CSA-Forecaster: Stacked Model for Forecasting Child Sexual Abuse

Saravanan Parthasarathy¹, Arun Raj Lakshminarayanan^{2*}, A. Abdul Azeez Khan³ and K. Javubar Sathick⁴

¹Research Scholar, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. saravanan_cse_2019@crescent.education, https://orcid.org/0000-0002-6229-8688

^{2*}Associate Professor, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. arunraj@crescent.education, https://orcid.org/0000-0001-8181-5022

³Associate Professor, Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. abdulazeezkhan@crescent.education, https://orcid.org/0000-0001-6960-752X

⁴Associate Professor, Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India. javubar@crescent.education, https://orcid.org/0000-0002-2248-8380

Received: November 11, 2023; Accepted: January 06, 2024; Published: February 29, 2024

Abstract

Child sexual abuse is a pervasive and distressing issue that poses serious threats to the well-being and development of children. Early identification and prevention of such incidents are crucial for ensuring child safety and protection. In this study, we investigate the application of stacked machine learning models for the forecasting of child sexual abuse cases. Data on child sexual abuse incidents were gathered from StatBank Denmark and used in this analysis. The geographical coordinates of the municipalities were incorporated as part of the descriptive analysis to examine the distribution and prevalence of child abuse cases. Our approach incorporates a stacked ensemble framework that combines the XGBoost, LSTM, and Random Forest algorithms. By leveraging the strength of individual models and capturing diverse patterns in the data, the stacked model aims to improve prediction performance. Our experimental results demonstrate that the CSA-Forecaster model outperforms individual models in forecasting child sexual abuse incidents. The proposed model achieved an RMSE of 0.094. MAE of 0.0712, MAPE of 0.1557, and R² of 0.8028, indicating robust performance. The outcomes of this research have significant repercussions for the creation of proactive interventions and support systems. Child protection agencies and experts might be equipped to more effectively allocate resources and potentially prevent future abuse instances by employing machine learning models.

Keywords: Crime Forecasting, Child Sexual Abuse, Predictive Policing, Stacked Regression, Random Forest, XGBoost, LSTM.

Journal of Internet Services and Information Security (JISIS), volume: 14, number: 1 (February), pp. 219-241. DOI: 10.58346/JISIS.2024.11.015

^{*}Corresponding author: Associate Professor, Department of Computer Science & Engineering, B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, India.

1 Introduction

All living beings engage in sexual activity to produce offspring and humans, in addition to fulfilling the reproductive function, are also interested in sexual activities to experience physical pleasure and assert power. When humanity emerged from the animal kingdom, society attempted to regulate basic instincts such as sexual activity. Sexual acts in a civilized society could only be performed with the consent of all parties involved. As a result, a sexual liaison without either party's explicit consent is considered abuse. It is presumed a barbaric act in terms of social norms and is punishable by law. Any sexual activity involving children is regarded as a heinous crime. Even if a child offered informed consent, performing a sexual act on him or her is illegal, and the action is classified as Child Sexual Abuse (CSA). An adult or another child could have committed exploitation against children. The indicted person would be prosecuted and sentenced by legal procedures.

Sexual exploitation of children occurs in every country regardless of socioeconomic conditions. CSA is defined as inappropriate touching of genitals or breasts, disrupting the child by displaying pornographic contents, penetrating the vagina, anal using genitals, fingers, or objects, oral sex, masturbation, exhibitionism, and inducing the child to perform sex (Christensen, 2017). However, depending on the cultural, social, and political backgrounds of each country or region, the definitions and limitations of CSA might differ. According to UNICEF, one in every ten girls under the age of twenty is a victim of sexual abuse (UNICEF, 2023). It also confirmed that CSA could occur at home, school, or in public venues. Many such incidents are committed by a well-known individual or a family member of the child. Transnational organized crime syndicates play a significant role in CSA by covertly orchestrating the pornographic industry. Pornographic content producers view children's bodies as another commodity. Child porn has become a multibillion-dollar industry because of the digital revolution (Naidu, 2020). Child traffickers use antisocial elements to force children into prostitution and pornography. Most serious crimes are committed by repeat offenders or organized groups. Most criminals followed a pattern that was defined by a comfort zone (Perry, 2013). In contrast to other types of crime, sexual abuse during childhood is linked to an increased likelihood of being sexually assaulted as an adult, a phenomenon called revictimization (Jaffe, 2023). The victim's life would be devastated, and he or she might have psychological problems as a result (Zarchev, 2022). The authorities were unable to suspect or investigate everyone in society. As a result, predictive policing would be useful in identifying crime patterns, planning patrols, and developing surveillance strategies (Andrejevic, 2017).

The first step in implementing predictive policing operations is to extract knowledge from historical crime data. Due to the exponential increase in crime, structured and unstructured crime data is accumulating all over the world. The massive size of records, differences in documentation methodology, legal variations between regions and countries, data format, insufficient or missing information, redundancy, lack of timewise tracking, and poor data quality are a few constraints in crime analysis. Because of these obstacles, analytical and strategic organizations struggled to develop a global perspective on crime. For the past few years, researchers and legal entities have attempted to use Machine Learning (ML) techniques to build optimal systems for processing massive amounts of data and determining crime patterns (Smith, 2018). The ML approaches are designed for data preprocessing, segmentation, analysis, and knowledge extraction. Expert systems are capable of accurately forecasting crimes and identifying hotspots (Saraiva, 2022).

Despite Denmark being one of the world's most peaceful countries, there were 9683 sexual offences registered in 2022 (Statista, 2022). According to the European Crime Prevention Network (EUCPN), nearly 500 rapes are reported in Denmark each year. It is believed the true number was around 2600 or

higher (EUCPN, 2022). In 2020, children were the victims of 9% of all reported sexual crimes in Denmark. Considering the EUCPN's pronouncements on sexual assault, the number of CSA instances might be substantially higher. The high prevalence of CSA in Denmark could be attributed to several factors, including socioeconomic status, familial background, gender disparities, children's perspective of sexual assault, and their tolerance of abusive adults (Helweg-Larsen, 2006). Over the past half-century in Denmark, there has been a shift in how childhood sexuality is perceived, and it has become more frequently associated with sexual abuse. This transition was probably caused by a growing fear and awareness of child sexual abuse, which became a major concern in Danish childcare facilities (Buch Leander, 2019; Leander, 2023). Forecasting child abuse using artificial intelligence (AI) could have helped alleviate unwarranted anxiety among parents and daycare providers (Dayi et al., 2022). It could also assist legal organizations in successfully planning and carrying out preventative measures. Despite the potential of machine learning in forecasting child abuse, significant ethical considerations arise. Predictions characterized by bias might be discriminatory, and the collection of sensitive data gives rise to concerns regarding privacy (Landau, 2022). In the case of false positives, unwarranted investigations could result in significant damage, whereas false negatives endanger children who are already vulnerable. The utilization of opaque models gives rise to apprehensions regarding equity and responsibility, while algorithmic bias exacerbates possible inequities. Diverse data, transparency, human supervision, rigorous evaluation, and robust privacy measures are essential for mitigating these risks (Lupariello, 2023). Active participation from the community and the establishment of stringent regulations are essential components of responsible implementation. In the end, ethical principles and safeguards must be meticulously considered when implementing machine learning for child protection to maximize benefits and minimize potential harm. This work employed ML methods to predict CSA incidents in Denmark considering the aforementioned information.

The remainder of the paper is structured as follows: Section 2 describes the current research accomplishments in ML methodologies. Section 3 describes the data set's characteristics as well as the proposed model. Section 4 presents the findings of the exploratory analysis. Section 5 lists the experimental outcomes. Finally, the paper concludes in Section 6 with the proposed research direction.

2 Related Work

Crimes are committed by different kinds of people from various kinds of backgrounds. Although each crime is distinct, there are hidden patterns among them. Criminal investigators are primarily concerned with pattern recognition and classification. The basic dimensions of a crime incident are space, time, law, offender, and target or victim details (Verma, 2002). However, Geo-politics, economy, education, housing, population, employment rate, and the availability of healthcare facilities are the few metrics influencing criminal events (De Nadai, 2020; Sugiharti, 2023). Since conventional methods of forecasting and investigating are inadequate in the modern era, Machine Learning strategies have emerged. In recent years, the limitations of manual investigation processes have been gradually overcome by AI-based approaches.

McClendon and Meghanathan examined the patterns in the UCI and Mississippi crime data sets. Data scaling was used because the dimensions of the data sets differed. The Linear Regression (LR) algorithm handled data randomness better than the Additive Regression and Decision Stump algorithms (McClendon, 2015). Using the LR model, Awal et al. forecasted murder, sexual repression, abductions, and theft in various regions of Bangladesh. It denoted the relationship between population growth and the occurrence of crime (Awal, 2016). Sukhija et al. investigated the role of social factors in rape events in Haryana, India, using Multiple Linear regression. The authors discovered a link between the urban

population, literacy rate, employment opportunities, police station density, and sexual offences (Sukhija, 2020). The economic status of the residents has a significant impact on the neighborhood's amenities. Xu et al. investigated the availability of streetlights at crime scenes in Detroit. According to the proposed model, the presence of streetlights must be considered an indicator of a secure location (Xu, 2018).

Safat et al. classified crime incidents in Chicago and Los Angeles. On the data sets, Decision Tree, KNN, Logistic Regression, Multi-layer Perceptron (MLP), Naive Bayes, Random Forest, Support Vector Machine (SVM), and XGBoost models were evaluated. The XGBoost achieved the highest accuracy (94%) in Chicago, while the k-nearest neighbor (KNN) narrowly outperformed XGBoost in Los Angeles, by achieving 89% accuracy. The five-year crime counts were projected using the LSTM and ARIMA models. The ARIMA model outperforms the LSTM in both cities, with lower RMSE and MAE values (Safat, 2021). The urban population fluctuates throughout the day because of people migrating. Poongodi et al. forecasted travel duration and fare by analyzing taxi trips and New York City climatic data. MLP and XGBoost algorithms were used to process the spatial-temporal attributes. The RMSE of the XGBoost model (0.44) was lower than that of the MLP model (Poongodi, 2022). Ratcliffe et al. tested a hotspot identification platform based on Gradient Boosting in Philadelphia. The predictive policing programme forecasted both property and violent crimes (Ratcliffe, 2021). The patrolling strategies were suggested based on the hotspot identified. As a result, 31% of the decrease was observed in the property crime segment; however, the application provided a subpar projection for violent crimes.

Tariq et al. proposed an RNN-LSTM model to forecast Philadelphia's daily and monthly violent crime rates. In comparison to the ARIMA, Simple Exponential Smoothing (SES), and Holt-Winters (HW) models, RNN-LSTM performed better, with RMSE values of 13.4 and MAPE values of 4.75 (Tariq, 2021). Devi et al. recommended an ensembled RNN-LSTM model for forecasting Sacramento's illicit activities. The proposed N-Beats RNN achieved the best results compared to the other models, with an MAE of 0.1407 (Devi, 2021). The impact of seasonal components on crime rate predictions was investigated by Cruz-Nájera et al. using a variety of statistical, machine learning, and deep learning models. The projections obtained via a simple moving average (SMA) using seasonal removal techniques are superior to those made using ARIMA, LightGBM, and ANN (Cruz-Nájera, 2022). Incorporating spatiotemporal correlations in crime data and making use of supplementary data to improve regional spatial characteristics, Dong et al. suggested a crime rate prediction model based on 2D convolution and long short-term memory neural network (2DCONV-LSTM). The study used extensive experiments to demonstrate that recording both temporal and spatial correlations in crime data, as well as employing auxiliary data to construct regional spatial features, improves prediction performance. Compared to Support Vector Regression and LSTM, the recommended framework reduced prediction error by 17.8% and 8.2%, respectively (Dong, 2022).

Approaches, methods, and models from other domains could also be used to improve the legacy performance of the system. The proposed approach, whether direct or transformative, should produce the desired result during practical implementation. The model's success could only be measured by its efficacy and practicality. To generate short-term forecasts at a reasonable cost, univariate methodologies would be most suitable (Weller, 1979). In this study, we attempted to predict CSA events using a stacked machine-learning model.

3 Materials

3.1. Data Set

The data set used in this study was obtained from StatBank Denmark and comprises information on sexual crimes against children, as well as the region, municipality, offence categories, and quarterly crime details (Statistikbanken, 2023). The derived data set contains 5430 crime incidents from January 1, 2007, to December 31, 2022. Denmark is divided into five provinces: Hovedstaden, Sjaelland, Syddanmark, Midtjylland, and Nordjylland. The regions collectively contain 98 municipal divisions and Christiansø. Meanwhile, the location of some offences was mapped under 'Unknown Municipality' in this data set. In the category of sexual crimes against children, there are four types of offences.

- 1. Heterosexual offences against a child under 12 (Repealed in 2013)
- 2. Homosexual offences against a child under 12 (Repealed in 2013)
- 3. Sexual offences against a child under 12 (New from 2013)
- 4. Sexual offences against a child under 15 (New from 2013)

Because the events occurred against children, the crime categories were removed from the data set. There are no missing values in this data set; however, the crimes classified as unknown municipalities were treated as a separate region and municipality to effectively analyze and classify the data. The municipalities were geolocated using latitude and longitude coordinates. The finalized Denmark Child Sexual Abuse (DCSA) data set includes 100 rows with 64 attributes. Table 1 details the attributes of the DCSA data set used for the exploratory analysis. The crime counts and timeframes were transposed, and a new data set with quarterly information (QoQ) was created. It was named DCSA-QoQ and was used to forecast CSA occurrences. Table 2 outlined the attributes of the same.

Attribute name	Description	Data
		Туре
Region	Name of the Region	String
Municipality	Name of the Municipality	String
Latitude	Latitude of the Municipality	Float
Longitude	Longitude of the Municipality	Float
Quarter on Quarter Crime	Year 2007 to 2022 – 64 Columns with number of CSA	Integer
Count	events	

Table 1: Attributes Details of the DCSA Data set

Table 2: Attributes Details of the DCSA-QoQ Data set

Attribute name	Description	Data Type
Timeline	Quarter on Quarter	Date
Crime Count	Number of CSA occurred in each Quarter	Integer

3.2. Proposed Methodology

The proposed methodology for this study is depicted in Fig. 1. It is divided into the following sections:

Phase I: Regions and municipalities with geographical details were segmented to perform the exploratory analysis. Using visualization techniques, the extracted knowledge was graphically represented.

Phase II: XGBoost, LSTM and Random Forest algorithms were combined to forecast CSA crimes. The stacking methodology aimed to improve the performance of the meta-regressor.

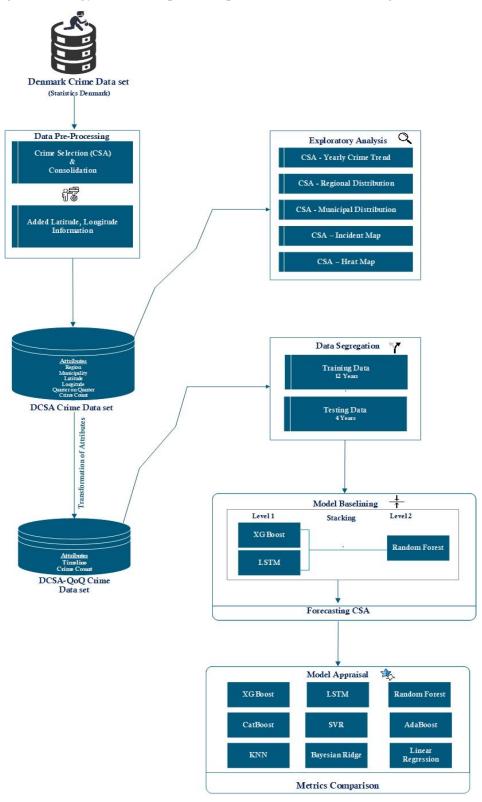


Figure 1: Proposed methodology

4 Exploratory Analysis

Exploratory analysis helps to decode the characteristics of data through visualization. The relationship between the attributes, data distribution, outliers, trends, and patterns could be determined by examining the graphical representations. Concurrently, it is advantageous to identify appropriate prediction methods and tools. The subsequent images depicted the characteristics of the Denmark Child Sexual Abuse (DCSA) data set.

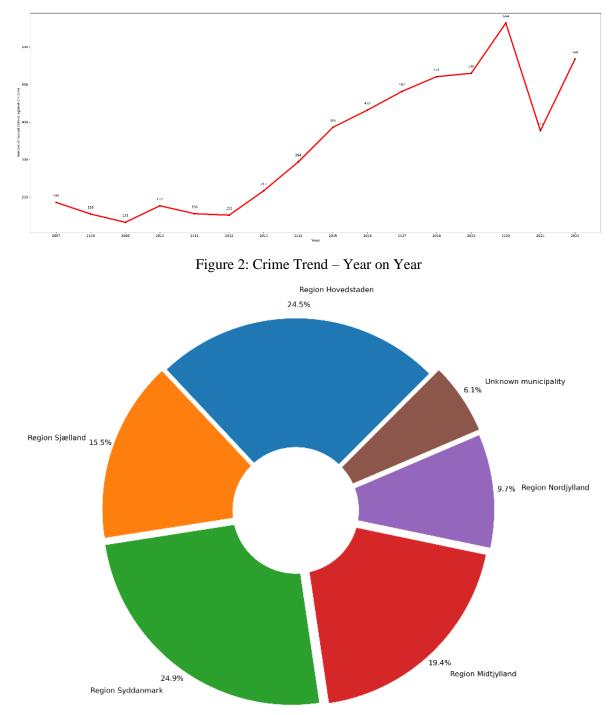


Figure 3: Percentage of CSA - Region wise

CSA-Forecaster: Stacked Model for Forecasting Child Sexual Abuse

The trend of child sexual abuse in Denmark from 2007 to 2022 is depicted in Fig. 2. During the years 2007 to 2009, there was a declining trend. In 2010, it increased before following a downward trend for several years. Although the number of CSA offences increased dramatically from 2013 to 2020, they reached their peak at 664 in 2020. However, the number of child abuse cases declined considerably to 377 in 2021, only to rise again the following year. Fig. 3 depicts the incidence rate of CSA in each Danish region. Regions Syddanmark, Hovedstaden, Midtjylland, Sjaelland, and Nordjylland accounted for 24.9%, 24.5%, 19.4%, 15.5%, and 9.7% of CSA incidents respectively. The events that occurred under the division of Unknown Municipality were 6.1%. Syddanmark is the third largest populated and the second least densely populated region in the country (Wikipedia, 2023). It is more acquiescent to sexual crimes against children. Hovedstaden, the most populous and densely populated region of Denmark which includes Copenhagen, reported a higher number of CSA offenses, following the Syddanmark region.

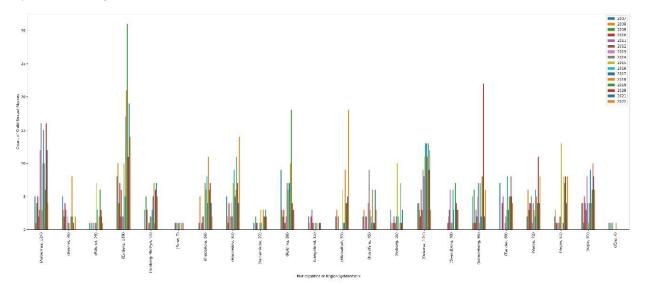


Figure 4: CSA - Region syddanmark

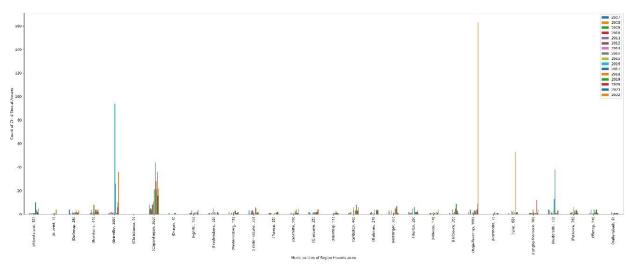


Figure 5: CSA - Region hovedstaden

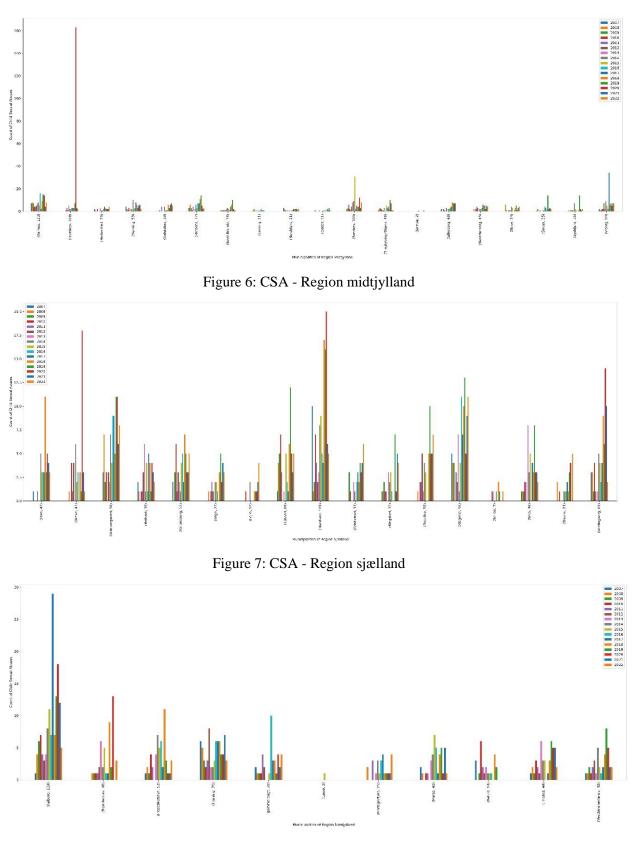


Figure 8: CSA - Region nordjylland

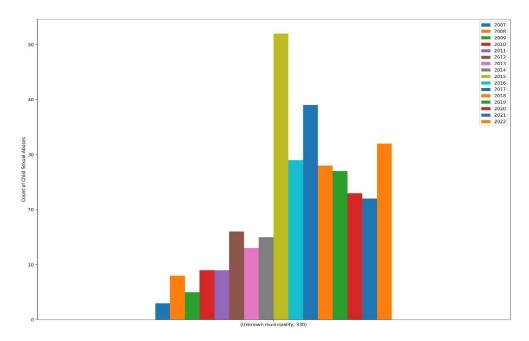


Figure 9: CSA - Unknown municipality

Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8 represent the municipalities listed under each region as well as their respective crime rates. In the Region Syddanmark, the municipalities of Esbjerg (169), Odense (124), and Aabenraa (124) attained the top three positions in terms of the number of CSA incidents, while Ærø had only four reported incidents (Refer to Fig. 4). The municipalities of Copenhagen (262), Høje-Taastrup (199) and Brøndby (180) reached the top positions in Region Hovedstaden. Christiansø was the best municipality for children with no CSA offences during the period (Refer to Fig. 5).

The municipalities of Favrskov (198), Aarhus (123), and Randers (104) were identified as the most vulnerable in the Region of Midtjylland. The municipality of Samsø (2) in the same region was comparatively safer for children (Refer to Fig. 6). In the region of Sjlland, Næstved (119) obtained the topmost place which was followed by Slagelse (96), and Guldborgsund (90). There were only seven such offences notified at the municipality of Solrød (Refer to Fig. 7). Aalborg (139), Hjørring (70), and Frederikshavn (53) emerged as the top locations with the highest number of incidents in the Region of Nordjylland. In the year 2015, one CSA offence occurred in the municipality of Læsø (Refer to Fig. 8). There were 330 documented CSA instances in the category Unknown Municipality (Refer to Fig. 9).

S.No	Municipality	Total	S.No	Municipality	Total
1	Unknown municipality	330	14	Slagelse	96
2	Copenhagen	262	15	Guldborgsund	90
3	Høje-Taastrup	199	16	Sønderborg	86
4	Favrskov	198	17	Kolding	86
5	Brøndby	180	18	Haderslev	83
6	Esbjerg	169	19	Vejle	81
7	Aalborg	139	20	Horsens	77
8	Odense	124	21	Rudersdal	73
9	Aabenraa	124	22	Varde	72
10	Aarhus	123	23	Hjørring	70
11	Næstved	119	24	Vordingborg	69
12	Randers	104	25	Ishøj	69
13	Viborg	99			

Table 3: CSA – Top 25 Municipalities

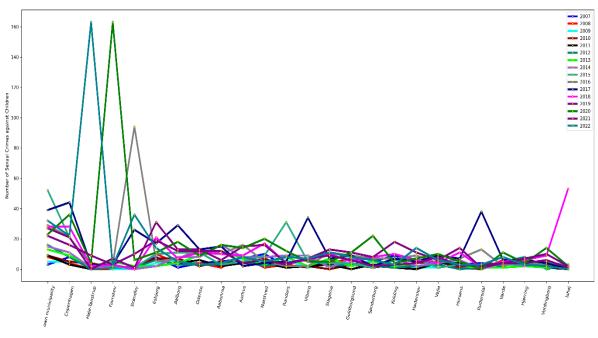


Figure 10: CSA - Top 25 municipalities - Year-wise trend

Table 3 listed the top 25 municipalities that were more vulnerable to CSA issues. Although the unknown municipality topped this section, the events in that segment occurred all over the country. Furthermore, Copenhagen was a perilous place for children. Fig. 10 reveals that Favrskov and Høje-Taastrup obtained the peak in the years 2020, 2022, respectively. The vicissitudes of socioeconomic-environmental factors might have caused the sudden increase in CSA incidents.

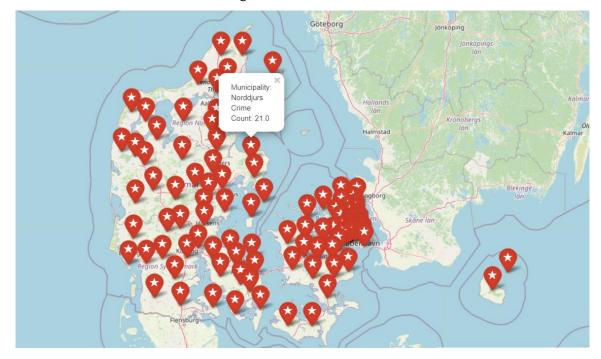


Figure 11: Denmark - CSA Incident Count - Map

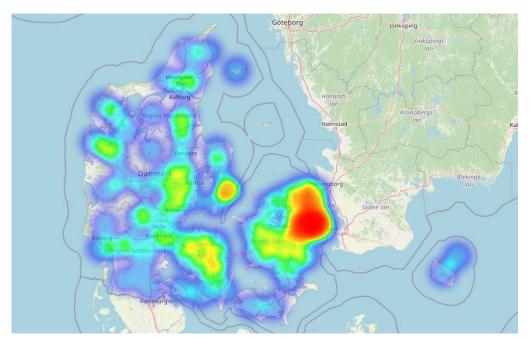


Figure 12: Denmark - CSA Heat Map

According to the social disorganization theory, geographical factors play a significant role in crime occurrence. Plotting crime events alongside geographic details would provide additional information. The folium-aided map in Fig. 11 includes crime data for each municipality based on their latitude and longitude coordinates. The markers were updated according to the zoom level. By clicking on the marker, the details about a particular municipality would be visible. The heat map embodied in Fig. 12 indicates the density of the CSA events. It confirms that the Region Hovedstaden is an alarmingly vulnerable location for child sexual exploitation.

5 Experiments and Results

The transformed DCSA-QoQ data set contains temporal data about CSA incidents in Denmark. Fig. 13 depicts the partitioning of the min-max normalized data into training and testing sets. The training data consists of the number of incidents from 2007 to 2018, while the testing data covers the years 2019 through 2022. Each node in the diagram represents the quarterly cumulative counts. In this study, the training and testing data were segmented using the rolling window technique.

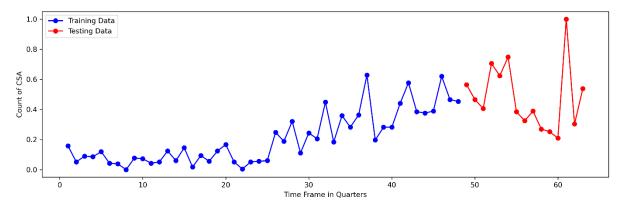


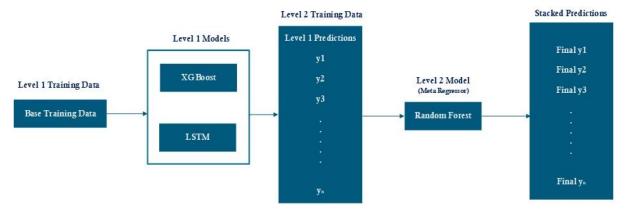
Figure 13: DCSA-QoQ - Training vs Testing data

5.1. Evaluation Metrics

The Machine Learning models' corollaries were evaluated by calculating the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) as depicted in Eq. (1 – 4).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} (1)$$
$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| (2)$$
$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)}{y_i} (3)$$
$$R^2 = 1 - \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} (4)$$

'n' indicates the number of samples, y_i specifies the actual value, \hat{y}_i represents the predicted value, \bar{y} represents the sample mean.



5.2. Stacking Model

Figure 14: CSA-Forecaster stacked model

CSA-Forecaster Stacked Model

Input: D(x, y)

Parameters:

n: Number of iterations for XGBoost,

- λ : Regularization parameter for XGBoost,
- m_t: Input at current timestep for LSTM,

 n_{t-1} : Hidden state from previous timestep for LSTM,

num_trees: Number of trees in the RF,

max_features: Maximum number of features considered for each split in RF,

min_samples_split: Minimum number of samples required to perform a split in RF,

max_depth: Maximum depth of each tree in RF,

random_state: Random seed for reproducibility in RF

Output: CSA – Forecaster (x)

- 1. Read the dataset D(x, y)
- 2. Initialize the parameters for XGBoost
- 3. For i = 1 to n
 - 3.1 Calculate the loss function, , $L(a_i, p_i) = \frac{1}{2}(a_i, p_i)$
 - 3.2 Build the trees, $X = \sum_{i=1}^{n} L(a_i, p_i) + \frac{1}{2}\lambda 0^2$

3.3 Determine the first order derivative (g_i) , second order derivative (h_i) and

plug the Taylor approximation, $g_i = \frac{d}{dp_i} \left(\frac{1}{2}(a_i, p_i)^2\right)$ and $h_i = \frac{d^2}{dp_i^2} \left(\frac{1}{2}(a_i, p_i)^2\right)$ 3.4 Estimate the output score for the enduring nodes, $0 = \frac{g_i}{h_i + \lambda}$

- 3.5 Compute the similarity scores, S = $\frac{\sum_{i=1}^{n} E^2}{N + \lambda O^2}$
- 3.6 Calculate the gain ratio, G_i = S_l + S_r S_R
- 3.7 Trim the tree, $X = G \gamma$
- 4. End For
- 5. Evaluate the base learners:
 - 5.1. Initialize the ensemble predictions $p_i(x) = 0$
 - 5.2 For each base learner T:
 - 5.2.1 Evaluate the base learner on the dataset D(x, y) to obtain O
 - 5.2.2 Update the ensemble predictions $p_i(x) = p_{(i-1)}(x) + 0$
- 6. Result : XGB(x)
- 7. Prepare the training and testing set for LSTM
- 8. Concatenate the input and hidden state: concat = $[m_t, n_{t-1}]$
- 9. Calculate the forget gate: $F = sigmoid(W_f * concat + b_f)$
- 10. Calculate the input gate: $I = sigmoid(W_i * concat + b_i)$
- 11. Calculate the candidate cell state: $CS = tanh(W_c * concat + b_c)$
- 12. Calculate the update to the cell state: $c_t = forget_gate * c_{t-1} + input_gate * CS$
- 13. Calculate the output gate: Out = sigmoid($W_0 * \text{ concat} + b_0$)
- 14. Calculate the updated hidden state: $h_t = out * tanh(c_t)$
- 15. Return the updated hidden state (h_t) and cell state (c_t)

- 16. Result: LSTM(x)
- 17. Prepare the meta regressor, Random Forest
- 18. Construct a new dataset as D(X, Y) by combining the XGB(x) and LSTM(x)
- 19. Initialize an empty list to store the trees: forest = []
- 20. For t = 1 to num_trees:
 - 20.1 Create a bootstrap sample B(t)
 - 20.2. If the max_depth is reached || the number of samples in $B(t) < \min_{s}$ split:
 - 20.2.1. Create a leaf node with the majority class in B(t)
 - 20.3. Else:
 - 20.3.1. Randomly select a subset of features of size max_features
 - 20.3.2. Identify the best feature and split point that maximize GI on B(t)
 - 20.3.3. Split B(t) into B_left(t) and B_right(t), based on the best split
 - 20.3.4. Create a decision node with the selected feature and split point
 - 20.3.5. Repeat steps 3.2.1 and 3.2.2 until reaching the max _depth or end
- 21. Add the trained tree T(t) to the forest
- 22. Return the RF model
- 23. Generate the ensemble predictions using the trained T(t)
- 24. Return the ensemble predictions
- 25. Result: CSA Forecaster (x)
- 26. End CSA Forecaster Algorithm

Stacking is a method for combining predictive models and improving their performance. The stacking regression approach could include 'm' regression algorithms aligned at multiple levels. The base-level models are trained with complete training data. The outcomes of the base models were infused as input to the meta-regressor at the next level. The bias identified at the base level would be helpful to improvise the outcome of the meta-regressor. The recurrent grouping of various ML models could be useful to find a better amalgamation. Since the stacking approach yields better denouements, it is exerted to build classification and regression models (Srinivas, 2020; Saha, 2021). Hence, the stacking approach was implemented in this study.

Two levels of the proposed model are depicted in Fig. 14. XGBoost and LSTM were deployed in Level 1 and Random Forest was employed as meta regressor at Level 2. Using training data as a foundation, the XGBoost and Long short-term memory (LSTM) algorithms predicted the possible count of CSA occurrences. XGBoost algorithm keeps all residuals at one leaf, computes the similarity score, and divides the population into two groups based on the score. The gain is used to validate the clustered entities. The leaves are continuously split until the end node is found, at which point a value known as gamma is assigned. If the difference between leaf gain and gamma is negative, the tree's particular branch would be pruned. The XGBoost algorithm constructs trees one at a time, learning from the errors discovered at each level. It reduces loss and builds trees by using the loss function. Eq. (5) represents the calculation of the loss function.

CSA-Forecaster: Stacked Model for Forecasting Child Sexual Abuse

$$L(\Phi) = \sum_{i} l(\hat{y}_{i}, y_{i}) + \sum_{k} \Omega(f_{k}) \text{ Where } \Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^{2} (5)$$

'i' indicates the number of samples, \hat{y}_i represents the predicted value, y_i specifies the expected value, $\Omega(f)$ represents the complexity, 'k' designates the number of trees, 'T' characterizes the number of leaves, γ and λ determines the degree of complexity, and 'w' denotes the weights of leaves.

In this study, XGBoost was employed at level 1 to estimate the occurrence of CSA offenses. The advantage of using XGBoost as a level 1 regressor is its ability to handle complex relationships and capture non-linear patterns in the data. The results were then combined with those of another Level 1 regressor, LSTM. For time series data, where the sequence of the data points across time is critical, LSTM models thrive. Typically, time series data is presented as a series of observations across time, with a timestamp attached to each observation. LSTM models are particularly effective at identifying recurring structures in time series data. LSTM has memory cells, gates, and secret states in its design. Memory cells allow the model to retrieve information from previous iterations, while gates regulate the flow of information into and out of the cells. By utilizing these principles, the LSTM effectively captured both the short- and long-term dependencies present in each time series. The model was trained to make predictions about future values by analyzing the data and minimizing a loss function that measures the deviation between predictions and actual values. After training, the LSTM model made predictions based on completely novel time series inputs. The model forecasted the future values by running the input sequence through LSTM layers and using the results. One fully linked dense layer was used to transfer the LSTM's hidden states to the anticipated values in the output layer. LSTM model successfully captured complicated patterns and produced accurate predictions in time series data because of their memory cells, gates, and capacity to capture temporal dependencies.

The input gate (i) of the LSTM determines how much information is added to the state of the cell at any given time. The forget gate (f) controls how much data from the previous cell state is forgotten. What data would be added to the cell state is decided in the Cell state update (g). The memory of an LSTM cell is stored in its cell state (denoted by C). The output gate (o) controls how much of the newly updated cell state is output as the hidden state. The hidden state (h) represents the output of the LSTM cell. (Refer to Eq. 6 - 11).

$$i(t) = \sigma(W_i [h(t-1), x(t)] + b_i)$$
(6)

$$f(t) = \sigma(W_f [h(t-1), x(t)] + b_f)$$
(7)

$$g(t) = tanh(W_g [h(t-1), x(t)] + b_g)$$
(8)

$$C(t) = f(t) * C(t-1) + i(t) * g(t)$$
(9)

$$o(t) = \sigma(W_o [h(t-1), x(t)] + b_o)$$
(10)

$$h(t) = o(t) * tanh(C(t))$$
⁽¹¹⁾

where:

i(t) is the input gate at time t.

 σ represents the sigmoid activation function.

 W_i is the weight matrix associated with the input gate.

h(t-1) is the previous hidden state.

x(t) is the input at time t.

 b_i is the bias term associated with the input gate.

f(t) is the forget gate at time t.

 W_f is the weight matrix associated with the forget gate.

 b_f is the bias term associated with the forget gate.

g(t) is the cell state update at time t.

tanh represents the hyperbolic tangent activation function.

 W_q is the weight matrix associated with the cell state update.

 b_q is the bias term associated with the cell state update.

C(t) is the cell state at time t.

C(t-1) is the previous cell state.

o(t) is the output gate at time t.

 W_o is the weight matrix associated with the output gate.

 b_o is the bias term associated with the output gate.

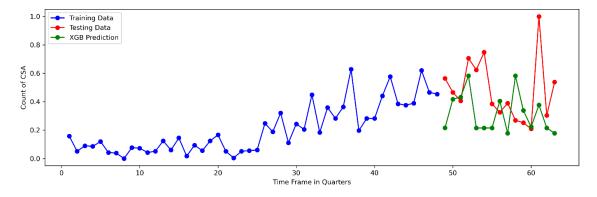
h(t) is the hidden state at time t.

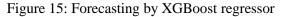
o(t) is the output gate at time t.

In this study, the Random Forest regressor was utilized at level 2 for the purpose of enhancing the prediction. Random Forest regressor is designed to aggregate the predictions of numerous decision trees to construct a robust and accurate regression model. It generates multiple decision trees simultaneously, each trained on a random subset of the training data. The randomness in the training process helps diversify the decision trees and reduce overfitting. Each decision tree independently predicts the target variable using its own randomly selected features at each branch. The predictions of all the decision trees are then combined to generate the final forecast. Typically, the final forecast is calculated as the mean or median of the individual predictions made by the ensemble of decision trees.

As a meta-regressor, Random Forest aggregated the predictions of the foundational models. It learned to weigh and combine predictions depending on their performance on training data. It reduced the biases of the base models by considering their individual biases and modifying their contributions to the final prediction. The aggregation procedure helped to balance out the biases and harness the diversity of the underlying models, resulting in a more balanced and accurate forecast. Additionally, it gained knowledge from the mistakes made by the XGBoost and LSTM models. Random Forest discovered patterns and made modifications to enhance overall prediction accuracy by assessing the gaps between the predictions of the base models and the true target values.

CSA-Forecaster: Stacked Model for Forecasting Child Sexual Abuse





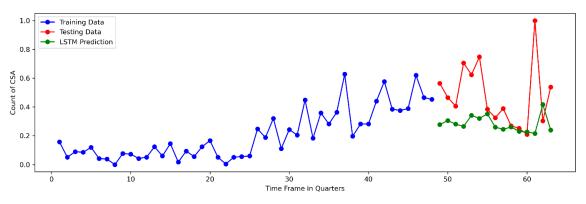


Figure 16: Forecasting by LSTM

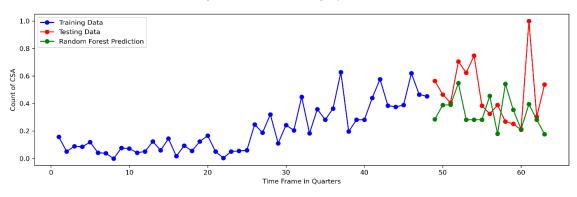


Figure 17: Forecasting by random forest

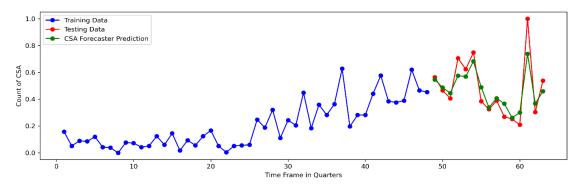


Figure 18: Forecasting by CSA-Forecaster

Model	RMSE	MAE	MAPE	R ²
CSA-Forecaster	0.094	0.0712	0.1557	0.8028
XGBoost	0.2946	0.2289	0.4475	-0.9378
LSTM	0.2969	0.2136	0.3579	-0.9675
Random Forest	0.2644	0.2025	0.3884	-0.5608
CatBoost	0.2754	0.218	0.4415	-0.69369
SVR	0.2798	0.194	0.3508	-0.74736
AdaBoost	0.2685	0.2123	0.4266	-0.60962
KNN	0.2471	0.1801	0.34	-0.36303
Bayesian Ridge	0.2808	0.1857	0.3562	-0.76018
Linear Regression	0.2816	0.1857	0.359	-0.77016

Table 4: KPI Evaluation – CSA-Forecaster vs State of art models

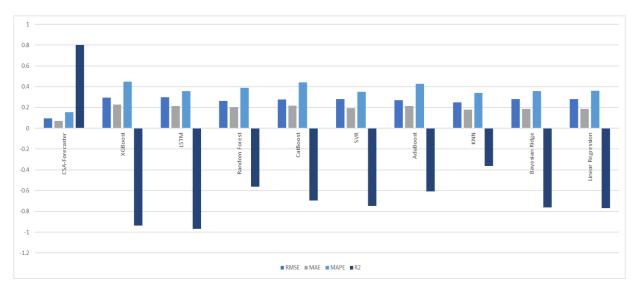


Figure 19: RMSE - MAE - MAPE - R2 of CSA-Forecaster vs State of art models

The predictions of the XGBoost, LSTM, and Random Forest models are illustrated in Figs. 15, 16 and 17, respectively. Fig. 18 denotes the propitious prognostication of the proposed CSA-Forecaster model. Table 4 compares the performance of the CSA-Forecaster model to that of other cutting-edge models such as XGBoost, LSTM, Random Forest, CatBoost, Support Vector Regression (SVR), AdaBoost, KNN, Bayesian Ridge, and Linear Regression. The base models, as well as other state-of-the-art models, exhibited a notable number of forecasting errors, as evidenced by the RMSE, MAE, and MAPE measures. However, the CSA-Forecaster model surpassed both the base models and other state-of-the-art models in terms of performance. The proposed approach substantially decreased the forecast error, reduced the standard deviation of residuals, and minimized the average magnitude of errors. The R-Squared value of the model indicated a lower variance, as depicted in Fig. 19. Considering these metrics collectively, it could be concluded that the CSA-Forecaster model is highly effective in predicting CSA incidents.

6 Conclusion and Future Work

Crimes are complex events influenced by a variety of social, economic, and environmental factors. Given the lasting impact of criminal offenses on both individuals and society, legal entities prioritize proactive measures. While all forms of crime should be eliminated, particular emphasis is placed on addressing the most severe ones, such as crimes against children. Child sexual abuse profoundly affects the physical, mental, and emotional well-being of the victims, leaving lasting scars that continue to haunt them throughout their lives. Confronting child sexual abuse necessitates the design and implementation of effective strategies to combat and prevent such heinous acts.

In this study, XGBoost, LSTM and Random Forest regression models were stacked together to forecast child sexual abuse in Denmark. The proposed bi-level stacked CSA-Forecaster model was evaluated against base models. The model's RMSE (0.094), MAE (0.0712), MAPE (0.1557), and R² (0.8028) indicate that it is superior to the others. This model could be useful to forecast CSA events and devise countermeasures. As a future work, we planned to create a classification model that could predict CSA occurrences by incorporating additional attributes, such as demographic details of offenders and victims, motivation of offences, place of occurrence, socioeconomic-environmental characteristics of locality, and status of recidivism.

Author Contributions: Study conception and design: Saravanan Parthasarathy; Data curation and visualization: Saravanan Parthasarathy, A. Abdul Azeez Khan, K. Javubar Sathick; Analysis and interpretation of results: Saravanan Parthasarathy; Draft manuscript preparation: Saravanan Parthasarathy; Supervision, Validation, Reviewing: Arun Raj Lakshminarayanan.

Availability of Data and Material: The dataset was derived from Statistikbanken. [Online]. Available: https://www.statbank.dk/STRAF11.

Funding: This research received no external funding.

Conflict of Interest: The authors declare no conflict of interest.

Acknowledgments: We would like to thank the 'Statistics Denmark' platform for providing publicly available data which was used in this study.

References

- [1] Andrejevic, M. (2017). Digital citizenship and surveillance To Pre-empt a thief. *International Journal of Communication*, *11*, 879–896.
- [2] Awal, M.A., Rabbi, J., Hossain, S.I., & Hashem, M.M.A. (2016). Using linear regression to forecast future trends in crime of Bangladesh. *In 5th international conference on informatics, electronics and vision (ICIEV)*, 333-338.
- [3] Buch Leander, E.M., Pallesgaard Munk, K., & Lindsø Larsen, P. (2019). Guidelines for preventing child sexual abuse and wrongful allegations against staff at Danish childcare facilities. *Societies*, 9(2), 42. https://doi.org/10.3390/soc9020042
- [4] Christensen, L. (2017). Victims of child sexual abuse: The psychology of victims. *In the Psychology of Criminal and Antisocial Behavior*, 419-438.
- [5] Cruz-Nájera, M.A., Treviño-Berrones, M.G., Ponce-Flores, M.P., Terán-Villanueva, J.D., Castán-Rocha, J.A., Ibarra-Martínez, S., & Laria-Menchaca, J. (2022). Short Time Series Forecasting: Recommended Methods and Techniques. *Symmetry*, 14(6), 1231. https://doi.org/10.3390/sym14061231
- [6] Dayi, A.K., Rodoplu, V., Nakip, M., Pehlivan, B., & Güzelis, C. (2022). Multi-Channel Subset Iteration with Minimal Loss in Available Capacity (MC-SIMLAC) Algorithm for Joint

Forecasting-Scheduling in the Internet of Things. Journal of Wireless Mobile Networks Ubiquitous Comput. Dependable Appl., 13(2), 68-95.

- [7] De Nadai, M., Xu, Y., Letouzé, E., González, M.C., & Lepri, B. (2020). Socio-economic, built environment, and mobility conditions associated with crime: a study of multiple cities. *Scientific reports*, *10*(1), 13871. https://doi.org/10.1038/s41598-020-70808-2
- [8] Devi, J.V., & Kavitha, K.S. (2021). Automating Time Series Forecasting on Crime Data using RNN-LSTM. *International Journal of Advanced Computer Science and Applications*, *12*(10). 458-463.
- [9] Dong, Q., Ye, R., & Li, G. (2022). Crime amount prediction based on 2D convolution and long short-term memory neural network. *ETRI Journal*, 44(2), 208-219.
- [10] EUCPN, Danish Policy on Sexual crime, 2022. https://eucpn.org/document/danish-policy-on-sexual-crime
- [11] Helweg-Larsen, K., & Larsen, H.B. (2006). The prevalence of unwanted and unlawful sexual experiences reported by Danish adolescents: Results from a national youth survey in 2002. *Acta Paediatrica*, 95(10), 1270-1276.
- [12] Jaffe, A.E., Brockdorf, A.N., Messman, T.L., Gratz, K.L., & DiLillo, D. (2023). Is the risk for sexual revictimization cumulative? A prospective examination. *Women's health issues*, 33(2), 208-214.
- [13] Landau, A.Y., Ferrarello, S., Blanchard, A., Cato, K., Atkins, N., Salazar, S., & Topaz, M. (2022). Developing machine learning-based models to help identify child abuse and neglect: key ethical challenges and recommended solutions. *Journal of the American Medical Informatics Association*, 29(3), 576-580.
- [14] Leander, E.M.B. (2023). Children's Sexuality and Nudity in Discourse and Images in a Danish Education and Care Journal over 50 Years (1970–2019): The Emergence of "The Child Perpetrator of Sexual Abuse" in an International Perspective. *Archives of sexual behavior*, 52(1), 49-78.
- [15] Lupariello, F., Sussetto, L., Di Trani, S., & Di Vella, G. (2023). Artificial intelligence and child abuse and neglect: a systematic review. *Children*, 10(10), 1659. https://doi.org/10.3390/children10101659
- [16] McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, 2(1), 1-12.
- [17] Naidu, K. (2020). This Kailash Satyarthi interview is a reminder that the worst form of child labour is child porn — an \$8 bill. Business Insider. https://www.businessinsider.in/how-big-isthe-child-porn-industry-explains-kailash-sathyarthi/articleshow/70443534.cms
- [18] Perry, W.L. (2013). Predictive policing: The role of crime forecasting in law enforcement operations. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233. pdf
- [19] Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., & Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. *International Journal of System Assurance Engineering and Management*, 1-12.
- [20] Ratcliffe, J.H., Taylor, R.B., Askey, A.P., Thomas, K., Grasso, J., Bethel, K.J., & Koehnlein, J. (2021). The Philadelphia predictive policing experiment. *Journal of Experimental Criminology*, 17, 15-41.
- [21] Safat, W., Asghar, S., & Gillani, S.A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. *IEEE access*, *9*, 70080-70094.
- [22] Saha, M., Santara, A., Mitra, P., Chakraborty, A., & Nanjundiah, R.S. (2021). Prediction of the Indian summer monsoon using a stacked autoencoder and ensemble regression model. *International Journal of Forecasting*, 37(1), 58-71.

- [23] Saraiva, M., Matijošaitienė, I., Mishra, S., & Amante, A. (2022). Crime prediction and monitoring in porto, portugal, using machine learning, spatial and text analytics. *ISPRS International Journal of Geo-Information*, 11(7), 400. https://doi.org/10.3390/ijgi11070400.
- [24] Smith, B.M. (2018). Can we predict when and where a crime will take place? BBC News. https://www.bbc.com/news/business-46017239
- [25] Srinivas, K., Rani, B.K., Rao, M.V.P., Patra, R.K., Madhukar, G., & Mahendar, A. (2020). Prediction of heart disease using hybrid linear regression. *European Journal of Molecular & Clinical Medicine*, 7(5).
- [26] Statista, Denmark: number of sex crimes, (2022). https://www.statista.com/statistics/576081/number-of-sexual-crimes-in-denmark/
 [27] Statista and Statistics of the second sec
- [27] Statistikbanken. https://www.statbank.dk/STRAF11
- [28] Sugiharti, L., Purwono, R., Esquivias, M.A., & Rohmawati, H. (2023). The nexus between crime rates, poverty, and income inequality: A case study of Indonesia. *Economies*, 11(2), 62. https://doi.org/10.3390/economies11020062
- [29] Sukhija, K., Singh, S.N., & Kumar, M. (2020). Using Linear Regression to investigate parameters associated with Rape crime in Haryana. In 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 107-111.
- [30] Tariq, H., Hanif, M.K., Sarwar, M.U., Bari, S., Sarfraz, M.S., & Oskouei, R.J. (2021). Employing Deep Learning and Time Series Analysis to Tackle the Accuracy and Robustness of the Forecasting Problem. *Security and Communication Networks*, 1-10.
- [31] UNICEF Sexual violence against children (2023). https://www.unicef.org/protection/sexual-violence-against-children
- [32] Verma, A., & Lodha, S.K. (2002). A typological representation of the criminal event. *Western Criminology Review*, *3*(2), 37-48.
- [33] Weller, D., Hoover Institution, Ctr for Econometric Studies of the Justice System, & United States of America. (1979). Forecasting Crime Rates: A Review of the Available Methodology. Center for Econometric Studies of the Justice System, Hoover Institution, Stanford University. https://www.ojp.gov/ncjrs/virtual-library/abstracts
- [34] Wikipedia contributors (2023,). Regions of Denmark. Wikipedia. https://en.wikipedia.org/wiki/Regions_of_Denmark
- [35] Xu, Y., Fu, C., Kennedy, E., Jiang, S., & Owusu-Agyemang, S. (2018). The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan. *Cities*, 79, 45-52.
- [36] Zarchev, M., Ruijne, R.E., Mulder, C.L., & Kamperman, A.M. (2022). Prevalence of adult sexual abuse in men with mental illness: Bayesian meta-analysis. *B J Psych open*, 8(1), e16. https://doi.org/10.1192/bjo.2021.1069

Authors Biography



Saravanan Parthasarathy obtained his bachelor's degree in computer science and engineering from the Madurai Kamaraj University, Madurai in 2003. He completed his master's degree in computer science and engineering from Anna University, Chennai in 2015. He obtained his Postgraduate degrees in Psychology, Medical Sociology and Business Management from TNOU, Madras University, Alagappa University and Annamalai University. He is currently pursuing his doctorate from B.S. Abdur Rahman Institute of Science and Technology in the field of Crime Informatics. He has 13 years of experience in the software industry. At present, he is affiliated to B.S. Abdur Rahman Crescent Institute of Science and Technology, as a full-time research scholar. He has published 15 research articles in international conferences and journals.



Arun Raj Lakshminarayanan embarked on his academic journey with a bachelor's degree in computer science and engineering from Anna University in 2006. He continued his pursuit of knowledge by completing his master's degree in computer science and engineering, also from Anna University, in 2010. His dedication to academia and research culminated in the successful completion of a Ph.D. in Information and Communication Engineering from Anna University in 2017. He currently serves as an "Associate Professor" at B.S. Abdur Rahman Crescent Institute of Science and Technology. The breadth of his knowledge and the impact he has had in academia highlight his dedication to computer science and engineering. His scholarly significance is further enhanced by the publication of 40 research articles in international conferences and journals.



A. Abdul Azeez Khan embarked on his academic journey by earning a bachelor's degree in computer science from the University of Madras in 2002. Building upon his foundation, he pursued a Postgraduate degree in Computer Science from the same institution in 2004. In 2008, he achieved his M.Phil degree from Periyar University. Driven by a thirst for knowledge, he went on to complete his doctorate from B.S. Abdur Rahman Institute of Science and Technology in 2018. With a cumulative experience of 13 years in both teaching and industry, Dr. A. Abdul Azeez Khan currently serves as an Associate Professor in the Department of Computer Applications at B.S. Abdur Rahman Crescent Institute of Science and Technology. His academic prowess is proved through his publication of 7 research articles in international journals and the presentation of 6 papers in international conferences.



K. Javubar Sathick received his Doctorate in Computer Science from B.S. Abdur Rahman Institute of Science and Technology in 2018. Prior to this, he completed his Postgraduate degree in Computer Applications from Anna University in 2008. With a robust background spanning 13 years in both teaching and industry, Dr.K. Javubar Sathick currently holds the position of Associate Professor in the Department of Computer Applications at B.S. Abdur Rahman Crescent Institute of Science and Technology. He has published 10 research articles in international journals. His research interests encompass Knowledge Management, E-Learning, Web Mining, Web Applications, Big Data, and Information Retrieval.