Application of Social Vulnerability Index in Gun-violence Modeling

Adam Sam
School of Computing and Data Science
Wentworth Institute of Technology
550 Huntington Ave, Boston, MA 02115
sama1@wit.edu

2nd Mehmet Ergezer
Department of Computer Science and Networking
Wentworth Institute of Technology
550 Huntington Ave, Boston, MA 02115
ergezerm@wit.edu

***Abstract*—Gun violence remains undoubtedly a critical health issue in the United States, with significant disparities observed in differing communities. This study investigates the relationship between the Centers for Disease Control and Prevention's (CDC) Social Vulnerability Index (SVI) and firearm-related death rates at the county level, employing machine learning to identify predictive patterns. We compare the performance of multiple regression models -Support Vector Regression (SVR), Random Forest, and XGBoost - trained on three datasets derived from the SVI: thematic scores, demographic measurements, and an extended dataset incorporating GDP per capita and educational attainment.**

Feature importance analysis suggests that socioeconomic status is the most influential thematic predictor of firearm death rates, while more fine-grained demographic measurements particularly highlight associations with disproportionately higher rates in African American communities, consistent with prior research on structural inequalities. These findings, however, are subject to the ecological fallacy.

XGBoost outperforms other models, achieving the lowest root mean squared error (RMSE = 9.188) on death rate predictions when trained on the extended demographic dataset; a 33.6% improvement over our baseline model (RMSE = 13.838).

Findings underscore the utility of the SVI dataset as a tool for understanding gun violence and emphasize the need for targeted interventions in socioeconomically vulnerable communities. Future research could enhance predictive power by integrating additional variables such as gun ownership rates and mental health statistics.

Our implementation of this study can be found at the following URL: https://github.com/asamn/SVI-MODELING

***Index Terms*—gun violence, social vulnerability index, machine learning, predictive modeling, regression, XGBoost**

I. Introduction

Gun violence remains one of the most pressing public health crises in the United States. Every day, an estimated 125 people are killed with firearms, and more than 200 more suffer injuries from gun-related incidents [1]. In 2021 alone, the United States accounted for 16.61% of all global gun deaths, ranking second in total firearm fatalities worldwide [2]. Gun violence is a uniquely American problem that has lingering consequences beyond the immediate loss of life.

Beyond the devastating toll on victims and their families, entire communities also experience heightened fear, reduced community engagement, and economic stagnation [3]. One study in New York State found that elementary schools situated in areas of concentrated gun violence were consistently associated with lower state standardized test scores [4]. This suggests that the consequences of gun violence are both immediate and long-term, negatively influencing our future generation's opportunities and well-being.

Political polarization and emotional biases combined with a lack of federal funding effectively obstruct progress towards finding meaningful solutions mitigating the underlying cause of firearm related violence [5]. A machine learning approach that uses objective statistics could provide an ideal solution to offering an unbiased analysis of the underlying factors that may be contributing to gun violence [6].

Existing studies of gun violence have identified factors related to community distress as potential predictors of gun violence [7], [8], but many studies are rather limited in terms of geographic scope, often limiting analysis to only a select few urban communities. Additionally, they do not provide a comparative analysis encompassing other various potential predictors. Particularly, very few studies have evaluated the predictive potential of the Centers for Disease Control and Prevention's (CDC) Social Vulnerability Index (SVI), which provides feature-rich, aggregated socioeconomic and demographic data on a national, county-level scale. Existing work using the SVI for predicting gun violence suffer the same flaws of being restricted to small geographic areas—for example, focusing only on a small selection of cities [9], [10]. Given that the SVI provides a standardized, federally maintained measure of social and economic vulnerability across all U.S. counties, its utility as a predictor of firearm violence on a national scale remains an underexplored opportunity.

The purpose of this study is to assess whether the Centers for Disease Control and Prevention's (CDC) Social Vulnerability Index (SVI) can serve as an effective predictor of national and county-level firearm deaths in the United States when used in combination with machine learning models. Originally designed to evaluate community resilience to hazards such as natural disasters and public health crises, the SVI aggregates a wide range of demographic and socioeconomic indicators—including poverty levels, household composition, and access to transportation—that may also correlate with patterns of gun violence [11].

Specifically, we aim to evaluate the effectiveness of various machine learning models in predicting U.S. county-level firearm deaths per capita using the SVI dataset. We ask the following questions:

1) To what extent is the utility of the Social Vulnerability Index for predicting county-level firearm death rates in the United States on a national scope?
2) Which demographic and socioeconomic factors within the SVI are most strongly associated with firearm mortality rates?
3) How does the predictive performance of various machine learning models compare when training on the SVI?

We compare multiple predictive models to identify the best-performing approach and conduct feature importance analysis to identify the demographic and socioeconomic conditions most strongly associated with gun violence. To further enhance model performance and improve the generalizability of our findings, as well as measuring the utility of the SVI dataset alone, we expand the dataset with additional variables, including educational attainment and GDP per capita.

II. Related Works

Preliminary applications of machine learning methods for predictive modeling have shown promising results such as in medical, environmental, and industrial domains [12]–[14]. Machine learning based modeling has often been demonstrated to outperform on complex datasets when compared to traditional statistical modeling methods [15], [16]. Novel attempts to model gun violence as an infectious disease were found to yield poor results, which emphasizes the complexity of the topic and the need for machine learning based methodologies to accomplish meaningful analyses [17].

Studies focusing on machine learning based gun violence modeling have particularly succeeded in the extraction of contributing features from socioeconomic datasets [18], [19]. However, these studies do not consider using the feature-rich SVI dataset in their analyses, and have therefore identified a lack in comprehensive features in their data, such as educational attainment and housing status, all of which are included in the SVI.

Besides the lack of features, previous studies have also identified a deficiency caused by the lack of data pertaining to the incidence of gun violence itself, most notably from governmental organizations such as the FBI and NYCPD, which, in comparison to non-governmental organizations such as The Violence Project, were found insufficient even in basic details, such as the date of incidents [20]. A proposal has been made to utilize natural language processing (NLP) techniques to form an extensive database of gun violence incidents by extracting unbiased data from news articles pertaining to gun-related incidents [21], further highlighting the need for objective data on a topic that is widely influenced by political biases and lacking in federal support.

Because of the underwhelming federal support on gun violence research, coupled with the lack of feature-rich data in related studies, we explore this topic using data collected by a more federally funded domain; that being the Centers for Disease Control and Prevention (CDC). Particularly, the CDC's Social Vulnerability Index (SVI) has promising potential as a strong predictor for U.S. gun violence incidents for it's geographically encompassing, feature-rich dataset. Similarly, we also obtain our gun violence incidents data from the CDC in the form of county-level death rates, providing an adequate source of data for our purposes.

The SVI is a scoring index that quantifies the impact of demographic and socioeconomic conditions that put certain communities at greater risk when facing hazards or external pressures ranging from natural disasters (hurricanes, tornadoes), human-made incidents (chemical spills, fires), or public health crises (COVID-19) [11]. In addition to providing a numerical score indicating social vulnerability, the SVI data includes the indicating features that contribute to said score. These features include demographic measurements falling under four categories - socioeconomic status, household characteristics, ethnic minority status, and housing and transportation quality.

Although intended to be used as an index for measuring community resilience towards disasters [22], its inclusion of various demographic features marks the index as a potentially useful source of predictors for gun violence, narrowing down which specific socioeconomic and demographic aspects are the most contributing towards violence.

Of the studies that do consider the SVI as a dataset, there still exists deficiencies in terms of geographical scope. In recent work, the SVI has been applied to a similar topic of pediatric firearm violence among five major U.S cities, and it was found that the index was indeed a strong indicator of pediatric violence [9]. This work was further expanded by the same researcher by including data outside of pediatric incidents and found the same conclusions among the same five cities [10]. Another study incorporating the SVI focuses only on the Chicago area [23]. The choice of selecting only a few U.S cities may be treated as a deficiency of biased data, as a small and exclusive sample of cities can not be considered representative of the whole U.S population. Therefore, we intend to contribute further to this work by expanding the scope to a nationwide, county-level analysis. This expanded scope will provide a more universal and applicable understanding of gun violence causes in the U.S.

III. Methods

The approach for this study is split into two main problems; the first being determining a best performing model trained on a training set that minimizes the scored error of the testing set. Additionally, the first problem will be split further into two sub-problems, which is explained in the "Dataset processing" subsection. The second problem is to analyze the best performing model and perform feature importance analysis to determine the main contributing feature on firearm-death rates. The overview of the approach is described further in the subsections below.

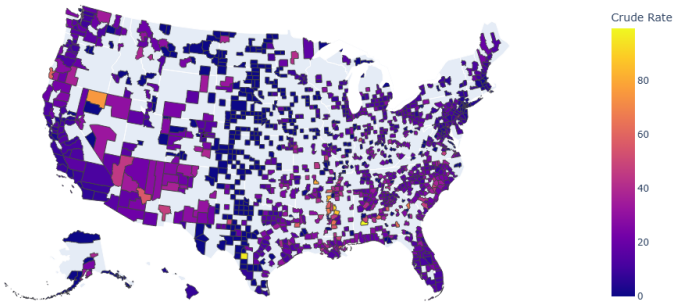

Fig. 1. Choroplethic rendering of U.S firearm deaths per 100,000 people. Gray areas represent counties with missing or redacted data.

A. Dataset

Both the county-level SVI and firearm-deaths dataset are obtained from the CDC in CSV format. All data is from the year 2022.

The firearm-deaths dataset uniquely identifies counties through FIPS codes and measures total firearm deaths, total population, and death rates per 100,000 people within a county. The CDC's WONDER database was used to query and obtain the firearm-deaths data. It is worth noting that the CDC allows further filtering of firearm-deaths categorized as "homicides", and while it may be reasonable to consider only incidents of homicides to represent gun-violence, a study investigating the accuracy of official governmental data has found inconsistencies in the CDC's reporting of unintentional firearm-deaths [24], hence we elect to use the CDC's broader category of firearm-deaths encompassing all the intents of death. Figure 1 is a chlorophethic rendering of this dataset after pre-processing. Upon initial observations of this choropleth, we noticed a few isolated outlier counties with abnormally high death-rates. It became apparent that these outliers align with tragic mass shootings that occurred throughout 2022 - particularly, the high death-rates in Uvalde County, Texas mirroring the Uvalde school shooting that claimed the lives of 21 people [25].

The SVI dataset similarly uses FIPS codes for counties. Both datasets use the same schema and are therefore easily joined on the basis of FIPS code. Columns within the SVI dataset fall under two main categories - thematic scores that rank the overall vulnerability of a particular county, and the demographic measurements in which the thematic scores are derived from.

Demographics are measured in various forms, including the raw total estimated amounts of a county's population falling under a demographic (denoted by an "E_" flag within the column name) the percentage of population under that demographic ("EP_" flag), and the percentile rankings based on that percentage ("EPL_" flag).

Thematic scores come in two forms; a summation of all percentile demographic measures related to the same theme, and a percentile ranking score of each summation (denoted by the "SPL_" and "RPL_" flags respectively). Of the thematic scores, there are four distinct themes outlined in Table I. Some measures are unthemed, serving as adjunct variables. The unthemed measures that are included in the dataset are: households with Internet access and estimates of specific minority groups. Table II shows all demographic features used, and how their column names appear within the dataset.

TABLE I
SOCIAL VULNERABILITY INDEX THEMES AND MEASURES (2022)

Theme	Demographic Measures
1. Socioeconomic Status	Below 150% Poverty, Unemployed, Housing Cost Burdened, No High School Diploma, No Health Insurance
2. Household Characteristics	Aged 65 & Older, Aged 17 & Younger, Civilian with a Disability, Single-Parent Households, English Language Proficiency
3. Racial and Ethnic Minority Status	Aggregated Minority Composition (African American, Hispanic, etc.)
4. Housing Type and Transportation	Multi-Unit Structures, Mobile Homes, Crowding, No Vehicle, Group Quarters
5. Unthemed	Households without Internet Access, Measures of Specific Minority Groups (African American, Hispanic, etc.)

Since the demographic measurements directly calculate the thematic scores, this introduces the presence of multicollinearity within the data. To address this, the SVI dataset will be split into two separate datasets, with one containing only the thematic scores, and the other containing only the demographic measurements. Separate models will be trained on both datasets - thematic scores provide a generalized sense of which theme contributes most to gun deaths, whereas the demographic measurements provide a more narrow analysis on which demographic features are the most contributing.

It should be noted that values exist within the firearm-deaths data where the death rates are labeled as "Unreliable." As described by the CDC's documentation, values that are associated with instances where gun deaths are measured to be less than 20 are not calculated. Despite this, the population and death measurements for these instances remain unredacted, allowing for manual calculation. Therefore, values that are "Unreliable" are manually calculated using the corresponding formula:

$$R = \frac{D}{P} * 100,000$$

where R is the death rate, D is the recorded amount of deaths in a county, and P is the population. Values where the rates are labeled "Suppressed" or "Not Available" indicate counties measurements that were redacted or not measured by the CDC, and therefore are dropped from the dataset. Table III presents a statistical overview of the firearm death-rate data after pre-processing.

After preparation, the total number of features for the demographic measurements dataset is 24, with 1309 samples, while our thematic scoring dataset consists of 4 features with 1309 samples, with each sample representing a county.

TABLE II
DEMOGRAPHIC PERCENTAGE MEASUREMENTS

Feature	Description	Theme
EP_POV150	% below 150% poverty line	1
EP_UNEMP	% unemployed (age 16+)	1
EP_HBURD	% with high housing cost burden	1
EP_NOHSDP	% without high school diploma (age 25+)	1
EP_UNINSUR	% uninsured (civilian, noninstitutionalized)	1
EP_AGE65	% aged 65 and older	2
EP_AGE17	% aged 17 and younger	2
EP_DISABL	% with a disability	2
EP_SNGPNT	% single-parent households with children	2
EP_LIMENG	% who speak English "less than well"	2
EP_MINRTY	% minority population	3
EP_MUNIT	% housing in large multi-unit buildings	4
EP_MOBILE	% mobile homes	4
EP_CROWD	% crowded households (more people than rooms)	4
EP_NOVEH	% households with no vehicle	4
EP_GROUPQ	% living in group quarters	4
EP_NOINT	% without internet subscription	5
EP_AFAM	% Black or African American (not Hispanic)	5
EP_HISP	% Hispanic or Latino	5
EP_ASIAN	% Asian (not Hispanic)	5
EP_AIAN	% American Indian or Alaska Native (not Hispanic)	5
EP_NHPI	% Native Hawaiian or Pacific Islander (not Hispanic)	5
EP_TWOMORE	% two or more races (not Hispanic)	5
EP_OTHERRACE	% other race (not Hispanic)	5

To measure the general utility of the SVI dataset as it stands in isolation, we expand the dataset with two additional county-level datasets that introduce adjunct demographic measures not covered in the SVI dataset. The first dataset being GDP per capita [26] and the other dataset being highest educational attainment categorized by the following; high school diploma, some college participation, and bachelor's degree or higher [27]. These measures are listed in Table IV. Both datasets cover the year 2022 and are joined to the SVI dataset using FIPS code, creating an additional four features to the dataset, totaling 28 features.

Figure 2 shows a correlation matrix for each thematic score in the dataset, displaying slightly positive correlations between firearm death rates and the thematic scores. The RPL_THEMES (aggregated scores) feature appears to be moderately correlated with each other feature, with a significant correlation with RPL_THEME1 (socioeconomic score), hence

TABLE III
DESCRIPTIVE STATISTICS OF FIREARM DEATH-RATES (PER 100K PEOPLE)

Statistic	Value
Sample Count	1309.000
Mean	14.173
Standard Deviation	13.714
Minimum	0.000
25% Percentile	0.000
50% Percentile (Median)	13.365
75% Percentile	21.027
Maximum	99.413

TABLE IV
ADJUNCT MEASUREMENTS

Feature	Description
PCT_HIGH	% with high school diploma
PCT_SOMECOL	% with some college or associate degree
PCT_BACH	% with bachelor's degree
GDP_PERCAP	estimated GDP per capita in USD

we exclude the RPL_THEMES feature from model training to account for collinearity. Figure 3 is a correlation matrix for each demographic feature in the dataset, which highlights presence of slight collinearities between the features; the most notable instance being the significantly negative correlation between high-school diploma and bachelor's degree attainments. Earning a bachelor's degree requires first obtaining a high school diploma. Therefore, as the percentage of the population that achieves a bachelor's degree increases, the percentage of the population whose highest attainment is a high school diploma will naturally decrease.

To address these collinearities, we consider tree-based models such as Random Forest and XGBoost for their demonstrated effectiveness on collinear datasets [28]. Unlike linear models, which can become unstable when features are highly correlated, tree-based models select features based on their ability to reduce impurity at each split, effectively ignoring the redundant features contributing to collinearity.

B. Python implementation

The Numpy, Pandas, and SciKit-Learn libraries are used to process the CSV data and perform model training. Choroplethic visualizations of county data are achieved using the Plotly Express library, which conveniently supports mapping county-level visualizations by providing FIPS codes. To efficiently save and load trained models, we employ the Joblib library, removing the need to retrain the models each time our code is run.

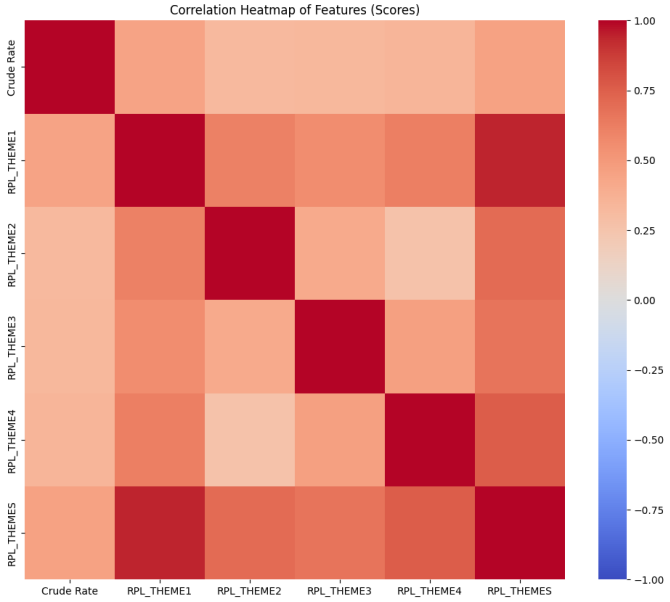

Fig. 2. Correlation heatmap of thematic scoring dataset

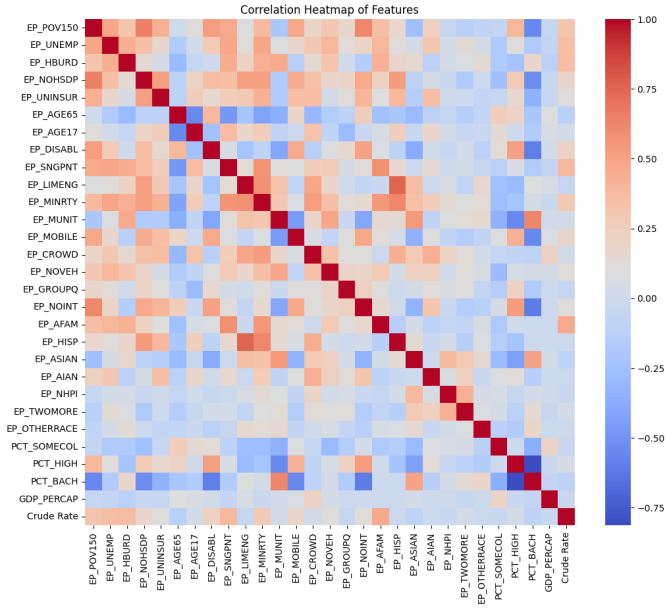

Fig. 3. Correlation heatmap of extended dataset

C. Training

The following models are considered for training:

- Support Vector Machines
- Random Forest
- XGBoost

Each model is chosen for its differing approaches to modeling data: SVR is able to model nonlinear relationships by determining a hyperplane that best fits the data within a specified tolerance while being computationally inexpensive to train [29]. Random Forest, an ensemble of decision trees, is known for its high performance and interpretability, particularly in datasets with noisy or correlated features, which is especially beneficial given the collinearity observed within the dataset [30]. XGBoost, a gradient boosting framework that sequentially builds ensemble trees, with each tree correcting past errors, was included due to its renowned effectiveness in regression tasks on complex datasets, although at the trade-off of its more meticulous hyperparameter tuning process [31].

To serve as a baseline, we construct a naive model to simply predict the mean value of the training targets for all instances in the testing set. Specifically, the baseline predicted the average firearm-death rate from the training data as a constant output for each test sample. This simple model serves as a reference point to evaluate the magnitude of improvement in predictive accuracy of the models.

To improve training performance, data values are normalized using SciKit's MinMaxScaler(), transforming each feature into a value between 0 and 1 while maintaining the original distribution of data. This scaling of data improves training performance especially for distance-based algorithms like SVM. The data is then split between 20% testing and 80% training. We create a training pipeline for each model consisting of distinct search spaces for cross-validation.

Our model set is trained separately on three datasets: the first dataset being only the SVI scores, the second being only the measures, and the third being the measures extended with GDP per capita and educational attainment features.

D. Cross-validation

To promote robust model evaluation and prevent overfitting, we implement a randomized 5-fold cross-validation for hyperparameter tuning using cross-validation functions provided by SciKit-Learn. Because of their relatively low tuning requirements, the Support Vector Regression (SVR) and Random Forest models were tuned using the Grid Search technique via GridSearchCV. For the XGBoost model, which requires more intricate hyperparameter tuning, we utilize Bayesian optimization via BayesSearchCV from the skopt library to more efficiently navigate a large and continuous hyperparameter space.

To construct search spaces, we employ an adaptive grid expansion strategy. Initially, a small grid is defined for each model, covering a broad range of values for each hyperparameter. After running the preliminary search, we inspected the best-performing parameters. If any selected hyperparameter value appeared at the boundary of its defined range — such as the minimum or maximum value — we interpret this as an indication to expand the search space for that particular parameter, extending it further in the direction of the boundary. Conversely, for values that fall within the middle of the boundary, we narrow the search space to focus more closely on that value. This process is repeated until the search space was fine-grained enough while maintaining that the optimal value no longer appeared on an edge, ensuring that we capture the optimal parameters without exhaustively searching an unnecessarily large parameter space. This approach allowed

for more precise tuning while reducing the computational costs that come with large search spaces.

For SVR, hyperparameters such as regularization (C), kernel type (linear, rbf, degree), and kernel coefficient (gamma) were searched over a predefined, grid. Random Forest tuning only needed to consider two parameters; the number of estimators (n_estimators) and maximum tree depth (max_depth).

In contrast, XGBoost's hyperparameter tuning leveraged Bayesian optimization across a much larger search space. The search included both continuous and discrete distributions over the number of estimators, learning rate, and subsample, among others. This technique enables faster convergence toward optimal configurations by using past evaluations to inform future sampling decisions.

During each fold, performance was using measured negative mean squared error as the scoring metric, ensuring minimization of prediction error. The best set of hyperparameters for each model was selected based on the lowest average MSE across folds, and the final model using the best parameters was retrained on the full training set before being evaluated on the test set.

E. Interpretation

Permutation importance is used to rank features by their influence on predicting firearm-death rates. This technique measures the decrease in model performance when the values of each feature are shuffled [32]. The features are ranked based on their importance to determine their influence in predicting firearm-death rates. It is important to note that while permutation importance can provide insight into the contribution of each feature, its results cannot be taken with absolute certainty, particularly in the presence of correlated features.

IV. Results

A. Performance

The performance of each tuned model was assessed using its root mean squared error (RMSE) score when evaluated on the testing set. The results for the three datasets are summarized in Tables V, VI, and VII. Figure 4 shows a residual KDE plot comparing each model. Residual points quantify differences between the predicted and actual value of each data point. Observed residuals are calculated with the formula:

$$e_i = y_i - \hat{y}_i$$

Where e_i is the residual for the i-th observation. y_i is the actual value for the i-th observation. $\hat{y}_i$ is the predicted value for the i-th observation.

The initial models trained on the four thematic SVI scores all exhibited similarly low performance, with RMSEs approximating to 11.5. This suggests that the thematic scores-while useful for summarizing overall social vulnerability-lack the granularity needed to capture county-level variations in factors influencing firearm death rates. Broad, aggregated measures can obscure the influence of specific socioeconomic components that may be more directly related to gun violence.

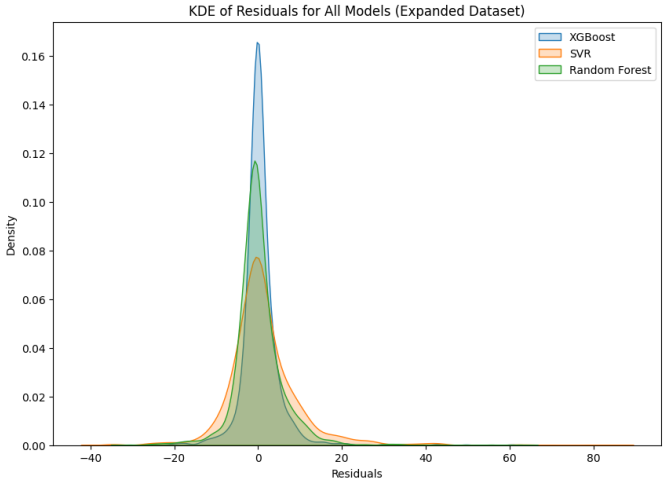

Fig. 4. KDE residual plot of each model trained on the extended dataset. XGBoost demonstrates highest accuracy, indicated by its steep bell-curve centered around zero.

TABLE V
Model Performance on Scores

Model	**RMSE**	**Training RMSE**
Baseline	13.354	13.354
Support Vector Regression	11.627	11.922
Random Forest	11.483	10.239
XGBoost	11.562	9.330

Upon training the models on the more feature-rich dataset including individual demographic measures, overall performance noticeably improved. This suggests that the individual demographic measures enable a more nuanced understanding of the social factors influencing gun violence compared to the aggregated thematic scores. The XGBoost model, in particular, begins to demonstrate its superior performance on larger datasets, as it is found to have the lowest RMSE value.

The inclusion of additional features in our extended dataset, incorporating additional educational attainment measures and GDP per capita, marginally improved performance across all models. This suggests that these additional socioeconomic indicators contribute incrementally to the models' ability to predict death rates, including these features adds further valuable context to the models.

Overall, XGBoost trained on our extended demographic dataset achieved the highest predictive performance (RMSE = 9.188) among all experiments, with a 33.6% improvement over the baseline model. Models trained on demographic variables consistently outperformed those trained on thematic scores, even more so when we extended the dataset with additional features. This demonstrates how higher-dimensional data yields greater predictive power by enabling more nuanced differentiation between the samples, and suggests that while the SVI dataset alone may be adequate for predicting firearm death rates, gains in predictive performance can be accumu-

TABLE VI
MODEL PERFORMANCE ON MEASUREMENTS

Model	RMSE	Training RMSE
Baseline	13.354	13.354
Support Vector Regression	9.996	9.127
Random Forest	9.639	3.934
XGBoost	9.561	0.944

TABLE VII
MODEL PERFORMANCE ON EXTENDED MEASUREMENTS

Model	RMSE	Training RMSE
Baseline	13.838	13.838
Support Vector Regression	9.701	8.750
Random Forest	9.255	3.934
XGBoost	9.188	0.897

lated through the addition of features.

B. Feature Importance

Using the best-performing model (XGBoost with demographic data), we extracted demographic feature importance based on the permutation importance technique. The most influential features included are in Table IX. Figure 5 shows a SHAP graph explaining the contribution of each feature to the model's predictions, using results from the XGBoost model.

TABLE VIII
FEATURE IMPORTANCE OF THEMATIC SCORES

Feature	Permutation Importance
RPL_THEME1	0.2374
RPL_THEME2	0.0905
RPL_THEME4	0.0901
RPL_THEME3	0.0557

TABLE IX
FEATURE IMPORTANCE OF DEMOGRAPHIC MEASURES

Feature	Permutation Importance
EP_AFAM	0.30387
EP_HBURD	0.092437
EP_ASIAN	0.057843
EP_DISABL	0.040408
EP_UNEMP	0.037692
EP_AGE17	0.031683
EP_NOVEH	0.030318
EP_AGE65	0.025054
EP_UNINSUR	0.014713
EP_OTHERRACE	0.012696
EP_HISP	0.01175
EP_GROUPQ	0.009342
EP_SNGPNT	0.009098
EP_AIAN	0.007818
EP_POV150	0.006996
PCT_SOMECOL	0.006073
EP_NOINT	0.005547
EP_CROWD	0.005495
PCT_BACH	0.005437
PCT_HIGH	0.005268
EP_MUNIT	0.003841
GDP_PERCAP	0.000657
EP_MOBILE	0.000261
EP_TWOMORE	-0.000094
EP_NOHSDP	-0.000506
EP_NHPI	-0.000948
EP_MINRTY	-0.001042
EP_LIMENG	-0.00428

Thematic scoring importances are similarly calculated using permutation importance. All models found the same order of rankings for the thematic scoring importances, with the most important theme being socioeconomic status. Thematic Score importances derived by XGBoost are provided in Table VIII - these provide a more generalized sense of which theme correlates to firearm deaths.

Most notably, the findings of the demographic measurement models suggest that African American communities are the most susceptible to gun violence, as indicated by the EP_AFAM feature having the highest permutation importance value. Referring to the SHAP graph in Figure 5, there is a distinction where samples with a high percentage of African American population (red data points) are skewed to the right of the origin point, indicating positive association with death-rate prediction, and samples with a low percentage (blue data points) are skewed to the left, indicating a negative impact. This confirms the feature to have a positive correlation, aligning with initial correlation matrix in Figure 3. This is a consequence of the historical mistreatment of African Americans within the U.S through systematic racism, which has disproportionately placed African Americans at a disadvantage [33], [34].

Despite the initial thematic models placing ethnic minority status as the least important theme may appear to contradict this finding, this can be explained by the very nature of the SVI dataset, which does not count the measurements of specific minority groups (EP_AFAM, EP_ASIAN, etc.)

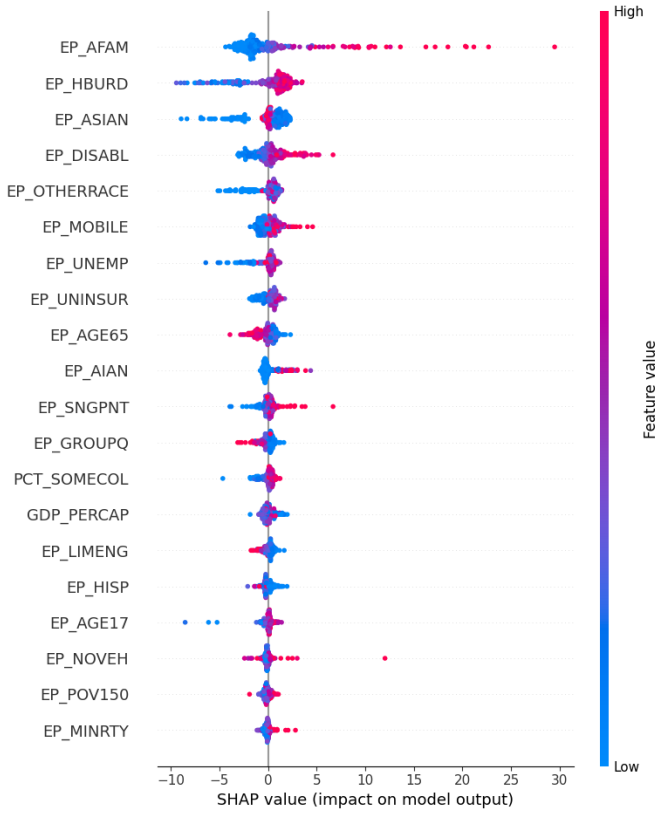

Fig. 5. SHAP Plot of features

into its aggregated calculation of the ethnic minority score (RPL_THEME3) - rather, these specific measurements are categorized as adjunct ”unthemed” features that do not contribute to any thematic score. The only measurement that does contribute to the minority status score (EP_MINRTY) only generalizes by aggregating minority groups into the same measurement and does capture the proportionality of specific minority groups.

The next most influential features include EP_HBURD (percentage of households with high housing cost burden) and EP_ASIAN (percentage of Asian residents). High housing cost burden appears to be positively correlated with firearm deaths, and has been associated with economic instability and displacement risk that may indirectly contribute to violent crime [35], [36]. It is unclear whether the ranking of EP_ASIAN is positively, or negatively correlated. Referring to the SHAP graph in Figure 5, there is a notable skew to the left, where samples with low Asian population are predicted to lower the death rate. Conversely, referring to the dense cluster at the origin point, at the left side of this cluster are samples with high Asian population, and at the right are samples with low population, indicating that samples with low population do not affect rates, contradicting the tail-like structure seen at the left of the graph. A possible explanation to this could be the geographical clustering of Asian populations in urban areas with other minority groups and socioeconomic features that do influence firearm death rates - suggesting that the tail-like structure represents areas with an overall low minority population, and the cluster represents urban areas with a dense minority population. This suggests that the consideration of minority status is more important than initially suggested through its low thematic ranking, and warrants further exploration.

Other features with moderate importance include EP_DISABL (percentage of residents with a disability), EP_UNEMP (unemployment rate), both appearing to be positiely correlated with firearm deaths. These predictors align with existing studies that have linked disability prevalence and limited employment opportunities to higher risks of community violence [37], [38].

However, these findings should be interpreted with caution, as our analysis is based on data from a single year (2022), meaning temporal dynamics of gun violence cannot be captured. The results are also subject to limitations in data quality, including possible reporting inconsistencies in firearm death counts and demographic measures. This is further exasperated by the amount of missing county data observed within our dataset. Finally, this study operates at an ecological level, and the observed associations between demographic characteristics and firearm death rates should not be assumed to imply causation at the individual level-avoiding the ecological fallacy is particularly important when interpreting demographic variables such as EP_AFAM.

V. Conclusion

This study investigated the potential of the CDC’s Social Vulnerability Index (SVI) as a predictive tool for modeling firearm-related deaths per capita at the county level across the United States. By training and comparing Support Vector Regression, Random Forest, and XGBoost models on multiple variations of the SVI dataset, we evaluated the utility of both the aggregated thematic scores and granular demographic measures. The results demonstrated that XGBoost, particularly when trained on the extended demographic dataset incorporating GDP per capita and educational attainment, achieved the highest predictive performance (RMSE = 9.188), with a 33.6% improvement over the baseline model.

Feature importance analysis suggest socioeconomic status as the most influential thematic predictor, while the percentage of a county’s population identifying as Black or African American (EP_AFAM) emerged as the single most important demographic feature. This aligns with prior research documenting the disproportionate burden of gun violence on historically marginalized communities, and contradicts with the initial finding of aggregated minority status being the least important thematic predictor. This underscores the value of disaggregated, group-specific demographic measures over broad composite measures. Additional predictors such as high housing cost burden, unemployment, disability rates, and youth population also suggested associations with firearm mortality.

While these findings highlight the promise of using the SVI for the predictive analysis of gun violence, several limitations must be acknowledged. Firstly, the study relied on cross-sectional data from a single year (2022), preventing assessment of temporal trends. Secondly, potential reporting inconsistencies and missing data in firearm deaths and demographic measures may affect the validity of the results. Thirdly, the analysis was conducted at the ecological level, meaning that assumptions from our findings are subject to the ecological fallacy, where observed associations with particular demographics cannot be assumed to hold true on an individual level.

Despite these limitations, this work supports the value of the SVI as a foundation for firearm violence modeling and demonstrates that predictive performance can be enhanced through the integration of supplementary indicators.

Future work for this project could be expanded to include integration with additional county-level datasets, such as gun laws, gun ownership, and mental health statistics. As observed through our expansion of the SVI dataset, incorporating more comprehensive demographic measurements can help improve predictive power. Expanding temporal analysis and incorporating time-series models could also enable forecasting trends and identification of early warning signs, making predictive models not just descriptive but proactive tools in violence prevention.

REFERENCES

[1] Everytown Research, "Gun violence in america," May 2020.

[2] World Population Review, "Gun deaths by country 2025."

[3] Y. Irvin-Erickson, M. Lynch, A. Gurvis, E. Mohr, and B. Bai, "A neighborhood-level analysis of the economic impact of gun violence," tech. rep., Urban Institute, 2017.

[4] R. H. K. D. A. L. A. P. N. S. T. J.-B. Dessa Bergen-Cico, Sandra D. Lane and R. A. Rubinstein, "Community gun violence as a social determinant of elementary school achievement," *Social Work in Public Health*, vol. 33, no. 7-8, pp. 439–448, 2018. PMID: 30427288.

[5] P. M. Carter, M. A. Zimmerman, and R. M. Cunningham, "Addressing key gaps in existing longitudinal research and establishing a pathway forward for firearm violence prevention research," *J. Clin. Child Adolesc. Psychol.*, vol. 50, pp. 367–384, May 2021.

[6] S. J. West, "Chapter 37 - applying data science to the study of gun violence," in *Handbook of Gun Violence* (N. D. Thomson, ed.), pp. 497–508, Academic Press, 2025.

[7] A. Ali, J. Broome, D. Tatum, J. Fleckman, K. Theall, M. P. Chaparro, J. Duchesne, and S. Taghavi, "The association between food insecurity and gun violence in a major metropolitan city," *Journal of trauma and acute care surgery*, vol. 93, no. 1, pp. 91–97, 2022.

[8] D. E. Goin, K. E. Rudolph, and J. Ahern, "Predictors of firearm violence in urban communities: A machine-learning approach," *Health Place*, vol. 51, pp. 61–67, 2018.

[9] A. M. Polcari, L. E. Hoefer, K. M. Callier, T. L. Zakrison, S. O. Rogers, M. C. Henry, M. B. Slidell, and A. J. Benjamin, "Social vulnerability index is strongly associated with urban pediatric firearm violence: An analysis of five major us cities," *Journal of Trauma and Acute Care Surgery*, vol. 95, no. 3, 2023.

[10] A. M. Polcari, M. B. Slidell, L. E. Hoefer, M. C. Henry, T. L. Zakrison, S. O. Rogers, and A. J. Benjamin, "Social vulnerability and firearm violence: Geospatial analysis of 5 us cities," *Journal of the American College of Surgeons*, vol. 237, no. 6, 2023.

[11] CDC, "Social vulnerability index," Oct. 2024.

[12] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F. Amenta, "Applications of machine learning predictive models in the chronic disease diagnosis," *Journal of personalized medicine*, vol. 10, no. 2, p. 21, 2020.

[13] A. Hamdan, K. I. Ibekwe, E. A. Etukudoh, A. A. Umoh, and V. I. Ilojianya, "Ai and machine learning in climate change research: A review of predictive models and environmental impact," *World Journal of Advanced Research and Reviews*, vol. 21, no. 1, pp. 1999–2008, 2024.

[14] J.-R. Ruiz-Sarmiento, J. Monroy, F.-A. Moreno, C. Galindo, J.-M. Bonelo, and J. Gonzalez-Jimenez, "A predictive model for the maintenance of industrial machinery in the context of industry 4.0," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103289, 2020.

[15] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, p. 455, 2020.

[16] A.-L. Boulesteix and M. Schmid, "Machine learning versus statistical modeling," *Biometrical Journal*, vol. 56, no. 4, pp. 588–593, 2014.

[17] C. Loeffler and S. Flaxman, "Is gun violence contagious? a spatiotemporal test," *Journal of Quantitative Criminology*, vol. 34, pp. 999–1017, Dec 2018.

[18] R. Singhal, "Predictive modeling of gun violence using machine learning: Understanding the role of demographic and socioeconomic factors at the county level," *International Journal of High School Research*, 2024.

[19] M. Rasool and M. Maphosa, "Exploring global gun violence prediction through machine learning models," in *International Conference on Artificial Intelligence and its Applications*, pp. 205–211, 2023.

[20] X. Lei, *Analyzing the Risk of Mass Shootings in the United States*. PhD thesis, Iowa State University, 2022.

[21] E. Pavlick, H. Ji, X. Pan, and C. Callison-Burch, "The gun violence database: A new task and data set for NLP," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (J. Su, K. Duh, and X. Carreras, eds.), (Austin, Texas), pp. 1018–1024, Association for Computational Linguistics, Nov. 2016.

[22] B. E. Flanagan, E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis, "A social vulnerability index for disaster management," *Journal of Homeland Security and Emergency Management*, vol. 8, no. 1, p. 000010220215477355 1792, 2011.

[23] C. Dirago, M. Poulson, J. Hatchimonji, J. Byrne, and D. Scantling, "Geospatial analysis of social vulnerability, race, and firearm violence in chicago," *Journal of surgical research*, vol. 294, pp. 66–72, 2024.

[24] C. Barber and D. Hemenway, "Too many or too few unintentional firearm deaths in official u.s. mortality data?," *Accident Analysis Prevention*, vol. 43, no. 3, pp. 724–731, 2011.

[25] M. I. Jowarder, "Aftermath of uvalde shooting: An account of psychological trauma of survivor children.," *Journal of Loss & Trauma*, vol. 28, no. 7, 2023.

[26] U.S. Bureau of Economic Analysis (BEA), "GDP by County, Metro, and Other Areas — U.S. Bureau of Economic Analysis (BEA) — bea.gov." https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas.

[27] U.S. Department of Agriculture, Economic Research Service, "County-level Data Sets - Documentation — Economic Research Service — ers.usda.gov." https://www.ers.usda.gov/data-products/county-level-data-sets/documentation, 2025.

[28] S. Chowdhury, Y. Lin, B. Liaw, and L. Kerby, "Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance," in *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, pp. 17–25, 2022.

[29] D. Basak, S. Pal, and D. Patranabis, "Support vector regression," *Neural Information Processing – Letters and Reviews*, vol. 11, 11 2007.

[30] A. Cutler, D. Cutler, and J. Stevens, *Random Forests*, vol. 45, pp. 157–176. 01 2011.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, p. 785–794, ACM, Aug. 2016.

[32] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," 2019.

[33] T. L. Gillum, C. J. Hampton, and C. Coppedge, "Using the socio-ecological model to understand increased risk of gun violence in the african american community," *Psychological Reports*, vol. 128, no. 1, pp. 126–148, 2025. PMID: 38804858.

[34] B. T. Johnson, A. Sisti, M. Bernstein, K. Chen, E. A. Hennessy, R. L. Acabchuk, and M. Matos, "Community-level factors and incidence of

gun violence in the united states, 2014–2017," *Social Science Medicine*, vol. 280, p. 113969, 2021.
[35] S. Shamsuddin and C. Campbell, "Housing cost burden, material hardship, and well-being," *Housing Policy Debate*, vol. 32, no. 3, pp. 413–432, 2022.
[36] R. Remeikiene, L. Gaspareniene, A. Fedajev, E. Raistenskis, and A. Krivins, "Links between crime and economic development: Eu classification.," *Equilibrium (1689-765X)*, vol. 17, no. 4, 2022.
[37] C. A. Winters, "Learning disabilities, crime, delinquency, and special education placement," *Adolescence*, vol. 32, no. 126, 1997.
[38] K. Edmark, "Unemployment and crime: Is there a connection?," *Scandinavian Journal of Economics*, vol. 107, no. 2, pp. 353–373, 2005.