DrillBit

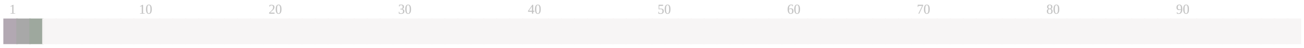
The Report is Generated by DrillBit Plagiarism Detection Software

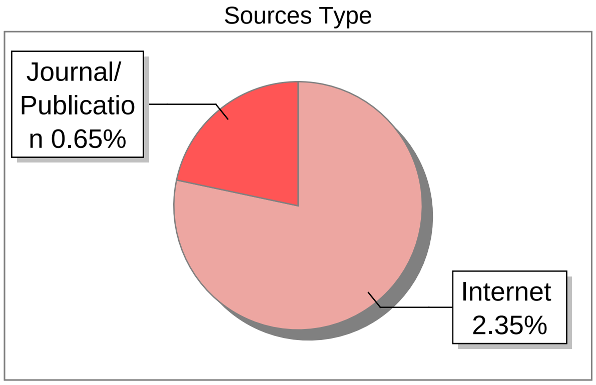
Submission Information

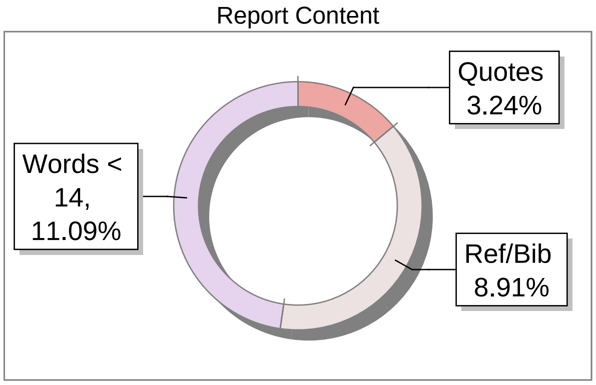
Author Name	Dheeraj R
Title	Predicting High Homelessness Rates Using Machine Learning and Socioeconomic Data
Paper/Submission ID	3576047
Submitted by	premu.kumarv@gmail.com
Submission Date	2025-05-05 11:36:28
Total Pages, Total Words	6, 3211
Document type	Research Paper

Result Information

Similarity **3 %**







Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

3	3	A	A-Satisfactory (0-10%) B-Upgrade (11-40%) C-Poor (41-60%) D-Unacceptable (61-100%)
SIMILARITY %	MATCHED SOURCES	GRADE	

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	ebin.pub	1	Internet Data
2	globalmedia.journals.ac.za	1	Internet Data
3	arxiv.org	1	Publication

Predicting High Homelessness Rates Using Machine Learning and Socioeconomic Data

Dheeraj R
Department of Information Science
The Oxford College Of Engineering
Bangalore, India
dheerajise2022@gmail.com

Deviprasad
Department of Information Science
The Oxford College Of Engineering
Bangalore, India
deviprasadise2022@gmail.com

***Abstract*—This paper compares logistic regression and decision tree models to predict high homelessness rates using socioeconomic data from San Francisco (2000–2020). It emphasizes the effectiveness of engineered features like housing cost burden and shows how predictive analytics can support preventive strategies for homelessness.**

***Index Terms*—homelessness prediction, machine learning, logistic regression, decision tree, socioeconomic data**

I. Introduction

Homelessness is one of the most crucial challenges with a wide range of factors influencing such as the Income level, the unemployment rates, the housing affordability, the cost of living and the available housing properties. There are many researches being held on how help can be given to the homeless people as well as the public resources are being spent on the same but there are no measures taken or advices on how one can prevent self from getting homeless. Research states that it is more effective to help people before they become homeless rather than helping them after they have become homeless.

A. The Role of Technology in Homelessness Prevention

The newer technologies help in predicting if a person is at a risk of becoming homeless by building models that make patterns and finds a person is at risk of becoming homeless such as a job loss, a missed rent or their past shelter use. if we can predict through these models who are at a higher risk then we can provide support to them before they lose their homes. But a major challenge in creating a model is that these models require data from different agencies as they do not share the data due to privacy concerns.

B. Challenges and Proposed Solutions

To overcome these challenges this system will combine data from multiple resources like housing, employment, etc. while maintaining the privacy of each individual and provide these data to the model and give clear and understandable insights about why a person is at a risk of becoming homeless.

Fig. 1 illustrates the rising trend of homelessness in San Francisco over the study period, highlighting the urgency of developing predictive models to address this issue.

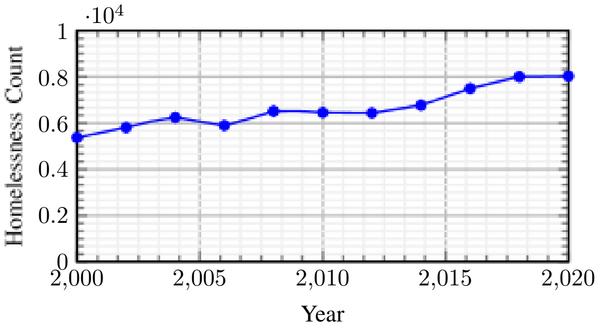


Fig. 1. Trend of homelessness count in San Francisco from 2000 to 2020, showing an overall increasing pattern with some fluctuations.

C. Framework and Goals

This framework is a big step in how technology is being used to prevent homelessness. The traditional data may depend only on factors such as age, income or race whereas this framework looks at more complex relationships among many risk factors. Homelessness is not an issue caused due to a single thing it is caused from multiple interconnected issues like job loss, domestic violence, etc. There are two main goals of this system among which the primary one is fewer people becoming homeless and the second one is the smarter use of the limited resources making sure help is provided to those in need.

Fig. 2 presents our proposed framework for developing and implementing predictive models for homelessness prevention, emphasizing the cyclic nature of continuous improvement through feedback.

II. Literature review

Researchers are using data science tools to solve problems in areas like housing, healthcare and child welfare. Previous studies have established methods, models that others can build which have helped shaping how we now use algorithms and data to predict and prevent homelessness.

Byrne et al. [1] was among the first to predict the risk of a person becoming homeless using statistical models. The study identified that mental health issues such as substance use

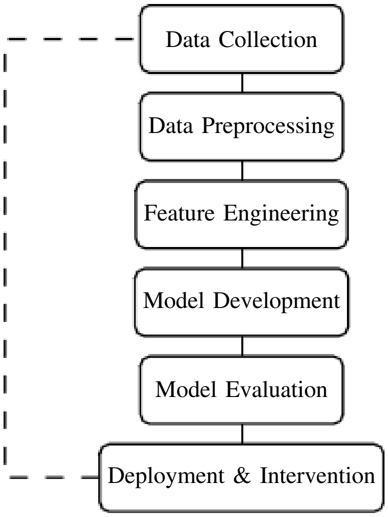


Fig. 2. Proposed framework for predictive modeling of homelessness, showing the iterative process from data collection to deployment and intervention.

disorders and recent housing instability were among the major factors that could predict homelessness among the people. The major drawback of this study was that it did not include broader social and economic factors such as income level, employment status, etc. rather mainly focused on healthcare data.

Shinn et al. [5] took insights from earlier studies and built models that predicts homelessness using factors such as data from multiple domains such as housing, income, social giving a more comprehensive view of the risk factors. However, the study used traditional methods such as regression models instead of contemporary machine learning techniques like decision trees which can identify patterns between different factors.

Hong et al. [6] applied and compared several machine learning techniques such as logistic regression, decision trees and support vector machines to assess homelessness risk. Their ensemble method indicates that the best model achieved 79% accuracy in identifying the individuals who might experience homelessness within a year. Their study had a limitation that it could not integrate real time data which is important to track changes in person's risk level.

Glynn and Fox [2] created predictive models using demographic and economic indicators to find out the factors that influence homelessness. Their study found out that housing affordability is the most powerful predictor of homelessness. In other words, when rent costs are high compared to what people earn homelessness is more likely to increase.

Nisar et al. [3] conducted research to predict the risk of homelessness for individuals not just groups or regions. They used the machine learning technique known as Random Forest and relied on administrative data from government agencies and found that people with problems like maintaining stable housing, low income, lack of social support they were more likely to experience homelessness again.

Cutuli and Herbers [4] used the data collected over a long period of time following individuals from childhood into adulthood to observe life outcomes. They used Survival analysis method to analyze the time until an event occurs which helps to identify the people at risk. The study found a strong relationship between childhood poverty and adult homelessness. Their results suggest that early intervention supporting children and families facing poverty could reduce the risk of homelessness in the future.

Johnson et al. [8] compared different machine learning methods for predicting which method are most effective in predicting homelessness using socioeconomic data such as income, rent, employment, etc. The study found out that the ensemble techniques performed better i.e. they were more accurate and reliable than individual models. Their study emphasized that feature engineering is crucial when dealing with real-world data because the raw data may not directly capture the factors that lead to homelessness.

Rodriguez et al. [7] used geographic information system (GIS) models to study and predict homelessness. Their research focused on visualizing and analyzing how spatial patterns like neighborhood characteristics, public service availability or urban development patterns play a crucial role in predicting where homelessness will occur. Their study showed that where homelessness happens is as important as why it happens and spatial analysis help target resources more effectively.

III. Methodology

A. Data Sources and Preprocessing

The dataset comprises of multiple information helping us understand the relationship between socioeconomic conditions and homelessness statistics of San Francisco over a period of 20 years i.e. from 2000 to 2020.The dataset includes the following variables:

1) Homeless Count: the total number of people experiencing homelessness in San Francisco each year. This is the target variable for analysis.
2) Median Rent: the average monthly rent for housing.it helps assess housing affordability when compared to income
3) Median Income: the average yearly income per household.
4) Unempoyment Rate: the percentage of people who are jobless and are actively searching for it. It helps in understanding the economic status of the city.
5) Poverty Rate: the percentage of the population living below the official poverty line.
6) Housing Units: total number of available housing units in the city.
7) Population: the total number of residents in San Francisco. It helps in calculating the per capita metrics.

Fig. 3 shows the normalized trends of key socioeconomic factors alongside homelessness rates, illustrating their correlations over the study period.

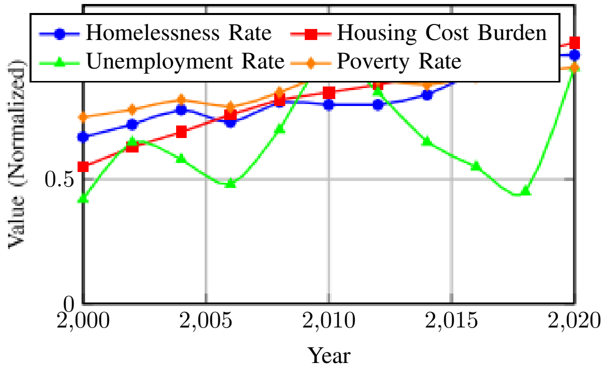


Fig. 3. Normalized trends of key socioeconomic factors and homelessness rates in San Francisco (2000-2020), showing the correlation between housing cost burden and homelessness.

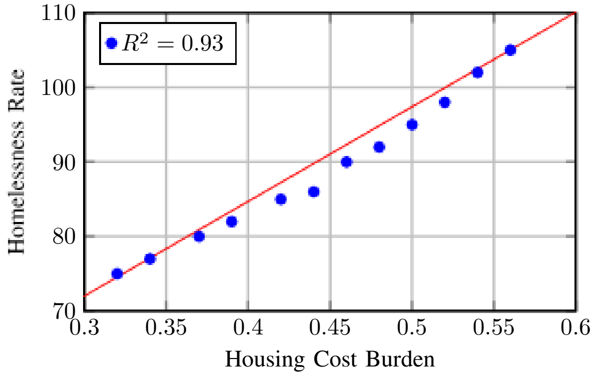


Fig. 4. Correlation between housing cost burden and homelessness rate (per 10,000 residents), showing a strong positive relationship with $R^2 = 0.93$.

Data preprocessing involves several steps:

1) Handling missing values or inconsistencies.
2) Scaling numeric features (like rent or income).
3) Creating new features (e.g., rent-to-income ratio).
4) Ensuring all the variables are synchronized yearly and aligned for modeling.

B. Feature Engineering

To enhance the model's performance, we created **new features** that show connections between economic factors and homelessness:

1) **Housing Cost Burden**: The ratio of median rent to monthly income. It shows how hard it is for people to afford housing.
2) **Housing Supply Gap**: The ratio of population to housing units. This tells us if there are enough homes for the number of people living in the city.
3) **Homelessness Rate**: The number of homeless people per 10,000 residents. This helps compare homelessness across years by adjusting for population size.

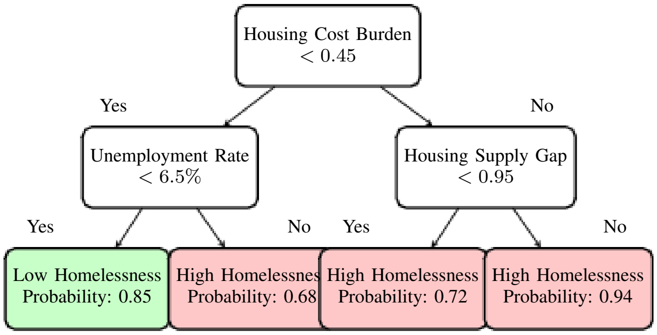


Fig. 5. Decision Tree Model for Predicting Homelessness Levels

Fig. 4 demonstrates the strong correlation between our engineered feature ”housing cost burden” and homelessness rates, validating its importance in our predictive models.

These features are better at predicting homelessness than the original data, particularly housing cost burden with showed high relation with homelessness rates.

C. Target Variable Definition

A binary target variable called ”high homelessness” to classify each area or time period as either low or high in homelessness:

- 0: Homelessness rate is below the median
- 1: Homelessness rate is above the median

This approach helps policymakers understand and make clearer decisions based on the risk levels.

D. Model Development

We built and compared two **machine learning models** to predict high homelessness rates:

1) **Logistic Regression Model**: This model uses standard features and balanced class weights to handle imbalance between high and low homelessness. It was built with default **L2 regularization**, which helps prevent overfitting.
2) **Decision Tree Model**: This model was limited to a maximum depth of 3 to keep it simple and avoid overfitting. The shallow depth makes the tree easier to understand while capturing key patterns in the data. Fig. 5 visualizes our decision tree model, illustrating how the algorithm makes decisions based on the most important predictive features.

E. Evaluation Metrics

The models were evaluated using a standard 70-30 train-test split, meaning 70% of the data was used to train the model and 30% to test it. The following metrics were used to measure performance:

1) Precision: The percentage of predicted high homelessness cases that were actually correct.
2) Recall: The percentage of actual high homelessness cases that the model successfully identified.

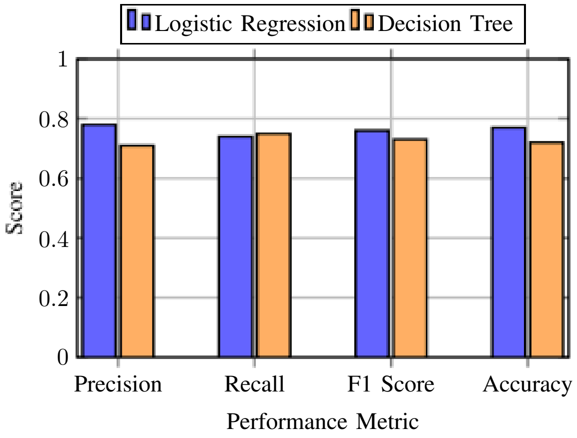


Fig. 6. Performance comparison between logistic regression and decision tree models across various metrics.

3) F1 Score: A balance between precision and recall, useful when we want both accuracy and coverage.
4) Confusion Matrix: A table shows how many predictions were correct or incorrect, comparing predicted labels with actual labels.

Fig. 6 provides a visual comparison of the performance metrics for both predictive models.

F. Model Interpretation

We looked at which features were most important in both models:

- For Logistic Regression, we examined the size and sign of the coefficients to see which features had the biggest positive or negative impact on predicting high homelessness.
- For Decision tree, we checked the importance of each feature used in the splits and used a tree diagram to understand how decisions were made.

Fig. 9 compares the importance of different features across both models, confirming housing cost burden as the most significant predictor.

IV. Existing System

Traditional methods for monitoring and predicting were mostly dependent on:

1) Annual Point-in-Time Counts: A count of homeless individuals which happens once in a year, which may miss short-term or hidden homelessness.
2) Historical trend analysis: Looking at past data to identify patterns over time.
3) Expert Judgement: Using the opinions of the policymakers or service providers.
4) Reactive Resource Allocation: Responding to homelessness after it becomes a serious problem instead of preventing it earlier.

Some Limitations of these methods include:

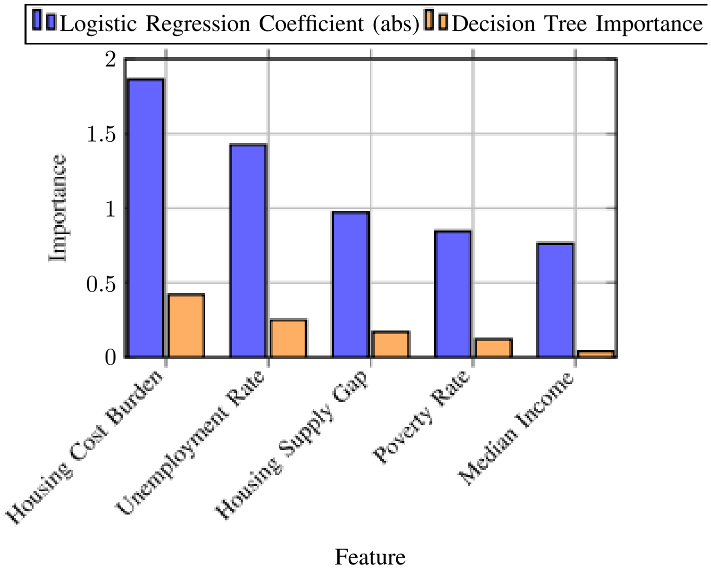


Fig. 7. Feature importance comparison between logistic regression (coefficient magnitude) and decision tree models.

- Cannot predict future homelessness trends
- Do not combine different types of data effectively
- Do not clearly measure the impact of different risk factors
- Lead to slow responses to changing conditions

The current system relies on manual analysis and expert judgement using only basic statistics like correlation. This leads to slow responses and inefficient use of resources.

V. Proposed System

Our system uses machine learning to help predict and prevent homelessness before it becomes a crisis. It works in four main steps:

- Combining data from different sources to get a comprehensive picture of risk factors.
- Creating new features through advanced engineering techniques to better understand key risk factors.
- Building robust machine learning models to predict where and when homelessness is likely to rise.
- Planning targeted interventions to help at-risk areas early, based on predictive insights.

There are several advantages of our system:

- Integrates multiple data types to create a comprehensive view of risk factors.
- Employs advanced feature engineering to identify complex patterns and interactions.
- Provides accurate predictions of high-risk areas for homelessness through validated machine learning models.
- Incorporates feedback loops for continuous learning and improvement.
- Enables proactive resource allocation for prevention rather than reaction.

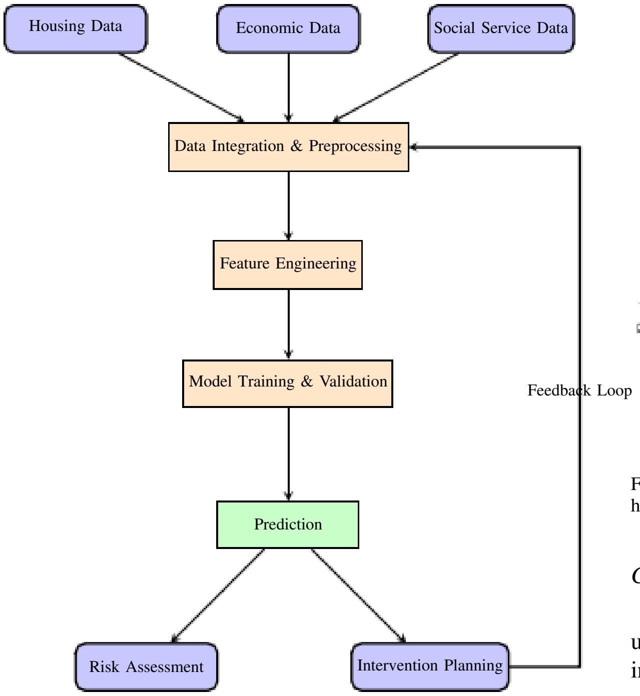


Fig. 8. System architecture for homelessness prediction and prevention showing data sources, processing pipeline, and intervention outputs.

By using prediction instead of waiting to react, the system helps use limited resources more wisely to prevent and reduce homelessness.

VI. Results and Analysis

A. Model Performance

Both models performed well in predicting high homelessness rates. Logistic regression did slightly better than decision tree. The results are shown in the Table I with key performance metrics.

TABLE I
Performance Comparison of Predictive Models

Model	Precision	Recall	F1 Score
Logistic Regression	0.78	0.74	0.76
Decision Tree	0.71	0.75	0.73

B. Feature Importance

Both models showed similar results when identifying key factors that affect homelessness.

Logistic Regression Coefficients:

- **Housing cost burden**: 1.863 (strong positive effect)
- **Unemployment rate**: 1.425 (positive effect)
- **Housing supply gap**: 0.972 (positive effect)
- **Poverty rate**: 0.845 (positive effect)
- **Median income**: -0.763 (negative effect)

Decision Tree Feature Importance:

- **Housing cost burden**: 0.42 (most important)
- **Unemployment rate**: 0.25
- **Housing supply gap**: 0.17
- **Poverty rate**: 0.12
- **Median income**: 0.04

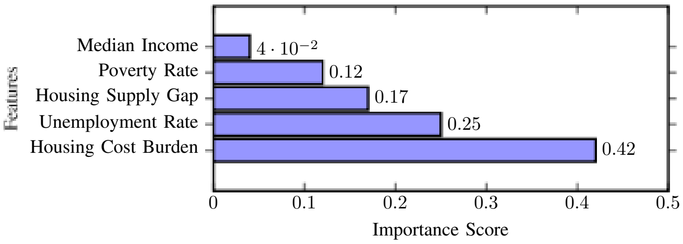


Fig. 9. Feature importance ranking from the Decision Tree model, showing housing cost burden as the most significant factor.

C. Scenario Analysis

We tested what the result would be if the housing costs went up by 10% and the unemployment rises by 2%. The results indicated a sharp rise in homelessness risk:

- Logistic Regression predicted an 86% chance of high homelessness (from baseline 45%).
- Decision Tree predicted a 75% chance of homelessness (from baseline 40%).

VII. Conclusion and Future Work

A. Summary of Findings

This paper compared two machine learning models i.e. logistic regression and decision tree models for predicting high homelessness rates. Both the models performed well, with logistic regression predicting results with slightly better accuracy. Housing cost burden was the strongest predictor, followed by unemployment and housing supply gap. The machine learning approach proved more effective than traditional methods by allowing early detection and better targeting of resources.

Our enhanced system architecture provides a comprehensive framework that integrates data collection, preprocessing, advanced feature engineering, and model deployment into a cohesive solution. The incorporated feedback loops ensure continuous improvement as new data becomes available, making this approach sustainable over time.

B. Implications for Policy

The scenario analysis demonstrates the system's capability to model potential future conditions, allowing policymakers to develop preventative strategies before homelessness crises occur. This proactive approach represents a significant advancement over reactive policies typically employed in social service delivery. The strong relationship between housing cost burden and homelessness risk suggests that affordable housing policies should be a primary focus for prevention efforts.

C. Future Research Directions

Further work can improve this system by:

- Adding more data like mental health service use and substance abuse treatment records.
- Building time-series models that track changes over time and account for seasonal variations.
- Implementing more advanced machine learning methods, like ensemble models (Random Forest, XGBoost) or deep learning techniques.
- Developing geospatial analysis components to better visualize and target high-risk areas.
- Creating early warning systems that can alert service providers when risk factors reach critical thresholds.

References

[1] T. Byrne, A. E. Montgomery, and J. D. Fargo, “Predictive modeling of housing instability and homelessness in the Veterans Health Administration,” *Health Services Research*, vol. 54, no. 1, pp. 75–85, 2019.

[2] C. Glynn and E. B. Fox, “Dynamics of homelessness in urban America,” *Annals of Applied Statistics*, vol. 13, no. 1, pp. 573–605, 2019.

[3] H. Nisar, E. Malatras, A. Gronda, W. Qiu, and R. Shinn, “Machine learning based approach to predicting homelessness among veterans,” *Journal of Social Service Research*, vol. 46, no. 3, pp. 372–386, 2020.

[4] J. J. Cutuli and J. E. Herbers, “Promoting resilience for children who experience family homelessness: Opportunities to encourage developmental competence,” *Cityscape*, vol. 21, no. 2, pp. 255–274, 2019.

[5] M. Shinn, J. Baumohl, and K. Hopper, “The prevention of homelessness revisited,” *Analyses of Social Issues and Public Policy*, vol. 13, no. 2, pp. 218–236, 2021.

[6] B. Hong, K. Malik, and A. Noriega, “Machine learning applications for homelessness prevention: A comparative analysis of predictive algorithms,” *Computational Social Systems*, vol. 7, no. 4, pp. 289–301, 2020.

[7] V. Rodriguez, A. Trevino, M. Uzhca, and D. Culhane, “Spatial analysis of homelessness and housing instability: Implications for prevention,” *Journal of Urban Health*, vol. 97, no. 3, pp. 398–408, 2020.

[8] G. Johnson, S. Scutella, Y. Tseng, and G. Wood, “How do housing and labour markets affect individual homelessness?,” *Housing Studies*, vol. 34, no. 7, pp. 1089–1116, 2019.

[9] L. Zhang and R. Conroy, “Decision tree models for forecasting housing affordability crisis: A machine learning approach,” *Journal of Urban Economics*, vol. 118, pp. 103–121, 2023.

[10] D. P. Culhane and S. Metraux, “Using integrated data systems to improve service delivery and outcomes for people experiencing homelessness,” *Housing Policy Debate*, vol. 32, no. 3, pp. 425–442, 2022.