RESEARCH REPORT

Institute of Transportation Studies | UNIVERSITY OF CALIFORNIA

Developing a Safety Effectiveness Evaluation Tool for California

Yanlin Qi, Graduate Research Assistant, Institute of Transportation Studies, University of California, Davis
Jia Li, Ph.D., Assistant Professor, Department of Civil and Environmental Engineering, Washington State University
Michael Zhang, Ph.D., Professor, Department of Civil and Environmental Engineering, University of California, Davis

September 2024

ITS UCDAVIS Institute of Transportation Studies

Report No.: UC-ITS-2022-04 | DOI: 10.7922/G2D50KB8

Technical Report Documentation Page

1. Report No.
UC-ITS-2022-04

2. Government Accession No.
N/A

3. Recipient's Catalog No.
N/A

4. Title and Subtitle
Developing a Safety Effectiveness Evaluation Tool for California

5. Report Date
September 2024

6. Performing Organization Code
ITS-Davis

7. Author(s)
Yanlin Qi http://orcid.org/0000-0001-6572-1093
Jia Li, Ph.D. http://orcid.org/0000-0002-2971-9268
H. Michael Zhang, Ph.D. http://orcid.org/0000-0002-4647-3888

8. Performing Organization Report No.
UCD-ITS-RR-24-14

9. Performing Organization Name and Address

Institute of Transportation Studies, Davis
1605 Tilia Street, Davis, CA 95616

10. Work Unit No.
N/A

11. Contract or Grant No.
UC-ITS-2022-04

12. Sponsoring Agency Name and Address
The University of California Institute of Transportation Studies
www.ucits.org

13. Type of Report and Period Covered
Final Report (October 2021 – September 2023)

14. Sponsoring Agency Code
UC ITS

15. Supplementary Notes
DOI:10.7922/G2D50KB8

16. Abstract

Crash modification factor (CMF) is an effectiveness measure of safety countermeasures. It is widely used by state agencies to evaluate and prioritize various safety improvement projects. The Federal Highway Administration (FHWA) CMF Clearinghouse provides CMFs for a broad range of countermeasures, but still, the existing CMFs often cannot meet the needs for characterizing the safety impacts of countermeasures in new scenarios. Developing CMFs, meanwhile, is costly, time-consuming, and requires extensive data collection. A more cost-effective way to provide preliminary CMF estimations is needed. To address this need, this study develops a low-cost and easily extendable data-driven framework for CMF predictions. This framework performs data mining on existing CMF records in the FHWA CMF Clearinghouse. To tackle the heterogeneity of data, interdisciplinary techniques to maintain model compatibility were created and used. The project also integrates multiple machine-learning models to learn the complex hidden relationships between different safety countermeasure scenarios. Finally, the proposed framework is trained against the CMF Clearinghouse data and performs comprehensive evaluations. The results show that the proposed framework can provide CMF predictions for new countermeasure scenarios with reasonable accuracy, with overall mean absolute errors less than 0.2. We also discuss an enhanced approach that leverages structured information in certain CMF descriptions, which can boost the CMF prediction accuracy, showing a mean absolute error less than 0.1 in a case study.

17. Key Words
Crash modification factors, highway safety, data mining, machine learning

18. Distribution Statement
No restrictions.

19. Security Classification (of this report)
Unclassified

20. Security Classification (of this page)
Unclassified

21. No. of Pages
35

22. Price
N/A

Form Dot F 1700.7 (8-72) Reproduction of completed page authorized

About the UC Institute of Transportation Studies

The University of California Institute of Transportation Studies (UC ITS) is a network of faculty, research and administrative staff, and students dedicated to advancing the state of the art in transportation engineering, planning, and policy for the people of California. Established by the Legislature in 1947, ITS has branches at UC Berkeley, UC Davis, UC Irvine, and UCLA.

Acknowledgments

This study was made possible with funding received by the University of California Institute of Transportation Studies from the State of California through the Road Repair and Accountability Act of 2017 (Senate Bill 1). The authors would like to thank the State of California for its support of university-based research, and especially for the funding received for this project.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the State of California in the interest of information exchange. The State of California assumes no liability for the contents or use thereof. Nor does the content necessarily reflect the official views or policies of the State of California. This report does not constitute a standard, specification, or regulation.

Institute of Transportation Studies | UNIVERSITY OF CALIFORNIA

Developing a Safety Effectiveness Evaluation Tool for California

Yanlin Qi, Graduate Research Assistant, Institute of Transportation Studies, University of California, Davis
Jia Li, Ph.D., Assistant Professor, Department of Civil and Environmental Engineering, Washington State University
Michael Zhang, Ph.D., Professor, Department of Civil and Environmental Engineering, University of California, Davis

September 2024

UCDAVIS Institute of Transportation Studies

Report No.: UC-ITS-2022-04 | DOI: 10.7922/G2D50KB8

Table of Contents

Table of Contents

List of Tables

List of Figures

Executive Summary

Executive Summary

This project aims to develop a machine-learning-based framework that can predict crash modification factors (CMFs) for different safety countermeasures. The framework will mine the data available in the CMF Clearinghouse to uncover hidden CMF relationships and provide a cost-effective solution to estimating CMFs not covered by the CMF Clearinghouse.

Road safety is a top priority for national and state transportation agencies in the United States. To make transportation infrastructure systems safer and reduce fatalities, different countermeasures can be applied to crash-prone locations. However, prioritizing different safety projects can be challenging given the different effectiveness, costs, and benefit-cost ratios of each project, and the limited budget available. CMFs play an essential role in this process, as they quantify the effectiveness of safety countermeasures under different scenarios.

While the CMF Clearinghouse provides practitioners with a list of reliable CMFs developed from individual studies, available CMFs are still not enough to cover all potential countermeasure scenarios of interest to state Departments of Transportation, because of the special infrastructure types or countermeasures to consider. Experimental or observational studies are the dominant approaches for estimating CMFs, but these approaches have limitations, such as requiring years of effort to collect data or adequate amounts of crash data.

To address these challenges, we proposed and implemented a machine-learning-based framework that will mine the data available in the CMF Clearinghouse to uncover hidden CMF relationships and provide a cost-effective solution to estimating CMFs not covered by the CMF Clearinghouse. The proposed framework is cost-effective, time efficient, and fully explores the existing CMF data. We train and test the proposed approach on the CMF Clearinghouse data using extensive experiments. The results show that the framework can predict CMFs with reasonable accuracy. The proposed framework flexibly tackles the heterogeneous data from the CMF Clearinghouse, including capturing the semantic contexts of different countermeasures and maintaining compatibility with the high-cardinality categories, missing rates, and noises.

The proposed approach is unique in that it will make the best use of available knowledge within the CMF Clearinghouse, reducing the data and time burden of estimating CMFs not covered by the CMF Clearinghouse. We also identify factors that limit the performance of our model and its future extensions and data needs. Additionally, we would like to make it clear that it does not replace the traditional cross-sectional or time series-based methods. Rather, it complements the other two when a quick first estimate is needed for a countermeasure scenario with no CMFs backed by these methods.

Contents

Introduction

In the United States, road safety is among the top priorities of national and state transportation agencies. To make transportation infrastructure systems safer and reduce fatalities, different countermeasures can be applied to crash-prone locations. This raises the question of how to prioritize different safety projects, given that each project may have different effectiveness, costs, and benefit-cost ratios, and the total budget available is limited. In this process, the crash modification factor (CMF) plays an essential role, as it quantifies the effectiveness of each safety countermeasure under different scenarios. This makes the decision process more data-informed and tractable. Roughly speaking, a CMF is a multiplicative factor or function used for computing the expected number of crashes after the implementation of a given countermeasure at a specific site. In conjunction with safety performance functions (SPFs), CMFs can be applied to estimate crash numbers after a countermeasure is applied when the past crash number is known.

Several sources provide archives of existing CMFs developed in previous studies. The Highway Safety Manual published by the American Association of State Highway and Transportation Officials provides practitioners with a list of reliable CMFs developed from individual studies, together with the quantitative methods for safety evaluation of facility decisions (1). While the HSM provides only the best available CMFs, the CMF Clearinghouse maintained by Federal Highway Administration (FHWA) serves as a comprehensive online repository and search engine for CMFs (2). This repository summarizes the published work for developing CMFs throughout the world and houses thousands of categorized CMFs and details of their methodologies. In addition to the efforts at the national level, many state Departments of Transportation develop state-specific CMFs to meet their own needs, e.g., evaluating the effectiveness of safety countermeasures with special infrastructure types (e.g., six-lane freeways) or specific driving conditions (e.g., long winters), which are not covered by the Highway Safety Manual or FHWA Clearinghouse. Several state practices have implemented CMF variation studies to narrow and reorganize their CMF lists (3, 4) or safety improvement projects to update CMF data based on jurisdiction-specific crash data (5).

Nevertheless, available CMFs are still not enough to cover all potential countermeasure scenarios of interest to state Departments of Transportation, because of the special infrastructure types or countermeasures to consider. For instance, a CMF for roundabouts may be available from the CMF Clearinghouse, but if that CMF is only applicable to the roundabout conversion of rural, stop-controlled intersections, then there are clear gaps in CMF coverage (e.g., a CMF applicable to the conversion of rural, signalized intersections is missing). Therefore, there is a constant demand for developing additional CMFs for different scenarios of applications.

There are two dominant approaches for estimating CMFs, namely experimental or observational studies, both of which can be based on before-after or cross-sectional data (6). In experimental studies, control-test experiments for the treatment of interest are designed. Studies based on this approach are more rigorous, but it usually takes years of effort to collect data. This long study-cycle hinders its wide applications. Observational studies, in contrast, have two advantages. One is that they bring fewer ethical concerns than experimentation;

another is they can leverage data collected retrospectively from sites with the already implemented treatment to estimate CMFs (7). There are also differences, for observational and experimental studies, between before-after and cross-sectional data requirements. Before-after studies normally require multi-year crash data before and after treatment to make a comparison of the safety performance and estimate CMFs (8–12). Cross-section studies mostly require crash inventory data collected from multiple sites with and without the treatment in a single period (13–15). Both types of studies are only applicable under conditions where an adequate amount of crash data is available.

To address the challenge of limited data, the alternative approaches that are often used for CMF estimation focus on the use of expert knowledge of CMFs. Methods adopted by these studies typically include meta-analysis, expert panels, and surrogate measures (6). The meta-analysis studies estimated an overall CMF by statistically combining CMFs from multiple previous studies (16, 17). Less formally, the expert panel studies leveraged expert knowledge in meetings to derive CMFs (18). The surrogate measure studies mainly focused on developing CMFs by establishing the relationship between changes in surrogates (e.g., vehicle speed) with changes in crashes (19, 20). However, these methods are not cost-effective, because deriving CMF for every single scenario would take a significant amount of expert time. In addition, the potential relevance between similar scenarios is not considered. More advanced approaches to deriving CMFs through knowledge mining are still needed.

Considering the above review, we propose to develop a machine-learning-based framework to mine the CMF Clearinghouse data and uncover hidden CMF relationships in the data. As a one-stop repository with considerable information on CMFs, the CMF Clearinghouse has long been used as a search database, providing safety practitioners with a good start in compiling the safety effectiveness of infrastructure decisions. With thousands of CMF records and future expansions, the CMF Clearinghouse has great potential as a knowledge base worth mining for deriving additional CMFs. However, no knowledge-mining scheme has been developed. Therefore, we sought to develop a knowledge-mining approach for deriving additional CMFs without requiring excess data or time. Furthermore, we illuminate the challenges and gaps this approach faces.

Machine learning (ML) advances have found broad applications in various fields, including safety analysis (21–25). However, studies have applied machine learning models only to predict CMFs under specific crash types, and such models usually require a large amount of crash data. For instance, Wen et al. applied the light gradient boosting machine (LightGBM) and the shapley additive explanation (SHAP) to derive explainable CMFs for run-off-road (ROR) crashes based on more than 28,000 crash records in Washington State (26). Meanwhile, the relations of existing CMF records in the FHWA CMF Clearinghouse are not yet fully understood and explored. The CMF Clearinghouse contains around 9,000 detailed CMF records. Many of these records are explicitly or implicitly related. For example, 'Widen paved shoulder from 6 ft to 8 ft' adds to, and could be related to, an initial description that states 'Widen paved shoulder from 4 ft to 6 ft.' As another example, 'Install chevron signs on horizontal curves' and 'Provide highway lighting' are both intended to improve roadway visibility conditions. Even though descriptions of these countermeasures are different, one may reasonably guess that they have similar effects, and their CMFs are close. Besides countermeasure descriptions, the relationships of different countermeasure scenarios may also be captured by one or more factors in the Clearinghouse (e.g., the countermeasure categories, site conditions, crash types, and others). Thus, the proposed knowledge-mining scheme can formulate the predictive modeling of CMFs into a data-driven framework. This framework can capture the complex underlying relationships among different countermeasure scenarios, which makes it able to predict CMFs for new scenarios.

In short, the main outcomes of this study are as follows:

- We formulate the CMF estimation problem as a knowledge-mining problem, which is new to the literature. A novel data-driven framework is proposed, which is cost-effective, time efficient, tailors to, and fully explores the existing CMF data.
- We train and test the proposed approach on the CMF Clearinghouse data using extensive experiments. The results show that the framework can predict CMFs with reasonable accuracy.
- The proposed framework flexibly tackles the heterogeneous data from the CMF Clearinghouse, including capturing the semantic contexts of different countermeasures and maintaining compatibility with the high-cardinality categories, missing rates, and noise.
- As an extension, we propose an enhanced approach that can predict CMFs with better accuracy by leveraging structured information in certain countermeasures.

We identify factors that limit the performance of our model and its future extensions and data needs. Additionally, this approach does not replace the traditional cross-sectional or time series-based methods, rather it complements the other two when a quick first estimate is needed for a countermeasure scenario with no CMFs backed by these methods.

Methodology

In this section, the data-driven CMF prediction framework is formulated. First, the data source is explored. Then, the machine-learning problem is defined with model input/output, data encoding approach, and regressive methodology introduced.

Data Description

The FHWA CMF Clearinghouse is an open-source and regularly updated repository. As of the completion of this report, this repository houses 8993 records of CMFs and corresponding countermeasures and site conditions. In the native spreadsheet downloaded, there are 59 data fields indicating the applicability of the CMFs (e.g., countermeasure names and categories, intersection types, area types, crash types, and crash severity types), the developing details (e.g., method type, study title, and publish year), and the statistical properties (e.g., standard error rating). Due to the intrinsic data characteristics and other limitations, this data collection consists of structured and unstructured data with missing values.

Data Exploration

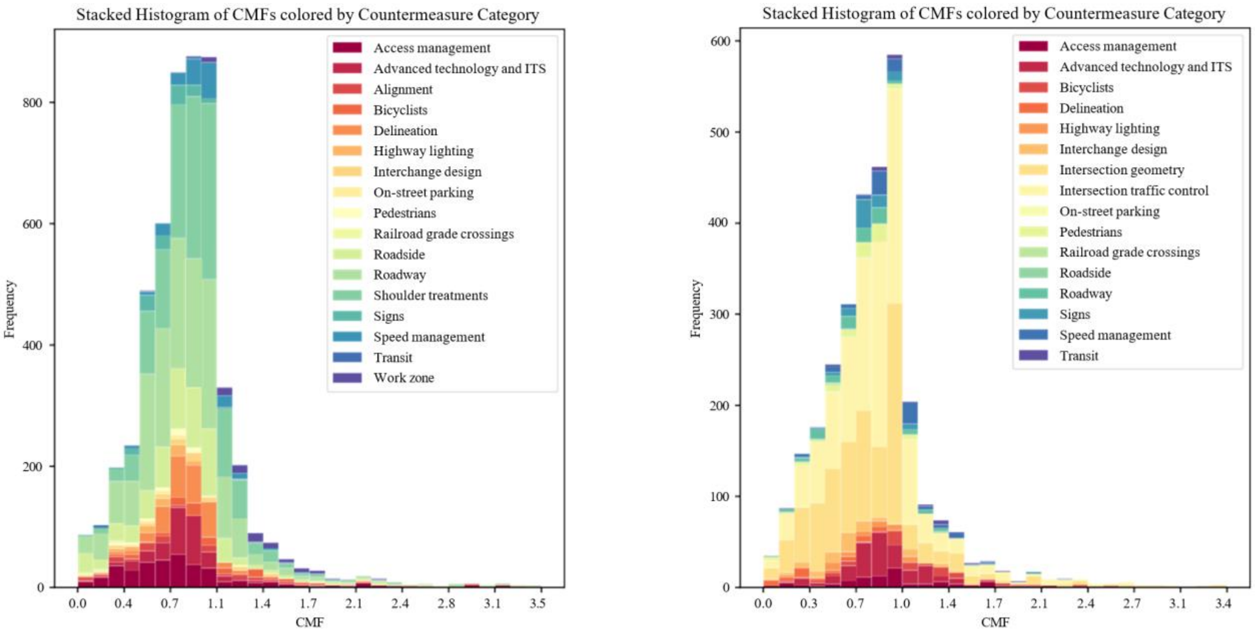
In this database, two main types of countermeasures, namely roadway segment and intersection, can be differentiated by an existing indicator. Countermeasures belonging to one of these two types were thereby reorganized into two separate groups. Such reorganization helps to reduce the deviation caused by missing items on unnecessary data fields (e.g., the intersection type of a roadway segment countermeasure tends to be missing). Meanwhile, it helps to avoid unnecessary input demands when applying the model. Based on this reorganization process, the models for countermeasures in the roadway segment type and the intersection type are developed independently. The results of these two facility types will also be discussed separately. Nonetheless, the proposed scheme is a general framework and works for both types of facilities in the same way.

Countermeasure structure. The countermeasure descriptions are worth further exploration since they carry the semantic contexts of the road safety treatment. As illustrated in Table 1, there are countermeasures that textually contain the quantitative changes due to the treatment (e.g., the structured examples listed in Table 1 and Table 3), which are thereby categorized as structured countermeasures. However, the structured countermeasures are the minority, and such countermeasures from different categories are still incomparable due to the semantic differences (e.g., the quantitative changes of road slope angles and shoulder widths are different metrics and therefore not comparable). By contrast, nearly 90 percent of the total countermeasures are textual descriptions without explicit quantitative indications, thus categorized as unstructured countermeasures (illustrated by the unstructured examples in Table 1). The structured countermeasures are only special subsets of the unstructured countermeasures, and all countermeasures can be regarded as unstructured if no further sub-feature extraction is implemented.

Table 1. Statistics of the structured and unstructured countermeasures categorized by the roadway segment type and the intersection type

Facility Type	Countermeasure Type	Countermeasure Example	Count	Percentage
Roadway	Structured	Widen paved shoulder from 4 ft to 6 ft	1045	11.62%
	Unstructured	Reduce lane width from 12 ft to less than 12 ft	4851	53.95%
Intersection	Structured	Change left turn phasing consistency from 61.9% to 31.6%	13	0.14%
	Unstructured	Implement systemic signing and marking improvements at stop-controlled intersections	3084	34.29%

Countermeasure diversity. For the data-driven approach, predictive modeling on homogeneous countermeasures may encounter failure in its generalization to new scenarios. Therefore, the diversity of countermeasures is checked. Figure 1 shows the frequency distribution of CMF values over different countermeasure categories. Among all the categories under the intersection type, intersection geometry and intersection traffic control are the two most studied categories in the literature, characterized by the highest proportions of CMF cases in the Clearinghouse. Similarly, the countermeasures under the roadway and shoulder width categories make up of most of the countermeasure cases for the roadway segment. While the absolute count of countermeasures in each category varies, the CMFs for most countermeasures approach 1.0. Specifically, the countermeasure frequencies of most categories reach the top at CMF values around 1.0 and gradually decrease to the left and right sides. Meanwhile, over 99% of the countermeasures reach CMF values within the range of $[0, 2.0]$. Such diversity in the countermeasure categories potentially provides good support for our knowledge-mining approach. Also, very few countermeasures have CMFs exceeding 2.0. These extreme values are treated as statistical outliers and therefore are not included in this study.



(a) Roadway segment **(b) Intersection**

Figure 1. The CMF frequency distribution over different countermeasure categories for the roadway segment (a) and intersection (b) types in the FHWA CMF Clearinghouse

Input Feature Selection

The data fields defined for countermeasures, site conditions, and crashes available in the CMF Clearinghouse were selected as the input features for the predictive analysis. As summarized in Table 2, the collected data fields are classified into five categories. The first three feature types demonstrate the characteristics of different countermeasures, crashes, and area types, which are generic influential factors for CMFs. The latter two feature categories indicate the sub-attributes of specific facility types (intersection and roadway segment). These two types of facility-related variables will be used separately when developing specific models for intersections and roadway segments since much less relevance exists across facility types.

Most of the data fields in Table 2 are categorical variables characterized by many unique values and high missing rates. As an exception, the countermeasure names are sentences with textual meanings. Moreover, there are semantic relationships between countermeasures. The CMF Clearinghouse holds data with obvious heterogeneity, which is the major barrier to using it for knowledge-gathering that could lead to an in-depth understanding. In this study, the proposed data-driven framework will introduce interdisciplinary techniques to address these problems.

Table 2. Summary of the selected data fields for CMF prediction

Feature Category	Variable	Missing Rate -Intersection	Missing Rate -Roadway	Variable Type	Number of Categories
Countermeasure	Countermeasure name	0.00%	0.00%	Text sentence	-
	Countermeasure category	0.00%	0.00%	Categorica l Variable	19
	Countermeasure subcategory	0.00%	0.00%		40
Crash	Crash type	9.10%	15.40%		100
	Crash time-of-day	17.70%	21.30%		5
	Crash severity	0.40%	16.50%		19
Local area	Area type	14.70%	23.20%		7
Intersection	Intersection type	6.80%	-		8
	Intersection geometry	12.80%	-		9
	Traffic control type	7.60%	-		8
Roadway	Roadway type	-	51.30%		13
	Road division type	-	16.50%		7

Data-Driven Predictive Modeling

System Overview

In this section, the machine-learning based CMF prediction framework is formulated, as shown in Figure 2. This framework consists of three main parts: the feature-embedding part, the regressive learning part, and the output part. First, we selected the variables most relevant to the CMFs from the CMF Clearinghouse (Table 2). In the following section, we explain how to tackle these unstructured non-numerical variables to generate the embedded model inputs and then explain the regressive modeling for CMF prediction.

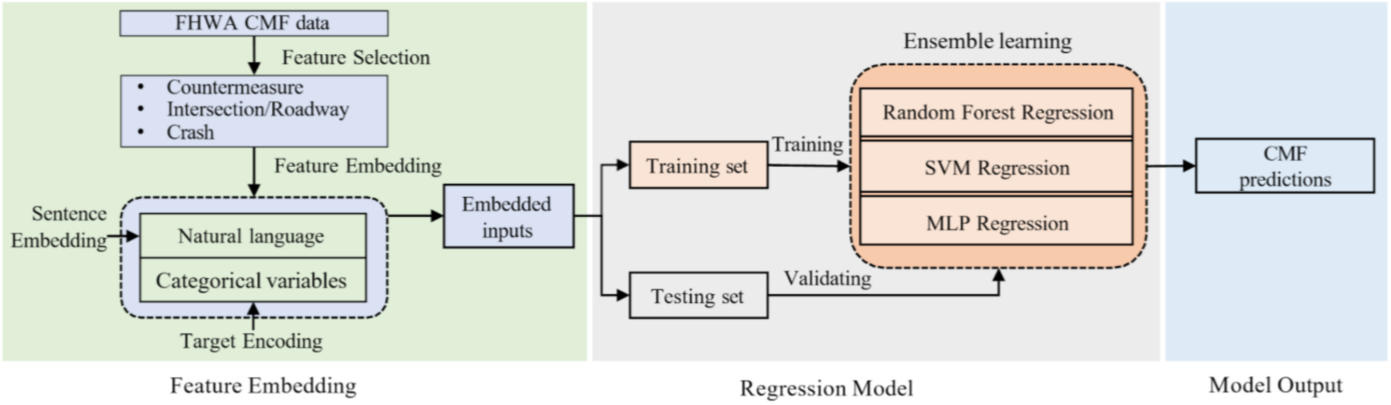


Figure 2. The architecture of the proposed data-driven framework for CMF prediction

Input Feature Encoding Processing Pipeline

Countermeasure embedding. The countermeasure embedding process is to capture the semantic contexts depicted by the countermeasure texts and embed them into the machine-processable features. As mentioned, most of the recorded countermeasures are in an unstructured format. However, the underlying semantic contexts among these textual descriptions are much less intuitive to capture. Drawing on the achievement in the natural language processing area, we introduce the Sentence-BERT model proposed by Reimers and Gurevych in 2019 (27), which is a deep-learning-based sentence-embedding model, to compute the semantically meaningful embeddings of each countermeasure. Furthermore, the pre-training process has been shown effective in improving model performance for diverse natural language processing tasks, including sentence embedding (28). Leveraging the models pre-trained on the large-scale corpora online, we can draw support from the state-of-the-art technology for the embedding computation of our limited countermeasure names. Therefore, rather than working from scratch, we adopted one of the pre-trained Sentence-BERT models, all-mpnet-base-v2, to be the embedding encoder in this study, given its overall best encoding quality.

A fine-tuning on our countermeasure descriptions was further implemented to fit into the specific context of this work, which is illustrated in Figure 3(a). To achieve this, a small sequence of countermeasure pairs with their semantic similarity is defined via expert domain knowledge. These countermeasure pairs are then used as fine-tuning inputs to slightly adjust the model parameters. For fine-tuning, the definition of the loss function is critical. The cosine similarity between different countermeasures is selected as the cost function and defined by Equation 1:

$$cos\theta = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} = \sum_{k=1}^{m} \frac{v_{i,k} v_{j,k}}{\sqrt{\sum_{k=1}^{m} \quad v_{i,k}^2}\sqrt{\sum_{k=1}^{m} \quad v_{j,k}^2}} \#(1)$$

where v_i and v_j are the embeddings yielded by the model from countermeasure pair of c_i and c_j, respectively. The similarity of these embeddings is evaluated using cosine similarity and compared with the gold similarity score. This fine-tuning process can benefit the similarity recognition of different countermeasures.

This process aligns with our aim that countermeasures with similar safety effects should be semantically close, and dissimilar countermeasures should be semantically far away. After fine-tuning, the pre-trained sentence-BERT model is finally capable of mapping individual countermeasures into a vector space such that semantically similar countermeasure descriptions are close. Through countermeasure embedding, each countermeasure can be transformed from a sentence c into a fixed-size dense vector $v \in R^{m\times1}$, which is illustrated in Figure 3(b).

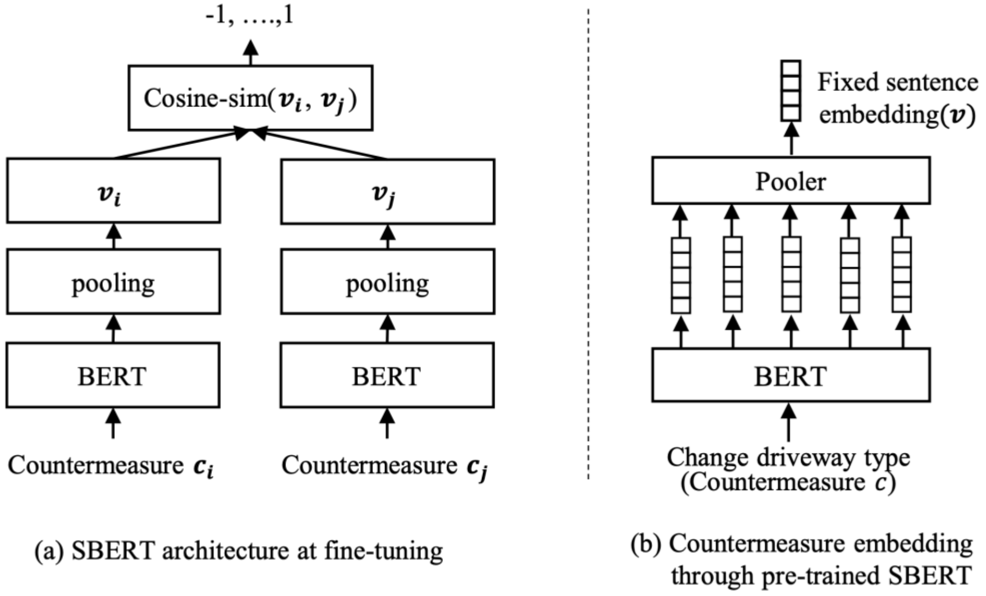

-1, ….,1
Cosine-sim($\boldsymbol{v_i}$, $\boldsymbol{v_j}$)
$\boldsymbol{v_i}$
$\boldsymbol{v_j}$
pooling
pooling
BERT
BERT
Countermeasure $\boldsymbol{c_i}$
Countermeasure $\boldsymbol{c_j}$
(a) SBERT architecture at fine-tuning
Fixed sentence embedding($\boldsymbol{v}$)
Pooler
BERT
Change driveway type (Countermeasure c)
(b) Countermeasure embedding through pre-trained SBERT

Figure 3. Sentence-BERT framework architecture at fine-tuning (a) and for computing countermeasure embedding (b)

Sub-feature extraction. The sub-feature extraction is an auxiliary method of the feature embedding approach for certain structured countermeasures. As shown in Table 3 (a), these structured countermeasures consistently include five types of sub-features about the shoulder-width treatments, despite the slight differences in their textual expressions. Considering this, the transformation and replacement of the countermeasure $ci \rightarrow vi \in R^{m\prime\times1}$ with m' as the number of sub-features can be performed. Then the sub-features, as shown in Table 3 (b), can also be part of the model inputs without using the sentence-embedding approach, either as numerical or categorical variables. The sentence-embedding strategy can be adopted as a general approach that works for both structured and unstructured countermeasures. The sub-feature extraction for structured countermeasures will only be used when customized predictive modeling is recommended, but not included in the main prediction framework of this work. This approach will be later discussed in conjunction with the complementary models for result analysis.

Table 3. Sub-feature extraction from structured countermeasure descriptions (an exemplary illustration)

(a) Structured countermeasures

Countermeasure Examples (Shoulder width)
Widen paved shoulder from 3 ft to 4 ft
Widen shoulder (paved) (from 0 ft to 4 ft)
Widen shoulder (unpaved) (from 0 ft to 4 ft)
Pave deteriorated shoulder (2 ft)
Reduce paved shoulder from 3 ft to 1 ft

(b) Sub-features extracted from structured countermeasures

Action	Prior Shoulder Type	Post Shoulder Type	Prior Shoulder Width (ft)	Post Shoulder Width (ft)
Widen	Paved	Paved	3	4
Widen	No shoulder	Paved	0	4
Widen	No shoulder	Unpaved	0	2
Pave	Deteriorated	Paved	2	2
Reduce	Paved	Paved	3	1

Categorical variable encoding and missing data handling. The high cardinality feature and considerable missing values make it difficult for the categorical data to be efficiently handled by the model. Therefore, the target encoding method is introduced for categorical variable encoding in this study (29). This method targets mapping the high-cardinality categorical attributes to continuous scalar variables. The encoding scheme is to map each high-cardinality categorical item to the probability estimate of the target variable. For the regressive purpose in this study, the numerical encoding corresponds to the expected value of the CMF values (y) pertaining to a specific category cell K_i. It should be noted that only the records in the training set are used in generating the encoder so that data leakage is avoided. That is, for each individual categorical value u_i of a high-cardinality categorical attribute $u \in R^{NTR \times 1}$ with NTR as the size of the training set, the scalar encoding value s_i is calculated as the mixture of two probabilities: the posterior probability of y given $u = u_i$ and the prior probability of y, which is calculated as Equation 2

$$s_i = \lambda(n_i) = \frac{\sum_{k \epsilon k_i} \quad y_k}{n_i} + \left(1 - \lambda(n_i)\right) \frac{\sum_{k=1}^{N_{TR}} \quad y_k}{n_i} \#(2)$$

where y_k is the CMF value at the observation cell k_i for which $u \ = \ u_i$ with size n_i, and $\lambda\ (n_i)\ \in\ [0, 1]$ is the weight factor monotonically increasing with n_i. In this study, $\lambda\ (n_i)$ is defined by a single parameter function as Equation 3 with the empirical control factor $l \ = \ 100$, the details of which can be found in (29).

$$\lambda(n_i) = \frac{n_i}{n_i + l} \#(3)$$

As for the missing values, the target encoding method treats them as any other value like $u \ = \ u_0$ and applies Equation 2 to them. Without additional treatment, this method maintains compatibility and advantages with these missing values. That is, if the occurrence of a missing value u_0 for the categorical attribute u potentially has relevance for the CMF value, then its scalar encoding s_0 will capture that information. If not, then s_0 will correspond to a "neutral" representation of the missing value (no extra benefits).

Input feature integration. Once the countermeasure descriptions and the other categorical variables are embedded or encoded, these input variables will be integrated as the input matrix X. Then the embedded numerical model input matrix $X \ \in \ R^{n \times r}$ with *n* as the sample size and *r* as the number of predictor variables in X, is suited for use in the subsequent regressive model.

Regressive Modeling

The relationships between the CMF values and the explanatory variables are complex and prone to be non-linear in practice. The machine learning algorithms can fit in well with such pattern mining problems. To improve the predicting performance, we tested out several most-used models in machine learning and used the ensemble learning approach to form an integrated model for the CMF prediction problem.

Regression models. For this study, the input feature dimension is relatively high, and the number of data samples is limited. Meanwhile, there is a large proportion of missing values in the data. The machine learning models are selected based on these properties, which include:

- Random forest regression (RF Regression). Random forest (30) is one of the most effective learning algorithms available for classification and regression problems when a considerable proportion of data are missing. The ensemble learning of a multitude of decision trees native to this algorithm enables it not to rely too heavily on any individual feature and therefore prevents over-fitting.
- Support vector machine regression (SVM Regression). SVM is a popular machine learning method proposed by Vladimir Vapnik and his colleagues (31, 32), which is especially effective for regression and classification with data in high dimensional spaces. The implementation in this study is based on LIBSVM (33), with the default kernel as linear and regularization parameter as 1.0.
- Multi-layer perception (MLP). MLP is one of the most used artificial neural networks (ANNs) with ability to solve problems stochastically (34). The MLP method suits well for our complex non-linear

regression problems, while not as prone to overfitting as other neural networks with deeper layers, like the deep learning models.

Ensemble learning. An ensemble technique is to combine multiple machine learning algorithms to obtain better predictive performance than any individual constituent model (35, 36). In this study, multiple machine-learning models are tested out for CMF predictions. However, the performance of these models fluctuated slightly, which introduces difficulties in selecting the best model. To simplify the model selection and stabilize the model performance, we introduce the simple bagging-based ensemble technique that uses the average CMF predictions of the three models as the final model prediction of CMF values which is

$$\widehat{y_\imath} = \frac{1}{R}\sum_{r=1}^{R} \quad \widehat{y_{\imath,r}}\#(4)\ ,$$

where $\widehat{y_{\imath,r}}$ is the prediction given by each base regression learner and R is the total number of base learners.

Experimental Setting and Model Evaluation

Train-Test Split

We use the train-test split to evaluate the proposed machine-learning algorithm for CMF predictions on the FHWA CMF Clearinghouse dataset. With this open-source dataset, we randomly selected 80 percent of the data as the training set and assigned the remaining 20 percent to the test set.

Evaluation Metrics

Once the model has been trained to learn the mapping pattern from the inputs to CMFs, the model performance will be evaluated on the test set and reported as an error in the predictions. Two commonly used metrics—the mean absolute error (MAE) and the root mean square error (RMSE)—are used to evaluate and report the overall performance of the proposed model. In practice, a CMF value smaller than 1.0 indicates an expected reduction in crashes, while a CMF larger than 1.0 represents an expected increase in crashes after the implementation of the target countermeasure. Therefore, setting 1.0 as the benchmark, we consider the consistency rate (CR) that calculates how many CMF predictions fall on the same side with the investigated CMFs as the third evaluation metric. The definitions of them are given by Equation 5–7:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \quad |y_i - \widehat{y_\imath}|\#(5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \widehat{y_\imath})^2\#(6)}$$

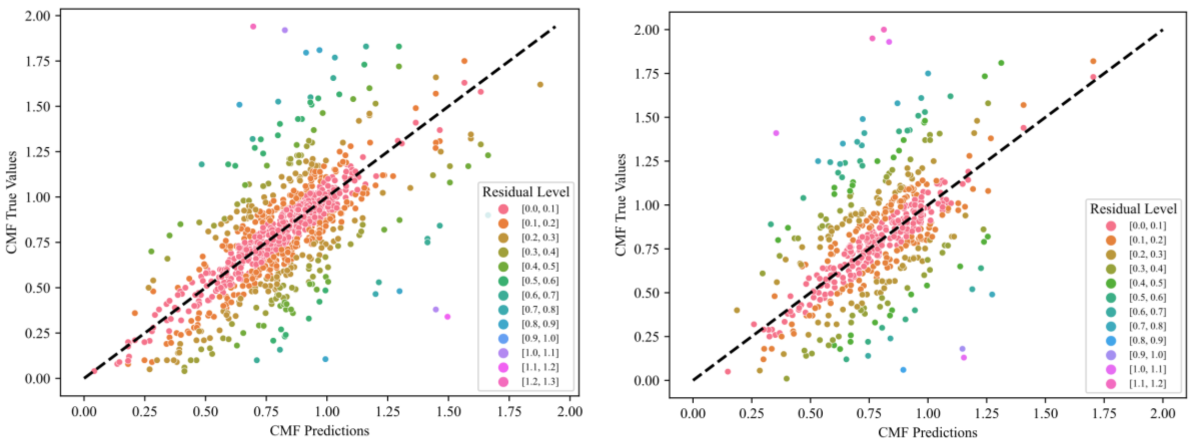
$$CR = \frac{\sum_{i=1}^{n} \quad c_i}{n} \#(7) \text{ ,}$$

where y_i and $\widehat{y_i}$ are the true CMF values listed in the CMF Clearinghouse and the predicted values respectively, n is the sample size of test data, and $c_i = \ 1 \ if \ (y_i - \ 1)(\widehat{y_i} - 1) \geq \ 0$, otherwise 0 is the consistency indicator with the true CMF value at countermeasure scenario i.

Result Analysis

Unstructured Countermeasure Scenarios

The relation of the CMF predictions and the investigated CMFs is shown in Figure 4. To demonstrate the model performance at different levels, the testing samples with different absolute residual levels are indicated by different colors. Whereas some outliers exist, the scatter plots show that the data points in general are symmetrically distributed and concentrated around the upward diagonal line. As summarized in Table 4, around 45% to 50% of the tested countermeasures have their predictions deviating less than 0.1 from the investigated CMFs. These results demonstrate the overall good consistency between the CMF predictions and the investigated CMFs in the test set. The RMSE, MAE errors and CR values under evaluation are further summarized in Table 4. As indicated, the MAEs (value range 0.1–0.2) and RMSEs (value range 0.*2*–*0.*3) for both types of countermeasures tend to be within a reasonable range, since our main objective is to provide a first guess of CMF values for additional countermeasures. The statistics in Table 4 also show that around 80% of the CMF predictions correctly reflect the positive or negative safety effects of the tested countermeasures. In addition, the model for the roadway segment subset outperformed the model for the intersection subset. This may be due to two reasons: first, the number of roadway-related countermeasures is nearly twice that of the intersection type, which provides a better knowledge source for data mining; second, the countermeasure descriptions in the roadway segment type are on average longer and therefore potentially more informative for the capture of semantic contexts.



(a) Roadway segment

(b) Intersection

Figure 4. The scatter plots of the CMF predictions and the CMF true values on the test set for countermeasures under the roadway segment type (a) and the intersection type (b)

Table 4. Statistics of the model performance evaluation on the test dataset

Facility Type	Evaluation Metric			
	MAE	RMSE	CR	Absolute Residual ≤ 0.1
Roadway	0.15	0.22	83.3%	49.3%
Intersection	0.18	0.26	79.3%	44.4%

We further grouped the test samples into subsets based on different site-condition types and countermeasure categories. This helps to evaluate the model performance on a more detailed scale. The MAE value of each subgroup is calculated and shown in Figure 5. The distribution of the subgroup MAEs shows that the MAEs for most subgroups are close to the overall test MAE, for both intersection and roadway segment types. The subgroup MAE values also shows that the developed model performs stably in different scenarios. Admittedly, there also exist some abnormal MAE values in these subgroups, which indicates significant deviations from the previously investigated CMF values. These abnormal MAEs have also elevated the overall MAE and RMSE values.

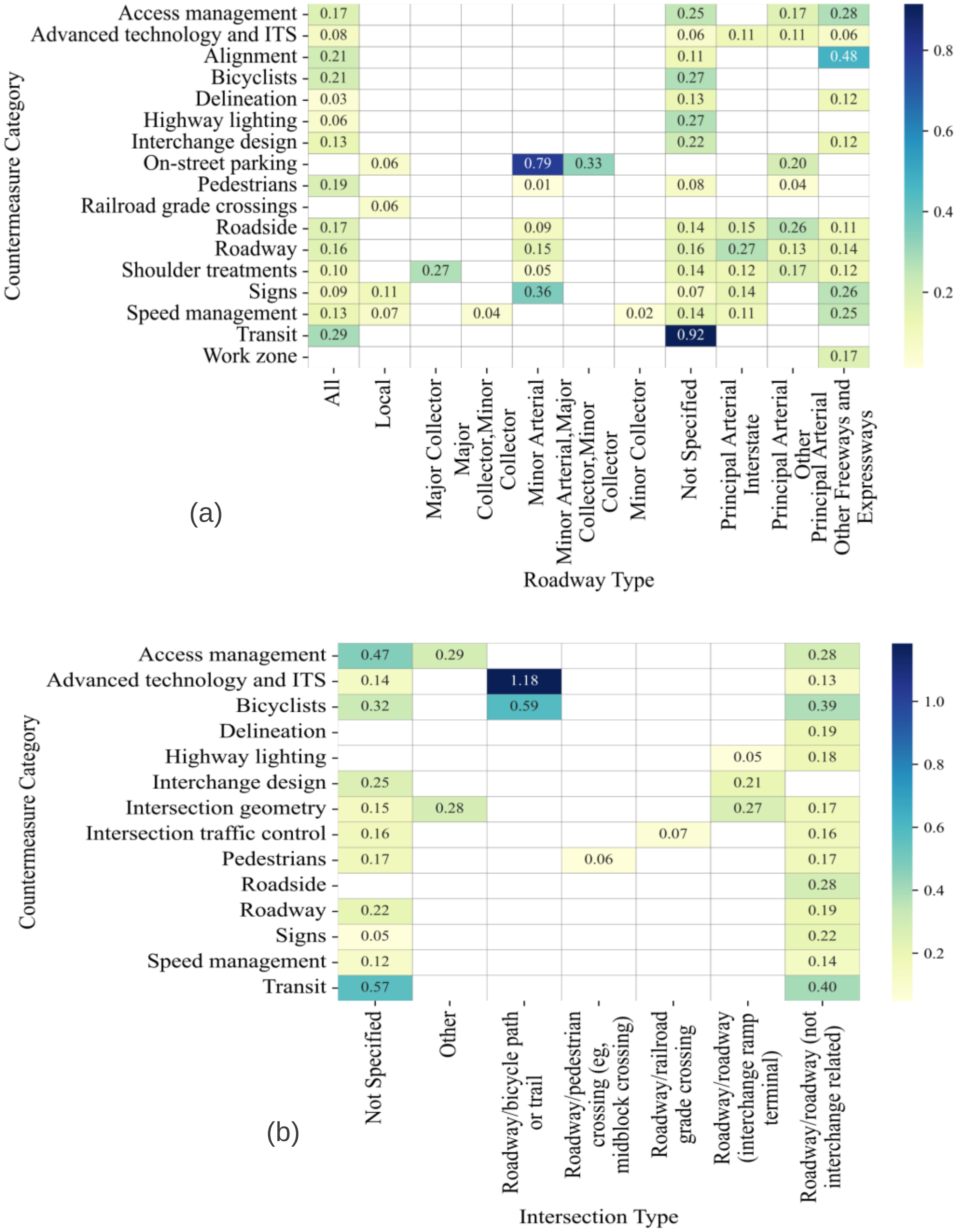


Figure 5. MAE distributions for subgroups based on different site-condition types and countermeasure categories for the roadway segment type (a) and the intersection type (b), respectively, with lighter colors indicating smaller MAEs and darker colors, the opposite.

To explore the inherent reasons for these abnormal CMF predictions, we count the training and testing sample sizes of each subgroup and analyze their relationships with subgroup test MAE values. The pairwise relationships among the test MAE value, training sample size, and the testing sample size for each subgroup are shown in Figure 6. As demonstrated, subgroups with larger training sample sizes generally have smaller MAEs and therefore more accuracy in CMF predictions. Subgroups with larger MAEs, which correspond to the cells with darker colors in Figure 5, are some of the subgroups with few or even no countermeasures on the training dataset. Without enough countermeasures for data mining, such subgroups have a greater chance to have poor MAEs and therefore less accurate CMF predictions. This is reasonable considering the basic data-mining rule: although our model is designed to provide reasonably quick predictions for new scenarios, there should be enough other similar countermeasures (like belonging to the same general category) for the model to learn from. This data gap will be reduced as CMF values for additional countermeasures that belong to such subgroups become available.

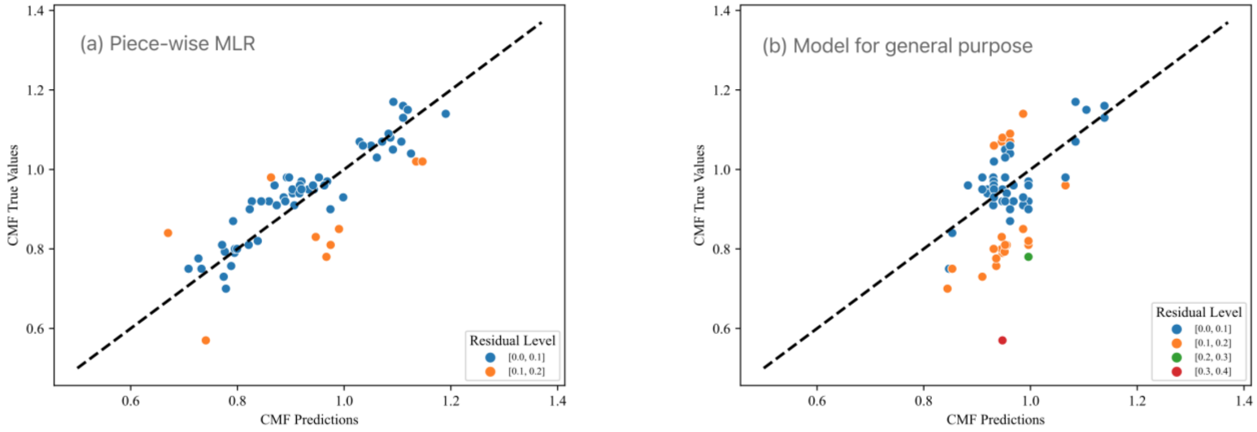


Figure 6. The corresponding relationships of the test MAE values, the training sample sizes (blue points), and testing sample sizes (orange points)

Table 5. Statistics of model performance evaluation on the test dataset

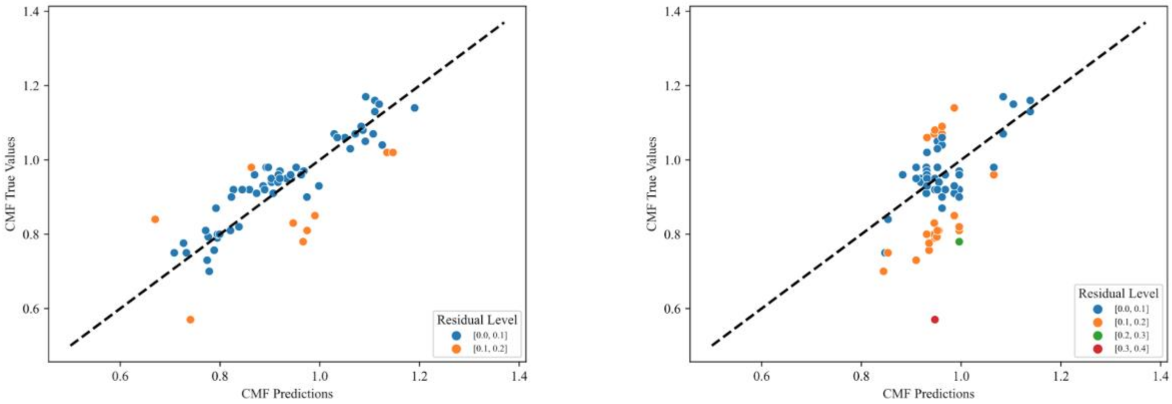
Model	Evaluation Metric			
	MAE	RMSE	CR	Absolute Residual ≤ 0.1
Piecewise MLR for shoulder width	0.05	0.07	100.0%	86.3%
The generic CMF prediction model	0.09	0.11	78.7%	63.6%

Structured Countermeasure Scenarios

While we propose the above main framework to provide CMF predictions for more generic scenarios, we also seek to boost the framework with a complementary approach. The complementary model is customized for the structured countermeasures, based on leveraging the sub-feature extraction process. As illustrated in Table 3 countermeasures that resemble the demonstrated structures can be transformed into several subfeatures. These subfeatures are more explicit and informative; therefore, a customized prediction model can be developed accordingly. Tailored to countermeasures under a specific structure, such a complementary model can be applied only to each individual type of structure independently. Furthermore, there should be enough countermeasure cases that share the same structure to support the customized regressive modeling.

To illustrate how this complementary scheme works, we use the structured countermeasures under the shoulder width subcategory as a case study. To develop the complementary model, 330 structured countermeasures under the shoulder width were collected from the CMF Clearinghouse repository. Given the reduced complexity in data and context, subfeatures extracted from these countermeasures are directly integrated with other categorical variables listed in Table 2. No further processing is involved. Given the limited sample data, a simple piece-wise multiple linear regression model (MLR) with a breakpoint at 1.0 is adopted. To test the model performance, the train-split rule is followed in the same way. Finally, there are 264 CMF cases on the training set with the rest 66 cases as the test set.

To evaluate the model performance, the distribution of data points characterized by CMFs for testing and corresponding CMF predictions by the customized model are shown in Figure 7 (a). For comparison, the relation of CMFs and corresponding predictions by the proposed generic model on the same test are shown in Figure 7 (b). In Figure 7 (a), the CMF predictions and investigated CMF values are closely distributed along the upward diagonal line, corresponding to the MAE and RMSE values (both less than 0.1) for piece-wise MLR in Table 5. Compared with the performance of the generic CMF prediction model, the customized model for the shoulder width type obtained better performance from all aspects. This is expected since the model for a general prediction purpose needs to adapt to various circumstances and provides overall reasonable predictions for multiple categories. The better performance and simpler structure of the customized model also enlighten us that a more accurate representation of countermeasure details, e.g., through subfeature extraction, would help to improve the prediction accuracy on CMFs under our knowledge-mining framework.



(a) CMF predictions v.s. CMF true values by the customized piece-wise MLR

(b) CMF predictions v.s. CMF true values by the model for a general prediction purpose

Figure 7. The true CMFs vs the predictions from the piecewise MLR model (a) and the general prediction model (b) on the test set for the shoulder width category

Discussion and Conclusions

CMF is an important measure of the effectiveness of safety countermeasures. However, developing CMFs through traditional approaches is usually costly and time consuming, but there is often a need to obtain a quick CMF estimate for new safety improvement scenarios. The proposed framework is intended to fill this gap.

This study developed a data-driven framework for CMF prediction by mining the FHWA's CMF Clearinghouse data. To our knowledge, this is the first work that enables predicting CMFs in a cost-effective, time efficient, and reproducible way. Aimed at making full use of the knowledge on CMFs from previous studies, our knowledge-mining scheme is less dependent on detailed crash inventory data and more extendable for a general prediction purpose. The technical novelties of this approach include the following. First, the proposed framework handled the heterogeneity of CMF Clearinghouse data in a flexible way, including the unstructured descriptions of countermeasures, high-cardinality categories, missing rates, and noise. This is achieved through introducing the cutting-edge natural language processing techniques and the target encoding method. Second, the proposed framework combined multiple machine learning methods to make predictive modeling of CMFs on different countermeasure scenarios. We trained and evaluated the model against FHWA Clearinghouse data, and the results show that there is a good consistency of tested model predictions and the publicly known CMFs with reasonable overall accuracy reported.

Still, several aspects of the current framework can be strengthened and a word of caution on its applicability should be noted. First, our approach is meant to provide a preliminary estimate of CMFs and is not to replace the traditional methods for developing CMF for specific scenarios. Second, the accuracy can be further improved if structured or domain knowledge is leveraged. In this research, we developed another model handling structured countermeasure descriptions and this model obtained a better accuracy. Third, in the current framework, the values of CMFs are predicted, but the confidence levels of these predictions are not provided. Such confidence levels would be helpful in determining the fidelity of CMF predictions.

We envision several possible extensions of the current work. First, combining the proposed approach with traditional approaches may help to further improve CMF estimation accuracy and reduce the cost. Second, domain knowledge may enhance the model's prediction capacity. One example is that different countermeasures may have similar effects but fall into different categories. For instance, 'highway lighting' and 'signs' are different categories in the Clearinghouse, but they both increase roadway visibility, and this information may be more explicitly incorporated in the model. Third, based on the pattern of existing CMFs (such as the clusters of countermeasures in high-dimensional vector space), we think it is possible to include the confidence level of predictions. Lastly, when CMF Clearinghouse includes more case-specific CMF records, our model can take advantage of the additional information and potentially achieve better performance.

References

AASHTO, *Highway safety manual*, Vol. 1. American Association of State Highway and Transportation Officials, 2010.

Jones, K., K. Yunk, and D. Carter, The CMF Clearinghouse: A Handy Safety Tool. *Accident Reconstruction Journal*, Vol. 20, No. 6, 2010.

McDaniel-Wilson, C., ODOT's HSIP Countermeasures and Crash Reduction Factors (CRF Appendix). *Oregon Department of Transportation*, 2018.

Smith, S. and R. A. Scopatz, *Roadway Safety Data and Analysis Case Study: North Carolina's State-Specific CMFs*, 2016.

Gan, A., J. Shen, and A. Rodriguez, Update of Florida crash reduction factors and countermeasures to improve the development of district safety improvement projects, 2005.

Gross, F., B. N. Persaud, and C. Lyon, *A guide to developing quality crash modification factors*. United States. Federal Highway Administration. Office of Safety, 2010.

Gross, F., K. Eccles, and D. Carter, *Crash Modification Factors Needs Assessment Workshop*, 2015.

Tribbett, L., P. McGowen, and J. Mounce, *An evaluation of dynamic curve warning systems in the Sacramento River Canyon*. Western Transportation Institute, Montana State University Bozeman, 2000.

Mamlouk, M. and B. Souliman, Effect of traffic roundabouts on accident rate and severity in Arizona. *Journal of Transportation Safety & Security*, Vol. 11, No. 4, 2019, pp. 430– 442.

Srinivasan, R., D. Carter, C. Lyon, and M. Albee, Before–After Evaluation of the Realignment of Horizontal Curves on Rural Two-Lane Roads. *Transportation research record*, Vol. 2672, No. 30, 2018, pp. 43–52.

Tang, H., V. V. Gayah, and E. T. Donnell, Crash modification factors for adaptive traffic signal control: An Empirical Bayes before-after study. *Accident Analysis & Prevention*, Vol. 144, 2020, p. 105672.

Hussein, N. A. and R. A. Hassan, Evaluating safety effectiveness of surface treatment at signalised intersections: a before and after study. *International journal of pavement engineering*, Vol. 19, No. 11, 2018, pp. 1034–1041.

Jang, K., K. Chung, D. R. Ragland, and C.-Y. Chan, Safety performance of highoccupancy-vehicle facilities: Evaluation of HOV lane configurations in California. *Transportation research record*, Vol. 2099, No. 1, 2009, pp. 132–140.

Avelar, R., K. Dixon, S. Ashraf, A. Jhamb, B. Dadashova, et al., *Developing Crash Modification Factors for Bicycle-Lane Additions While Reducing Lane and Shoulder Widths*. United States. Federal Highway Administration. Office of Safety Research and Development , 2021.

Al-Marafi, M. N., K. Somasundaraswaran, and R. Ayers, Developing crash modification factors for roundabouts using a cross-sectional method. *Journal of traffic and transportation engineering (English edition)*, Vol. 7, No. 3, 2020, pp. 362–374.

Elvik, R., Introductory guide to systematic reviews and meta-analysis. *Transportation research record*, Vol. 1908, No. 1, 2005, pp. 230–235.

Williamson, M. R., R. N. Fries, Y. Qi, and P. Mandava, Identifying the safety impact of signal coordination projects along urban arterials using a meta-analysis method. *Journal of Traffic and Transportation Engineering*, 2018.

Washington, S. P., D. Lord, and B. N. Persaud, Use of expert panels in highway safety: A critique. *Transportation research record*, Vol. 2102, No. 1, 2009, pp. 101–107.

Chen, Y., B. Persaud, and C. Lyon, *Effect of speed on roundabout safety performance: implications for use of speed as surrogate measure*, 2011.

Davis, G. A., Mechanisms, mediators, and surrogate estimation of crash modification factors. *Accident Analysis & Prevention*, Vol. 151, 2021, p. 105978.

Li, X., D. Lord, Y. Zhang, and Y. Xie, Predicting motor vehicle crashes using support vector machine models. *Accident Analysis & Prevention*, Vol. 40, No. 4, 2008, pp. 1611– 1618.

Park, J. and M. Abdel-Aty, Assessing the safety effects of multiple roadside treatments using parametric and nonparametric approaches. *Accident Analysis & Prevention*, Vol. 83, 2015, pp. 203–213.

Pu, Z., Z. Li, R. Ke, X. Hua, and Y. Wang, Evaluating the nonlinear correlation between vertical curve features and crash frequency on highways using random forests. *Journal of transportation engineering, Part A: Systems*, Vol. 146, No. 10, 2020, p. 04020115.

Zhang, X., S. T. Waller, and P. Jiang, An ensemble machine learning-based modeling framework for analysis of traffic crash frequency. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 35, No. 3, 2020, pp. 258–276.

Zeng, Q., H. Huang, X. Pei, and S. Wong, Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. *Analytic methods in accident research*, Vol. 10, 2016, pp. 12–25.

Wen, X., Y. Xie, L. Jiang, Y. Li, and T. Ge, On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development. *Accident Analysis & Prevention*, Vol. 168, 2022, p. 106617.

Reimers, N. and I. Gurevych, Sentence-bert: Sentence embeddings using siamese bertnetworks. *arXiv preprint arXiv:1908.10084*, 2019.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Micci-Barreca, D., A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, Vol. 3, No. 1, 2001, pp. 27–32.

Breiman, L., Random forests. *Machine learning*, Vol. 45, No. 1, 2001, pp. 5–32.

Boser, B. E., I. M. Guyon, and V. N. Vapnik, A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

Drucker, H., C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, Support vector regression machines. *Advances in neural information processing systems*, Vol. 9, 1996.

Chang, C.-C. and C.-J. Lin, LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, Vol. 2, No. 3, 2011, pp. 1–27.

Sarle, W. S., Neural networks and statistical models. In *Proceedings of the Nineteenth Annual SAS Users Group International Conference*,, Citeseer, 1994.

Opitz, D. and R. Maclin, Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, Vol. 11, 1999, pp. 169–198.

Polikar, R., Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, Vol. 6, No. 3, 2006, pp. 21–45.