**Proceedings of the 2024 ASME Joint Rail Conference**
**JRC2024**
**May 13–15, 2024, Columbia, SC, USA**

**JRC2024-130036**

# KERNEL RIDGE REGRESSION IN PREDICTING RAILWAY CROSSING ACCIDENTS

**Ethan Villalobos[1], Constantine Tarawneh[1], Jia Chen[2], Evangelos E. Papalexakis[2], Ping Xu[1,*]**

[1]University of Texas Rio Grande Valley, Edinburg, TX
[2]University of California Riverside, Riverside, CA

## ABSTRACT

*Expanding on the insights from our initial investigation into railway accident patterns, this paper delves deeper into the predictive capabilities of machine learning to forecast potential accident trends in railway crossings. Focusing on critical factors such as "Highway User Position" and "Equipment Involved," we integrate Kernel Ridge Regression (KRR) models tailored to distinct clusters, as well as a global model for the entire dataset. These models, trained on historical data, discern patterns and correlations that might elude traditional statistical methods. Our findings are compelling: certain clusters, despite limited data points, showcase remarkably Root Mean Squared Error (RMSE) values between predictions and real data, indicating superior model performance. However, certain clusters hint at potential overfitting, given the disparities between model predictions and actual data. Conversely, clusters with vast datasets underperform compared to the global model, suggesting intricate interactions within the data that might challenge the model's capabilities. The performance nuances across clusters emphasize the value of specialized, cluster-specific models in capturing the intricacies of each dataset segment. This study underscores the efficacy of KRR in predicting future railway crossing incidents, fostering the implementation of data-driven strategies in public safety.*

## 1. INTRODUCTION

The rapid evolution of railway systems worldwide necessitates an equally dynamic approach to ensuring their safety. This evolution manifests in various forms, such as the adoption of advanced signaling technologies, the integration of real-time monitoring systems, and the implementation of automated and intelligent control mechanisms. In our preceding research [1], we embarked on an exploratory journey into the realm of railway accidents. Utilizing graph mining techniques, we unearthed intricate patterns and relationships between railway companies, centered around shared accident localities. This initial foray into the data laid a robust foundation, highlighting potential patterns ripe for predictive analysis. However, recognizing patterns is merely the first step; the pivotal challenge lies in forecasting future trends and incidents, a task that demands sophisticated predictive tools and methodologies. It is this challenge that our current paper aims to address.

In transitioning from pattern recognition to predictive analysis, our research shifts gears towards leveraging the power of machine learning. Machine learning, with its remarkable ability to decipher and anticipate complex patterns from voluminous datasets, presents itself as the ideal choice to accomplish this task. Our current study is driven by the critical factors identified in [1], notably "Highway User Position" and "Equipment Involved." These factors emerged as significant influencers in railway crossing accidents, shaping our approach to model development.

Our methodology involves constructing models specifically tailored to distinct data clusters identified in our initial research, as well as developing a comprehensive global model that encompasses the entire dataset. This dual approach is designed to harness insights from a global implementation perspective and a cluster-specific conceptual connection. By doing so, we aim to capture both the overarching trends and the nuanced variations within each cluster. Central to our predictive models is the application of Kernel Ridge Regression (KRR) [2], which is adept at handling the multifaceted nature of the factors influencing railway crossing accidents by utilizing the universality of nonlinear kernels.

The models are meticulously trained on historical data from a broad spectrum of railway operators in the United States, including major freight companies, national passenger rail services, and regional transit authorities, with a focus on uncovering patterns and correlations that might elude traditional statistical methods. This diversity reflects the varied conditions under which railway accidents occur, from urban commuter lines to long-distance freight routes. Our findings present a diverse spectrum of results: some clusters, despite limited data points, showcased remarkably low Root Mean Squared Error (RMSE) values, indicating

*Corresponding author
Documentation for `asmeconf.cls`: Version 1.37, February 13, 2025.

high model accuracy. Conversely, certain clusters hinted at overfitting, as evidenced by discrepancies between model predictions and actual occurrences. Intriguingly, larger clusters, despite their wealth of data, did not perform as well as anticipated, especially in comparison to the global model. This phenomenon underscores the complexity within the data-set and the potential limitations of the models in capturing all underlying dynamics. The disparate outcomes across various clusters underscore the necessity of developing specialized models for each cluster. These models, fine-tuned to the unique characteristics of their respective datasets, are more adept at capturing the intricacies that a universal model might overlook.

Our exploration with KRR marks a significant stride in the predictive analysis of railway crossing incidents. It not only enhances our understanding of the factors contributing to these accidents but also sets the stage for data-driven strategies to improve public safety. As railways continue to advance, ensuring their safety remains a paramount concern. While this paper lays the groundwork for predictive analysis in railway systems, our journey in this field is far from over. In our subsequent research, we plan to delve into more efficient machine learning methods, aiming to leverage the full extent of the dataset for a more comprehensive analysis. Our series of papers not only document our progress but also reflect our commitment to employing cutting-edge technology and methodologies in fostering a safer and more efficient future for railway transportation.

## 2. PRELIMINARIES

### 2.1 Spectral Clustering for Data Segmentation

Prior to regression analysis, we utilized spectral clustering to detect inherent groupings within the dataset, based on connectivity principles and distance measures between data points. Spectral clustering excels at identifying clusters with non-linear boundaries, outperforming traditional clustering methods by leveraging the eigenvalues of the similarity matrix.

This process began with the creation of a similarity matrix using the Gaussian (RBF) kernel to compute the similarity between data points. The RBF kernel was chosen for its compatibility with the KRR technique used later in our analysis. We converted the similarity matrix into a graph representation, where nodes represent individual data entries and edges indicate the level of similarity. The normalized Laplacian of this graph is then computed, from which eigenvalues and eigenvectors are extracted to facilitate data projection into a dimensionally reduced space that is conducive to clustering. To determine the optimal number of clusters, we perform grid search and evaluate the clustering quality using the silhouette score.

### 2.2 Kernel Ridge Regression

Consider a dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^m$ denotes the features of $i$-th data sample and $y_i$ denotes the output values. Denote $\boldsymbol{y} = [y_1, \ldots, y_n] \in \mathbb{R}^n$ and $\mathbf{X} = [\boldsymbol{x}_1; \ldots; \boldsymbol{x}_n] \in \mathbb{R}^{n\times m}$. Assume there exists a linear relationship between $\boldsymbol{y}$ and $\mathbf{X}$ such that

$$\hat{\boldsymbol{y}} = \mathbf{X}\boldsymbol{\beta}, \tag{1}$$

where $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_n] \in \mathbb{R}^n$ denotes the predicted output and $\boldsymbol{\beta} \in \mathbb{R}^m$ is the vector of coefficients. Then, traditional ridge regression seeks to solve the following optimization problem:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{2}$$

where $\lambda > 0$ signifies the regularization parameter. The L2-norm $\|\cdot\|_2^2$ penalizes the magnitude of the coefficients, thereby controlling model complexity and preventing overfitting [3].

However, it is usually impractical to have linear assumptions in reality. Instead, we seek to find a nonlinear function $f$ that best describes the relationship between $y_i$ and $\boldsymbol{x}_i$ such that $y_i = f(\boldsymbol{x}_i) + e_i, \forall i$, where $e_{i,t}$ is minimized accordingly to certain optimality metric. Directly optimizing $f$ in the functional space is infeasible since there are infinitely many possible solutions. To address this challenge, kernel based methods are usually considered [4]. Suppose that the nonlinear function $f$ belongs to the reproducing kernel Hilbert space (RKHS) $\mathcal{H} := \{f | f(\boldsymbol{x}) = \sum_{i=1}^{\infty} \alpha_i \kappa(\boldsymbol{x}, \boldsymbol{x}_i)\}$ induced by a positive semidefinite kernel $\kappa(\boldsymbol{x}, \boldsymbol{x}_i) : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ that measures the similarity between $\boldsymbol{x}$ and $\boldsymbol{x}_i$. By the Representer Theorem [5], the nonlinear function $f$ can be expressed by a weighted kernel expansion over the data samples as

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\alpha}^\top \boldsymbol{\kappa}_{\mathbf{X}}(\boldsymbol{x}), \tag{3}$$

where $\boldsymbol{\kappa}_{\mathbf{X}}(\boldsymbol{x}) \in \mathbb{R}^n$ collects all $\kappa(\boldsymbol{x}_i, \boldsymbol{x})$, and $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the coefficient vector to be learned. Among various kernel functions, the RBF kernel is frequently used, defined as:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}\|^2), \tag{4}$$

where $\gamma$ is a scale parameter that adjusts the kernel's sensitivity.

KRR that uniquely blends the regularization principles of ridge regression with the functional mapping capabilities of kernel methods is then adopted [2]. The objective function thus becomes

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \mathbf{K}\boldsymbol{\alpha}\|_2^2 + \lambda\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha}. \tag{5}$$

Here, $\mathbf{K}$ is the kernel matrix and $\boldsymbol{\alpha}$ is the coefficient vector in the kernel-transformed space.

For the specific case of predicting railway crossing accidents, KRR allows for a nuanced modeling of the intricate and non-linear relationships between various factors and accident occurrences. This aligns with the patterns identified in our previous research. The judicious selection of a kernel and a suitable regularization parameter enables the development of a model that is both predictive and generalizable, offering reliable forecasts for future accident trends.

## 3. MATERIALS AND METHODS

### 3.1 Data Acquisition and Preprocessing

The foundational data for this study is sourced from the "US Highway Rail Road Crossing Accident" dataset, available on Kaggle and compiled by the US Department of Transportation. This dataset spans from January 1, 1975, to February 28, 2021, and includes a multitude of variables reflecting the complex nature of railway crossing incidents in the United States. To ensure computational efficiency while preserving data integrity, we randomly

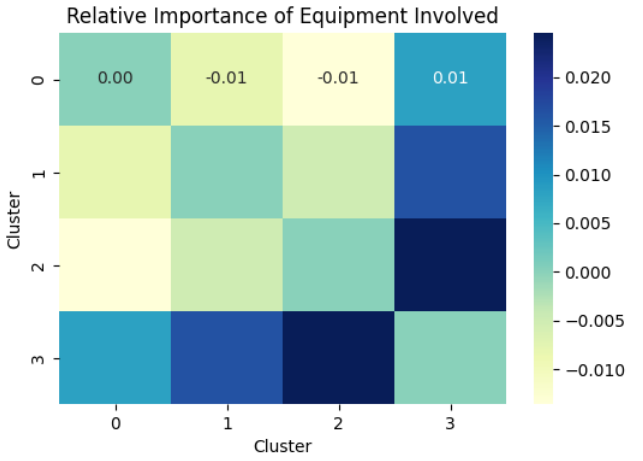


**FIGURE 1:** This heatmap illustrates the varying impact of Equipment Involved on accident prediction across clusters, reinforcing the hypothesis that the equipment involved significantly influences the frequency and severity of accidents.

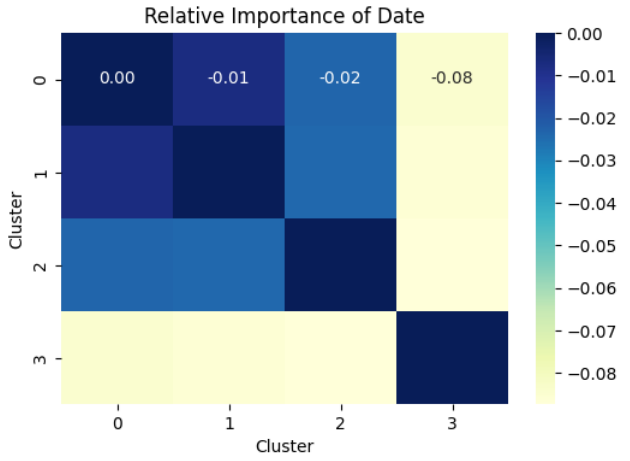


**FIGURE 2:** This heatmap captures the relative importance of the Date factor, reinforcing the hypothesis that while Date has an impact, it's less definitive than physical and environmental conditions.

sample 10% of the original dataset. The chosen subset underwent meticulous preprocessing, with a focus on pivotal features such as "Highway User Position", "Equipment Involved", and "Incident Year", alongside the critical target variable, "Accident Count". These features were prioritized based on their known significance in accident causation and prediction. For instance, "Highway User Position" provides crucial information about the relative positions of vehicles and other users in the vicinity of the railway crossing, which is essential for understanding collision dynamics. Similarly, "Equipment Involved" offers insights into the types of equipment present during accidents, shedding light on potential hazards and safety vulnerabilities. "Incident Year" serves as a temporal factor, capturing trends and variations in accident rates over time. This stage of processing included thorough cleaning to prepare the data for the advanced analytical techniques that followed. Furthermore, Figure 1, depicting a heatmap of the relative importance of "Equipment Involved," visually demonstrates its higher significance compared to other factors such as "Date" in Figure 2. These heatmaps are provided here for illustration, offering visual confirmation of the prioritization of certain features over others in accident prediction.

### 3.2 Feature Preparation and Model Training

We partition the dataset into training and testing sets with an 80-20 split to ensure robust training and reliable testing. The categorical features are one-hot encoded to be compatible with machine learning algorithms, while all features are standardized to a common scale to eliminate bias from data scale differences. The processed dataset is then used for KRR with an RBF kernel, which is known for its effectiveness in capturing non-linear relationships within data. We apply KRR to develop both a global model for the entire dataset and cluster-specific models for the data clusters identified through spectral clustering.

### 3.3 Model Evaluation and Comparative Analysis

Model performance is evaluated using the RMSE, which measures the average magnitude of prediction errors, providing a clear indicator of model accuracy. We conduct a comparative analysis between the global model and the cluster-specific models to underscore the unique predictive capacities of each segment within the dataset. This analysis is instrumental in demonstrating the efficacy of customized models tailored to the distinct characteristics of each cluster.

## 4. RESULTS

### 4.1 Global Model Performance

The global KRR model, encompassing the entirety of the dataset, record a scaled RMSE of 0.17. This metric serves as a baseline for subsequent model comparisons. The model's performance is visually represented through a line plot contrasting the actual versus predicted yearly accident counts, as shown in Figure 3. This visualization provides a macroscopic view of the model's ability to capture the overarching trends and fluctuations in railway accidents over an extended period.

### 4.2 Cluster-specific Model Performance

In assessing the performance of cluster-specific models, the following results were observed:

- **Cluster 0:** Reported a scaled RMSE of 0.22, which is indicative of a solid model performance overall. This value, however, points to certain instances where the model faced predictive challenges, providing clear direction for targeted enhancements. See Figure 4.

- **Cluster 1:** Showed a scaled RMSE of 0.16, reflecting a highly accurate model performance that effectively captures this cluster's internal variability, marking it as a robust predictive tool for this specific data subset. See Figure 5.

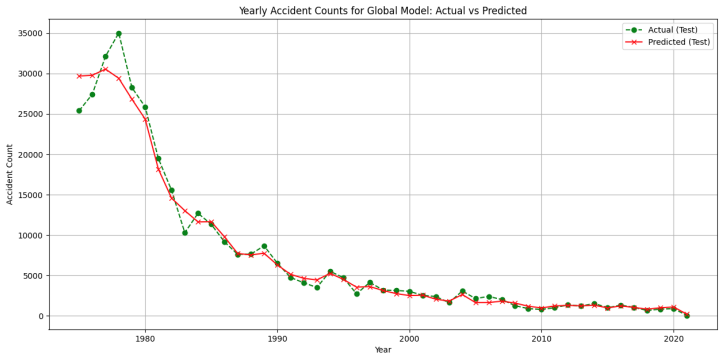


**FIGURE 3: Yearly Accident Counts for Global Model: Actual vs. Predicted.** The graph presents the global model's overarching performance, capturing a long-term declining trend in accidents which corresponds with the actual data. Notable is the model's adeptness at reflecting major trends across the timeline while also revealing areas of overestimation and underestimation, suggesting opportunities for further model refinement.

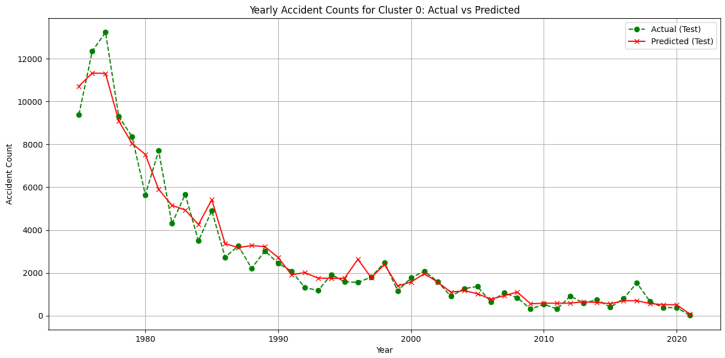


**FIGURE 4: Yearly Accident Counts for Cluster 0: Actual vs. Predicted.** The graph demonstrates the model's difficulty in accurately predicting the highly variable accident data of this cluster, with apparent over-estimations during certain periods.

- **Cluster 2:** Demonstrated a scaled RMSE of 0.19, denoting a competent level of accuracy. While the predictions are generally reliable, there is potential for incremental improvements to achieve even closer alignment with observed outcomes. See Figure 6.

- **Cluster 3:** Reported the highest scaled RMSE at 0.23. Despite this, the model's performance is considered reasonable, as it navigates the multifaceted and complex interactions inherent to this cluster's accident data, maintaining a commendable degree of predictive reliability. See Figure 7.

These results highlight the distinctive characteristics of each cluster, underlining the value of customized models for accurate predictions in railway safety analysis.

### 4.3 Interpretation and Insights

The variation in RMSE values across the global and cluster-specific models highlights the intricate nature of railway accident data. The global model, with a scaled RMSE of 0.17, reflects

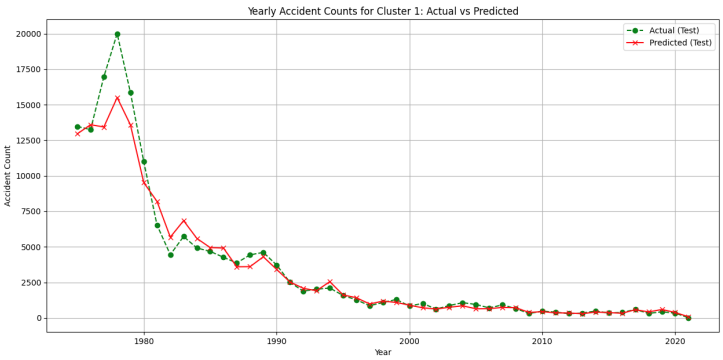


**FIGURE 5: Yearly Accident Counts for Cluster 1: Actual vs. Predicted.** The predictive performance for Cluster 1 exhibits a moderate level of accuracy, capturing the general declining trend of accidents over time but undershooting the peak accident counts, indicative of the model's challenges with abrupt fluctuations.

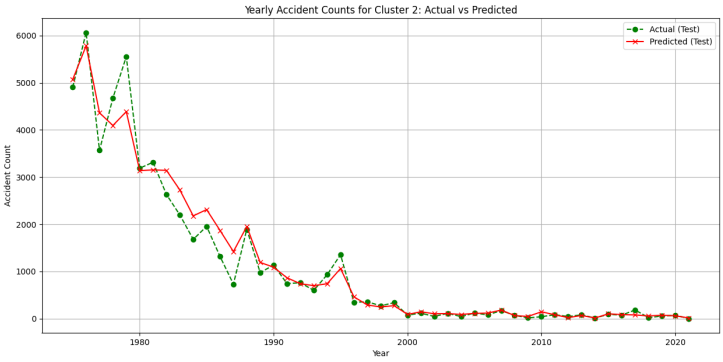


**FIGURE 6: Yearly Accident Counts for Cluster 2: Actual vs. Predicted.** The model for Cluster 2 closely tracks the actual data, including the early peak in accidents, though it tends to slightly overestimate the magnitude of these peaks.

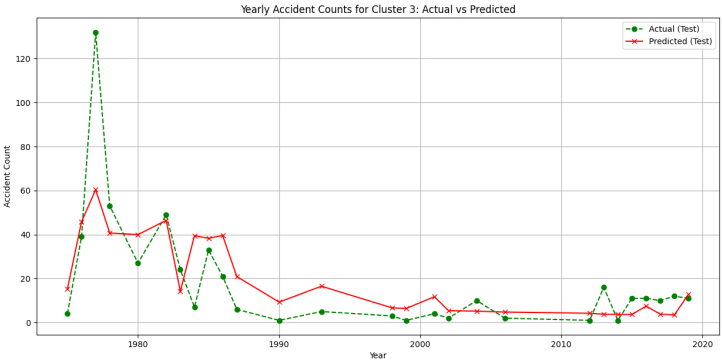


**FIGURE 7: Yearly Accident Counts for Cluster 3: Actual vs. Predicted.** Cluster 3's model shows respectable accuracy, closely following the actual data's trend, with a notable initial overestimation during the peak period.

a broad analysis across various railway companies, encompassing both freight and passenger services, without being confined to specific geographic regions. In contrast, the cluster-specific models, with scaled RMSE values from 0.16 to 0.23, offer insights into more localized patterns and trends. These clusters are derived from relationships between railway companies sharing accident localities within the same state and year, as constructed in a graph network. The clusters are not strictly location-specific; they represent shared accident attributes between companies, which may include accident frequencies and environmental conditions. This nuanced view highlights the need for predictive models that can accommodate the unique characteristics of each cluster, underscoring the multifaceted requirements of railway safety.

The visual representations of model outcomes, depicted in line plots contrasting actual versus predicted accident counts, serve as an intuitive means to gauge the models' performance. These visualizations facilitate a deeper interpretation of complex data patterns and are instrumental in identifying areas that may benefit from further analysis and model refinement.

The predictive outcomes and corresponding graphs are invaluable tools for hypothesis testing, enabling us to evaluate the validity of our hypotheses. They demonstrate how well our theoretical understanding of the influential factors aligns with actual data. While these outcomes do not offer definitive proof, they provide a directional guide for further investigation and hypothesis refinement. It is through the confluence of model performance, domain expertise, and analytical interpretation that we derive insightful conclusions. For example, the cluster-specific models with lower RMSE values, such as Cluster 1 with 0.16 and Cluster 2 with 0.19, support the hypothesis that factors like "Highway User Position" and "Equipment Involved" have a significant influence on accident occurrences. The higher RMSE values observed for Cluster 0 (0.22) and Cluster 3 (0.23) may suggest a need to consider additional variables or more complex interactions not previously accounted for in our models. The predictive models thus act as a lens through which we view the data, with each model's performance either reinforcing or challenging our preconceived notions regarding accident causation and prevention.

Ultimately, the true value of our models lies not just in their predictive accuracy but also in their ability to inform and refine our hypotheses about railway safety, leading to more effective safety measures and policies.

**4.3.1 Reinforcement of Previous Hypotheses.** Similarity matrices for each factor provide in-depth insights into how various attributes influence the formation of clusters. Please see section 3 of our preceding research [1] for a more in depth explanation of how these values are calculated and interpreted. See Figure 1 and Figure 2 for a visualization of the most and least influential factors based on our results. Please refer to the legend adjacent to each heatmap for precise values, as reliance on color intensity alone may not accurately convey the relative importance of each factor.

The high accuracy of predictions within certain clusters aligns with our previous hypothesis that "Highway User Position" and "Equipment Involved" are critical factors in accident occurrences. Models demonstrating lower RMSE in clusters with pronounced accident patterns indicate that these features indeed play a significant role in shaping accident trends. This finding reinforces the need for heightened safety measures and targeted interventions focusing on these particular elements to mitigate risks.

**4.3.2 New Insights into Railway Safety.** On the other hand, the varying model performances across clusters suggest that factors such as "Visibility" and "Weather Condition," which showed moderate similarity scores previously, might exhibit a more complex relationship with accident rates than initially surmised. Clusters with higher RMSE values may indicate the presence of additional, less obvious variables that interact with visibility and weather conditions, affecting the predictability of accidents. This revelation points to the potential benefits of developing more granular safety protocols that account for the interplay of environmental conditions with human and equipment-related factors.

**4.3.3 Temporal Factors and Accident Prediction.** The limited impact of "Date" and "Time" on accident causation, as hypothesized in our previous paper, is further corroborated by the cluster-specific models. These models indicate that while temporal factors may influence accident occurrences, their role is not as definitive as the physical and environmental conditions present at the crossing. This insight shifts the focus of railway safety measures from time-based to condition-based strategies, potentially leading to more effective risk management practices.

**4.3.4 Implications for Equipment Design and Policy.** The moderate importance of "Equipment Type" and "Equipment Struck" in accident causation, as indicated by our earlier hypotheses, has been further nuanced by the predictive models. Clusters with better predictive outcomes suggest that where certain types of equipment are involved, and specific parts are struck, there are identifiable patterns that could be critical for designing safer railway equipment and formulating policies that address these specific scenarios.

**4.3.5 Forward-Looking Safety Enhancements.** The results of our predictive modeling provide a robust validation of our initial hypotheses while also paving the way for new inquiries into railway crossing safety. The insights gained underscore the importance of developing specialized predictive models for each cluster, considering the unique characteristics inherent in different segments of the data. As we continue to unravel the complex fabric of railway accident causation, these models become invaluable in formulating forward-looking, data-driven safety enhancements. The transition from recognizing patterns to actively predicting and preventing accidents marks a paradigm shift in railway safety management, with the ultimate goal of safeguarding lives and improving the resilience of railway infrastructure.

## 5. CONCLUSION

This research represents a substantial leap in understanding and predicting railway crossing accidents through the integration of machine learning techniques. Building on our initial exploration with graph mining, we have now ventured into the predictive realm, employing KRR to forecast potential accident trends.

Our dual approach, comprising both global and cluster-specific models, has provided a comprehensive and nuanced understanding of the data.

1. **Efficacy of KRR:** The application of KRR has proven to be effective in capturing the complex relationships within our data. The variance in RMSE values across different clusters and the global model underscores KRR's capability to handle diverse data patterns.

2. **Insights from Cluster-specific Analysis:** The disparate performances of cluster-specific models highlight the importance of tailored predictive strategies. This approach is crucial for addressing the unique characteristics and challenges posed by each cluster.

3. **Global Model vs. Cluster-specific Models:** Our analysis draws attention to the strengths and limitations of a universal modeling approach versus more focused, cluster-specific models. This comparison provides valuable insights for future predictive modeling in railway safety.

4. **Implications for Future Research and Practice:** The methodologies and findings from this study offer a solid foundation for future research, especially in the application of more advanced machine learning techniques for comprehensive accident prediction.

However, our study is not without limitations. The reliance on a subset of the data for computational feasibility may affect the completeness of our analysis. Additionally, while the global model provides a broad overview, it may not capture the intricate details evident in the cluster-specific models. Future research could address these limitations by incorporating larger, more diverse datasets and employing more sophisticated machine learning algorithms. Such advancements would allow for a more accurate and holistic understanding of railway crossing accidents.

As we progress, we eagerly anticipate further exploring the predictive capabilities of machine learning in our subsequent paper. This next phase will utilize the extensive insights gained thus far, applying advanced algorithms for a deeper and more predictive analysis. Our goal is to transition from understanding past patterns to effectively anticipating and mitigating future risks, thereby enhancing railway safety at a broader scale. In our pursuit of this goal, we will explore techniques that enable us to approximate complex, high-dimensional feature spaces more effectively. By employing methods that enrich our feature set and capture the nuanced interactions within our data, we can construct models that offer greater precision and reliability. This strategy, intrinsic to our next phase of research, aligns with our objective of shifting towards a more predictive analytical framework—a transformative step in our continuous efforts to bolster railway safety. This progression from descriptive to anticipatory analytics signifies a crucial leap in our ongoing quest. It embodies our commitment to harnessing the full potential of machine learning, paving the way for identifying key predictors of accidents and, ultimately, formulating more efficacious preventative strategies.

## 6. ACKNOWLEDGMENTS

This study was made possible by funding support provided by the NSF CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) under Grant No. 2112650 and the University Transportation Center for Railway Safety (UTCRS) at UTRGV through the USDOT UTC Program under Grant No. 69A3552348340.

## REFERENCES

[1] Villalobos, Ethan, Lugo, Hector, Chen, Biqian, Gutierrez, Miguel, Tarawneh, Constantine, Xu, Ping, Chen, Jia and Papalexakis, Evangelos. "Spectral Clustering in Railway Crossing Accidents Analysis." *submitted to 2024 ASME/IEEE Joint Rail Conference* (2024).

[2] Welling, Max. "Kernel Ridge Regression." *Max Welling's classnotes in machine learning* (2013): pp. 1–3.

[3] Evgeniou, Theodoros, Pontil, Massimiliano and Poggio, Tomaso. "Regularization networks and support vector machines." *Advances in computational mathematics* Vol. 13 (2000): pp. 1–50.

[4] Hofmann, Thomas, Schölkopf, Bernhard and Smola, Alexander J. "Kernel methods in machine learning." (2008).

[5] Schölkopf, Bernhard, Herbrich, Ralf and Smola, Alex J. "A generalized representer theorem." *International conference on computational learning theory*: pp. 416–426. 2001. Springer.