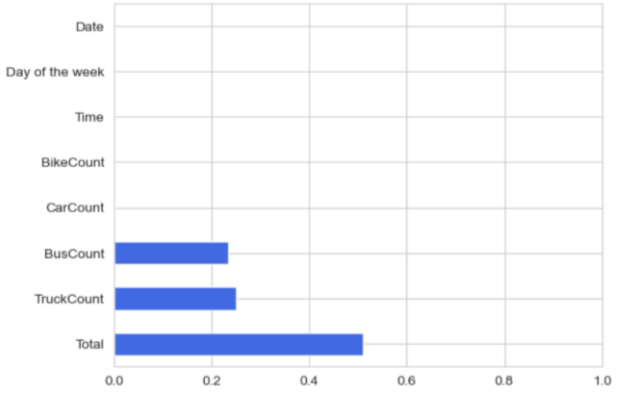

Figure 2. The importance of numerical features in the XGBoost model (Picture credit: Original)

3.2. Classification Discussion

In order to analyze the influence of different numerical features on different models, this paper uses the above-mentioned 14 models to analyze the complete dataset. The results are shown in the Table 3.

Table 3. The accuracy of 14 models under the complete dataset

Accuracy:	0.61(+/-0.00)	[ADA]
Accurac:	0.84(+/-0.02)	[SGD]
Accuracy:	1.00(+/-0.00)	[XGB]
Accuracy:	1.00(+/-0.00)	[RF]
Accuracy:	0.98(+/-0.00)	[ET]
Accuracy:	0.96(+/-0.01)	[KN]
Accuracy:	0.89(+/-0.01)	[LG]
Accuracy:	0.76(+/-0.01)	[RM]
Accuracy:	1.00(+/-0.00)	[HGB]
Accuracy:	1.00(+/-0.00)	[BM]
Accuracy:	1.00(+/-0.00)	[GB]
Accuracy:	0.86(+/-0.01)	[GNB]
Accuracy:	0.61(+/-0.00)	[DM]
Accuracy:	0.96(+/-0.00)	[SV]

It can be observed that using this dataset on most models can achieve accurate classification. On the XGboost, HBC, BC and GBC models, the classification accuracy rate has reached 100%.

Research the impact of eliminating numerical features with low correlation and low importance on the accuracy of different models. The results are shown in the Table 4.

Table 4. Th e outcomes after the separate elimination of Low correlation and low importance numerical features

		Delete 'date'	Delete 'day of the week'	Delete 'time'
[ADA]	Accuracy:	0.79(+/-0.09)	0.79(+/-0.09)	0.79(+/-0.09)
[SGD]	Accuracy:	0.84(+/-0.02)	0.85(+/-0.01)	0.84(+/-0.02)
[XGB]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)	1.00(+/-0.00)
[RF]	Accuracy:	1.00(+/-0.00)	0.99(+/-0.00)	1.00(+/-0.00)
[ET]	Accuracy:	0.97(+/-0.01)	0.97(+/-0.01)	0.97(+/-0.01)
[KN]	Accuracy:	0.92(+/-0.00)	0.92(+/-0.01)	0.91(+/-0.01)
[LG]	Accuracy:	0.88(+/-0.01)	0.88(+/-0.01)	0.88(+/-0.01)
[RM]	Accuracy:	0.77(+/-0.01)	0.77(+/-0.01)	0.76(+/-0.01)
[HGB]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)	1.00(+/-0.00)
[BM]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)	1.00(+/-0.00)
[GB]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)	1.00(+/-0.00)
[GNB]	Accuracy:	0.81(+/-0.01)	0.81(+/-0.01)	0.82(+/-0.01)
[DM]	Accuracy:	0.60(+/-0.00)	0.60(+/-0.00)	0.60(+/-0.00)
[SV]	Accuracy:	0.94(+/-0.01)	0.94(+/-0.01)	0.94(+/-0.01)

After eliminated the ‘time’, ‘date’, and ‘day of the week’ numerical features from the dataset, accordingly. It was found that the model's prediction accuracy did not changes considerably by elimination of these features. This indicates that these time features are not the main factors determining the traffic flow status, which is consistent with the previous conclusions of correlation and importance analysis.

The two numerical features, ‘BikeCount’ and ‘CarCount’, were respectively removed. The results are shown in Table 5.

Table 5. The outcomes after the separate elimination of High correlation and low importance numerical features

		Delete 'BikeCount'	Delete 'CarCount'
[ADA]	Accuracy:	0.79(+/-0.09)	0.79(+/-0.09)
[SGD]	Accuracy:	0.85(+/-0.02)	0.84(+/-0.01)
[XGB]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)
[RF]	Accuracy:	0.99(+/-0.00)	0.99(+/-0.00)
[ET]	Accuracy:	0.97(+/-0.00)	0.96(+/-0.01)
[KN]	Accuracy:	0.90(+/-0.01)	0.88(+/-0.01)
[LG]	Accuracy:	0.88(+/-0.01)	0.88(+/-0.01)
[RM]	Accuracy:	0.77(+/-0.01)	0.77(+/-0.01)
[HGB]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)
[BM]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)
[GB]	Accuracy:	1.00(+/-0.00)	1.00(+/-0.00)
[GNB]	Accuracy:	0.82(+/-0.01)	0.83(+/-0.01)
[DM]	Accuracy:	0.60(+/-0.00)	0.60(+/-0.00)
[SV]	Accuracy:	0.93(+/-0.01)	0.92(+/-0.01)

These two variables did not significantly impact accuracy in almost all of models, despite having a comparatively strong correlation with the labels in the correlation research. The overall model accuracy is nearly the same. This could be because these two features include a lot of redundant information, and the information in other features can be used to augment the information eliminated from these two elements. It's also possible that the majority of models don't rely much on these two numerical characteristics to foresee traffic flow conditions. The results are obvious, deleting these two features did not cause a significant decrease in accuracy, indicating that in this dataset, if numerical features only have the characteristic of being highly correlated with other numerical features, they cannot play a decisive role in the final traffic flow condition prediction results.

As shown in Table 6, the removal of the three features—'BusCount', 'TruckCount', and 'Total'—significantly reduced the accuracy of most models, indicating their importance in classification. Different models, however, prioritize these features differently. For instance, removing 'BusCount' notably decreased accuracy in the ADA and GNB models (by over 10%), suggesting its higher weight in these models. The GNB model's accuracy dropped sharply when 'TruckCount' was removed, but the ADA model's accuracy only slightly decreased. Removing the 'Total' feature had little impact on GNB's performance, highlighting that GNB relies more on 'BusCount' and 'TruckCount', while ADA emphasizes 'BusCount'

Table 6. The outcomes after the separate elimination of High correlation and high importance

		Delete 'BusCount'	Delete 'TruckCount'	Delete 'Total'
[ADA]	Accuracy:	0.67 (+/- 0.08)	0.76 (+/- 0.08)	0.74 (+/- 0.06)
[SGD]	Accuracy:	0.84 (+/- 0.02)	0.85 (+/- 0.02)	0.85 (+/- 0.02)
[XGB]	Accuracy:	0.97 (+/- 0.01)	0.96 (+/- 0.01)	0.97 (+/- 0.01)
[RF]	Accuracy:	0.94 (+/- 0.01)	0.94 (+/- 0.00)	0.96 (+/- 0.00)
[ET]	Accuracy:	0.92 (+/- 0.01)	0.92 (+/- 0.01)	0.94 (+/- 0.01)
[KN]	Accuracy:	0.85 (+/- 0.01)	0.82 (+/- 0.01)	0.89 (+/- 0.01)
[LG]	Accuracy:	0.87 (+/- 0.01)	0.87 (+/- 0.01)	0.88 (+/- 0.01)
[RM]	Accuracy:	0.77 (+/- 0.01)	0.77 (+/- 0.01)	0.77 (+/- 0.01)
[HGB]	Accuracy:	0.97 (+/- 0.01)	0.96 (+/- 0.00)	0.97 (+/- 0.00)
[BM]	Accuracy:	0.95 (+/- 0.01)	0.94 (+/- 0.00)	0.95 (+/- 0.00)
[GB]	Accuracy:	0.96 (+/- 0.01)	0.95 (+/- 0.00)	0.96 (+/- 0.00)
[GNB]	Accuracy:	0.73 (+/- 0.01)	0.71 (+/- 0.02)	0.82 (+/- 0.01)
[DM]	Accuracy:	0.60 (+/- 0.00)	0.60 (+/- 0.00)	0.60 (+/- 0.00)
[SV]	Accuracy:	0.90 (+/- 0.01)	0.87 (+/- 0.01)	0.93 (+/- 0.01)

The influence of the three main features on the classification accuracy of the model was verified again. The prediction results with and without the three main features retained are shown in Table 7.

Table 7. The prediction results with and without the three main features

	Only three main features remain	Only three main features were eliminated
[ADA]	Accuracy:	0.60 (+/- 0.00)
[SGD]	Accuracy:	0.84 (+/- 0.01)
[XGB]	Accuracy:	1.00 (+/- 0.00)
[RF]	Accuracy:	1.00 (+/- 0.00)
[ET]	Accuracy:	0.98 (+/- 0.01)
[KN]	Accuracy:	0.96 (+/- 0.01)
[LG]	Accuracy:	0.88 (+/- 0.01)
[RM]	Accuracy:	0.96 (+/- 0.01)
[HGB]	Accuracy:	1.00 (+/- 0.00)
[BM]	Accuracy:	1.00 (+/- 0.00)
[GB]	Accuracy:	1.00 (+/- 0.00)
[GNB]	Accuracy:	0.86 (+/- 0.01)
[DM]	Accuracy:	0.60 (+/- 0.00)

It can be observed that, apart from the ADA model, the data classification accuracy of the model that retains only the main features is basically the same as that of the full-feature model. This indicates that retaining only these three key features is sufficient to achieve a relatively good classification accuracy. However, after removing the three main features, the classification accuracy of the 14 models dropped significantly. This indicates that before using the model's prediction results, principal component analysis can be conducted on the numerical features to select the numerical features with high correlation and high importance, which can ensure the stability of the prediction accuracy.

4. Conclusion

This study analyzed historical traffic flow data to identify numerical features that significantly influence prediction accuracy. The results showed that PCA and selection of high-correlation, high-importance features can maintain stable prediction accuracy. Reducing the dataset to essential features enhances predictive efficiency and reduces computational cost, benefiting both real-time and long-term traffic forecasting. Key features such as BusCount, TruckCount, and Total were found to be crucial for accurate predictions across 12 models, excluding AdaBoost and Dummy models. The Dummy model, which uses preset strategies instead of learning from data, serves as a performance baseline, while AdaBoost underperformed despite retaining important features. Further investigation is required to explore the AdaBoost model's limitations. Overall, eliminating key numerical features leads to a significant decrease in accuracy, validating the importance of these features for accurate traffic flow prediction.

References

[1] X. Yin, G. Wu, J. Wei, Y. Shen, H. Qi and B. Yin, Transactions on Intelligent Transportation Systems, 23 (6), 4927-4943 (2021).

[2] B. Gomes, J. Coelho and H. Aidos, Intelligent Systems with Applications, 20, 200268.

[3] X. Liu, X. Qin, M. Zhou, H. Sun and S. Han, Transactions on Intelligent Transportation Systems, 2508-2521 (2022).

[4] J. Liu, F. Zheng, X. Liu and G. Guo, Intelligent Transportation Systems Magazine, 221-236 (2009).

[5] T. Lan, X. Zhang, D. Qu, Y. Yang and Y. Chen, Sustainability, 15 (2), 1374 (2023).

[6] C. Xiu, S. Zhan, J. Pan, Q. Peng, Z. Lin and S. C. Wong, Transportmetrica A: Transport Science, 1-37 (2024).

[7] S. Zhang, J. Ma, B. Geng and H. Wang, Electronic Research Archive, 32 (2) (2024).

[8] M. Berlotti, G. S. Di and S. Cavalieri, S. Sensors, 24 (7), 2348 (2024).

[9] L. Liu, C. Li, Y. Yang and J. Wang, J. Sustainability, 16 (23), 10216 (2024).

[10] X. Qi, G. Mei, J. Tu, N. Xi amd F. Piccialli, F. Transactions on intelligent transportation systems, 8687-8700 (2022).

[11] J. Ou, J. Xia, Y. J. Wu and W. Rao, Transportation Research Record, 2645 (1), 157-167 (2017).

[12] Kaggle-Traffic Prediction Dataset, 2024, available at https://www.kaggle.com/code/guanlintao/100-ensemble-traffic-prediction-dataset.

[13] Q. D. Dinh, D. Kunk, T. Son, et al. PloS one, 20 (4), e0319484 (2025).

[14] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, Artificial Intelligence Review, 54, 1937-1967 (2021).

[15] X. Wu, J. Wang. International Journal of Environmental Research and Public Health, 20 (6), 4977 (2023).

[16] R. Shwartz-Ziv and A. Armon, Information Fusion,81, 84-90 (2022).

[17] M. Alamri, M. Ykhlef. IEEE Access, 12, 14050-14060 (2024).