Examining the Impact of Light Brightness on Suicide Rates: A Five-Year Case Study of New York Counties (2018–2022)

Darren Deng, Joya Wheatfall-Melvin, Cassie Zhang

December 17, 2024

Abstract

Light brightness, particularly artificial illumination, is an often-overlooked environmental factor that may have a significant impact on mental health and suicide rates. Existing research predominantly explores the national effects of light brightness on human circadian rhythms, sleep patterns, and animal migration, but limited attention has been given to its influence on mental health outcomes such as suicide rates at the local level. Understanding how variations in artificial light brightness correlate with suicide rates is crucial, as localized patterns may reveal insights missed by broader studies. Our research aims to address this gap by investigating the relationship between light brightness and suicide rates in New York counties.

1 Introduction

With rapid urbanization and technological advancement, human-caused environmental changes like artificial light brightness have become increasingly common. Artificial light brightness is widely recognized as a byproduct of urban expansion, industrialization, and human excess. Other research has concluded light brightness is damaging to the human body, even affecting our nuerohormones. For instance, individuals with a genetic predisposition to bipolar disorder may be at greater risk of developing the condition due to the impact of light pollution on their biorhythms[1]. This study investigates the relationship between artificial brightness and suicide rates across New York counties, incorporating key mental health indicators such as depression rates, suicide rates, and post-traumatic stress disorder (PTSD) prevalence. To address this, we ask: How does light brightness affect suicide rates at the local level in New York counties? By analyzing the interplay between environmental, psychological, and social variables, this study provides critical insights about how the globalized world may be causing us harm.

2 Data Acquisition

Light Brightness Map

We acquired our data for light brightness from the VIIRS Light brightness Map. Visible Infrared Imaging Radiometer Suite (VIIRS) is a measurement tool created by NASA that measures night-time light emission per day[2]. Although VIIRS is one of the main sources used to capture artificial light emission, this data also takes into account cloud visibility and hourly forecasting to produce aggregated light emission calculations for the each latitudinal and longitudinal point on the globe. Utilizing the select tool on the map interface, we select the state boundaries of New York to capture light brightness of the entire area. With the New York state range selected, we download the data as a TIF file and convert to CSV file.

Indicators

We obtained many of the mental health indicators and natural indicators from public health open databases. Utilizing the built-in dashboards for each website, we selected a data range from 2018 - 2022 for all New York counties. All data sets are downloaded from respective websites, and exported into CSV files.

- `Real_GDP` is the aggregated GDP per county provided by the New York Regional Economic Analysis Project.
- `Age-Adjusted Death Rate (Deaths per 100,000)` is aggregated suicide and self-inflicted injury death rates for New York by county from 2018-2022. Collected by the NIH National Institute on Minority Health and Health Disparities.
- `Depression Rate (per 100k)`, `PTSD rate (per 100k)`, `Trauma Rate (per 100k)` are aggregated rates of depression, PTSD, and trauma via mental health screenings at the county level from Mental Health America's (MHA) Online Screening Program. This program provides free, anonymous, confidential, and clinically validated mental health screening tools.
- `tree_cover_loss` measures international tree cover loss down to the county level per state. Global Forest Watch uses an algorithm to map tree cover loss from Landsat satellite images.
- `gross_emissions_co2e_all_gases_Mg` is the forest-related greenhouse gas fluxes measured by New York County per year. Global Forest Watch uses a geospatial monitoring framework to estimate global forest carbon fluxes.

Limitations

It is important to acknowledge several limitations within collecting the variables for this research. First, the use of aggregated data at the county level prevents us from capturing individual-level variations in mental health outcomes or suicide rates, which could mask significant within-county disparities. Second, the data provided does not contain

Table 1: Mental Health Indicator Source Information

Variable Label	Source	Collection Range
`Real_GDP`	New York Regional Economic Analysis Project	2018–2022
`Age-Adjusted_DeathRate(per100k)`	National Institutes of Health	2018–2022
`Depression Rate (per100k)`	Mental Health America	2018–2022
`PTSD rate (per100k)`	Mental Health America	2018–2022
`Trauma Rate (per100k)`	Mental Health America	2018–2022
`tree_cover_loss`	Global Forest Watch	2018–2022
`gross_emissions_co2Mg`	Global Forest Watch	2018–2022

the latitude and longitude measurements for each county, which we manually had to account for. To eliminate potential misalignment, the data for `tree_cover_loss` and `gross_emissions_co2e_all_gases_Mg` are not used in our analysis, due to collected date range.

3 Methods

3.1 Data Cleaning and Merging

While the raw datasets used in our analysis were provided in a tidy, tabular format, several columns contained irregularities and inconsistencies that required cleaning to ensure the reliability of our analysis. For that reason, we conduct a thorough cleaning process on key columns essential for our analysis. We had one main essential process to complete before cleaning out fluff: (1) Linking the Brightness data to all 62 counties in New York State; and (2) Cleaning the mental health columns for irregularities and inconsistencies; and (3) Merge all clean datasets to create one analytic dataset.

3.1.1 Linking Brightness Index to county

Extracting brightness data: The TIF file containing light brightness intensity was read using the `rasterio` library. We accessed the first data band and traversed through each pixel to retrieve the brightness value and its corresponding geographical coordinates (latitude and longitude). Only valid pixels with brightness values greater than zero were retained.

Saving the raw data: The extracted brightness values, along with their latitude and longitude, were stored in a Pandas DataFrame. The data was then exported to a CSV file named light-brightness-data.csv.

Data transformation: The brightness data was read into Pandas for further analysis. We calculated the log-transformed brightness values using log(1 + Brightness) to normalize the distribution and created visualizations, such as histograms, to analyze the data distribution.

Grid-based aggregation: The latitude and longitude were divided into 100 bins each to create a grid. For each grid cell, the average brightness was calculated using groupby operations. The center coordinates of each grid cell were also computed, and the final aggregated grid data was saved.

Mapping brightness to counties: To link the brightness data to county-level information, we matched each pixel's latitude and longitude to the nearest county centroid using the geodesic distance function from the `geopy` library. This process iteratively found the closest county for each pixel.

Aggregating by county: Once the brightness data was matched to counties, we summed the brightness values for each county and stored the final aggregated results in a CSV file named county-light-brightness.csv.

3.1.2 Cleaning Real GDP

To clean and prepare the GDP data, we first loaded the dataset new-york-GDP.csv while skipping the first three rows to handle metadata, ensuring the file was parsed correctly using a comma delimiter. To focus on relevant data, we filtered the rows where the Description column equaled "Real GDP (thousands of chained 2017 dollars)". From the filtered data, we extracted the columns GeoName and 2019, renaming them to County-Name and Real-GDP-2019, respectively.

To standardize the county names, we cleaned the County-Name column by removing the text , NY to ensure consistency in naming. Additionally, the first row was removed to eliminate any residual artifacts from the earlier extraction. Finally, the cleaned data, containing only the County-Name and Real-GDP-2019 columns, was saved as Filtered-Real-GDP-2019.csv. We repeated this process several to get the range from 2018 - 2022.

3.1.3 Cleaning Age-Adjusted Death Rate

To clean and prepare the suicide rate data, we first loaded the dataset HDPulse-data-filtered.csv using a comma delimiter while skipping the first three rows to handle metadata and ensure clean data import. Any problematic rows were managed with on-bad-lines='skip' to avoid interruptions.

To further refine the data, we conditionally removed rows from index 70 to 93 if the dataset contained sufficient rows (more than 70), ensuring that unnecessary or corrupted rows were excluded. From the cleaned dataset, we selected only the relevant columns: County (county identifier), Average Annual Count (the average number of suicide cases), and Age-Adjusted Death Rate (deaths per 100,000) (standardized suicide rate). The resulting filtered dataset was then saved as HDPulse-data-filtered.csv in the specified directory.

3.1.4 Cleaning Depression Rate (per 100k), PTSD Rate (per 100k), and Trauma Rate (per 100k)

From the depression dataset (Depression-County-Bar-Full-Data.csv), PTSD dataset (KPI-PTSD-County-Full-Data.csv), and trauma dataset (KPI-County-Trauma-Full-Data.csv), we selected the columns County Name and per 100K All Years, renaming to Depression Rate (per 100k), PTSD Rate (per 100k), and Trauma Rate (per 100k), respectively. Rows with invalid or null values were removed to ensure data quality.

The cleaned datasets were then sequentially merged with the master dataset **merged-data.csv** using County Name as the key. The depression dataset was first merged using left-on='County-Name' from the master dataset and right-on='County Name' from the depression data, and the redundant County Name column was dropped. This process was repeated for the PTSD dataset and trauma dataset, ensuring consistency and removing duplicate columns after each merge.

3.1.5 Cleaning Tree Cover Loss, Gross Emissions CO2 Gasses

To clean and prepare the Tree Cover Loss and Gross Emissions CO2 Gases data, we began by loading the dataset treecover-loss-by-region-ha.csv and extracting the relevant columns: adm2, umd–ree-cover-loss-year, umd-tree-cover-loss-ha, and gfw-gross-emissions-co2e-all-gases-Mg. This subset of data was saved as cleaned-treecover-los-data.csv for further processing. Next, we loaded the metadata file adm2-metadata.csv and standardized its column names by renaming adm2-id to adm2 and name to county-name. To ensure consistency, the data type of the adm2 column in both datasets was converted to strings. The cleaned tree cover loss data was then merged with the metadata file using a left join on the adm2 column, replacing adm2 with the corresponding county-name for clarity.

3.1.6 Filtering to prepare final analytic data set

After merging all the variables we mentioned above, we get the final version of the dataset. We exported the merged dataset merged-data to a CSV file using the to-csv() function from the Pandas library. The output file was named merged-data-final.csv, and the parameter index=False was set to exclude the DataFrame's index column, ensuring a clean file for further analysis.

Table 2: Sample of Merged Dataset

County Name	**Brightness**	**Real GDP (2019)**	**Avg Count**	**Death Rate**	**Depression Rate**	**PTSD Rate**	**Trauma Rate**	**Tree Cover Loss**
Allegany	2694.3	1,723,115	7	17.3	23.16	59.6	57.15	21,474.35
Bronx	81936.7	44,654,229	77	5.2	23.62	33.4	0.76	211.85
Broome	18621.5	9,235,160	21	11.0	38.44	78.9	29.72	11,696.57
Cattaraugus	6915.9	2,912,501	11	15.8	27.90	72.6	146.85	58,194.15
Cayuga	7831.0	2,901,102	11	14.5	36.16	57.6	49.03	17,966.59
Chautauqua	13941.1	4,826,497	17	12.3	30.51	71.9	95.34	39,610.98

3.2 Empirical Strategy

To analyze the relationship between environmental brightness and suicide rates, holding mental health indicators constant(depression rate, trauma rate, and PTSD rate), we employed a combination of statistical modeling and machine learning techniques.

3.2.1 Regression Analysis

First, we estimated the effect of environmental brightness and mental health indicators on age-adjusted suicide rates using a Generalized Linear Model (GLM) with a Poisson distribution. This choice was appropriate because the dependent variable, the suicide rate (age-adjusted deaths per 100,000 people), represents count-based data averaged over the years 2018-2022. The regression equation takes the following form:

$$\text{Model 1:} \quad \ln\left(\mathbb{E}(\text{Age Adjusted Death Rates (per 100k)}_i)\right) = \beta_0 + \beta_1 \text{Brightness}_i \\ + \beta_2 \text{Depression Rate (per 100k)}_i) + \beta_3 \text{Trauma Rate (per 100k)}_i) + \epsilon_i \quad (1)$$

Here, the dependent variable is the age-adjusted suicide rate per 100,000 people for $county_i$, and the independent variables include Brightness, Depression Rate, and Trauma Rate. The Poisson model is appropriate given the nature of the strictly positive dependent variable; suicide rates cannot be negative. We assume that controlling for depression and trauma rates, we will find correlative relationship between suicide rates and brightness To further understand variable relationships, we used a correlation matrix to examine linear associations between predictors and the outcome variable.

To deepen our understanding of how mental health indicators collectively influence suicide rates, we incorporate an interaction term between depression and trauma rates into our empirical model. The rationale for creating this interaction variable stems from the hypothesis that the combined effects of depression and trauma may heighten mental health burdens in a way that is not captured when these variables are analyzed independently. In other words, individuals experiencing both higher depression rates and trauma rates may face compounded risks that amplify their impact on suicide rates.

As a result, we constructed an interaction variable as the product of the depression rate per 100,000 and the trauma rate per 100,000 for each county. We added this to our our second Generalized Linear Model (GLM). The updated model is expressed as:

$$\text{Model 2:} \quad \ln\left(\mathbb{E}(\text{Age Adjusted Death Rates (per 100k)}_i)\right) = \beta_0 + \beta_1 \text{Brightness}_i \\ + \beta_2 \text{Depression Rate (per 100k)}_i) + \beta_3 \text{Trauma Rate (per 100k)}_i) + \beta_4 (\text{Depression} \times \text{Trauma}) + \epsilon_i \quad (2)$$

3.2.2 Machine Learning

We also employed a series of machine learning models to analyze the relationship between brightness and the mental health indicators (depression rates, trauma rates, PTSD rates) to understand the predictive likelihood of high vs. low suicide rates across New York counties. We employed four machine learning models to predict suicide rate classifications and evaluate their predictive performance:

- **Logistic Regression**
- **Support Vector Machine (SVM)**
- **Random Forest**
- **K-Nearest Neighbors (KNN)**:

For each model, we calculated the **accuracy** and **F1-score** on the test set to evaluate predictive performance. A **confusion matrix** was generated for best fitting model (Logistic Regression) to illustrate the classification of counties into high and low suicide rate categories.

4 Results

In this section, we present the results of our analysis. We will first outline several key findings from our first two regression models. In the last subsection, we will cover predictive modeling results and accuracy pertaining to each of the four machine learning models used.

4.1 Regression Outcomes

In applying a stepwise regression using the Akaike Information Criterion (AIC) for variable selection, this stepwise method identified *Brightness*, *Trauma_rate*, and *Depression_Rate* as the optimal predictors.

Table 3: GLM Regression Results (Model 1)

Variable	Coefficient	Standard Error	z-value	P-value
Brightness	-4.929e-06	1.6e-06	-3.073	0.002
Depression Rate	-0.0145	0.010	-1.434	0.152
Trauma Rate	0.0094	0.004	2.133	0.033
Constant (Intercept)	2.4212	0.232	10.458	0.000
				Pseudo R-squared: 0.4100
				Log-Likelihood: -136.48

Table 4: Model 1 Correlation Matrix of Key Variables

Variable	Depression Rate (per 100k)	Trauma Rate (per 100k)	Brightness
Depression Rate (per 100k)	1.00000	0.70551	-0.17183
Trauma Rate (per 100k)	0.70551	1.00000	-0.45973
Brightness	-0.17183	-0.45973	1.00000

Although *Depression_Rate* had a relatively high p-value, we retained it in the model because it is a theoretically significant variable related to mental health and suicide. In addition, we identified a strong correlation (0.705) between *Depression_Rate* and *Trauma_rate*, indicating potential multicollinearity. To resolve this, we included an interaction term between the two variables observed in Model 2 (Table 5).

Based on the regression analysis results, we observed the relationship between age-adjusted death rate and multiple independent variables. In the analysis, we used a Generalized regression model that included depression rate, trauma rate, brightness, and the interaction between depression and trauma as independent variables. The model's R-squared value is 0.4289, indicating that these variables explain approximately 42.89% of the variation

Table 5: GLM Regression Results (Model 2)

Variable	Coefficient	Std. Error	z-value	P-value
Intercept	1.2261	0.920	1.332	0.183
Depression Rate (per 100k)	0.0295	0.034	0.863	0.388
Trauma Rate (per 100k)	0.0273	0.014	1.937	0.053
Brightness	-4.852e-06	1.59e-06	-3.051	0.002
Depression × Trauma Interaction	-0.0006	0.000	-1.342	0.180
Log-Likelihood				-135.55
Pseudo R-squared				0.4289

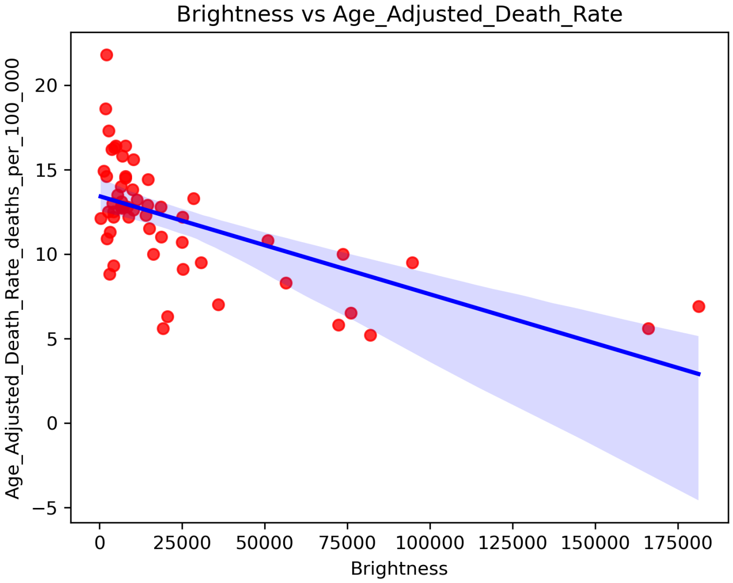

Figure 1: Brightness vs Age Adjusted Death Rate

in death rate. While the model does not fully explain all the variation, the deviance value of 26.284 suggests a good model fit. Specifically, the depression rate was not significantly related to the suicide rate (p-value = 0.388), meaning depression rate did not significantly affect the suicide rate.

The trauma rate had a coefficient of 0.0273, with a p-value of 0.053, indicating that trauma rate may have a positive influence on the death rate, although the effect is only marginally significant. *Brightness (light brightness) showed a significant negative correlation with the death rate (p-value = 0.002), suggesting that higher light brightness is associated with lower death rates.* This result may imply that areas with higher brightness are linked to certain health interventions or socioeconomic factors. The result also might be related to limited data with small scales. Finally, the interaction between depression and trauma did not significantly affect the suicide rate (p-value = 0.180), indicating that this interaction

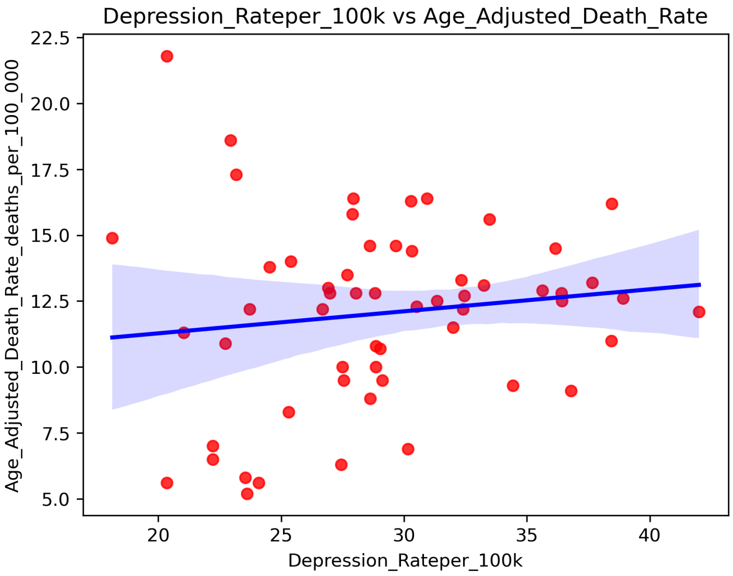

Figure 2: Depression Rateper 100k vs Age Adjusted Death Rate

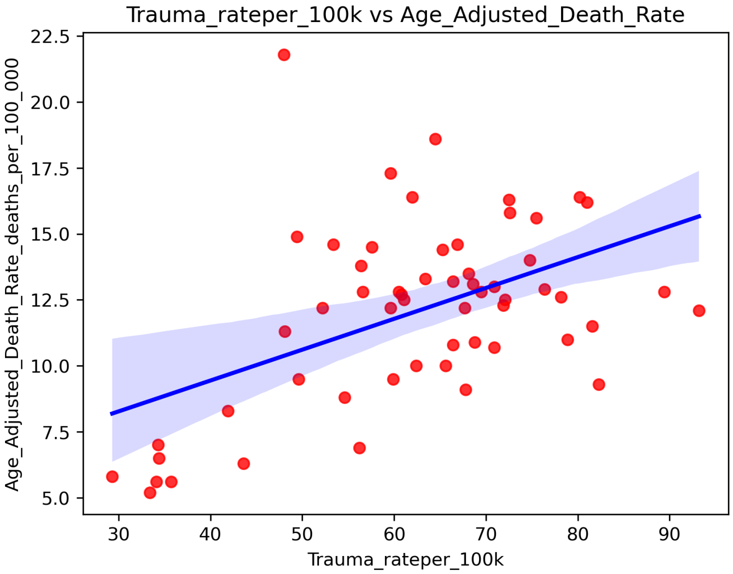

Figure 3: Trauma Rateper 100k vs Age Adjusted Death Rate

does not have a significant impact on suicide rate.

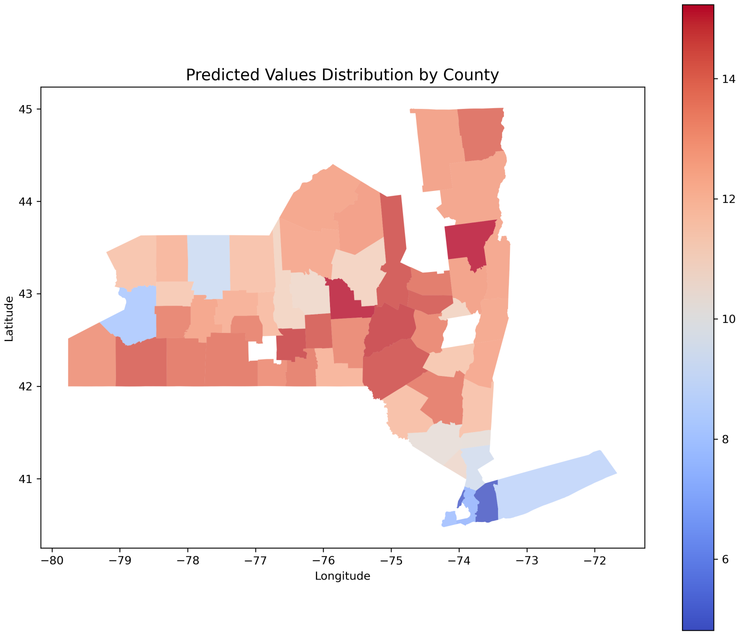

Figure 4: Spatial Predicted Values Distribution

4.2 Predictive Modeling Accuracy

Table 6: Performance comparison of machine learning algorithms

Algorithm	Accuracy	F1-Score
Logistic Regression	0.666667	0.649351
SVM	0.611111	0.541818
Random Forest	0.611111	0.609907
KNN	0.500000	0.498452

The results showed that a standardized logistic regression model was the most accurate, achieving 67% accuracy in predicting whether a county had higher or lower suicide rates. The model also yielded an F1 score of 65%, which reflects a good balance between precision and recall. However, other models performed less effectively. For instance, a Support Vector Machine (SVM) model, designed to account for nonlinearity, achieved 61% accuracy, while a K-Nearest Neighbors (KNN) model performed the lowest with 50% accuracy, essentially equivalent to random guessing.

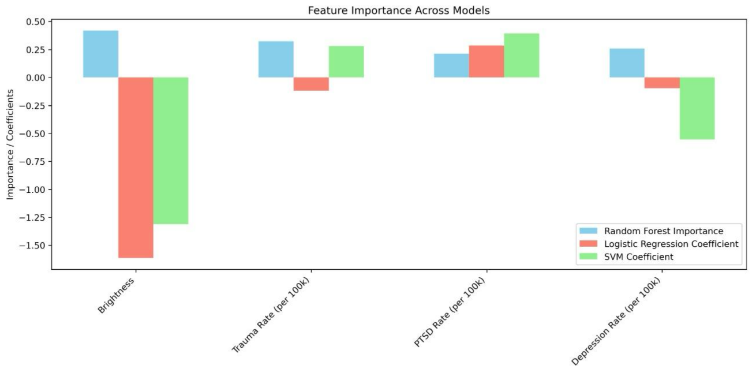

Figure 5: Feature Importance Across Algorithms

To determine which factors—brightness, depression rate, or trauma rate—have the most significant impact on predicting suicide rates per county, we examined feature importance across several models. K-Nearest Neighbors (KNN) model does not provide feature coefficients due to its distance-based algorithm.

Our analysis revealed that the Random Forest model highlights the weight of each coefficient and its associated uncertainty but does not measure the nature of the relationship between features and suicide rates. In contrast, the Logistic Regression and SVM models provided clearer insights into how features influence outcomes.

The results show that ”brightness” has the most significant and negative influence on predicting suicide rates. Specifically, as brightness increases, the likelihood of higher suicide rates tends to decrease. This finding underscores the protective role that increased brightness may play in reducing suicide risks at the county level.

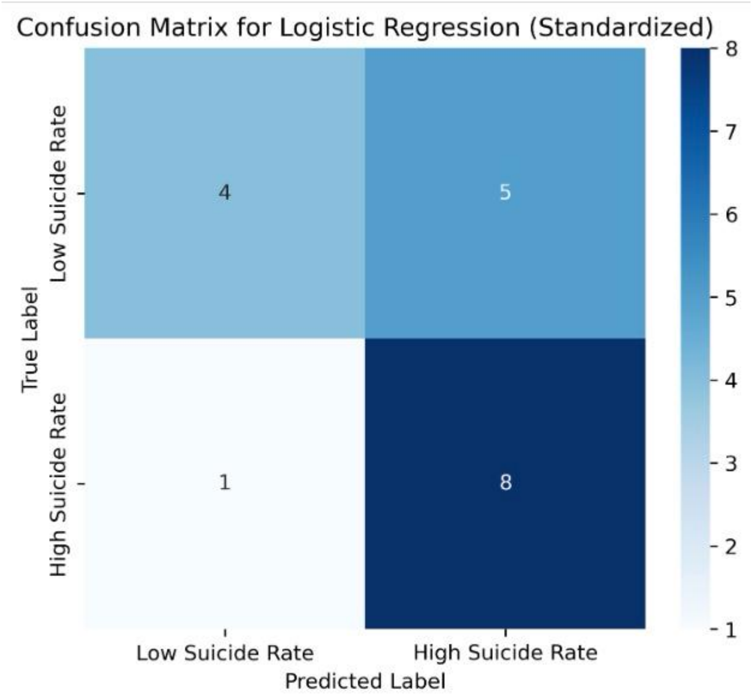

Figure 6: Logistic Regression Accuracy

The confusion matrix further illustrated the model's performance. It identified 8 true positives and 4 true negatives, but also produced 5 false positives and 1 false negative. This indicates that while the model performs well at predicting high suicide rates, it struggles to accurately identify counties with low suicide rates. These findings suggest that while brightness and mental health indicators can moderately predict suicide rates at the county level, the models require further refinement to improve their predictive power, especially for lower suicide rates.

5 Discussion

One of the main limitations of this study is the small sample size. New York State comprises only 62 counties, and due to the lack of data availability, our analysis was conducted on 57 counties. This relatively small dataset reduces the statistical power of our models, limiting our findings. A larger sample size across multiple states or regions would allow for more robust conclusions and better model accuracy. With a limited number of observations, even small variations in the data can disproportionately affect results, potentially contributing to unstable estimates and reduced precision in predicting suicide rates.

Additionally, regional disparities in brightness levels across New York State significantly influenced our findings. New York City, as a major metropolitan area, dominates the higher end of the brightness spectrum due to extensive artificial light brightness, while suburban and rural counties exhibit much lower brightness volumes. This uneven distribution may have masked the true relationship between brightness and suicide rates across different contexts, as the rural regions' low brightness levels may not fully capture the effects of artificial light exposure. The lack of variability in brightness outside of New York City further challenges our results, resulting in weaker correlation to suicide rates and other predictors. Expanding the analysis to include states with more diverse brightness patterns, such as California or Texas, could provide a more balanced perspective.

Lastly, the issue of aggregated indicators and the limited number of control variables constrain the depth of our analysis. The mental health variables (depression, trauma, and PTSD rates) are aggregated at the county level, potentially obscuring variations at the individual level that could better explain the relationship between mental health indicators and suicide rates. Moreover, the absence of additional control variables, such as economic factors, healthcare access, and demographic indicators, may have left out critical confounders factors that potentially affect suicide rates. Future research could benefit from more granular, disaggregated data and the inclusion of additional explanatory variables to build a more comprehensive and accurate model.

References

[1] Carta, M., Preti, A., & Akiskal, H. (2018). Coping with the New Era: Noise and Light Pollution, Hyperactivity and Steroid Hormones. Towards an Evolutionary View of Bipolar Disorders. *Clinical Practice and Epidemiology in Mental Health: CP & EMH*, 14, 33–36. https://doi.org/10.2174/1745017901814010033

[2] NASA VIIRS Land Science Investigator-Led Processing System. (2021). *VIIRS/NPP Lunar BRDF-Adjusted Nighttime Lights Yearly L3 Global 15 arc-second Linear Lat Lon Grid [Dataset]*. NASA Level 1 and Atmosphere Archive and Distribution System Distributed Active Archive Center. `https://doi.org/10.5067/VIIRS/VNP46A4.001`