Using Supervised Machine Learning to Predict the Status of Road Signs

Halil İbrahim ÇETİN

Abstract

Abstract

Traffic road signs play a critical role in ensuring road safety and efficient transportation systems. These signs communicate essential rules and conditions of the road to drivers, enabling them to make informed decisions. Failure to adhere to traffic signs or inadequate visibility of these signs can lead to severe consequences, including accidents, injuries, and fatalities. Despite their importance, many road signs suffer from issues such as wear and tear, low visibility, or inadequate retroreflectivity, making it difficult for drivers to recognize them, especially in low-light conditions.

The retroreflectivity of road signs is a crucial factor that enhances their visibility, particularly at night or under poor lighting conditions. Proper manufacturing standards and material selection are essential to ensure that road signs meet visibility regulations. Retroreflective sheeting materials are commonly used to improve the legibility of signs, allowing drivers to identify them from a safe distance.

This study explores the application of machine learning techniques to predict the status of road signs—whether they meet visibility and regulatory standards. Three algorithms—Random Forest, Artificial Neural Network (ANN), and Support Vector Machines (SVM)—were tested, and the impact of preprocessing techniques, including data scaling and Principal Component Analysis (PCA), on their performance was analyzed.

Data scaling methods such as normalization and standardization were applied to prepare the dataset, and their effects on prediction accuracy were investigated. PCA was utilized to reduce the dimensionality of the dataset, with varying impacts on the algorithms. While PCA improved the accuracy of ANN, it reduced the accuracy of SVM and had no significant effect on Random Forest when scaling was applied. The findings revealed that data scaling enhances the accuracy of all three models, with standardization proving more effective than normalization.

The study underscores the importance of maintaining high-quality road signs with adequate retroreflectivity to ensure road safety. By leveraging machine learning models, road signs can be evaluated efficiently, providing a cost-effective and reliable solution for ensuring compliance with visibility standards.

1. Introduction

Road safety is a critical aspect of transportation systems and road signs play a vital role in ensuring both the safety and efficiency of road networks. Low visibility of road signs significantly increases the risk of accidents, potentially leading to material damage, human injuries, and even fatalities. Therefore, road signs must be easily recognizable, and the information they convey should be legible from an appropriate distance.

The colors of road signs aid drivers in distinguishing between different types of signs during the daytime. However, the importance of recognizing and reading road signs becomes even greater at nighttime when lighting conditions are poor or nonexistent. During these conditions, retroreflective sheeting material is used to enhance the visibility of road signs. For optimal performance, the retroreflective material on road signs must meet established standards of retroreflectivity. The type of retroreflective material and its color are crucial factors that influence how well the material resists degradation under varying environmental conditions .

Accurately predicting the condition of road signs is essential for several purposes, including determining whether a road sign meets visibility requirements and deciding when replacement or maintenance is necessary. However, information about the status, placement, and condition of road signs is often limited. A traditional approach involves manually evaluating road signs by measuring retroreflectivity and color coordinates and comparing these measurements with regulatory standards. Such manual processes are time-consuming and prone to errors, leading to inconsistent evaluations.

This paper proposes an alternative approach that leverages machine learning algorithms for efficient and accurate evaluation of road signs. The aim is to predict the status of road signs using three different machine learning models: Random Forest, Artificial Neural Network (ANN), and Support Vector Machines (SVM). These models are compared based on their performance and accuracy in predicting whether a road sign meets the required visibility standards.

The dataset used to train the machine learning models includes measurements of key features such as sign color values, retroreflection levels, age, placement, and orientation of the signs. By utilizing these features, the models aim to classify road signs as either approved or not approved based on visibility criteria. This machine learning-based approach provides a faster and more reliable alternative to traditional manual evaluations, ultimately contributing to enhanced road safety and maintenance efficiency.

2. Literature Review

Recent studies have highlighted the critical role of retroreflectivity and luminance in enhancing the detectability and legibility of road traffic signs. These studies emphasize that retroreflectivity is influenced by a combination of factors, including vehicle-related aspects (such as headlight color and illumination angle), environmental conditions (weather and ambient lighting), and sign-specific characteristics (such as the type and color of the retroreflective sheeting material).

The development of retroreflective material dates back to the 1930s, and its use for road signs became widespread in the 1970s. Today, nearly all road signs incorporate retroreflective materials, which have undergone significant advancements in recent years, leading to the adoption of new manufacturing techniques.

Previous research has shown challenges in accurately predicting the service life of road signs based on minimum retroreflectivity levels. Poor correlations and unrealistic predictions were often attributed to significant variability in environmental conditions during data collection. Robust models for predicting retroreflectivity based on in-service measurements remain difficult to achieve due to the influence of these external factors.

Some studies have attempted to model the relationship between traffic sign retroreflectivity and factors such as age using statistical approaches like binary logistic regression. While these models demonstrated that sign age significantly contributes to the predictive accuracy, they also concluded that sign age alone is insufficient to determine the functional service life of a road sign. This limitation highlights the need for more comprehensive approaches that account for multiple variables to improve the reliability of predictions.

The findings from existing research underscore the complexity of predicting the condition and longevity of road signs, suggesting the importance of incorporating multiple factors and advanced methodologies, such as machine learning, to enhance prediction accuracy and support effective road sign management.

3. Data Pre-Processing

The dataset utilized in this study includes various parameters essential for analyzing the performance and condition of road signs. Key features in the dataset comprise retroreflectivity values, color measurements taken at different locations on the signboard, geographical coordinates of the signs, the year of manufacturing, and additional geographic and environmental details.

Before applying machine learning algorithms, data preprocessing steps are crucial to ensure the quality and usability of the dataset. These steps typically involve:

4. Data Scaling

Data scaling is a crucial step in data preprocessing that significantly impacts the performance of machine learning algorithms. Many machine learning methods perform more effectively when the features in the dataset have the same scale. To achieve this, two commonly used scaling methods are normalization and standardization.

Normalization

Normalization is the process of transforming features into a specific range, typically between 0 and 1, which enhances the compatibility of input variables within a model. This technique ensures that the magnitude of variables does not disproportionately influence the model. A commonly used method for normalization is the Min-Max Scaler, where each feature is scaled according to the following formula:

$$\mathrm{X}_{normalize} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Here:

- x_i_i is the original value of the feature.
- $\mathrm{X}_{normalize}$ is the normalized value.
- $\min(x)$ and $\max(x)$ are the minimum and maximum values of the feature, respectively.

By applying normalization, the range of features is confined, making it particularly suitable for algorithms sensitive to absolute magnitudes, such as neural networks.

Standardization

Standardization is another scaling technique that transforms data such that the mean of each feature is zero and the standard deviation is one. This method is especially useful when the dataset includes features with mixed units or scales. The standardization formula is as follows:

$$\mathrm{X}_{standardize} = \frac{x_i - \mathrm{mean}(x)}{\sigma(x)}$$

Where:

- x_ii is the original value of the feature.
- $\mathrm{X}_{standardize}$ is the standardized value.
- $\mathrm{mean}(x)$ is the mean of the feature.
- $\sigma(x)$ is the standard deviation of the feature.

Standardization is widely applied in machine learning models, particularly those relying on gradient descent, as it helps optimize convergence by ensuring that all features contribute equally to the model's learning process.

Both normalization and standardization are essential preprocessing techniques, and the choice of method depends on the specific requirements of the machine learning algorithm being used. Implementing these scaling techniques appropriately can lead to improved model performance and more reliable predictions.

5. Data Analysis

Before using the data to train and test the machine learning algorithms, there will be a number of preliminary steps to be applied to prepare this data. This includes data splitting and correlation analysis.

4.1. Data Splitting using K-fold Cross Validation (CV)

Cross-validation is a fundamental concept in machine learning that ensures the data used for training is separate from the data used for testing the model. This approach helps to evaluate the model's performance more reliably and reduces the risk of overfitting.

In this study, the dataset was split using K-fold cross-validation, where the dataset is divided into KKK equal subsets. For each iteration, K−1K-1K−1 subsets are used for training, and the remaining subset is used for testing. This process is repeated KKK times, with each subset being used as the test set exactly once. The performance metrics are calculated for each iteration, and the final performance metrics are obtained by averaging the results across all iterations.

In the original methodology, K=3K=3K=3 was selected. However, in this study, K=5K=5K=5 was chosen to provide a more robust evaluation of the model's performance. This adjustment ensures a more comprehensive assessment by using smaller test sets in each fold, thereby increasing the number of training examples per iteration.

By implementing 5-fold cross-validation, the evaluation process achieves greater reliability and ensures that the model's performance is measured consistently across multiple subsets of the data.

4.2. Correlation Analysis

A correlation analysis is performed to find the dependent features and extract the possible relationships between them. The correlation between the features in the dataset is represented in a heatmap, as depicted in Fig. 1. The heatmap is used to give a primary idea about correlation between the input variables listed in Table 2. Only two features were found to be correlated which are the color values in the CIE color system and the GPS coordinates of the road sign. These correlations are natural and did not affect the results. The features in the dataset were uncorrelated, therefore, the PCA can be used to find the representation of the data.

4.3. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a commonly used dimensionality reduction technique that can speed up machine learning algorithms and help classifiers make more accurate decisions. PCA works by applying an orthogonal transformation to convert a set of correlated features into a smaller number of linearly uncorrelated variables, known as principal components (PCs).

The principal components are ranked in descending order based on the amount of variance they explain in the dataset. The maximum number of principal components is limited by the number of original features in the dataset. The selection of the number of PCs to retain is a critical decision, as it directly impacts the accuracy and efficiency of the model.

One of the key criteria for selecting the number of PCs is the cumulative variance explained by the components. PCA ensures that the majority of the variance in the data is captured within the chosen number of components, allowing the model to operate effectively with reduced dimensionality.

In this study, PCA was applied, and different numbers of principal components were used to train classification algorithms. By reducing the dimensionality of the dataset while preserving the most important information, PCA facilitates faster processing and improves the performance of machine learning models.

4.4. Machine learning algorithms

Three algorithms are used in this study to predict the status of the road sign:

Artificial Neural Network (ANN)

A neural network with an input layer, a hidden layer, and an output layer was implemented in this study. The architecture of the neural network is represented as 10-30-2, where the input layer consists of 10 nodes corresponding to the features, the hidden layer has 30 neurons, and the output layer has 2 nodes representing the two classes (approved/disapproved).

The parameters used for training the neural network include a learning rate of 0.01, the logistic activation function, and Stochastic Gradient Descent (SGD) as the optimization method. In the original methodology, the maximum number of iterations was set to 1000. However, in this study, the maximum number of iterations was increased to 2000 to further refine the model's performance and allow more time for convergence during training.

This modification aims to enhance the model's accuracy by providing additional iterations for the neural network to optimize its weights effectively.

Support Vector Machines (SVM)

SVM is a supervised machine learning model that used in classification problems. It defines a hyperplane between the two classes and extends the margin in order to maximize the distinction between these classes, which results fewer close miscalculations. The performance of SVM largely depends on the kernel and choosing the right kernel can improve the performance of the classifier. In this study, Radial Basis Function (RBF) kernel is used which was found to give the best accuracy. The two parameters to be considered with RBF kernel function are C which is selected to be 2 and gamma which was auto selected.

Random forest

Random forests are a widely used ensemble method based on decision trees for supervised learning tasks. The algorithm operates by constructing multiple decision trees during training and aggregating their outputs to generate predictions. This approach enhances model accuracy and reduces the risk of overfitting by leveraging the collective decision-making of multiple trees.

In the original methodology, the number of trees in the forest was set to 100. However, in this study, the number of trees was increased to 200 to improve the robustness and accuracy of the model. Increasing the number of trees allows the algorithm to capture more patterns in the data, resulting in more reliable predictions while maintaining the inherent benefits of the ensemble approach.

This adjustment aims to ensure better generalization and performance across a variety of datasets and conditions.

.

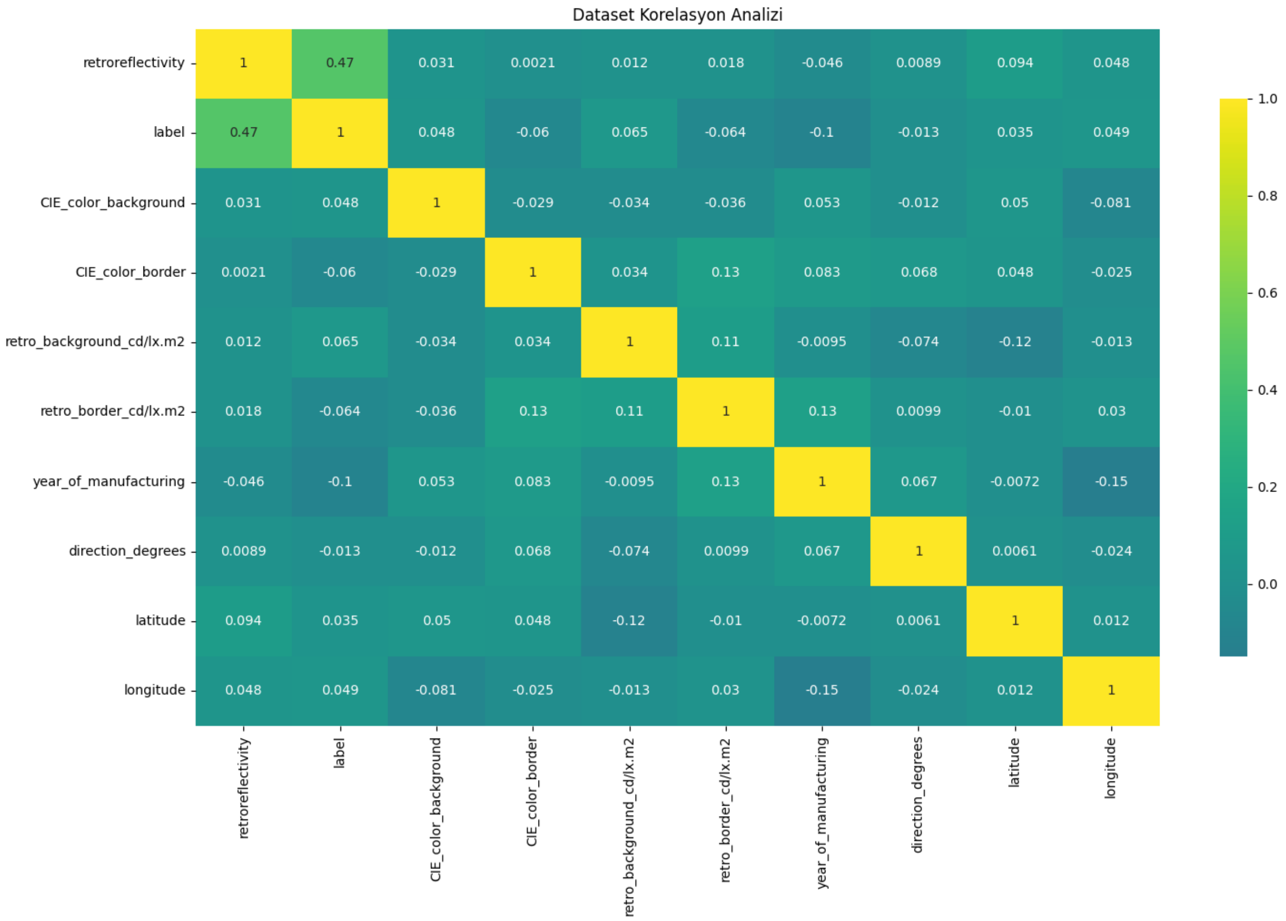

Fig. 1. Correlasion Analysys heatmap of Dataset

6. Results and discussion

As mentioned before, three classification algorithms were tested to predict the status of the road signs and different measurements are used to evaluate the performance of them:

- Accuracy: The correct number of predictions made by the model over all the observed values. The accuracy is calculated by Equation (3), where TP refers to true positive, TN refers to true negative, FP refers to false positive and FN refers to false negative:

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

- Precision: Precision gives how many of the correctly predicted cases actually turned out to be positive. The proportion is calculated with the formula shown in Equation (4):

$$Precision = \frac{TP}{TP + FP}$$

- Recall: Recall gives how many of the actual positive cases the model is able to predict correctly. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN}$$

- F1-score: F1-score is a metric which takes into account both precision and recall and is defined as follows:

$$F1\ Score = 2 * \frac{Precision *\ Recall}{Precision +\ Recall}$$

To evaluate the performance of the three classification algorithms, the predicted values are compared with the ground truth values in the test data. The performance results from the ANN, SVM, and random forest with/without scaling and with/without PCA are presented in Table 4.

Table 4. Classification results.

Algorithm	Scaling	PCA	Accucarcy	Precision	Recall	F1
ANN	No Scaling	No	0,9625	0.960526	1,0000	0,9786
ANN	Normalization	No	0,9625	0,972973	0,9863	0,9759
ANN	Standardization	No	0,9375	0,972973	0,9863	0,9664
ANN	No Scaling	3	0,9750	0,973333	1,0000	0,9864
ANN	Normalization	3	0,9125	0,912500	1,0000	0,9542
ANN	Standardization	3	0,9125	0,923077	0,9863	0,9536
ANN	No Scaling	4	0,9125	0,912500	1,0000	0,9795
ANN	Normalization	4	0,9125	0,912500	1,0000	0,9542
ANN	Standardization	4	0,8750	0,920000	0,9452	0,9324
SVM	No Scaling	No	0,9625	0,960526	1,0000	0,9786
SVM	Normalization	No	0,9625	0,960526	1,0000	0,9786
SVM	Standardization	No	0,9375	0,960526	1,0000	0,9786
SVM	No Scaling	3	0,9125	0,912500	1,0000	0,9542
SVM	Normalization	3	0,9125	0,912500	1,0000	0,9542
SVM	Standardization	3	0,9125	0,960526	1,0000	0,9542
SVM	No Scaling	4	0,9125	0,912500	1,0000	0,9542
SVM	Normalization	4	0,9125	0,912500	1,0000	0,9542
SVM	Standardization	4	0,9250	0,924051	1,0000	0,9605
Random Forest	No Scaling	No	0,9625	0,986486	1,0000	0,9931
Random Forest	Normalization	No	0,9625	0,986486	1,0000	0,9931
Random Forest	Standardization	No	0,9625	0,986486	1,0000	0,9931
Random Forest	No Scaling	3	0,9125	0,947368	0,9863	0,9664
Random Forest	Normalization	3	0,9125	0,912500	1,0000	0,9542
Random Forest	Standardization	3	0,9125	0,911392	0,9863	0,9473
Random Forest	No Scaling	4	0,9125	0,947368	0,9863	0,9964
Random Forest	Normalization	4	0,9125	0,911392	0,9863	0,9473
Random Forest	Standardization	4	0,9250	0,924051	1,0000	0,9605

K-Fold

ANN	0,9675	0,9813	0,9837	0,9824
SVM	0,9425	0,9484	0,9919	0,9669
Random Forest	0,9275	0,9436	0,9973	0,9853

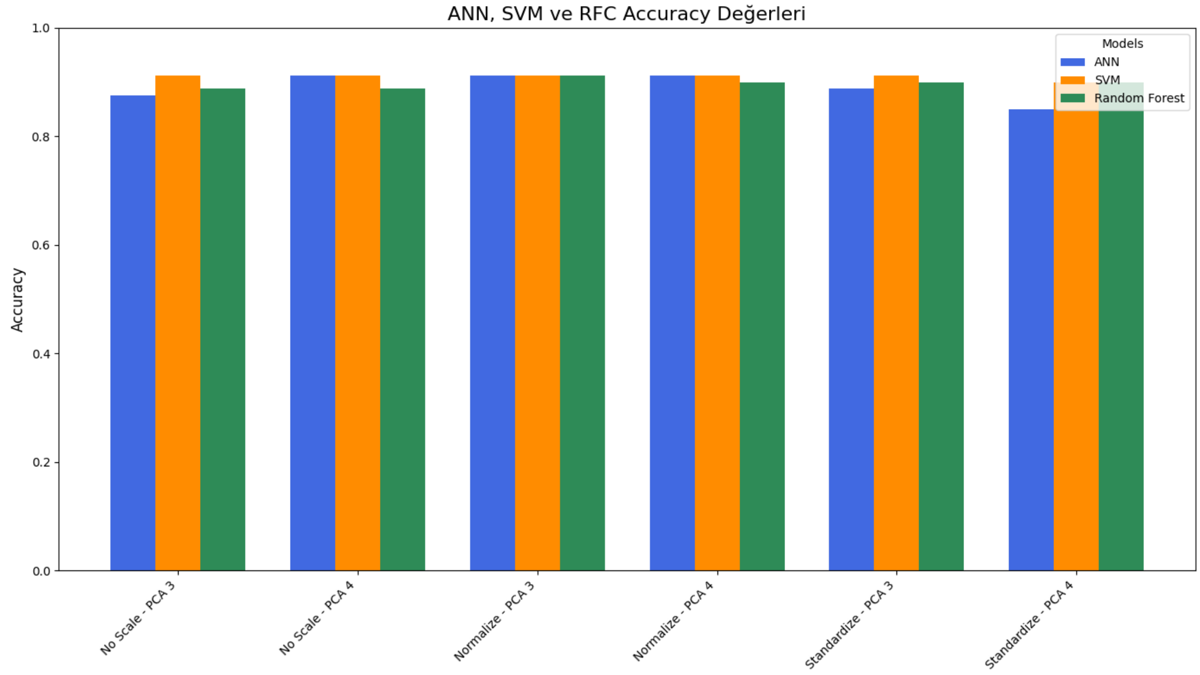

Fig. 2. Performance Result Graphs

7. Conclusion

This paper demonstrates that the status of road signs can successfully be predicted using machine learning algorithms. Three prediction algorithms—ANN, SVM, and Random Forest—were tested and evaluated based on their accuracy, precision, recall, and F1 scores. Among these, the Artificial Neural Network (ANN) achieved the highest overall performance in the K-Fold evaluation, with an accuracy of 96.75%, precision of 98.13%, recall of 98.37%, and F1 score of 98.24%.

Scaling of the dataset played a significant role in improving the performance of the models. Standardization generally provided better results compared to normalization, enhancing the ability of the algorithms to identify patterns in the data. PCA (Principal Component Analysis) was also implemented to reduce the number of features, which not only improved prediction accuracy in some cases but also contributed to speeding up the learning process. However, due to the relatively small dataset used in this study, the impact of PCA on computational speed was not evaluated.

The Random Forest algorithm, which initially showed high performance, achieved an accuracy of 92.75% in the K-Fold evaluation, with precision, recall, and F1 scores of 94.36%, 99.73%, and 98.53%, respectively. The SVM algorithm achieved an accuracy of 94.25%, a precision of 94.84%, a recall of 99.19%, and an F1 score of 96.69%. These results confirm the potential of machine learning models for predicting road sign status with high reliability.

In future work, a larger dataset will be collected and tested to validate the findings and further enhance the models. Additionally, advanced techniques such as transfer learning will be explored, utilizing a minimal number of features for classification to optimize both accuracy and efficiency.