Journal of information and
communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 21(3): 225-232, Sep. 2023 Regular paper

Performance Comparison of Machine-learning Models for Analyzing Weather and Traffic Accident Correlations

Li Zi Xuan and **Hyunho Yang***

School of Software, Kunsan National University, 54150, Republic of Korea

Abstract

Owing to advancements in intelligent transportation systems (ITS) and artificial-intelligence technologies, various machine-learning models can be employed to simulate and predict the number of traffic accidents under different weather conditions. Furthermore, we can analyze the relationship between weather and traffic accidents, allowing us to assess whether the current weather conditions are suitable for travel, which can significantly reduce the risk of traffic accidents. In this study, we analyzed 30000 traffic flow data points collected by traffic cameras at nearby intersections in Washington, D.C., USA from October 2012 to May 2017, using Pearson's heat map. We then predicted, analyzed, and compared the performance of the correlation between continuous features by applying several machine-learning algorithms commonly used in ITS, including random forest, decision tree, gradient-boosting regression, and support vector regression. The experimental results indicated that the gradient-boosting regression machine-learning model had the best performance.

Index Terms: machine learning, decision tree, gradient-boosting regression (GBR), support vector regression (SVR)

I. INTRODUCTION

Transportation networks are integrated into our daily lives. Modern cities boast highly complex transportation networks, the state of which influences the extent of urban development.

Consequently, to address issues related to transportation networks and increase their efficiency, countries worldwide have begun constructing intelligent transportation systems (ITS) that integrate a wide array of IT technologies, including data analytics, computer systems, data communications, sensors, artificial intelligence, and advancements in the transportation sector, such as service control and vehicle manufacturing.

Transportation networks often consist of intricate routes, with roads and highways serving as primary arteries. As road mileage and traffic flow increase, road accidents can cause significant damage. Adverse weather conditions are among the major causes of traffic accidents. Complex and variable weather conditions such as heavy rain, fog, clouds, and strong winds can severely affect high-speed vehicles. According to a statistical report, an average of six million crashes occur annually in the United States, with approximately one-fifth being weather-related. According to a study conducted by safety regulators, traffic accidents in the United States cost $340 billion a year, which is equivalent to $1000 per person for 328 million people [1].

This article is informed by a meticulous review and analysis of a vast body of relevant material from the past three years. Machine-learning models developed by various scholars, such as Zeroual et al. [2], Ahmed et al. [3], Liu [4], Tahir and Rashid [5], Sajan and Kumar [6], and Zhou et al. [7], have been explored in depth. We employed a diverse array of machine-learning models in our experiments.

By better analyzing and predicting the relationships

Received 12 June 2023, **Revised** 25 August 2023, **Accepted** 29 August 2023

***Corresponding Author** Hyunho Yang (E-mail: hhyang@kunsan.ac.kr, Tel: +82-63-469-4857)
School of Software, Kunsan National University, 54150, Republic of Korea

Open Access https://doi.org/10.56977/jicce.2023.21.3.225 print ISSN: 2234-8255 online ISSN: 2234-8883

between various weather factors and the number of traffic accidents, we can reduce the economic and human losses resulting from weather-related automobile accidents. In this study, a dataset of the traffic flow at nearby intersections in Washington, D.C., USA. from 2012 to 2017 was used, which encompassed various characteristics, such as air pollution, humidity, wind speed, wind direction, weather type, traffic flow, and traffic accidents. We primarily focused on analyzing the correlations between different weather factors and traffic accidents and comparing the prediction accuracy and performance evaluation of different machine-learning models for traffic accidents. The machine-learning models used in this study are outlined in the following sections.

The remainder of this paper is organized as follows. Section II discusses prior studies pertinent to our experiment, including their strengths and weaknesses. It also outlines the enhancements and innovations introduced by our study in comparison with previous work. Section III describes the parameters associated with machine-learning models, datasets, and data features. It also presents a predictive framework and process based on various machine-learning models, along with the ultimate data interpretation and evaluation metrics for the prediction results. In Section IV, we discuss the final prediction results obtained via different machine-learning models and compare them to identify the most effective models. Finally, Section V concludes the paper.

II. RELATED WORKS

The use of machine-learning models to simulate the correlation between weather conditions and traffic accidents is a data-driven methodology. This approach requires the collection, preprocessing, and application of substantial weather and traffic accident data for model training. It presents numerous challenges, such as obtaining high-quality, representative data, aligning weather and traffic accident data temporally and spatially, and selecting and fine-tuning machine-learning models to reveal the inherent relationship between the two factors. Zeroual et al. [2] proposed a machine-learning-based approach for predicting road traffic density, in which multiple data sources, such as historical traffic data, weather conditions, and temporal factors, are leveraged to build predictive models. Although this method improves forecasting accuracy, the model construction and training require substantial data and computational resources, requiring high precision and data integrity. Ahmed et al. [3] evaluated various machine-learning algorithms, such as decision trees, random forests (RFs), support vector machines (SVMs), and neural networks, for predicting the severity of road accidents. Each algorithm has unique advantages and limitations, necessitating careful consideration and selection based on the specific application scenarios. Liu [4] proposed a short-term traffic flow prediction method based on support vector regression (SVR). Although it provides accurate traffic flow forecasts, this method requires professional knowledge and time for parameter adjustment and model training. Furthermore, the accuracy and reliability of weather data can significantly affect the forecasting results. Tahir and Rashid [5] presented an approach to aid road weather and traffic services using ITS, in which machine-learning models are used to analyze and predict road weather and traffic conditions. However, the effectiveness of this approach is contingent on the quality and availability of ITS data, and considerable computational resources may be needed, depending on the complexity of the models and algorithms used. Sajan and Kumar [6] focused on predicting train delays and analyzing the impact of weather data using machine-learning techniques. The accuracy of their method, which helps improve train scheduling and delay predictions, depends on the quality and availability of the historical data. Furthermore, the performance of the method may be affected by factors beyond weather data.

In summary, the referenced studies present diverse methodologies for predicting road traffic density, road accident severity, short-term traffic flow, road weather, traffic services, and train delays using machine learning. Building on these methodologies, in this study, we analyzed the principles, characteristics, and applicability of different machine-learning models, innovatively encoded various types of weather data, and employed machine-learning models to analyze traffic accident frequency. Four distinct machine-learning models were compared, with different evaluation methods applied to analyze the final data, yielding promising results.

III. DATA PREPARATION PROCESS AND METRICS

A. Descriptive Statistical Distribution of Continuity Characteristics of Dataset

The dataset employed in this experiment was derived from traffic flow data at an intersection in Washington, D.C., USA spanning from October 2, 2012 to May 17, 2017 [8]. It encompassed important characteristics, such as air pollution, humidity, wind speed, wind direction, weather type, traffic flow, and traffic accidents. To better analyze the correlation between successive eigenvalues, we utilized Pearson correlation coefficients and heat maps, which are widely employed in the natural sciences to depict linear correlations between two variables. The formula for the Pearson correlation coefficient is as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} . \tag{1}$$

Table 1 presents a descriptive statistical heatmap illustrating the relationships between the continuous features of the dataset used and the two functional groups employed.

1. Time characteristics (e.g., holidays, weekends, hourly intervals, years, months)
2. Weather characteristics (e.g., rain, snow, temperature, and wind direction)

The heatmap, which is a visual representation of data showing the levels of correlation, reveals a clear and strong linear relationship between traffic accidents and traffic flow.

This implies that as traffic flow increases or decreases, the number of traffic accidents increases or decreases correspondingly, indicating a direct proportional relationship. With more vehicles on the road, accidents are more likely to occur. Additionally, an interesting pattern that emerged from the analysis was the linear relationships among temperature, traffic flow, and the number of accidents within specific hourly segments. This indicates that the variation in temperature, whether rising or falling, affects the traffic flow and the occurrence of accidents during these hours. For instance, during warmer or colder hours, traffic flow may increase because people prefer to travel in their private vehicles for comfort. Consequently, the increased traffic flow lead to an increase in accidents. This paper also presents a fascinating comparison between the likelihoods of collisions on rainy and snowy days.

Surprisingly, rainy conditions have a higher probability of collision than snowy conditions. This can be due to various factors, such as the more frequent occurrence of rainy days compared with snowy days or drivers being more cautious while driving in snow owing to the known hazardous conditions, leading to fewer accidents. By carefully analyzing the dataset, which included variables such as traffic flow, weather conditions, and temperature, we paved the way for improved predictive modeling in the subsequent step. The insights obtained from this analysis can be used to increase the accuracy of the predictive model, enhancing its ability to forecast traffic conditions and accident scenarios. The proposed model can benefit city planners, traffic management authorities, and drivers, contributing to the overall safety and efficiency of road transportation.

B. Machine-Learning Models

The machine-learning model employed in this experiment is a self-guided regression model that predicts road traffic accidents through simulation under given weather conditions. In machine learning, regression models comprise mathematical functions and associated mapping relationships for studying unique patterns and degrees of influence between independent and dependent variables. These models are commonly used for forecasting and time-series analyses. After analyzing the dataset and creating a regression model, regression analysis was performed to examine how well the straight lines fit the data points. Finally, the best-fitting regression model with the smallest deviation value was determined. To achieve this, two regression algorithms were used: decision tree and linear SVR [3]. In addition, two ensemble learning algorithms were used for model training: RF and gradient boosting.

1) Decision Tree

For handling continuous data, decision-tree regression models are widely used. A decision-tree regression is a binary tree created by identifying the best feature value *j* of the dataset and the optimal partition point *s* to divide it during processing (2). The output value (*j*, *s*) obtained after the initial partitioning of the region is evaluated by applying a loss function. The recursion continues until the specified conditions are satisfied.

$$\mathrm{m}_{j,s}[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2] \tag{2}$$

The region is divided by the selected pair (*j*, *s*), and the corresponding output values are determined.

$$R_1(j,s) = x|x^{(j)} \leq s, R_2(j,s) = x|x^{(j)} > s$$
$$\hat{c}_m = \frac{1}{N} \sum_{x_1 \in R_m(j,s)} y_i, x \in R_m, m = 1,2 \tag{3}$$

2) Integration Algorithms; Random Forest (RF) and Gradient-Boosting Regression (GBR)

Ensemble algorithms perform training tasks by organizing and aggregating multiple machine-learning algorithms to produce the best learning outcomes and performance metrics. Ensemble methods are commonly employed for regression, classification, and feature-extraction problems. Ensemble learning algorithms can be divided into two types: bagging

Table 1. Data for October 2, 2012

Air pollution index	Humidity	Wind speed	Weather type	Weather description	Traffic volume (Vehicles)	Accidents
121	89	2	Clouds	Scattered clouds	5,545	56
178	67	3	Clouds	Broken clouds	4,516	51
113	66	3	Clouds	Overcast clouds	4,767	43

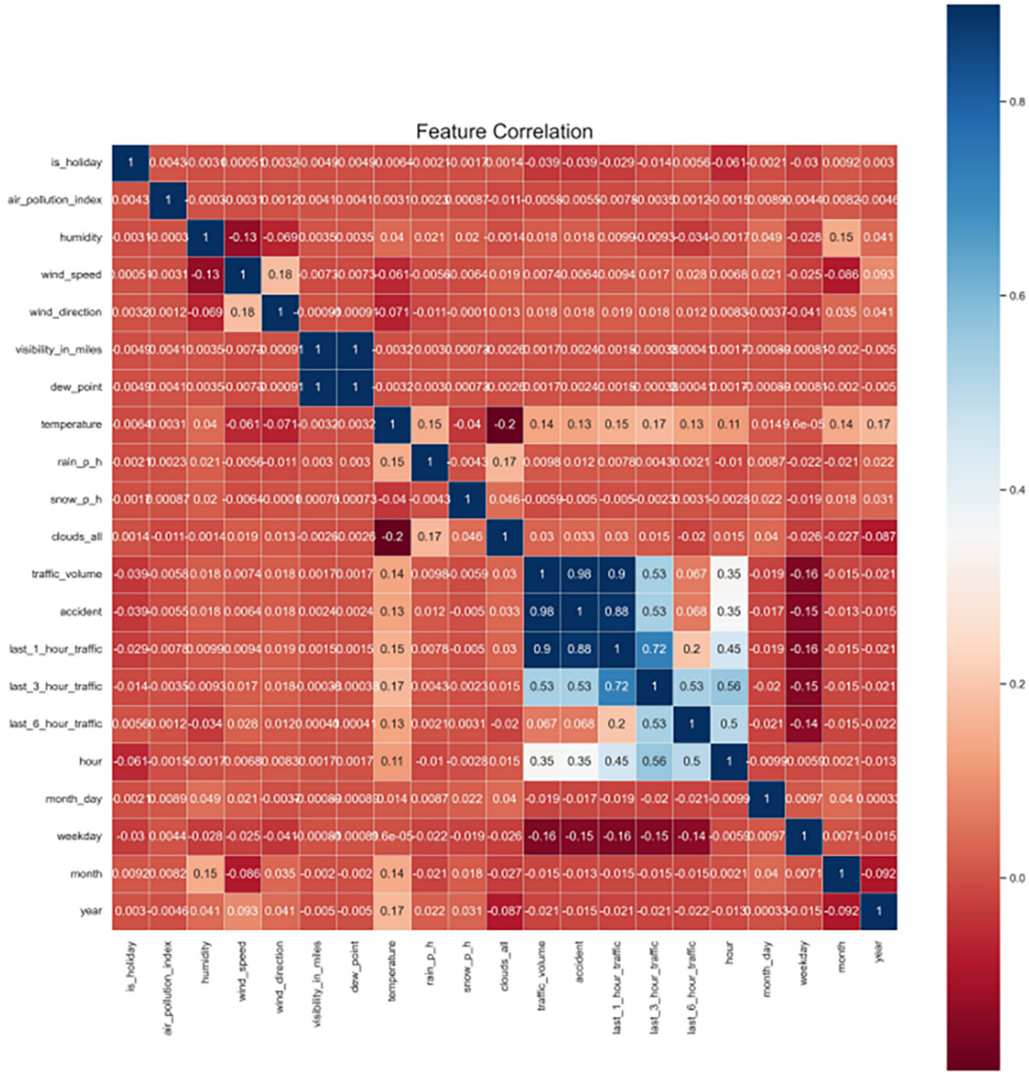


Fig. 1. Continuous eigenvalue heat map.

and boosting. In this experiment, the most representative bagging models were utilized: the RF and gradient-boosting regression (GBR) models.

a) RF

RF is an ensemble algorithm based on decision trees. The weak generalization ability of the decision tree is addressed by the RF, as the decision tree has only one tree with a single decision stream and a limited generalization capacity. An RF with multiple decision streams significantly enhances generalization. Simultaneously, an RF exhibits remarkable randomness, with regard to both random feature selection and random sampling. Given N and M features for each data output, the RF randomly selects X data samples and Y features, subsequently forming multiple decision trees.

RF classifies dataset samples by inputting them randomly, replacing them with multiple decision trees, obtaining the prediction results of all decision trees in parallel, then, taking the average as the final prediction result. Among the parameters for accurate RF prediction, the number of trees in the forest and the number of base estimators are essential; with the larger values of these parameters, the better the performance of the model. Out-of-bag data refer to the data used for testing the model without dividing the dataset into test and training sets when employing an RF.

b) GBR

The GBR algorithm is an ensemble algorithm that enhances the performance of the base algorithm. First, a dataset is input, and different training methods are selected according to the dataset. A weak base learner is trained for each training method, and the residual values of the different weak learners are computed. The process of iteratively adding weak learners is repeated, gradually transforming them into strong learners and reducing the loss function. In the gradient-boosting implementation, the prediction function after *m* iterations is assumed to be $F_m(x)$, and the corresponding loss function is $L(y, F_m(x))$. To achieve the fastest reduction in the loss function, the $(m + 1)$ iteration sub model function should be constructed along the gradient direction of the loss function. The direction of the gradient descent at this point is

$$-g_m(x) = -[\frac{\partial L(y,F(x))}{\partial F(x)}]_{F(x)=F_m(x)}. \tag{4}$$

Additionally, to solve the overfitting problem that tends to arise in GBR, regularization is generally used simultaneously to ensure the accuracy of the final prediction results.

3) Linear SVR

In machine learning, SVMs are utilized not only to address classification problems but also to tackle regression problems; this is referred to as SVR.

One of the models generated by support vector classification (SVM) relies solely on the training dataset, because the cost function for constructing the model disregards learning points that are out of bounds. Similarly, the models produced by SVR are characterized by a model-creation cost function that ignores any training data close to the model prediction. There are three types of SVR: regular SVR, Nusselt SVR, and linear SVR.

Generally, different kernel functions are employed to construct various models. The following kernel functions are commonly used.

Linear kernel function

$$kernel = < x, x' > \tag{5}$$

Polynomial kernel functions

$$kernel = (\gamma < x, x' > + r)^d \tag{6}$$

RBF kernel function

$$kernel = exp(-\gamma \| x - x' \|^2) \tag{7}$$

In this experiment, the linear SVR method was selected for model construction, according to the variable characteristics of the dataset.

C. Specific Process

The specific procedure of the experiment was as follows:

(i) The collected raw dataset was imported, and data cleaning and correlation analysis were performed.

(ii) The original dataset was encoded and the predictors were defined to assess the accuracy of the model.

(iii) The dataset was divided into test and validation datasets and passed onto different machine-learning mode

(iv) Line graphs of measurements and predictions based on training datasets under different machine-learning models were plotted, and scatter plots and bar charts were created for comparison of the coefficient of determination (R^2) and root-mean-square error (RMSE), along with box plots of the prediction error values.

(v) The prediction metrics were calculated for different machine-learning models.

(vi) The final results were statistically analyzed to determine the optimal model.

A flowchart of the experiment is shown in Fig. 2

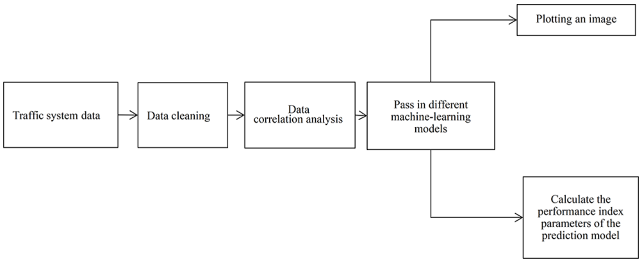

Traffic system data
Data cleaning
Data correlation analysis
Pass in different machine-learning models
Plotting an image
Calculate the performance index parameters of the prediction model

Fig. 2. Flowchart of the experiment.

D. Predicted Performance Indicators

To evaluate the prediction performance, we employed four of the most widely used evaluation indicators in the field of regression learning: the mean squared error (MSE), RMSE, mean absolute error (MAE), and R^2. The MSE is sensitive to outliers in the data; a larger value indicates a more significant error. The RMSE is a typical prediction indicator for regression models; a larger error corresponds to a greater impact. The MAE is used to test the magnitude of the error between the predicted and actual values; a smaller value is better. R^2 is the default metric used by the scikit-learn Python module when implementing linear regression; a value closer to 1 corresponds to better performance. The formulas for these metrics are as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \in [0, +\infty) \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \in [0, +\infty) \tag{9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |, \in [0, +\infty) \tag{10}$$

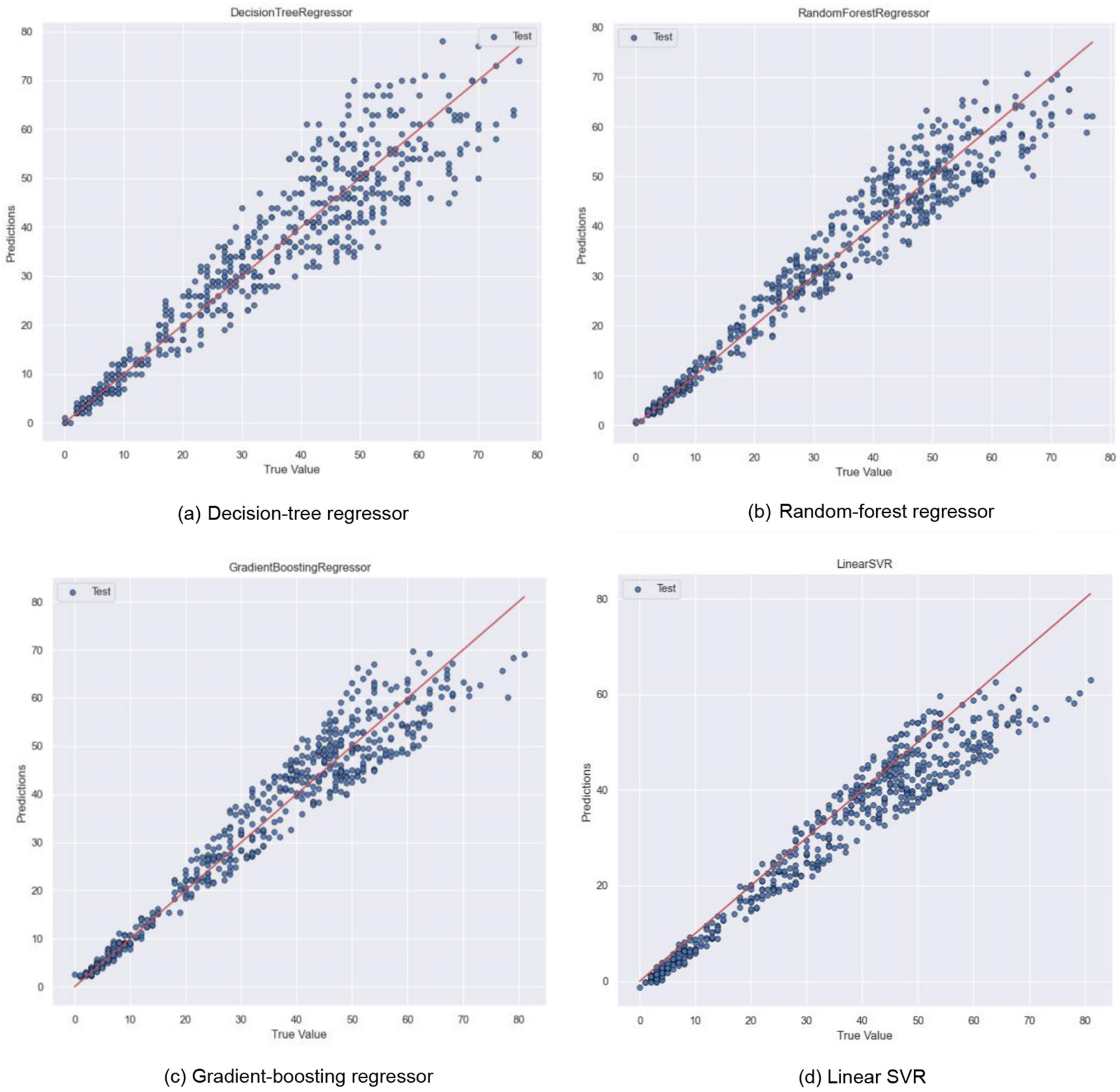
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y)^2} \in [0,1]. \qquad (11)$$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To test and validate the prediction results of various machine-learning models, the dataset was divided into test and training sets. The training set was first used for training, and the test set was then used for prediction validation. Table 2 presents the prediction performance metrics of various machine-learning models. As shown, the prediction metrics derived using the decision-tree algorithm were significantly worse than those for the other three algorithms. Therefore, the decision-tree algorithm was not applicable to the prediction task in this experiment. Among the remaining three algorithms, both the RF and GBR algorithms involve ensemble learning. Therefore, the two algorithms were compared separately. The GBR algorithm had slightly better

Table 2. Model evaluation based on the MSE, RMSE, MAE, and R^2 metrics

	Decision tree	RF	GBR	Linear SVR
MSE	44.56	21.62	20.92	27.23
RMSE	6.24	4.90	4.76	5.13
MAE	4.76	3.4419	3.3980	3.8354
R^2	0.91	0.94	0.95	0.93



(a) Decision-tree regressor

(b) Random-forest regressor

(c) Gradient-boosting regressor

(d) Linear SVR

Fig. 3. Scatter plots of the predicted weather type versus the number of car accidents for the test datasets.

In addition, linear SVR exhibited good prediction metrics.

As shown in Fig. 3, the scatter plots of the decision-tree regression algorithm exhibited a higher degree of scattering than those of the other three algorithms.

The scatter plots of the remaining three algorithms aligned better with the actual scenarios. The scatter plots obtained using the GBR algorithm were the most evenly distributed on both sides, indicating the best prediction performance.

Table 3 presents the ratio of correctness for the number of traffic accidents; under real conditions vs. the predicted values of the four models. From Fig. 4 and Table 3, we can see that RF and GBR outperform other algorithms, with the GBR algorithm showing the best prediction performance. Fig. 4 presents box plots of the prediction error values for the four machine-learning algorithms. As shown, the GBR algorithm converged close to 0, exhibiting superior performance to the other three machine-learning algorithms.

Table 3. Similarity ratio between the real situation and the line chart for the four models

	Decision tree	RF	GBR	Linear SVR
Prediction vs. Real	79.83	85.41	85.80	83.20

Figs. 5 and 6 present bar graph comparisons of the R^2 and RMSE, respectively, of the four machine-learning algorithms. As shown in Fig. 5, the GBR algorithm had the best results for R^2, with a value close to 1. Similarly, as shown in Fig. 6, the GBR algorithm had the smallest RMSE, i.e., the best results.

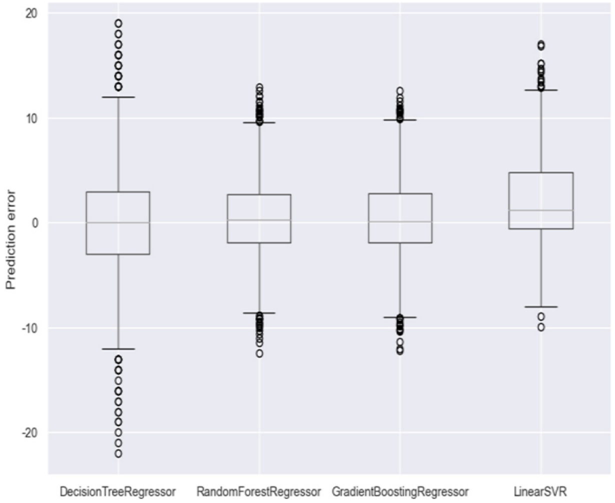


Fig. 4. Box plot of the prediction error values.

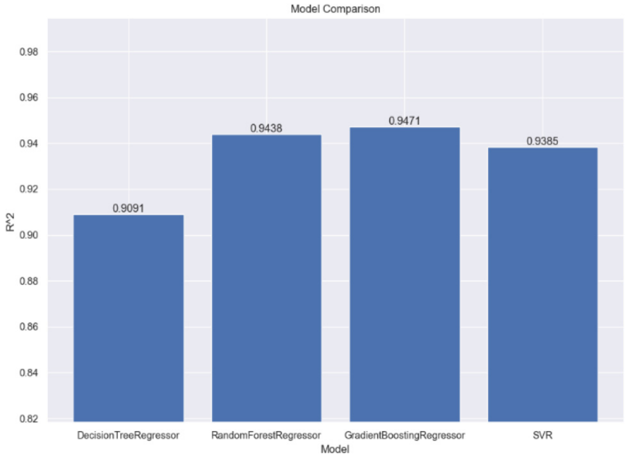


Fig. 5. R^2 bar chart.

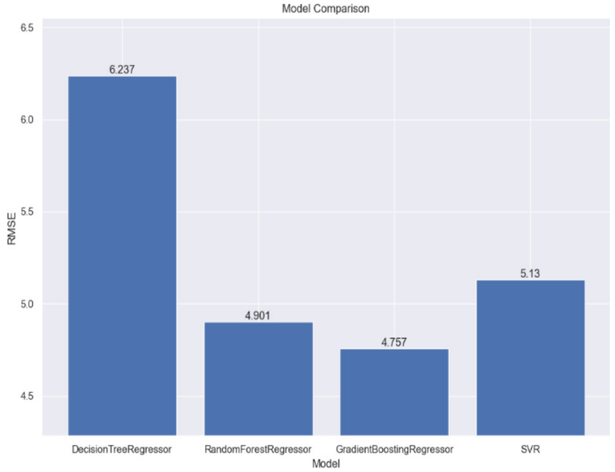


Fig. 6. RMSE bar chart.

V. CONCLUSION

By simulating and predicting the relationship between weather types and the number of traffic accidents using accurate machine-learning algorithms, we can effectively avoid losses, including personal injury, economic damage, and property destruction, caused by weather factors and ensure travel safety. In this study, we compared four machine-learning algorithms: decision tree, RF, GBR, and linear SVR. The GBR model exhibited the best performance in predicting the relationship between weather types and traffic accidents. In future work, we will use machine-learning models to predict the relationship between more complex and integrated weather factors and traffic accidents. Additionally, existing machine-learning models will be optimized to increase the prediction accuracy.

ACKNOWLEDGEMENTS

This study was supported by an industry-academic cooperation R&D program funded by the LX Spatial Information Research Institute (LXSIRI, Republic of Korea) [Project Name: Prediction of water storage rate against drought in agricultural reservoirs using spatial information-based artificial intelligence analysis/Project Number: 2023-501).

REFERENCES

[1] US News & World Report. (2023, January 10). US Study: One Year of Road Crashes Cost Society $340 Billion. [Online] Available: https://www.usnews.com/news/business/articles/2023-01-10/us-study-one-year-of-road-crashes-cost-society-340-billion.

[2] A. Zeroual, F. Harrou, and Y. Sun, "Predicting road traffic density using a machine learning-driven approach," in *Proceedings of 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Cape Town, South Africa, pp. 1-6, 2021. DOI: 10.1109/ICECET52533.2021.9698639.

[3] S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan, and S. K. Ray, "A comparative study of machine learning algorithms to predict road accident severity," in *Proceedings of 2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, London, United Kingdom, pp. 390-397, 2021. DOI: 10.1109/IUCC-CIT-DSCI-SmartCNS55181.2021.00069.

[4] L. Liu, "A Short-term traffic flow prediction method based on S. V. R.," in *Proceedings of 2021 2nd International Conference on Urban Engineering and Management Science (ICUEMS)*, Sanya, China, pp. 1-4, 2021. DOI: 10.1109/ICUEMS52408.2021.00008.

[5] M. N. Tahir and U. Rashid, "Demo: Intelligent transport system (ITS) assisted road weather & traffic services," in *Proceedings of 2020 IEEE Vehicular Networking Conference (V.N.C.)*, NewYork, USA, pp. 1-2, 2020. DOI: 10.1109/VNC51378.2020.9318377.

[6] G. V. Sajan and P. Kumar, "Forecasting and analysis of train delays and impact of weather data using machine learning," in *Proceedings of 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. 1-8, 2021. DOI: 10.1109/ICCCNT51525.2021.9580176.

[7] W. Ming Zhou, T. Shan, C. Chao, and Z. Jian dan, "Short-time traffic flow forecast with weather characteristics," in *Proceedings of 2020 International Conference on Computer Communication and Network Security (CCNS)*, Xi'an, China, pp. 142-145, 2020. DOI: 10.1109/CCNS50731.2020.00039.

[8] Jacob, "Traffic-Accident-Data Dataset for machine learning" [Online] Available: https://www.kaggle.com/datasets/bobaaayoung/trafficaccidentdata.

[9] S. Zhang, "Urban traffic accident prediction research based on meteorological data," in *Proceedings of 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE)*, Guilin, China, pp. 130-133, 2022, DOI: 10.1109/MLKE55170.2022.00031.

[10] M. Hellweg, J.-W. Acevedo-Valencia, Z. Paschal Idi, J. Nachtigall, T. Kretsch, and C. Stiller, "Using floating car data for more precise road weather forecasts," in *Proceedings of 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, Antwerp, Belgium, pp. 1-3, 2020. DOI: 10.1109/VTC2020-Spring48590.2020.9129401.

Li Zi Xuan

Li Zi Xuan received a B.E. degree from Xiamen University of Technology. He is currently pursuing an M.S. degree at the School of Software at Kunsan National University. His research interests include Big Data and machine learning.

Hyunho Yang

Hyunho Yang is currently a Professor at the School of Software, Kunsan National University, Kunsan, South Korea. His research interests include deep learning, machine learning, ubiquitous and pervasive computing, and Big Data.