IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 May 2024, accepted 2 August 2024, date of publication 7 August 2024, date of current version 2 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3439721

RESEARCH ARTICLE

A Hybrid Machine Learning and Regression Approach for Validating a Multi-Dimensional Crime Index in the Context of Crime Against Women

POONAM K. SARAVAG AND B. RUSHI KUMAR
Department of Mathematics, School of Advanced Sciences, Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India
Corresponding author: B. Rushi Kumar (rushikumar@vit.ac.in)

This work was supported by the Vellore Institute of Technology, Vellore, Tamil Nadu, India.

ABSTRACT Violence targeting women has endured since ancient times, encompassing a spectrum of offenses ranging from psychological anguish to physical and sexual assault. This study introduces a crime index rooted in diverse categories that directly or indirectly contribute to fostering criminal intentions. A composite weighted index, comprising four sub-indexes focusing on Health, Socioeconomic status, Education, and Judiciary, was created. The stability and homogeneity of the index were assessed using reliability testing. Validation of the proposed index was carried out through comparative analysis of baseline and ensembled models. The hybrid model was proposed by combining multiple linear regression and robust regression techniques with random forest and stochastic gradient descent as meta-regressors. The models were assessed by the evaluation metrics MAE, RMSE, and MAPE. The findings indicate that the Index demonstrates strong reliability, supported by a significantly high correlation. The ensemble hybrid model approach, effectively captures the variance of the model, with less error when compared with the baseline models but the homogeneous ensemble approach proved to be the best with minimum error. Women in less developed regions with extreme geographical conditions are at a higher risk of falling prey to victimization. The Index and its statistics conclude that social factors significantly contribute to the occurrence of violence against women. These findings hold the potential for informing enhanced strategies to curtail the menace.

INDEX TERMS Crime index, ensemble hybrid model, robust regression, reliability, validation.

I. INTRODUCTION

This Crimes, or criminality in general, have accompanied human civilizations since their inception. Criminal behavior is described as aberrant human conduct that transgresses the laws in effect in a specific location. The term ''criminality'' refers to the overall amount of crimes committed over a specified period of time in a particular nation or environment, or by a particular gang of criminals. Within the context of urban governance, crime poses a notable challenge in ensuring the security of residents, alongside challenges of a more scientific nature. Ensuring safety is a fundamental necessity, greatly influencing both the well-being of the locality's inhabitants and its overall reputation and economic appeal. Through the identification and resolution of the underlying factors behind the criminal activity, it becomes possible to devise proactive and corrective strategies, effectively managing, though not entirely eliminating, crime [1], [2]. The global worry about criminal activities is not unique to India. Throughout history, it has posed a substantial challenge to the progress of societies. In recent times, both the scale and prominence of this problem have undergone a substantial increase.

A. CRIME AGAINST WOMEN (CAW)

CAW or gender-related violence pertains to any detrimental or unlawful acts or conduct directed specifically at females

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Li.

due to their gender. CAW encompasses sexual, psychological, emotional, and financial abuse in addition to physical assault. Discrimination, domestic abuse, sexual assault, and other types of gender-based violence affect women disproportionately [3], [4]. Extreme acts of violence against women, such as acid assaults and honor killings, do occur at times. Beyond the harm done to the victim's physical and mental health, these crimes have a significant negative economic and social impact on the country. Patriarchal views and a desire to keep control over women frequently drive these crimes. The Indian authorities have implemented various measures to address the issue of violence against women. These include the enactment of the Protection of Women from Domestic Violence Act in 2005, the passage of the Criminal Law Amendment Act in 2013, and the establishment of the National Commission for Women. Despite these efforts leading to some positive changes in our community, there is no 100% success in preventing such incidents from happening in our vicinity. A complex issue requiring a deep understanding of its root causes is evident. India's ranking on the World Peace and Security Index declined from 133rd among 167 countries in 2019–20 to 148th among 170 countries in 2020–21. This highlights the status of women in the country compared to the global context. While crime may not be uniformly distributed across the nation, it is essential to pinpoint the fundamental factor that is disproportionately affecting specific regions [5]. The occurrence of CAW is impacted by a range of elements. The country's limited rate of successful prosecutions enables wrongdoers to evade consequences and engage in the same actions again in the future. This study suggests introducing a CAW index as a novel method for better understanding the incidence of such crimes [6]. This holds special significance for the government as it supervises state law enforcement to ensure the well-being of its citizens. The purpose is to identify trends in unlawful conduct and assist law enforcement agencies in distributing their assets effectively. The index is constructed by leveraging different elements that are highly probable to impact criminal attitudes toward women.

B. CAW IN RAJASTHAN

The northern Indian state of Rajasthan is the subject of the investigation, which holds the status of being the nation's eighth-most populous state and the largest in terms of land area in India. The sex ratio and literacy rate are 928 and 66.11%, with male and female literacy rates as 79.19% and 52.12%, respectively, marking the lowest female literacy rate in the country according to the 2011 census [7]. The sex ratio is lower than the country's average, 943. In recent years, there has been a troubling surge in the occurrence of offenses targeting women within the state. The statistics from the National Crime Records Bureau (NCRB) reveal that Rajasthan secured sixth rank in 2000 and 2005, fourth rank in 2010 and 2015, second position in 2019, third in 2020, and again second in 2021. Also, it recorded the highest count of rape cases in the year 2021. With the state being eighth in population appearing as the second highest in CAW cases becomes a concern of the study.

C. CRIME AGAINST WOMEN INDEX

There does not exist a specific ''Crime Against Women Index'' recognized at a global level. However, various organizations and countries compile statistics and indexes related to crimes against women, particularly focusing on issues like violence, harassment, and discrimination. Countries may also have their national indexes or reports that track crimes against women, such as India's National Crime Records Bureau (NCRB) which releases an annual report on crime statistics, including crimes against women.

It's worth acknowledging that constructing a relevant and comprehensive index necessitates meticulous deliberation on data sources, methodology, and cultural context to effectively capture the intricate nature of crimes perpetrated against women [8]. The implementation of such an index has the potential to make a substantial impact in terms of increasing societal awareness, catalyzing legislative reforms, and eventually striving toward the mitigation of violence and the promotion of the safety and welfare of women within the community.

In the process of mitigating and preparing for catastrophes, it is vital for state and local authorities to identify populations that are socially vulnerable. This identification is crucial in order to allocate more support to the people of these communities during the occurrence of a disaster. Despite being the most suitable entities to identify vulnerable areas, local authorities often face challenges such as inadequate funding, insufficient staffing, and overwhelming obligations in the realms of health and social services. In contrast, state and national agencies, although having enough staffing and funding, may encounter deficiencies in their resource allocation mechanisms.

Substance use, particularly alcohol, and drugs, can impair judgment, alter behavior, and reduce inhibitions. This can lead to an increased likelihood of aggressive or violent behavior, including sexual harassment, assault, domestic violence, and other crimes against women. Perpetrators under the influence of substances may be more likely to engage in criminal activities, often involving women as victims [9], [10].

Despite significant advancements in recent decades, the substantial and enduring challenge of infant mortality continues to exist on a global scale. Throughout the world, there has been a consistent reduction in infant mortality rates for over 30 years. Nevertheless, nearly four million children lose their lives before reaching their first birthday, primarily due to preventable reasons. In societies where female lives are undervalued due to high infant mortality rates, law enforcement, and justice systems might be less likely to take crimes against women seriously. This could result in a lack of adequate response to violence and abuse, emboldening perpetrators and contributing to a culture of impunity.

The female infant mortality rate can cause significant gender imbalance which leads to violence against women [11].

Women with disabilities often face unique challenges and vulnerabilities, making them more susceptible to various forms of violence and abuse. They tend to have limited access to resources, barriers to reporting, stigmatization, discrimination, legal and policy gaps, etc. [12]. Higher female work participation rates are generally associated with increased economic independence and empowerment for women. This, in turn, can contribute to a reduction in CAW as economic empowerment may lead to improved social status and a stronger voice against various forms of violence and discrimination. Conversely, areas with lower female work participation rates witness higher incidences of CAW due to factors such as financial dependence, lack of agency, and limited opportunities to escape abusive situations [13]. A low conviction rate in cases of CAW is a concerning issue that reflects the challenges and shortcomings in the criminal justice system's response to such crimes [14]. The crime-related economic theory proposes that factors influencing deterrence play a crucial role in shaping crime rates within a community. Viewing criminals through an economic lens, they are seen as logical individuals who assess the chances of getting caught and facing legal consequences before engaging in criminal activities. When the probability of getting arrested and caught rises, it acts as a deterrent for criminals, as it elevates the anticipated expenses associated with committing offenses [15], [16].

II. RELATED WORK

Research conducted in the area of CAW has revealed that these crimes are often under-reported and that victims encounter substantial obstacles in pursuing legal recourse. Also, these crimes have significant economic and social impacts on the entire nation. Therefore, it is essential to understand the role of various factors that influence or control these activities.

Kierepka [17] introduced an indicator to figure out where and how frequently a crime incident may happen annually in Wroclaw City. The cartographic studies of the aggregate hazard of reported offenses have been carried out, and the spatial distribution of the probability of crime has been examined in relation to the Land Use Conditions and Directions for Wroclaw. Kwan et al. [18] examined and assessed the Hong Kong crime index by comparing the relative severity of fifteen crime categories using Thurstone's scale and calculating each crime category's weight. They then devised a weighted index of crime intensity based on a time series approach. Chaudhari et al. [19] constructed an economic model to explain the fluctuations in crime rates using stochastic frontier analysis. Their analysis makes it possible to identify the error term in a predictable component under a predetermined set of hypotheses. In the case of a production function, the deterministic term, known as technical inefficiency, indicates the separation of the recorded crime rate and the frontier. This distance represents the proportion of unreported offenses. In India, their study of a crime index including murder, attempt to murder, homicide, rape, kidnapping and abduction, armed robbery, and robbery revealed an average under-reporting rate of 27%. Nau et al. [20] analyzed the correlation between major crime categories and the accuracy of AGS indexes in determining if crime rates for 1069 tracts in the jurisdiction of the Los Angeles Police Department from 2010 to 2014 would be above or below the median and in the highest or lowest quartile. He concluded that the personal crime index is the more reliable indicator of urban crime. Also found that five of the ten AGS indexes had a moderately strong correlation with LAPD crime. The c-statistics for robbery, homicide, aggravated assault, motor vehicle larceny, and personal crime ranged from 0.81 to 0.90 in unadjusted regressions and increased by up to 0.13 points when ACS variables were included. Ledingham et al. [21] researched and found that women with disabilities reported a higher prevalence of disclosing instances of sexual assault throughout their lives compared to non-disabled women, with those experiencing multiple impairments being at the greatest risk. A notable disparity exists between women with cognitive or multiple impairments and non-disabled women regarding their exposure to physical or non-physical coercion during their initial sexual experience [22].

Tate [23] developed a social vulnerability indicator and conducted an analysis to gauge the reliability of the indicator's rankings when different setups were considered. It was determined that the hierarchical approach was the most accurate, whereas the inductive method was the most precise. Additionally, a sensitivity analysis was carried out to identify the key factors impacting the stability of ranking results. This analysis revealed that the deductive index rankings were most affected by the choice of transformation method, hierarchical models were sensitive to the weighting scheme, and inductive indexes were influenced by the choice of indicators and the scale of analysis. Neupane et al. [24] developed a sustainable employability index for older workers by conducting a survey in 2016 among postal employees aged 50 or older. They followed up in 2018 and used the multi-variable log-binomial regression to calculate the employability index. They also calculated the area under the curve (AUC) to gauge how well the index could differentiate between different groups. The likelihood of having continued workability increased as the SE index quintiles rose. The SE index effectively distinguished between employees who maintained workability over two years. This scoring approach has the potential to assess the future employability of middle-aged postal workers.

A multifactor index was created through multilevel factor analysis using Mplus software. This index identified distinct and unrelated factors, along with their corresponding weights, within a nested dataset comprising data from 1096 individuals across 19 villages in nine domains. The study's findings indicated that higher levels of education, income, and

occupational status were linked to improved quality of life among individuals in this tribal population, both at the household and village levels. These factors enabled them to live healthier lives and allocate resources for leisure activities [25].

After exploring the review of existing literature, we found that indexes have been developed for the measurement of various fields like human development, economy, sustainability, and crime at different geographical locations. However, the Crime Against Women index, particularly for India, remains undiscovered. The main objectives of the present study are:

- To develop a Crime against women index for the states of India and the districts of Rajasthan state
- To check the reliability of the Index
- To employ an ensemble hybrid technique for validation
- Highlighting the factors hindering the safety of women

III. METHODOLOGY

A. DATA SOURCES AND PROCESSING

The issue of CAW is a topic of great sensitivity, hence necessitating the use of trustworthy and accurate statistics in order to make informed assessments and conclusions. To uphold precision, the majority of the data pertaining to the variables under consideration has been sourced from the 2011 census of India, the official website of the National Crime Records Bureau (NCRB), as well as several state and district websites. Given that the official publication of the population census for the year 2021 is now underway and not yet available on the government's website, the research relied on predicted statistics for the aforementioned year. Though the data regarding some variables was available on various government websites, it was employed as actual data rather than predicted data in that particular instance. Data was gathered for all the states and union territories of India, as well as for the districts of Rajasthan, distinctly. The data was categorized into four distinct typologies: Health, Socio-Economic, Educational, and Judicial, each of which depends on various significant factors. In the Health category, aspects such as women with disabilities, the mortality rate of female infants, and alcohol consumption rate were taken into account. The Socio-Economic segment was developed using indicators such as per-capita income, women's workforce participation rate, unemployment rate, and urbanization level. The Education category utilized the Education Index, a component of the Human Development Index, as a basis for its assessment. The Judicial parameter considered factors like conviction rates, the number of filed charge sheets, and the ratio of police personnel per one hundred thousand population.

For India, exponential projection was employed to estimate figures for the disabled female population, substance use, and female infant mortality rate. Data regarding the unemployment rate, urbanization, per-capita income, and female work participation rate was accessible for the year 2021. However, the Education Index for 2021 was not accessible, therefore it was also forecasted using the exponential projection method [26]. As for the Charge Sheet Rate, Conviction rate, and Number of police personnel per one lakh population, the government provided actual data through the NCRB website, hence these figures were utilized as-is.

Similarly, the exponential projection was employed to estimate various factors for the state of Rajasthan, including the disabled female population, substance usage, female infant mortality rate, urbanization, and per-capita income. While data for the unemployment rate and female workforce participation rate existed for the year 2021, the Education Index wasn't accessible for the same year. Consequently, the Education Index was also predicted using the exponential projection method. On the other hand, the Charge Sheet Rate, Conviction rate, and the Number of police personnel per one lakh population were directly obtained from the government's NCRB website, and thus, real data was utilized for these variables.

Because Telangana had not been established in 2011, there was no data available from that year for projecting into 2021. Consequently, Telangana wasn't taken into account when creating the Index; instead, the entire state of Andhra Pradesh was considered as a single entity. Furthermore, although Jammu and Kashmir was a state in 2011 and later became a Union territory in 2019, it was treated as a state for consistency when calculating figures for 2021.

The data was then cleaned and prepared for further analysis.

Examining the count of crimes without considering the population can result in unclear outcomes. Population size significantly influences the frequency of events. Thus the ratio of the number of crimes occurring in a particular area to the population gives the rate of criminal activities. The crime rate serves as a metric for gauging unlawful incidents occurring within a specific region during a particular time frame, presenting a representation of criminal occurrences relative to the population size. The crime rate of CAW is calculated by taking the ratio of crime cases to the population (Eqn. (1)).

$$Crime\ Rate\ (CR) = \frac{n}{p} \tag{1}$$

where n is the number of CAW cases occurring in a particular area and p is the total population of that place.

Graphs were generated to visually depict the rate of CAW cases across all states in India and the districts within the Rajasthan state. This graphical representation of the rate of crime aimed to offer a clearer understanding compared to raw numerical data.

B. CRIME AGAINST WOMEN INDEX (CAWI)

The purpose of the crime index is to comprehend criminal behavior beyond the count of reported occurrences, considering that population density and incident numbers can differ significantly across various regions. The index is divided into four sub-indexes: Health Index (HI), Socio-Economic Index (ESI), Educational Index (EI), and Judicial

Index (*JI*) [27] (Fig.1) The dimension index was employed to normalize the values of each factor. The dimension index was calculated by taking the ratio of the difference between the actual number of cases and the minimum cases to the difference between the maximum and the minimum cases (Eqn. (2)) [28].

$$Dimension\ Index\ (DI) = \frac{A_v - M_i}{M_x - M_i} \tag{2}$$

where A_v is the actual observation, M_i is the minimum value and M_x is the maximum value.

The composite Index of CAW was developed by combining the four sub-indexes [29].

The health sub-index was calculated by averaging the disabled female population (df), female infant mortality rate (imr), and substance use (su) indicators. The socio-economic sub-index was derived by averaging the unemployed population (un), urbanization rate (ur), per capita income (pci), and female work participation rate (fwpr). The education sub-index was determined by considering the education index (ei) from the human development index. The judicial sub-index was calculated by averaging the conviction rate (cr), average charge sheet rate (csr), and the number of police personnel per one lakh population (ppl) (Table 1).

TABLE 1. Sub-indexes.

Sub-Indexes	Formula
Health Index (HI)	$\frac{1}{3}(DI_{df} + DI_{imr} + DI_{su})$
Socio-Economic Index (SEI)	$\frac{1}{4}(DI_{un} + DI_{ur} + DI_{pci} + DI_{fwpr})$
Education Index (EI)	DI_{ei}
Judicial Index (JI)	$\frac{1}{3}(DI_{cr} + DI_{csr} + DI_{ppl})$

1) ORDER WEIGHTED GEOMETRIC AVERAGE (OWGA)

OWGA operator of dimension n represents a mapping $R^{+n} \rightarrow R^{+}$ corresponding to a weighting vector $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^T$ with $\eta_i \in [0, 1]$ and $\sum_{i=1}^{n} \eta_i = 1$ such that

$$OWGA(\alpha_1, \alpha_2, \ldots, \alpha_n) = \prod_{j=1}^{n} \psi_j^{\eta_j} \tag{3}$$

where ψ_j is the j^{th} largest of the α_i, $i = 1, 2, \ldots, n$ [30].

2) WEIGHTED CRIME INDEX

Though equal weights can be assigned to the sub-indexes to form an Index showing equal importance of the variables included in the development of the Index, the differential weighted index will prove to be more influential. The order weighted geometric average was employed to develop a weighted geometric average Index (Eqn. (3)). Thus the weighted index has been developed in a way assigning varying weights to the sub-indexes. As the economic and social status impacts more to criminal activities, the *SEI* was assigned a weight of 0.4, the *HI* was assigned 0.3, the *JI* as 0.2, and the *EI* as 0.1 (Eqn. (4)).

$$CAWI_{\eta} = (HI^{0.3})(SEI^{0.4})(EI^{0.1})(JI^{0.2}) \tag{4}$$

C. RELIABILITY

Reliability pertains to the consistency and dependability of a system, procedure, or measurement. It encompasses the degree to which one can have confidence in the execution of a designated task or the generation of steady outcomes across different circumstances and durations.

The procedures for estimating reliability are classified into two categories:

(i) External Consistency procedure

(ii) Internal Consistency procedure

The process of External Consistency evaluates stability and similarity using the Test-Retest Correlation method. On the other hand, the Internal Consistency process gauges uniformity through techniques. It is achieved using the Split-half correlation [31]. Between 2011 and 2021, the Test Re-test method was employed to assess the stability of the index. To evaluate its internal stability, the split-half technique was utilized for equivalence.

The weighted indexes from 2011 and 2021 were compared using a half-split method, between the sub-index values [32].

D. VALIDITY

Validity refers to the extent to which a concept, conclusion, or measurement is well-founded and accurately represents the intended meaning or reality [33]. We employed predictive validity to validate the model. The validation of the Index developed was carried out by fitting machine learning models. The primary aim was to develop a reliable model that fits the data giving the best possible outcome but in reality, it often fails to perform well on the real-time data. We used Python software for the data analysis and model fitting. The dataset is divided into two portions: a larger portion for training and a smaller portion for testing. We divided the dataset in a ratio of 75% to 25%.

1) ENSEMBLE METHODS

Ensemble methods are widely employed in machine learning. It involves combining multiple baseline models to create a more robust and comprehensive supervised model [34]. The fundamental concept behind ensemble learning is that when one weak predictor makes an incorrect prediction, other weak predictors can help rectify the errors, leading to overall improved performance [35]. In crafting ensemble regression models, three key factors demand thoughtful attention: (1) Choosing the most appropriate technique from the myriad of regression methods presents a challenge. (2) Finding the ideal number of individual regressors to improve accuracy. (3) Finding appropriate fusion techniques to efficiently integrate the outputs of diverse individual regressors for the final prediction [36].

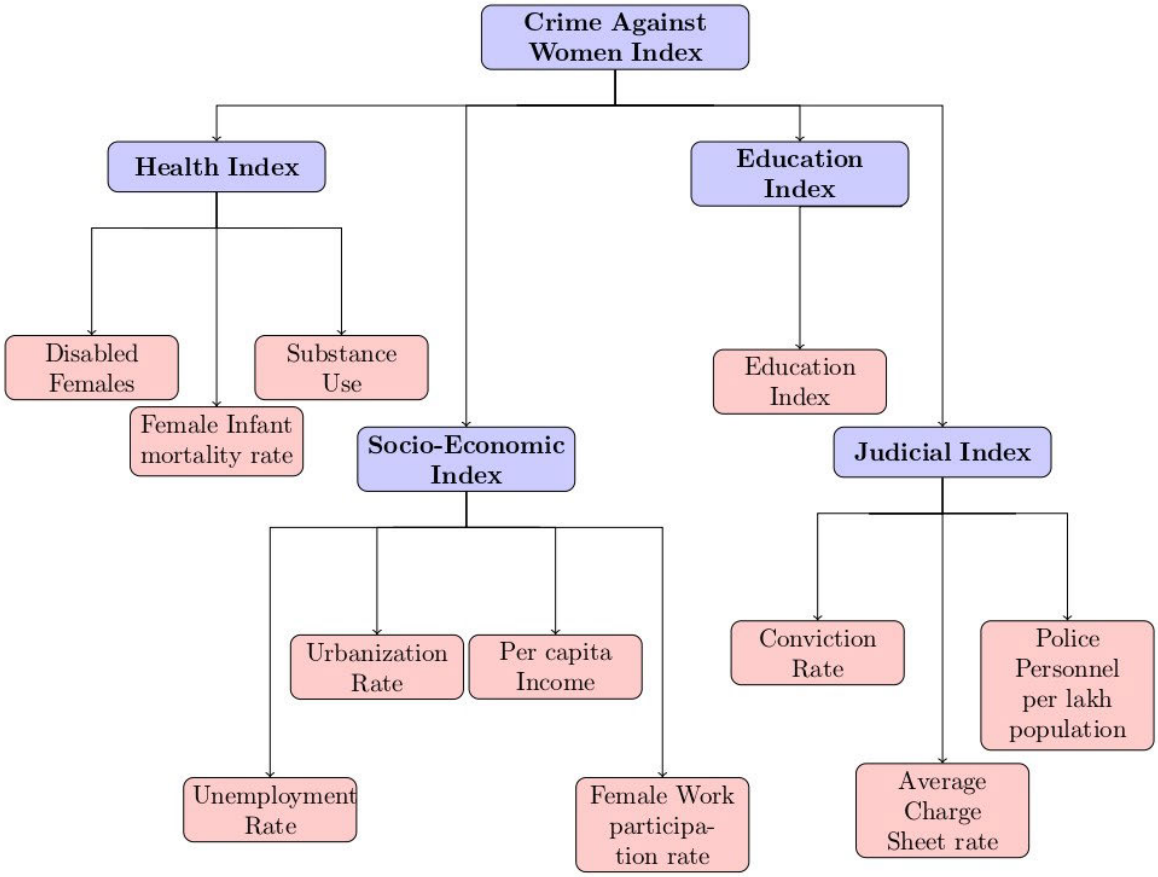


FIGURE 1. Crime against women index.

We employed homogeneous and heterogeneous ensemble techniques. We constructed eight homogeneous ensemble models using bagging and boosting and four heterogeneous ensemble models. These twelve models were built for the indices of India in the years 2011 and 2021 and also for the state of Rajasthan for the years 2011 and 2021.

The ensemble approach combined multiple linear regression and Robust regression models like Huber regression, RANSAC regression, and Theil Sen Regression. The model was built with the CAWI as the dependent variable and the sub-indexes on which the index depends as the explanatory variables.

The homogeneous models were built using the bagging and boosting techniques using the MLR, HR, RSR, and TSR. Bagging, also referred to as bootstrap aggregation, merges various weak learners to enhance accuracy. It partitions the training data into subsets and trains weak learners on each subset, amalgamating their results. This technique reduces the variance in the model. Conversely, boosting trains diverse weak learners on the training set and iteratively corrects errors from preceding models whose outcomes are inadequate. This process mitigates the model's bias [36].

The heterogeneous model was created by combining the baseline models using meta-regressors to generate a hybridized model, enhancing prediction accuracy. Random Forest Regressor and Stochastic Gradient Descent were employed as the meta-regressors for the hybridization of the models. The hybrid model's performance was evaluated by comparing its metrics with those of the individual models, assessing its effectiveness in comparison to them [37], [38] (Fig 2). Fig. 3 illustrates the configuration of the hybrid model suggested in the proposal. It trains the multiple linear regression and Robust regression models and ensembles the results of these models using Random Forest and Stochastic Gradient Descent as the meta-regressors for better performance.

2) MULTIPLE LINEAR REGRESSION

Multiple linear regression is a statistical technique employed for modeling the correlation between a dependent variable and two or more independent variables. This approach is an extension of simple linear regression, where only one independent variable is considered. In the context of multiple linear regression, the analysis expands to encompass multiple independent variables.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \\ &= \beta_0 + \sum_{i=1}^{n} (\beta_i x_i + \epsilon_i). \end{aligned} \tag{5}$$

where y_i is the CAW cases, β_0 is the intercept value, x_i's, $i = 1, 2, 3, 4$, are the four explanatory variable, β_i is the estimated regression coefficients of respective explanatory variables and ϵ is the error i.e. the difference of the estimated value of y_i with respect to the original value [39].

3) ROBUST REGRESSION

Robust regression models are used for modeling relationships between variables when there are outliers or influential data points in the dataset. These models provide an alternative to least squares regression by requiring less restrictive assumptions. These models help to overcome the limitations of least squares regression models and to provide more reliable parameter estimates. Traditional least squares regression, which minimizes the sum of squared residuals, can be highly influenced by outliers and may produce biased parameter estimates.

- Huber Regression
  It combines the least square regression and mean absolute deviation (L1 norm) regression. It handles the variability in the data with a loss function different than the least squares. The loss function (Eqn. 6) needs to be minimized as

$$\underset{\beta}{\text{minimize}} = \sum_{i=1}^{m} \phi(y_i - {x_i}^T \beta) \tag{6}$$

  for variable $\beta \in R^n$, where the loss function ϕ (Eqn. 7) is the Huber function with threshold $\tau > 0$,

$$\phi(\delta) = \begin{cases} \delta^2 & |\delta| \leq \tau \\ 2\delta\tau - \tau^2 & |\delta| > \tau \end{cases} \tag{7}$$

  τ serves as the parameter for enhancing robustness, striking a balance between bias and robustness. The loss function exhibits a quadratic response when dealing with smaller residuals ($|\delta| \leq \tau$) and a linear response for larger residuals ($|\delta| > \tau$).
- RANSAC (Random Sample Consensus)
  It is an iterative algorithm used for robust regression in the presence of outliers. The basic idea behind RANSAC is to fit a model to a subset of the data (inliers) and then evaluate the quality of the fit by counting the number of inliers within a certain threshold. The number of inliers is calculated by checking the vertical distance between each data point (x_i, y_i) and the line $y = mx+b$. The point is considered an inlier if this vertical distance d_i is less than the threshold value set as ϵ (Eqn.8).

$$d_i = |y_i - (mx_i + b)| < \epsilon \tag{8}$$

  The process is repeated until the highest number of inliers are selected with the applied condition of termination.
- Theil Sen Regression
  It is a non-parametric method for estimating the slope of a linear relationship between two variables. It does not assume any specific distribution for the data. The basic idea behind Theil-Sen regression is to calculate the median of all the slopes between pairs of points in the dataset (Eqn. 9).

$$Slope_{ij} = \frac{y_j - y_i}{x_j - x_i} \tag{9}$$

  The Theil-Sen slope is this median slope. It represents the best estimate of the slope of the underlying linear relationship in the presence of outliers. The Theil Sen slope (Eqn. 10) is

$$\zeta = Median(Slope_{ij}) \tag{10}$$

  The intercept of the Theil-Sen regression line can be calculated using the median of the intercepts of all lines passing through pairs of points.

4) META REGRESSOR

A meta-regressor is the statistical or machine learning model employed in the meta-regression analysis. It models the relationship between the study-level characteristics (covariates) and the effect sizes observed in individual studies.

- Random Forest Regressor
  Random Forest, an ensemble technique, fits both regression and classification models. It produces hundreds of decision trees, each functioning as an independent regression function. The result of the random forest regression is determined by averaging the outputs of all the individual decision trees. As the model is non-parametric, it doesn't make assumptions about prior parameters for class densities or fix the tree structure in advance. Instead, the tree evolves during the learning process based on the input data. Each decision tree comprises decision nodes and leaf nodes. Decision nodes assess each input sample through a test function, directing it along different branches according to the sample's features.
  Let X denotes the input vector that comprises n features, where $X = x_1, x_2, \ldots, x_n$, Y the output scalar and S_m (Eqn. 11) the training set consisting of m observations,

$$S_m = (X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m), X \in R^n, Y \in R \tag{11}$$

  During the training phase, the algorithm divides the input data at each node, optimizing the split function parameters to align with the set S_m. In the initial step, the decision tree strives to make the most effective split across all variables, commencing from the root. Subsequently, each node employs its split function on the new input X with this process iteratively recurring until a terminal node is reached. It is customary to halt tree growth either when a maximum level is attained or when a node contains fewer than a predefined number of observations. Following this training process, a prediction function $\bar{h}(X, S_m)$ is formulated over S_m. The random forest regression model can enhance

prediction performance. The model is less sensitive to outliers in the data. The combination of multiple trees helps to reduce overfitting compared to a single decision tree.
Random Forest is employed as a meta-regressor in the model combining the MLR, HR, RNR, and TSR regression techniques. It ensembles these regression models to build a hybrid model to validate the Index.

- Stochastic Gradient Descent Regressor (SGD)
Stochastic Gradient Descent (SGD) is an optimization algorithm used to minimize (or maximize) a function by iteratively moving in the direction of the steepest rapid decrease (or increase) in the function. The mathematical form of the Stochastic Gradient Descent update rule for updating the parameters θ of a model in each iteration is represented in Eqn.12.

$$\theta = \theta - \eta \nabla f_i(\theta) \tag{12}$$

where θ represents the parameters (weights) of the model that are being optimized, $f_i(\theta)$ is the objective function to be minimized (or maximized). In machine learning, this function is often the loss function, which measures the difference between the predicted values and the actual values, $\nabla f_i(\theta)$ represents the gradient of the objective function with respect to the parameters θ. It is a vector that points in the direction of the steepest increase of the function at the current point θ, and η is the learning rate, which determines the size of the steps taken during the optimization. It's a positive scalar value that is usually set in advance. The learning rate controls how much we are adjusting the weights of our network with respect to the gradient of the loss function.
The model is built for the 2011 and 2021 index of the country as well as for the state of Rajasthan.

E. EVALUATION METRICS

The evaluation metrics are used to measure and compare the accuracy of the models employed. The mean absolute error, root mean squared error, and mean absolute percentage error were used to check the validation [40]. These metrics offer a clearer insight into a model's performance and aid in making informed decisions about selecting the most suitable model. In this research, table 2 offers the employed evaluation metrics to assess the effectiveness of the regression models and the hybrid model.

1) MEAN ABSOLUTE ERROR (MAE)

It represents the average absolute difference between predicted and actual values, providing a straightforward measure of the model's accuracy. In machine learning, MAE is often used as a loss function during the training of regression models. The goal during training is to minimize the MAE, which means the model is trying to reduce the average absolute difference between predicted and actual values. MAE equally penalizes both overestimation and underestimation errors.

TABLE 2. Evaluation metrics.

Evaluation metrics	Formula
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Percentage Error (MAPE)	$\frac{100\%}{n}\sum_{i=1}^{n}(\frac{y_i - \hat{y}_i}{y_i})$

MAE is not affected by the scale of the data. MAE is less sensitive to outliers compared to other metrics [41].

2) ROOT MEAN SQUARED ERROR (RMSE)

RMSE measures the average magnitude of the errors between predicted and observed values. It is obtained by evaluating the square root of the mean of the squared deviations between predicted and observed values. A lower RMSE indicates better accuracy, as it means the model's predictions are closer to the actual values. RMSE is frequently used in regression analysis to assess the goodness of fit of a predictive model. RMSE provides a standardized way to compare the performance of different models.

3) MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

MAPE takes the average of absolute percentage differences between predicted and actual values. It is used to compare different models, as it is normalized to a scale of 0-100 and remains unaffected by the magnitude of the actual values. It is more sensitive to outliers than MAE.

The challenge in comparing models for predicting variables with varying mean values arises from the reliance of MAE and RMSE on the mean value of the variable. In contrast, MAPE offers a more intuitive and straightforward approach to assessing models. Although RMSE is commonly used to measure the accuracy of regression models, it is particularly favored in situations with outliers due to its heightened sensitivity compared to MAE and MAPE, making it a superior choice for datasets containing outlier data points.

IV. RESULTS AND DISCUSSION

A. TRENDS AND PATTERNS OF CRIME AGAINST WOMEN

The figures 4 and 5 depict the escalation in incidents of crimes against women in India and the state of Rajasthan during the years 2011 and 2021. Both graphs indicate a notable surge in the occurrence of such crimes over the decade. Though some Indian states show a minor decline, there is an overall upward trend in crime rates. In Rajasthan, the rate of cases of crimes against women surged by over 60% during this period. The district-wise scenario of Rajasthan shows that the rate of CAW has almost doubled. Districts including Jaisalmer, Barmer, Sikar, Alwar, Jalore, Karauli, Bhilwara, Dholpur, Ajmer, Jodhpur, Bharatpur, Sawai Madhopur, Nagaur, and Jhunjhunu witnessed an increase of more than the doubled rate which is of immediate concern and needs to be addressed as early as possible.

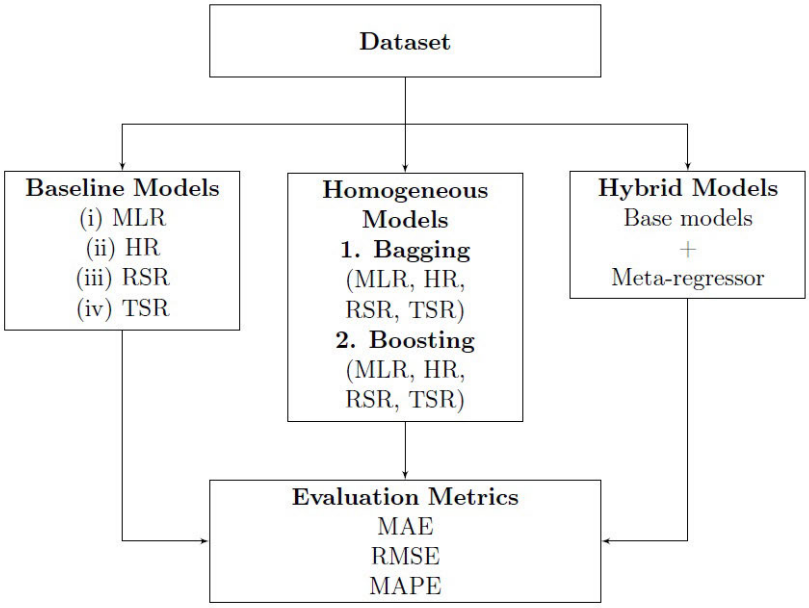


FIGURE 2. Models for validation of the developed index.

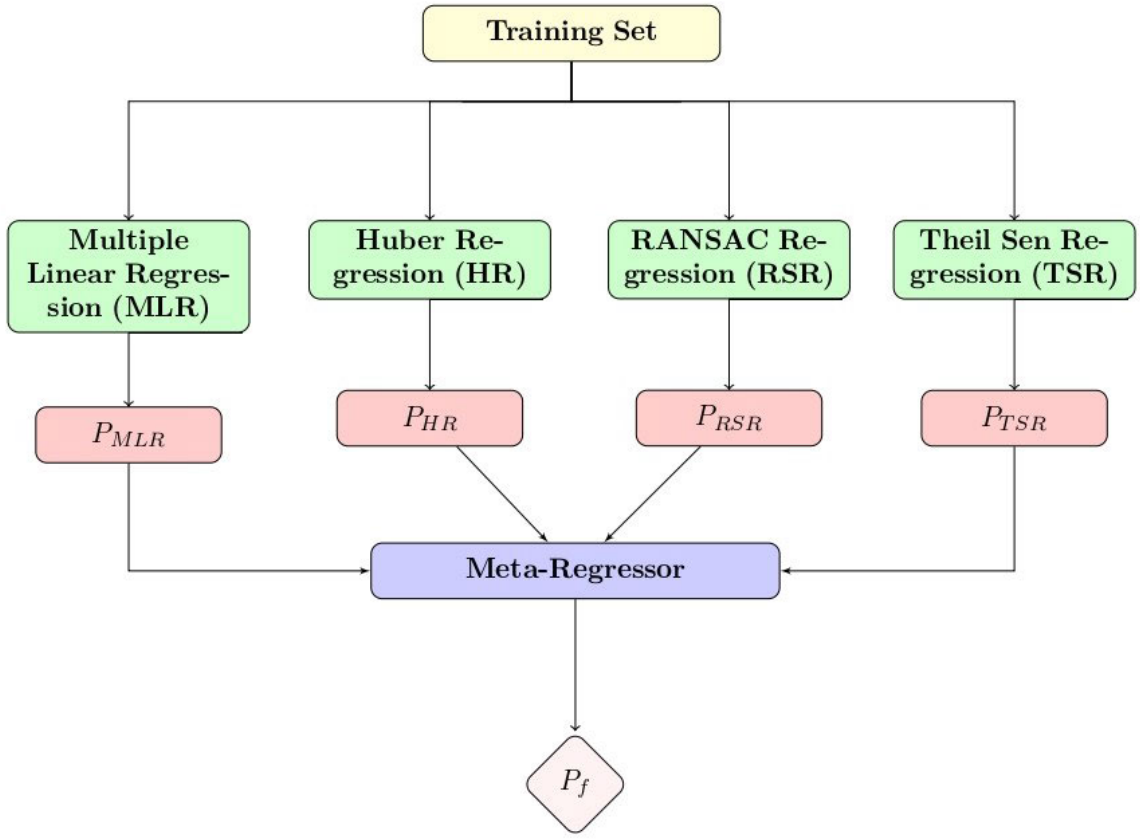


FIGURE 3. Flowchart of the ensemble hybrid model.

B. CRIME AGAINST WOMEN INDEX

The development of the Crime Against Women Index aimed to shed light on the contemporary societal factors that contribute to and promote CAW. This index comprises four sub-indexes, each influenced by a variety of factors. CAWI is determined by a nation's health, socioeconomic status,

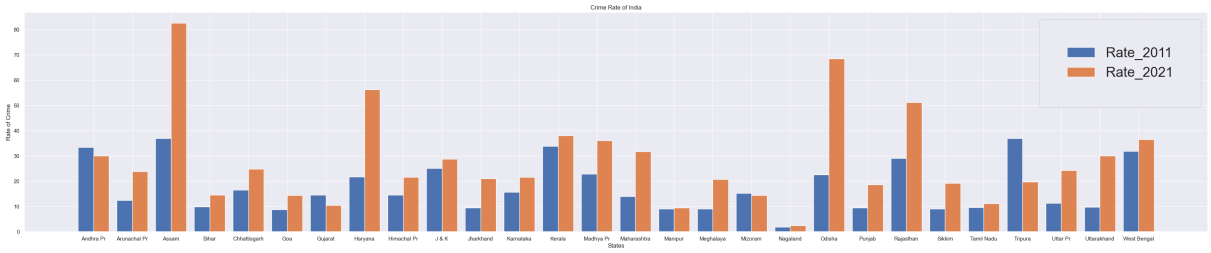


FIGURE 4. Comparative graph of crime rates in India in the years 2011 and 2021.

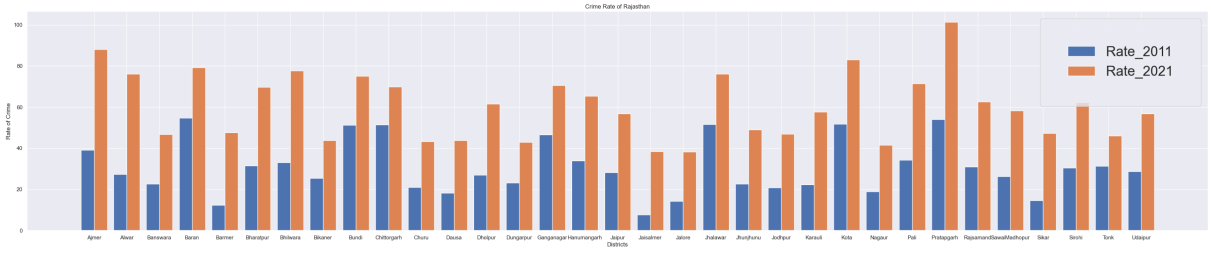
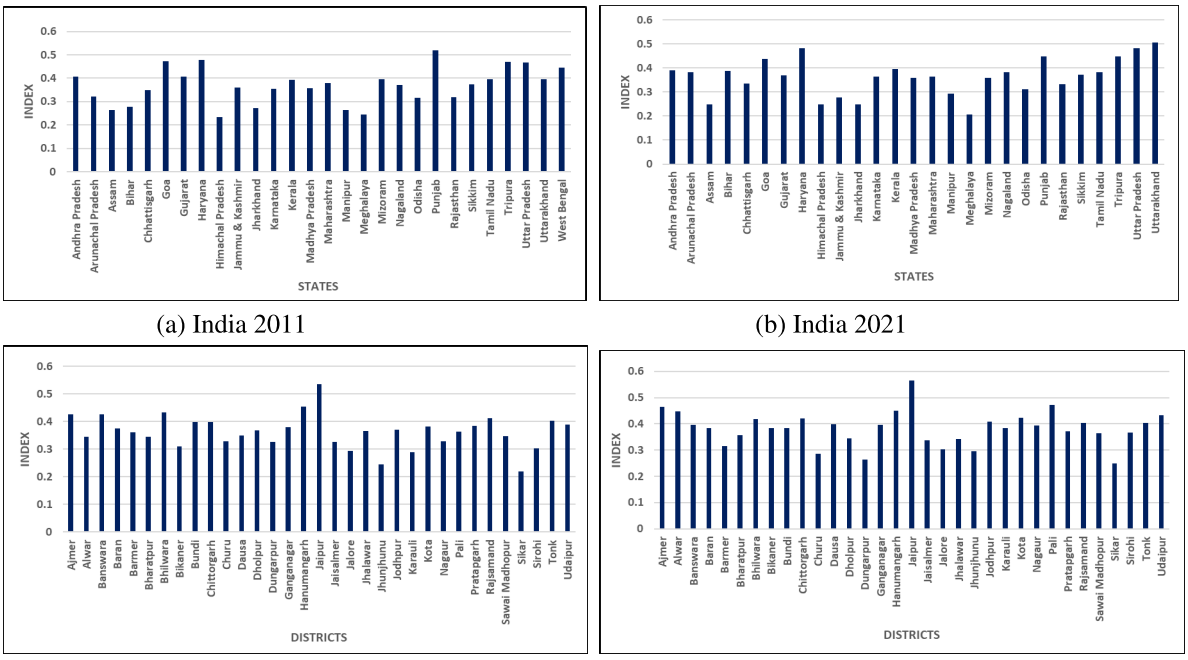


FIGURE 5. Comparative graph of crime rate in Rajasthan in the years 2011 and 2021.



(a) India 2011 (b) India 2021

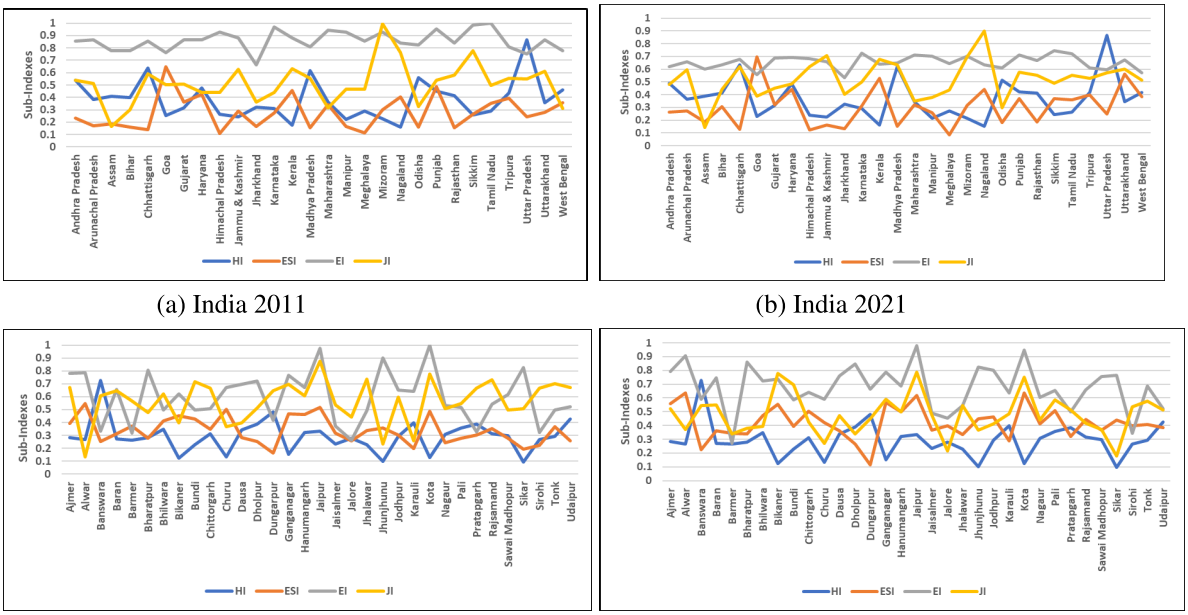
(c) Rajasthan 2011 (d) Rajasthan 2021

FIGURE 6. Trend of CAW index.

education system, and judicial framework. It operates on a scale from 0 to 1, where 0 signifies a high prevalence of such crimes and 1 indicates minimal incidents of CAW. The index was calculated for the years 2011 and 2021, covering all states across the country and districts within the state of Rajasthan.

1) INDIA-2011

The health subindex, which considers factors like disability in females, female infant mortality rate, and substance use, reveals that over half of the country's states have a Health Index (HI) exceeding 0.35 score. The Socio-Economic subindex, determined by urbanization, unemployment, per capita income, and female work participation rate, indicates only one state with a Socio-Economic Index (SEI) greater than a score of 0.5. The Education subindex, derived from the Human Development Index, shows values above 0.7 for all states except Chattisgarh. The Judicial subindex, dependent on conviction rate, average charge sheet rate, and police



(a) India 2011 (b) India 2021

(c) Rajasthan 2011 (d) Rajasthan 2021

FIGURE 7. Trend lines of the sub-indexes, HI, ESI, EI, and JI.

personnel per lakh population exceeds 0.5 for over half of the states (fig 7). In the overall weighted Index, the top five states with the lowest CAWI are Himachal Pradesh, Meghalaya, Manipur, Assam, and Jharkhand (Table 3). This suggests that less developed states in challenging geographical locations exhibit higher chances of occurrence of CAW.

C. INDIA-2021

The CAWI exhibits a pattern akin to that of 2011 in the year 2021. Around 50% of the states have Health Index (HI) scores exceeding 0.35, while over 50% of the states score below 0.5 in Social Empowerment Index (SEI). Education Index (EI) seems promising, with all states scoring 0.5 or higher. Assam has the lowest Judicial Index (JI) score; however, more than half of the states in the country have a JI of 0.5 or more (fig 6). According to the weighted Index of the subindexes, the top five states with the lowest ranks are Meghalaya, Jharkhand, Assam, Jammu & Kashmir, and Manipur (Table 3).

The indexes for 2011 and 2021 both indicate that northern and northeastern states in the country are at elevated risks of CAW. Rajasthan is among the top ten states in the CAWI. Consequently, a detailed analysis of the factors specific to the districts in Rajasthan has been conducted to create individual CAWI scores for each district.

D. RAJASTHAN-2011

The CAWI for the year 2011 in districts of Rajasthan state used similar criteria as those for the states, with the exception of the education index. The Education subindex was calculated using district literacy rates since specific education

TABLE 3. CAW index India.

States	2011 Rank	States	2021 Rank
Himachal Pradesh	1	Meghalaya	1
Meghalaya	2	Jharkhand	2
Manipur	3	Assam	3
Assam	4	Jammu & Kashmir	4
Jharkhand	5	Manipur	5

TABLE 4. CAW index Rajasthan.

Districts	2011 Rank	Districts	2021 Rank
Sikar	1	Sikar	1
Jhunjhunu	2	Dungarpur	2
Karauli	3	Churu	3
Jalore	4	Jhunjhunu	4
Sirohi	5	Jalore	5

index data for each district in the human development index was unavailable. About half of the districts, around 50%, scored above 0.3 in the Health Index (HI). In the Socio-Economic Index (SEI), all districts except three scored below 0.5. The Education Index (EI) appeared favorable for over 70% of districts, scoring 0.5 or higher. In the Job Index (JI), more than 60% of districts scored 0.5 or higher (fig 6). Combining these subindexes, the top five districts were Sikar, Jhunjhunu, Karauli, Jalore, and Sirohi (Table 4).

E. RAJASTHAN-2021

The CAWI data for 2021 follows a pattern similar to that of 2011. About half of the districts, particularly those with a Health Index (HI) of around 50%, have scores exceeding 0.3. In the Socio-Economic Index (SEI), with the exception of

TABLE 5. India 2011: Comparison of evaluation metrics.

Models	Regression	MAE	RMSE	MAPE
Baseline	MLR	0.0753	0.0910	0.2201
	HR	0.0677	0.1101	0.1606
	RSR	0.0753	0.0910	0.2201
	TSR	0.0680	0.1114	0.1491
Homogeneous Bagging	MLR_Bag	0.0267	0.0303	0.0705
	HR_Bag	0.0215	0.0307	0.0544
	RSR_Bag	0.0194	0.0277	0.0496
	TSR_Bag	0.0178	0.0276	0.0439
Homogeneous Boosting	MLR_Boost	0.0215	0.0261	0.0553
	HR_Boost	0.0193	0.0315	0.0469
	RSR_Boost	0.0201	0.0333	0.0490
	TSR_Boost	0.0192	0.0315	0.0463
Hybrid	MLR+HR+RSR+TSR-RFR	0.0417	0.0517	0.1394

TABLE 6. India 2021: Comparison of evaluation metrics.

Models	Regression	MAE	RMSE	MAPE
Baseline	MLR	0.0889	0.1254	0.2446
	HR	0.0921	0.1336	0.2552
	RSR	0.0515	0.0700	0.1429
	TSR	0.0779	0.1402	0.1928
Homogeneous Bagging	MLR_Bag	0.0272	0.0402	0.0753
	HR_Bag	0.0301	0.0469	0.0846
	RSR_Bag	0.0267	0.0447	0.0727
	TSR_Bag	0.0226	0.0397	0.0613
Homogeneous Boosting	MLR_Boost	0.0267	0.0375	0.0744
	HR_Boost	0.0276	0.0400	0.0772
	RSR_Boost	0.0362	0.0676	0.0899
	TSR_Boost	0.0231	0.0424	0.0623
Hybrid	MLR+HR+RSR+TSR-RFR	0.0466	0.0580	0.2384

TABLE 7. Rajasthan 2011: Comparison of evaluation metrics.

Models	Regression	MAE	RMSE	MAPE
Baseline	MLR	0.0780	0.0992	0.1751
	HR	0.0662	0.0854	0.1474
	RSR	0.0533	0.0675	0.1179
	TSR	0.0555	0.0778	0.1191
Homogeneous Bagging	MLR_Bag	0.0184	0.0277	0.0505
	HR_Bag	0.0170	0.0330	0.0436
	RSR_Bag	0.0187	0.0371	0.0489
	TSR_Bag	0.0162	0.0279	0.0456
Homogeneous Boosting	MLR_Boost	0.0165	0.0259	0.0420
	HR_Boost	0.0167	0.0319	0.0432
	RSR_Boost	0.0232	0.0412	0.0604
	TSR_Boost	0.0180	0.0327	0.0475
Hybrid	MLR+HR+RSR+TSR-SGD	0.0547	0.0657	0.1152

TABLE 8. Rajasthan 2021: Comparison of evaluation metrics.

Models	Regression	MAE	RMSE	MAPE
Baseline	MLR	0.0532	0.0591	0.1673
	HR	0.0507	0.0672	0.2009
	RSR	0.0307	0.0476	0.0942
	TSR	0.0747	0.0845	0.2648
Homogeneous Bagging	MLR_Bag	0.0206	0.0382	0.0529
	HR_Bag	0.0224	0.0455	0.0567
	RSR_Bag	0.0203	0.0313	0.0538
	TSR_Bag	0.0228	0.0342	0.0604
Homogeneous Boosting	MLR_Boost	0.0209	0.0377	0.0521
	HR_Boost	0.0230	0.0455	0.0586
	RSR_Boost	0.0204	0.0409	0.0537
	TSR_Boost	0.0200	0.0411	0.0511
Hybrid	MLR+HR+RSR+TSR-SGD	0.0338	0.0420	0.1246

9 districts, all others score below 0.5. For more than 60% of the districts, the Environmental Index (EI) appears promising, with scores reaching 0.5 or higher. In terms of the Justice Index (JI), approximately half of the districts achieve a score

of 0.5 or more. When considering the composite weighted Index, the top five districts are Sikar, Dungarpur, Churu, Jhunjhunu, and Jalore (Table 4).

F. RELIABILITY

The test-retest applied to the weighted index of 2011 and 2021 for India gave a correlation of 0.9119 and for the state of Rajasthan as 0.7828. The value of correlation coefficients shows a positive correlation. Thus, the weighted index for both years achieved external consistency. For internal consistency, split-half tests applied between the sub-indexes of the years 2011 and 2021 appeared to have a high correlation coefficient.

G. VALIDITY OF THE INDEX

The validation of the developed index was assessed through a comparative analysis of baseline, homogeneous, and a hybrid machine learning model. For India and Rajasthan state, the analysis was conducted on indexes for the years 2011 and 2021. In the hybrid approach for India, a combination of multiple linear regression and robust regression models such as Huber, RANSAC, and Theil Sen were integrated, along with Random Forest serving as the meta-regressor. Similarly, for the Rajasthan state, the hybrid model included multiple linear regression, Huber, RANSAC, and Theil Sen robust regression models, coupled with Stochastic Gradient Descent as the meta-regressor. The statistical findings indicated that while the hybrid model effectively mitigates the constraints inherent in multiple linear regression, it does not surpass the performance of homogeneous models. The hybrid models demonstrated superior performance compared to baseline models, manifesting lower error rates in almost all four cases. However, the homogeneous models exhibited superior performance over the developed hybrid model. The mean absolute error (MAE), the root mean squared error (RMSE), and the mean absolute percentage (MAPE) values of all the models built for India and Rajasthan Indexes of the years 2011 and 2021 were compared with the hybrid model in the tables 5, 6, 7 and 8. This shows the minimum values of MAE, RMSE, and MAPE for the homogeneous models when compared to the baseline and hybrid models, which signifies that the performance of the homogeneous models is better than the others. Thus the validation of the index was established with the help of the baseline, homogeneous, and hybrid models.

V. CONCLUSION

This study developed a crime against women index, which encompasses diverse typologies. The aim is to comprehensively examine the influence of these typologies on women's safety. The identified typologies encompass the health sector, judicial system, economic growth, and education system. The states of Himachal Pradesh, Meghalaya, Manipur, Assam, Jharkhand, and Jammu & Kashmir witnessed to have the lowest Index values, suggesting a higher likelihood of crimes against women. Most of these states lie on the border areas sharing international borders with China, Myanmar, Bangladesh, Bhutan, Nepal, Pakistan, and Afghanistan. Moreover, these areas are least urbanized and have challenging geographical conditions, which impede their overall development. Additionally, the districts Sikar, Jhunjhunu, Dungarpur, Karauli, Churu, Jalore, and Sirohi in Rajasthan also exhibit low Index values, indicating an increased potential for crimes against women. Even most of these districts appear in the bordering areas sharing borders with Haryana, Uttar Pradesh, Madhya Pradesh, and Gujarat. Also, these districts exhibit low levels of development. The index reveals that both states of India and districts within Rajasthan that exhibit significant urbanization and advancement tend to have comparatively lower crime. This suggests that women generally experience greater safety and security in the more developed parts of the country, in contrast to areas where basic life necessities are deficient. The districts characterized by lower levels of development and elevated crime rates occupy the top ranks of the index. This deduction emphasizes that regions, where women lack awareness of their fundamental legal rights, are more susceptible to victimization.

The validation of the index was conducted through a comparative analysis of machine learning approaches. A hybrid model was developed. This approach utilized MLR, Huber, RANSAC, and Theil Sen Regression, with Random Forest Regressor and Stochastic Gradient Descent serving as meta-regressors. The findings indicate that the hybrid model performed better than the baseline but the homogeneous models outperformed both demonstrating increased accuracy. Consequently, the homogeneous models are regarded as the most effective for validating the Index. Women's safety is a major challenge to the society as a whole. As CAW is not uniformly distributed throughout the globe, the reasons behind it need to be known as well as dealt with utmost importance.

The major contributions of the study are as follows:

- A weighted Crime Index developed for India and the state of Rajasthan, covering the years 2011 and 2021, concludes that women in highly urbanized and developed regions are less likely to become victims.
- The techniques employed to assess both the external and internal reliability of the index demonstrate a strong correlation, affirming the index's reliability.
- For the validation of the Index, a comparative analysis between baseline, homogeneous, and hybrid machine-learning models was performed.
- The homogeneous models exhibit the lowest error rates for the Index across both the years 2011 and 2021.
- The index points out states and districts with extreme geographical locations and low development as the ones occupying the highest positions.

As CAW is a sensitive issue, many cases go unreported or underreported. This is due to various factors like social

stigma, victim blaming, lack of awareness about the laws and rules, lack of economic independence, lack of trust in the legal process, etc. Consequently, the index established here can be a valuable tool to assist national and state governments in directing resources toward development and management efforts. Thus, we can conclude that the Index formed will be beneficial for the government, legislators, and law enforcement, enabling them to formulate effective laws and strategies to diminish and ultimately eliminate the issue of violence against women.

PRACTICAL IMPLICATIONS AND APPLICATIONS

The CAW index is a vital tool for understanding the prevalence and nature of factors influencing and fostering criminal activities.

- The CAW Index provides policymakers and stakeholders with comprehensive insights into the data on the frequency, types, and locations.
- The Index assists in the allocation of resources by highlighting areas where interventions are most needed.
- Law enforcement agencies can enhance safety measures by increasing patrols in high-risk areas, bolstering victim support services, and providing comprehensive training for officers in sensitively and effectively handling the situations.
- The Index can serve as a valuable tool for shaping public awareness campaigns by shedding light on the prevalence and repercussions of such offenses. These campaigns can effectively educate and mobilize communities.

LIMITATIONS

The limitations of the work include the non-availability of the census data for the year 2021 which would have added more reasonable factors in the development of the crime index.

FUTURE SCOPE

An in-depth study of the states and districts which are more vulnerable to crime against women can be done at the block level which would help the government and the policymakers to frame rules and laws to help the country overcome this menace.

DATA AVAILABILITY

The datasets generated and analyzed during the study are freely available on the National Crime Records Bureau (NCRB), the Census of India website, and various government portals.

CONFLICTING INTERESTS

The authors declared no conflict of interest.

CONTRIBUTION

All the authors contributed equally.

ABBREVIATIONS

CAW	Crime against women.
CAWI	Crime against women index.
EI	Education index.
HI	Health index.
HR	Huber regression.
JI	Judicial index.
MAE	Mean absolute error.
MAPE	Mean absolute percentage error.
MLR	Multiple linear regression.
RMSE	Root mean squared error.
RSR	RANSAC regression.
SEI	Socioeconomic index.
TSR	Theil sen regression.

ACKNOWLEDGMENT

This research was supported by the Department of Science and Technology (DST) India under FIST grant- SR/FST/MS-II/2023/139 - VIT Vellore.

REFERENCES

[1] K. Sukhija, S. N. Singh, and J. Kumar, "Spatial visualization approach for detecting criminal hotspots: An analysis of total cognizable crimes in the state of Haryana," in *Proc. 2nd IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol. (RTEICT)*, pp. 1060–1066, May 2017.

[2] A. Anjali and B. R. Kumar, "Spatial analysis of multivariate factors influencing suicide hotspots in urban Tamil Nadu," *J. Affect. Disorders Rep.*, vol. 16, Apr. 2024, Art. no. 100741.

[3] Á. González-Prieto, A. Brú, J. C. Nuño, and J. L. González-Álvarez, "Hybrid machine learning methods for risk assessment in gender-based crime," *Knowl.-Based Syst.*, vol. 260, Jan. 2023, Art. no. 110130.

[4] G. V. Manish, Simran, J. Kumar, and D. K. Choubey, "Identification of hotspot of rape cases in NCT of Delhi: A data science perspective," in *Proc. Int. Conf. Inf. Syst. Manage. Sci.*, vol. 521. Cham, Switzerland: Springer, 2021, pp. 485–496.

[5] V. Ceccato and A. Loukaitou-Sideris, "Fear of sexual harassment and its impact on safety perceptions in transit environments: A global perspective," *Violence Against Women*, vol. 28, no. 1, pp. 26–48, Jan. 2022.

[6] C. M. Spencer, S. M. Stith, and B. Cafferky, "What puts individuals at risk for physical intimate partner violence perpetration? A meta-analysis examining risk markers for men and women," *Trauma, Violence, Abuse*, vol. 23, no. 1, pp. 36–51, Jan. 2022.

[7] P. K. Saravag and B. R. Kumar, "An application of scan statistics in identification and analysis of hotspot of crime against women in Rajasthan, India," *Appl. Spatial Anal. Policy*, vol. 17, no. 3, pp. 963–982, Sep. 2024.

[8] M. Flood and B. Pease, "Factors influencing attitudes to violence against women," *Trauma, Violence, Abuse*, vol. 10, no. 2, pp. 125–142, Apr. 2009.

[9] S. N. Ogden, M. E. Dichter, and A. R. Bazzi, "Intimate partner violence as a predictor of substance use outcomes among women: A systematic review," *Addictive Behav.*, vol. 127, Apr. 2022, Art. no. 107214.

[10] S. Srivastava, P. Kumar, R. Rashmi, R. Paul, and P. Dhillon, "Does substance use by family members and community affect the substance use among adolescent boys? Evidence from Udaya study, India," *BMC Public Health*, vol. 21, no. 1, pp. 1–10, Dec. 2021.

[11] K. Shorette and R. Burroway, "Consistencies and contradictions: Revisiting the relationship between women's education and infant mortality from a distributional perspective," *Social Sci. Res.*, vol. 105, Jul. 2022, Art. no. 102697.

[12] J. Chopin, E. Beauregard, and N. Deslauriers-Varin, "Less exposed, more vulnerable? Understanding the sexual victimization of women with disabilities under the lens of victimological theories," *Int. Rev. Victimology*, vol. 30, no. 1, pp. 109–129, Jan. 2024.

[13] Z. Rodriguez, "The power of employment: Effects of India's employment guarantee on women empowerment," *World Develop.*, vol. 152, Apr. 2022, Art. no. 105803.

[14] T.-A. Craigie, V. Taraz, and M. Zapryanova, "Temperature and convictions: Evidence from India," *Environ. Develop. Econ.*, vol. 28, no. 6, pp. 538–558, Dec. 2023.
[15] T. Das and D. T. B. Roy, "More than individual factors; Is there any contextual effect of unemployment, poverty and literacy on the domestic spousal violence against women? A multilevel analysis on Indian context," *SSM-Population Health*, vol. 12, Dec. 2020, Art. no. 100691.
[16] P. Raj and M. M. Rahman, "Revisiting the economic theory of crime a state-level analysis in India," *Cogent Social Sci.*, vol. 9, no. 1, Dec. 2023, Art. no. 2170021.
[17] A. Lisowska-Kierepka, "How to analyse spatial distribution of crime? Crime risk indicator in an attempt to design an original method of spatial crime analysis," *Cities*, vol. 120, Jan. 2022, Art. no. 103403.
[18] Y. K. Kwan, W. C. Ip, and P. Kwan, "A crime index with Thurstone's scaling of crime severity," *J. Criminal Justice*, vol. 28, no. 3, pp. 237–244, May 2000.
[19] K. Chaudhuri, P. Chowdhury, and S. C. Kumbhakar, "Crime in India: Specification and estimation of violent crime index," *J. Productiv. Anal.*, vol. 43, no. 1, pp. 13–28, Feb. 2015.
[20] C. Nau, M. Sidell, K. Clift, C. Koebnick, J. Desai, and D. Rohm-Young, "A commercially available crime index may be a reliable alternative to actual census-tract crime in an urban area," *Preventive Med. Rep.*, vol. 17, Mar. 2020, Art. no. 100996.
[21] E. Ledingham, G. W. Wright, and M. Mitra, "Sexual violence against women with disabilities: Experiences with force and lifetime risk," *Amer. J. Preventive Med.*, vol. 62, no. 6, pp. 895–902, Jun. 2022.
[22] R. Kmet and Z. Dvorak, "Crime index as one of the main indicators of safety," *MEST J.*, vol. 8, no. 1, pp. 57–64, Jan. 2020.
[23] E. Tate, "Social vulnerability indices: A comparative assessment using uncertainty and sensitivity analysis," *Natural Hazards*, vol. 63, no. 2, pp. 325–347, Sep. 2012.
[24] S. Neupane, P. Kc, S. Kyrönlahti, A. Siukola, H. Kosonen, K. Lumme-Sandt, P. Nikander, and C. H. Nygård, "Development and validation of sustainable employability index among older employees," *Occupat. Med.*, vol. 73, no. 1, pp. 19–25, Feb. 2023.
[25] M. Bagavandas, "Development of multifactor index for assessing quality of life of a tribal population of India: Multilevel analysis approach," *BMC Public Health*, vol. 21, no. 1, pp. 1–14, 2021.
[26] G. R. Aryal, "Methods of population estimation and projection," *J. Population Develop.*, vol. 1, no. 1, pp. 54–61, Nov. 2020.
[27] S. Mukherjee, D. Chakraborty, and S. Sikdar, "Three decades of human development across Indian states: Inclusive growth or perpetual disparity," *Nat. Inst. Public Finance Policy*, vol. 15, no. 2, pp. 97–122, 2014.
[28] Y. R. Salama, R. Hamed, and M. Rashwan, "Modified human development index using data envelopment analysis approach," *J. Math. Statist.*, vol. 18, no. 1, pp. 115–133, Jan. 2022.
[29] F. Tobaigy, M. Alamoudi, and O. Bafail, "Human development index: Determining and ranking the significant factors," *Int. J. Eng. Res. Technol.*, vol. 12, no. 3, pp. 231–245, 2023.
[30] J. M. Merigó, M. Guillén, and J. M. Sarabia, "The ordered weighted average in the variance and the covariance," *Int. J. Intell. Syst.*, vol. 30, no. 9, pp. 985–1005, Sep. 2015.
[31] H. K. Mohajan, "Two criteria for good measurements in research: Validity and reliability," *Ann. Spiru Haret Univ. Econ. Ser.*, vol. 17, no. 4, pp. 59–82, Dec. 2017.
[32] V. Tripathi, C. Stanton, D. Strobino, and L. Bartlett, "Development and validation of an index to measure the quality of facility-based labor and delivery care processes in sub-Saharan Africa," *PLoS ONE*, vol. 10, no. 6, 2015, Art. no. e0129491.
[33] G. V. M. Reddy, Iswarya, J. Kumar, and D. K. Choubey, "Time series analysis of national stock exchange: A multivariate data science approach," in *Soft Computing for Problem Solving*, vol. 547. Singapore: Springer, 2023, pp. 691–707.
[34] W. Fang, H. Zhu, and Y. Mei, "Hybrid meta-heuristics for the unrelated parallel machine scheduling problem with setup times," *Knowl.-Based Syst.*, vol. 241, Apr. 2022, Art. no. 108193.
[35] T. Roshni, E. Mirzania, M. H. Kashani, Q.-A.-T. Bui, and S. Shamshirband, "Hybrid support vector regression models with algorithm of innovative gunner for the simulation of groundwater level," *Acta Geophysica*, vol. 70, no. 4, pp. 1885–1898, Aug. 2022.
[36] I. K. Nti, A. F. Adekoya, and B. A. Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction," *J. Big Data*, vol. 7, no. 1, pp. 1–40, Dec. 2020.
[37] D. K. Choubey, P. Dubey, B. P. Tewari, M. Ojha, and J. Kumar, "Prediction of liver disease using soft computing and data science approaches," in *6G Enabled Fog Computing in IoT: Applications and Opportunities*. Cham, Switzerland: Springer, 2023, pp. 183–213.
[38] W. Fang, S. Zhang, and C. Xu, "Improving prediction efficiency of Chinese stock index futures intraday price by VIX-Lasso-GRU model," *Expert Syst. Appl.*, vol. 238, Mar. 2024, Art. no. 121968.
[39] A. Anjali, B. R. Kumar, and J. Kumar, "Spatio-temporal aspect of suicide and suicidal ideation: An application of SaTScan to detect hotspots in four major cities of Tamil Nadu," *J. Sci. Res.*, vol. 65, no. 9, pp. 7–18, 2021.
[40] S. Kumar, B. Kumar, V. Deshpande, and M. Agarwal, "Predicting flow velocity in a vegetative alluvial channel using standalone and hybrid machine learning techniques," *Expert Syst. Appl.*, vol. 232, Dec. 2023, Art. no. 120885.
[41] A. Anjali and B. R. Kumar, "Exploring cause-specific strategies for suicide prevention in India: A multivariate Varma approach," *Asian J. Psychiatry*, vol. 92, Feb. 2024, Art. no. 103871.

POONAM K. SARAVAG received the bachelor's degree in mathematics and the master's degree in applied mathematics from The Maharaja Sayajirao University of Baroda, Gujarat, India, in 2011 and 2013, respectively. She is currently a Research Scholar with the Department of Mathematics, Vellore Institute of Technology, Tamil Nadu, India. With more than seven years of experience, she was an Assistant Professor in mathematics. Her research interests include applied statistics and machine learning in the field of criminology. In addition, she is a Life Member of Indian Mathematical Society.

B. RUSHI KUMAR is currently a Distinguished Professor of mathematics with the School of Advanced Sciences, Vellore Institute of Technology. His extensive research spans CFD, mathematical modeling, computational data science, and soft computing, supported by funding from DST, India, and the Royal Society's CSC follow-on grants. A recognized personality in applied mathematics, he holds membership in various mathematical societies. He has authored more than 140 research papers for national and international journals and conferences. He contributed to research articles and delivered invited talks at conferences, seminars, and workshops. In addition, he has played a pivotal role in organizing academic events at VIT and has successfully supervised 13 Ph.D. students, earning acclaim in academic circles for his expertise in applied mathematics and computing. He has received numerous research and teaching awards.