IDENTIFICATION OF PATTERNS IN CRIME RECORDS USING ENSEMBLE LEARNING APPROACH

MINI PROJECT REPORT

Submitted by

Kanaga Shanmugam P	**210701103**
Karan Balaji R S	**210701105**
Venkatesh V	**210701520**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Identification of Patterns in Crime Records using Ensemble learning Approach**" is the bonafide work of **"Karan Balaji R S (210701105), Kanaga Shanmugam P(210701103) and Venkatesh V(210701520)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Rahul Chiranjeevi. V
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on ________________

Internal Examiner **External Examiner**

ABSTRACT

The unsolved cases in the records of police are often referred to as Cold Cases. The statistics of the crimes and attempt to crimes happening in India are eye opening. Close to 60,00,000 crime cases are registered in India every year and majority of these cases go unsolved and fall under the category of Cold Cases due to a number of factors like lack of evidence, lack of witness and lack of information. By using machine learning algorithms, the gap between the cases that are solved and the cases which are unsolved can be bridged. Taking in the important features of case files into consideration, the solved cases can be fed as an input to a model which is trained by clustering algorithms like K Means, DBSCAN etc. When the unsolved cases are fed to the trained model, the case files of the cluster can be suggested. These case files can then be used as references to solve the cases. The use case of this model can prove to be extremely beneficial in the context of solving age old cases where the records of those cases might be misplaced to lost or the evidences pertaining to those cases might not have been preserved for very long. While the project's scope is not restricted to solving crimes, if extended to the area of cybersecurity can have boundless applications as the threat level in the domain of cybersecurity is very high and if extrapolated, can help in protecting various cybercrimes from happening.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time. We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, **Rahul Chiranjeevi V** Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Karan Balaji R S-210701105
Kanaga Shanmugam P-210701103
Venkatesh V-210701520

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

OC	Organized Crimes
SVM	Support Vector Machines
CMS	Crime Monitoring System
KNN	K Nearest Neighbours
FIR	First Information Report
JSON	Java Script Object Notation

CHAPTER 1
INTRODUCTION

1.1 GENERAL

Crime is considered an action which constitutes an offence and unlawful, which can be punished by law. There are different types of crimes which are punishable by law. Some of them, not limited to, are, Organized Crimes (OC), Cyber Crimes, Arson, Burglary and many more.

1.2 OBJECTIVE

The main aim of this project is to solve many unsolved (ie) Cold Cases that are existing within the jurisdiction of law enforcement agencies. The number of such cases alarming and the number of such cases are on the rise. This presents a challenge that the law enforcement agencies and by using Machine Learning algorithms and retrieving patterns out of the historically solved cases, relationships can be made and thereby this can prove to be an investigative lead for the unsolved cases.

1.3 EXISTING SYSTEM

While there have been many researches in the domain of solving issues in the law enforcement using the recent technological advancements, there is no significant system which has deployed the researches made in this domain. The researches has mainly revolved around analysing the crimes around the world and gathering insights from it.

1.4 PROPOSED SYSTEM

This project takes in the historical case files that have been solved and converts the handwritten case files into digital format. By using these case files as a base, patterns are identified when a new/unsolved case file is given to the system. By using clustering algorithm and ensemble learning, similar case files which were solved are given as an output, this will help by providing a new investigative lead.

CHAPTER 2

LITERATURE SURVEY

This paper [1] discusses the importance and benefits of using ensemble learning methods in crime prediction and the advantages it holds over the conventional machine learning models when used as a single classifier. Due to the dynamic nature of crimes, it is difficult to find the right configuration for the datasets to feed as an input to the ensemble model. The paper proposes a model called assemble-stacking based crime prediction method (SBCPM) which applies SVM model to achieve domain specific configurations. This model achieves 99% classification accuracy on training data.

One of the main types of crime, organized crimes have been studied in this paper [2] by identifying organized crimes through social media analysis. By analysing language cues in social media as indicators, this paper classifies crime type and location. The system generates Organized crime concepts to alert analysts of potential criminal activity by grouping information sources. Analysts can also investigate organized crime concepts through a prototype software system featuring social media scanning and map-based visualization. The system is illustrated using human trafficking and modern slavery.

This paper [3]describes a Crime Monitoring System (CMS) designed to detect crimes in real-time using camera surveillance. It aims to overcome human limitations like slow reactions by combining CCTV cameras with deep learning techniques. The system operates in three stages: detecting weapons, violence, and recognizing faces. It achieves high accuracy rates in each stage: over 80% for weapons, 95% for violence, and 97% for faces. Real-world testing shows the system effectively detects crime and alerts authorities promptly, improving overall security and safety measures.

The study conducted in the area of cross domain learning on crime prediction [4] gives an insight about the data insufficiency problem faced in small cities. As the number of researches on crime prediction are on the rise, the urban data has become more accessible to the researchers. But this paper discusses on how to overcome data insufficiency in small cities with respect to Canada. By using ensemble learning as it generalizes new data and the classified data is compared against the baseline models.

This review paper [5] gives a take on Artificial Intelligence being used in the field of crime prediction. By intensively analysing various criteria, the models are evaluated. Intensive research is carried out after reviewing 120 research papers that were published between 2008 and 2021. The research has concluded that crimes and spatial are the most applied categories in analysing crimes. The various ML models used across the 120 research papers were noted and supervised learning models were found to be the major contributors with 31% while a combination of supervised and unsupervised learning models contributed with 22% and unsupervised learning models alone contributed with 10%.

This study [6] draws attention towards the importance of crime prediction and forecasting to enhance urban safety which are considered as hotspots of crime. As an improvement to the existing studies which lack accuracy on learning models, this study uses various machine learning algorithms like SVM, XGBoost, KNN and ARIMA model to better fit the crime data. The findings suggest a moderate increase in Chicago's overall crime rate while Los Angeles experiences a decline. The study concludes that these predictive models can aid law enforcement in directing patrols and developing effective strategies to combat crime.

Aimed at aiding the Police Department with proper crime forecasting, this study [7] concentrates on machine learning algorithms for crime forecasting. By using Folium for data visualisation, the year wise trends of crimes were discovered. The machine learning algorithms used were Random Forest, K-Nearest-Neighbours, AdaBoost and Neural network out of which Neural Network provided promising results when tested with Chicago Police Department's records with an accuracy of 90.77%.

This review paper [8] focuses on the growing interest among researchers in using machine learning and deep learning techniques to predict crime by analysing over 150 articles in the process. It examines the diverse algorithms employed and datasets utilized for crime prediction, exploring emerging trends and factors influencing criminal behaviour. By providing a overview of the research in crime prediction, it serves as a valuable resource for both academics and law enforcement agencies.

This paper [9] compares machine learning algorithms for crime prediction using historical data of public property crime from a coastal city in China between 2015 and 2018. It finds that LSTM model outperforms other algorithms like KNN, SVM, Random Forest and that incorporating environmental factors improves prediction accuracy compared to the model which only uses historical crime alone. Thus the paper concludes that crime prediction techniques should use both environmental factors and historical crime to maximize the accuracy.

This study [10] specifically focuses on classification of Crime category in the United States of America by using the data collected from socio-economic data from the US Census and crime data from FBI UCR. By using supervised classification algorithms like Naïve Bayes and Decision Tree, the study has concluded that the Decision Tree algorithm outperforms Naïve Bayes by having an accuracy of 83.9519% whereas the latter has an accuracy of 70.8124% in predicting crime of different states of the country.

This paper [11] deals with a unique type of crime called economic crime, which generally takes a lot of time for the law enforcement officers to solve. This paper develops an algorithm that detects fictitious enterprise using a classification algorithm called the Support Vector Machine. This model proved to be efficient as it resulted with a 99.7%

accuracy in the testing data which consisted of the economic activities of 1100 companies in Ukraine out of which 355 were defined as fictious.

Various data mining algorithms and ensemble learning techniques applied in crime data analysis and prediction have been discussed in this paper [12]. It deals with the significance of crime forecasting to reduce criminal activities based on historical data. With the increasing rate of crime cases, accurate crime prediction becomes crucial. Data mining methods helps in finding out patterns. The study aims to analyse and discuss the effectiveness of different methods applied in crime prediction, which can be used as a foundation in solving further crimes.

Data Mining is one of the best practices that can be used to find out patterns and relationships within the dataset. This paper deals with [13] analysing each data mining technique extensively by finding out the pros and cons of each technique. This technique is mainly used in Crime Detection as the patterns which generally go unnoticed can be detected can be found out by applying Data Mining techniques. This survey is intended to serve as a state-of-the-art crime detection guide.

Criminology is a field which identifies crime characteristics. This study [14] uses data mining techniques to identify crime characteristics by using Decision Tree(J48). Decision Tree algorithm is considered to be most efficient among all the machine learning algorithms as experimental results have proved that Decision Tree(J48) holds an accuracy of 94.25% which can be considered as a safe score to be relied on.

As crime rates continue to rise, it poses a severe challenge to the law enforcement officers. To address this issue, this paper [15] proposes extracting data from crime records and applying data mining techniques, including classification and regression algorithms, to predict future crime trends. The law enforcement agencies can allocate their resources effectively by using this system since the model is trained on historical data and using this, future trends can be found out. The system also suggests visualizing predicted outcomes using clustering algorithms like K-means, providing a user-friendly interface for understanding and interpreting the data.

This study [16] looks at how people connect in social networks and predicts behaviour using fuzzy systems. It uses colors to show different levels of possible criminal behaviour based on factors like background and habits. By studying these connections, it helps spot unusual behaviour and adjusts the network to keep things safe, using fuzzy logic methods.

This research [17] explores using data mining and machine learning to predict violent crime patterns. It compares crime data from a dataset with actual statistics from Mississippi. Three algorithms are tested: Linear Regression, Additive Regression, and Decision Stump. The study finds that Linear Regression performs the best. It emphasizes the importance of using data mining in law enforcement for tasks like identifying crime hotspots and understanding trends. Despite challenges, the study highlights the value of these methods in improving public safety.

The changing nature of crime is making traditional approaches to crime ineffective. This paper [18] discusses on a growing pattern which is the combined use of computer vision and machine learning technology to deal with these limitations. It makes use of machine learning algorithms that use historical crime data to predict crimes in the future. In public places, computer vision algorithms are used simultaneously for anomaly detection and real-time surveillance. The capacity of these integrated techniques to improve law enforcement operations and reduce the negative effects of criminal activity on society is examined.

This paper [19] reveals the importance of predicting crimes based on various factors like weather, geographic location, literacy rate of the location and so on. These features creates a base for these crimes. Moreover, the paper suggests some us to use Hotspots based on geographic location to predict and stop it. The paper proposes a K-Means for clustering the data and creating the hotspot. This paper also proves the vitality of these kind of applications in police department. Plotting of 2D Hotspot in location of crimes are based on historic crime data.

This research [20] assumes that traditional algorithms use concepts which are stationary and expect them to be stationary. Using such algorithms in real world forecasting where the concepts and scenarios would change could cause a real problem as the machine is not future proof and it is susceptible to errors. Further, the paper deals with the predicting vulnerable victims of crimes occurred in large cities. It mentions that a significant number of types of victims are changed based on police countering.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Intel Core i5
RAM	4GB or above (DDR4 RAM)
GPU	Intel Integrated Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

FRONT END	HTML, CSS, Bootstrap, JavaScript
BACK END	Python, Django
FRAMEWORKS	Pytorch, Tensor Flow
SOFTWARES USED	Visual Studio, Jupyter Notebook

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

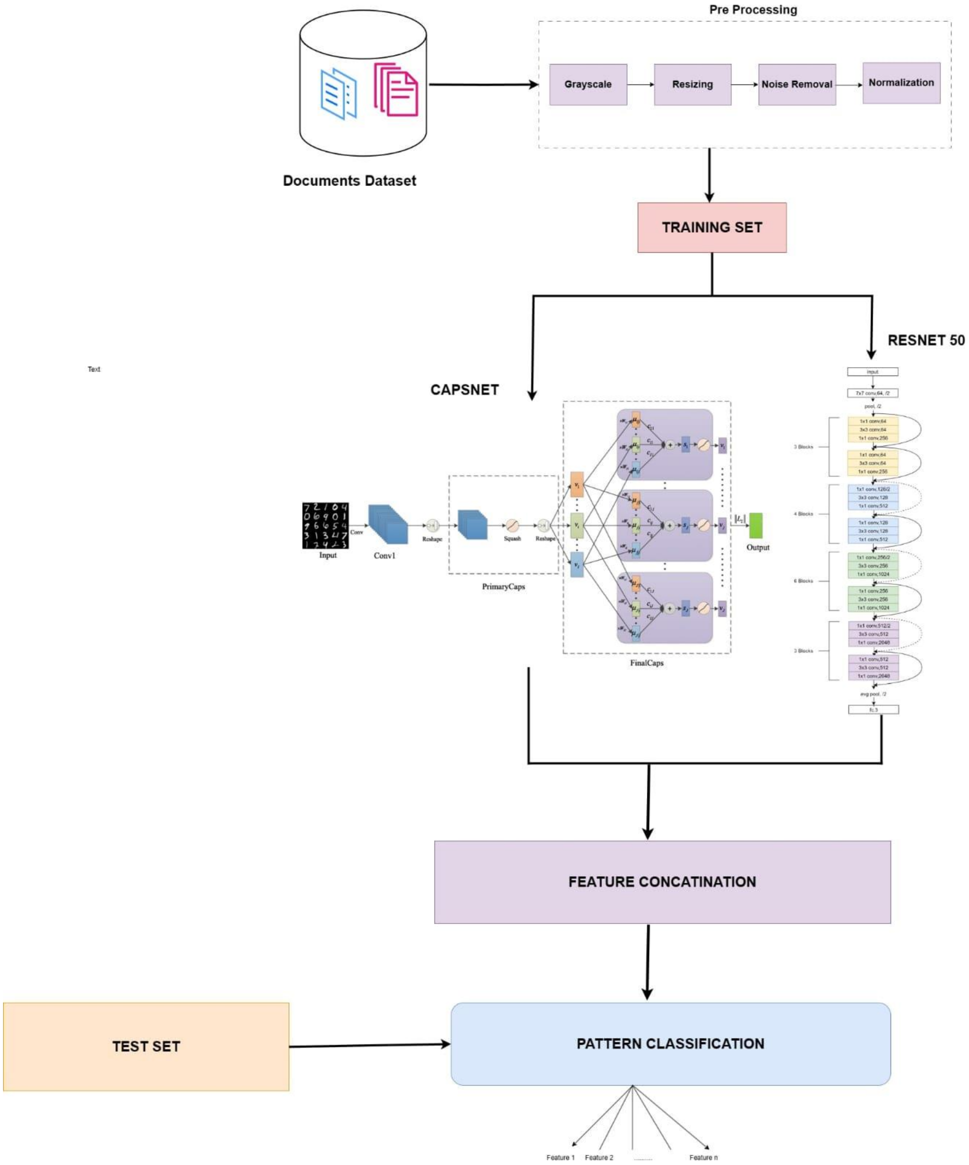


Fig 3.2.1 Architecture Diagram

PRE-PROCESSING:

Digitalizing Tamil and Hindi text from FIR (First Information Report) comes with a challenge of ignoring background noises as the paper's age can be decades. Before feeding the image into the models it has to be cleaned and uniformly resized. To reduce the number of dimensions of the data the image is grey scaled using binarization for separation of foreground and background which further helps in removal of image noise and text classification. The background is changed to black and the text and noise is converted to white. The colors values of 0 to 255 is normalized to 0 to 1. The ununiform images of the dataset is being resized into 64x64 size in order to reduce the training and working time of the system implemented along with that the zoom, height shift and other shifts are applied to make the character centered. The morphological operation of images are applied into the dataset in order to convert the text and background more clear and less vulnerable to errors. These helps in identifying shapes of different letters and strokes and erosion operations are done to diminish the sizes of boundaries. The data is then processed using Vott in order to convert it into JSON to feed it into CAPSNET and RESNET 50 the image is resized into 9x9 after Vott. Each image is given its own JSON file and then their coordinates are stored accordingly. The JSON can be further used in formation of classes for CAPSNET.

TRAINING SET:

In the domain of computer vision, there are two important techniques which greatly help in the application for the problem statement discussed here: CASPNET for handwritten regional text recognition and ResNet50 for image classification in crime scenes. CASPNET, an advanced convolutional neural network, specializes in deciphering handwritten text in various languages and styles, aiding in document digitization and analysis. Meanwhile, ResNet50, a deep learning model, excels in recognizing objects and patterns within crime scene images, aiding in forensic investigations. By deploying CASPNET, researchers can gather valuable information from handwritten documents, enhancing archival and investigative processes. On the other hand, ResNet50 plays a crucial role in identifying crucial elements within crime scenes, aiding law enforcement agencies in identifying suspects and reconstructing events accurately. Together, these technologies offer great tools for enhancing efficiency and accuracy in document analysis and crime scene investigation, contributing to the advancement of forensic science and law enforcement practices.

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATA PRE-PROCESSING:

This module mainly consists of the preliminary step of collecting crime files from various states and conversion of cleaning the image size by rescaling it to a proper and standard size. To reduce the number of dimensions of the data the image is turned to black and with using binarization for separation of background with foreground which contributes to removal of noise.

4.1.2 TRAINING SET:

By using the comprehensive models of Capsnet and Resnet, the handwritten case files are converted to a digital format which will later be fed to a machine learning algorithm for clustering. This is a very important step as this step involves changing handwritten text of various regional languages to digitalized format.

4.1.3 TRAINING MODEL:

The system is then trained with an ensemble model of various clustering algorithms like K Means, DBSCAN and Hierarchical clustering. Based on the various features of crime, clusters are formed. The dataset given to this model is the digitalized case files that have been solved, collected across various states.

4.1.4 PATTERN MATCHING:

When a new case file or an existing unsolved case file is fed to the system, the case file associates itself to the closest possible cluster and based on the output of pattern matching of certain key features, the cluster is recognized and the case files from that cluster is returned which can serve as a breakthrough for the investigation.

CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION

5.1.1 CAPSNET:

Data from various case files that were documented in the regional language where the crime was reported are collected together. The languages can be in English, Tamil or Hindi.

Initially, images are segmented individual characters. This is carried out using the Vott tool, an open-source software designed for image annotation. Vott provides the annotation process where users can define bounding boxes around objects of interest, in this case, characters within the case files. These boxes bounded by the users provide a visual representation of the segmentation process, outlining the boundaries of each character on the image. After processing the input image, Vott generates a JSON file for each annotated image, containing detailed coordinates of the bounding boxes, including various characteristics like width, height, and the coordinates of the upper-left corner. This enables precise identification and separation of characters for further analysis. Vott is capable of saving segmented character images, further helping the segmentation process. Following segmentation, all isolated characters were standardized to a size of 9 by 9 pixels to prepare them for input into the developed model. This step proved crucial in preprocessing case files written in regional language, enabling segmentation into lines, words, and individual characters for subsequent analysis.

CapsNet Model:

CapsNet is a type of neural network that uses a group of neurons to represent different parts of an object, like an object's characteristic or a specific part of it. Two convolutional layers are combined with a fully connected layer called RecCaps . The first layer converts the character image into blocks of activity. The second layer acts like a primary capsule and turns single output neurons into vectors with 8 dimensions. Then, RecCaps is used to capture the spatial relationships between all the local features from the primary capsule, and fed all these features into a higher dimensional capsule with 16 dimensions. The network's first layer works like a typical convolution layer in a CNN, using the ReLU activation function. In further layers, a special "squashing" activation function to shrink short vectors to zero and longer ones to a number close to 1 is used. This helps represent the probability of certain features being present. To run the network, segmented characters is fed from the case files into it. The first layer, extracts lower-level features from the characters. Then, the PrimaryCaps layer applies convolutional operations to get a 3-D matrix. This matrix has dimensions of 18 * 16 * 128, which is then split into 16 capsules, each with dimensions 18 * 16 * 8. We use a dynamic routing algorithm to connect primary capsules with advanced capsules, helping the model understand how different features relate to each other. This improves the model's ability to recognize and interpret handwritten characters. The algorithm calculates coupling coefficients for each iteration, which are used to compute input vectors for parent capsules and output vectors for capsules in the next layer. Finally, the sum of agreements between all capsules returns the final result.

5.1.2 RESNET50:

ResNet50 is a deep residual network designed to address the problem of network degradation. It introduces cross-layer connections to construct residual blocks, which learn the difference between input and output. This approach helps protect information integrity and simplifies the learning process. ResNet's structure accelerates network training by enabling faster loss reduction. The ResNet50 model consists of an initial independent convolutional layer, followed by pooling and four distinct convolutional residual modules. Each residual block comprises multiple convolutional layers and cross-layer connections, concluding with a pooling layer. The pool5 and fc1 features are extracted from the ResNet50 network for experimentation.

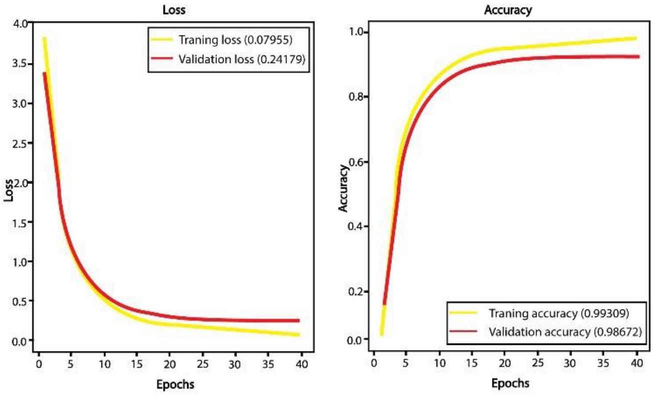
The Faster R-CNN object detection algorithm is utilized for semantic annotation of images, followed by content-based image retrieval. This method involves using ResNet50 for feature extraction, where the ResNet50 model is enhanced with a multi-scale pooling technique at the ROI-Pooling layer to improve feature representation. Specifically, the ROI-Pooling layer, responsible for mapping candidate regions back to the original image, undergoes modification by replacing single-scale pooling with multi-scale pooling. This alteration involves employing multiple pooling panes for maximum pooling at various scales (8x8, 4x4, 2x2, and 1x1) on the feature maps generated by the last convolutional layer, resulting in 85-dimensional feature vectors. This enhancement aims to overcome limitations in feature representation encountered with single-scale pooling. The improved multi-scale features are then utilized for object classification, contributing to enhanced accuracy in the classification process. Additionally, the semantic information obtained from Faster R-CNN object detection is integrated with the content-based image retrieval process, leading to improved retrieval rates and accuracy. This integrated approach involves utilizing the semantic information derived from object detection for image filtering, thereby reducing the search space for subsequent content-based retrieval based on depth features.

5.2 OUTPUT SCREENSHOTS

The analysis shows that strokes, edges, and curves are identified in caspnet to identify the specific regional language and also preprocessing removed all the unnecessary noises. The analysis on text recognition in Tamil gives around 90.12 +or - 2.24% and Hindi gives around 95.32 + or - 4.34%. The resnet50 used to identify crime scenes gives an sensitivity analytical score of 88.43 + or - 2.23% .This is a huge score in crime scene image classification. ResNet-50 exhibited the highest accuracy of 95.39% in digitizing handwritten text, surpassing alternative methods such as HWT, CSO, and BBO. ResNet-50's effectiveness is attributed to its utilization of residual modules and convolution layers, enabling superior feature extraction and training efficiency.

Table 5.2.1 Precision Table

Characters	precision	recall	f1-score	support
ka	0.95	0.96	0.98	300
cha	0.95	0.97	0.97	300
ta	0.96	0.95	0.95	300
pa	0.97	0.95	0.93	300
ya	0.93	0.94	0.94	300
ra	0.94	0.91	0.91	300
va	0.91	0.93	0.95	300
na	0.95	0.97	0.97	300
gna	0.98	0.95	0.96	300
zha	0.97	0.98	0.95	300
Accuracy			0.951	3000
Macro Average	0.95	0.95	0.95	3000
Weighted Average	0.95	0.95	0.95	3000



(a) Loss curve (b) Accuracy curve

Fig 5.2.1 Accuracy Curve

The graph is plotted for Accuracy and Loss. The graph is plotted against the number of Epochs. The Loss Curve results in having a training loss of 0.07955 and a validation loss of 0.24179. The accuracy curve increases exponentially having a training accuracy of 0.951 and validation accuracy of 0.98.

CHAPTER 6
CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

In conclusion, this paper has concentrated two important areas of research: handwritten regional language text classification using CapsNet and crime scene image classification using ResNet50 which provides a foundation for prediction of crime using historical crime data. In the domain of handwritten regional language text classification, the implementation of CapsNet proved to be promising. By taking advantage of the unique architecture of CapsNet, the model showed excellent performance in accurately categorizing handwritten text of various regional languages. This achievement helps greatly in digitalizing old case files which can help to solve cases which remain unsolved in that period. On the other hand, in the domain of crime scene image classification, the utilization of ResNet50 showcased remarkable capabilities in accurately identifying and classifying objects and scenes within crime scene images. This can be of great help in forensic investigation. The robustness and efficiency of ResNet50 make it a valuable tool for law enforcement agencies and forensic experts in analyzing and interpreting visual evidence, ultimately aiding in criminal investigations and ensuring justice.

Moreover, this research emphasizes the importance of using state-of-the-art deep learning techniques in addressing complex real-world problems. However, it's important to acknowledge the limitations and areas for future exploration. While CapsNet and ResNet50 show promising results, further research is needed to enhance their performance, scalability, and applicability across different datasets and scenarios. Additionally, the ethical implications of deploying such technology in sensitive domains like law enforcement warrant careful consideration as there might be a potential data leak issue and if done, it might have a serious impact.

In summary, this study lays a solid foundation for future researches aimed at advancing the fields of handwritten text classification and crime scene image analysis. This lays a solid foundation on which crime prediction is done. By incorporating regional language handwritten recognition, the scope and application of the project is increased multifold.

6.2 FUTURE ENHANCEMENTS

This project has a wide range of scope. While we have restricted the recognition of regional languages to a limited number, the number of languages can be increased. This project when applied in the domain of Cyber Security will have tremendous scope as the number of cyber crimes as increasing day by day and the pattern recognition can easily classify the patterns that happen in the domain of cyber security and crimes that are to happen in the future can be defended.

REFERENCES

[1] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," *IEEE Access*, vol. 9, pp. 67488–67500, 2021, doi: 10.1109/ACCESS.2021.3075140.

[2] S. Andrews, B. Brewster, and T. Day, "Organised crime and social media: a system for detecting, corroborating and visualising weak signals of organised crime online," *Secur Inform*, vol. 7, no. 1, Dec. 2018, doi: 10.1186/s13388-018-0032-8.

[3] M. M. Mukto *et al.*, "Design of a real-time crime monitoring system using deep learning techniques," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, doi: 10.1016/j.iswa.2023.200311.

[4] F. K. Bappee, A. Soares, L. M. Petry, and S. Matwin, "Examining the impact of cross-domain learning on crime prediction," *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00489-9.

[5] F. Dakalbab, M. Abu Talib, O. Abu Waraga, A. Bou Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," *Social Sciences and Humanities Open*, vol. 6, no. 1. Elsevier Ltd, Jan. 01, 2022. doi: 10.1016/j.ssaho.2022.100342.

[6] W. Safat, S. Asghar, and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," *IEEE Access*, vol. 9, pp. 70080–70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[7] A. Tamir, E. Watson, B. Willett, Q. Hasan, and J.-S. Yuan, "Crime Prediction and Forecasting using Machine Learning Algorithms," 2021. [Online]. Available: https://www.researchgate.net/publication/355872171

[8] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," *IEEE Access*, vol. 11, pp. 60153–60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[9] X. Zhang, L. Liu, L. Xiao, and J. Ji, "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*, vol. 8, pp. 181302–181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[10] R. Iqbal *et al.*, "An Experimental Study of Classification Algorithms for Crime Prediction." [Online]. Available: www.indjst.org

[11] A. Krysovatyy, H. Lipyanina-Goncharenko, S. Sachenko, and O. Desyatnyuk, "Economic Crime Detection Using Support Vector Machine Classification."

[12] A. Almaw and K. Kadam, "Survey Paper on Crime Prediction using Ensemble Approach." [Online]. Available: http://www.ijpam.eu

[13] S. Qayyum and H. Shareef Dar, "A Survey of Data Mining Techniques for Crime Detection," 2018.

[14] E. Ahishakiye, D. Taremwa, E. O. Omulo, and I. Niyonzima, "Crime Prediction Using Decision Tree (J48) Classification Algorithm," 2017. [Online]. Available: www.ijcit.com188

[15] V. Pande Student, C. Engg, V. Samant Student, B. E. Computer Engg, and S. Nair Asst Professor, "Crime Detection using Data Mining." [Online]. Available: http://www.ijert.org

[16] S. Gupta and S. Kumar, “Crime Detection and Prevention using Social Network Analysis,” 2015.

[17] L. McClendon and N. Meghanathan, “Using Machine Learning Algorithms to Analyze Crime Data,” *Machine Learning and Applications: An International Journal*, vol. 2, no. 1, pp. 1–12, Mar. 2015, doi: 10.5121/mlaij.2015.2101.

[18] N. Shah, N. Bhagat, and M. Shah, “Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention,” *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1. Springer, Dec. 01, 2021. doi: 10.1186/s42492-021-00075-z.

[19] G. Hajela, M. Chawla, and A. Rasool, “A Clustering Based Hotspot Identification Approach for Crime Prediction,” in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 1462–1470. doi: 10.1016/j.procs.2020.03.357.

[20] A. J. De Souza, A. P. Borges, H. M. Gomes, J. P. Barddal, and F. Enembreck, “Applying ensemble-based online learning techniques on crime forecasting,” in *ICEIS 2015 - 17th International Conference on Enterprise Information Systems, Proceedings*, SciTePress, 2015, pp. 17–24. doi: 10.5220/0005335700170024.

PLAGIARISM REPORT (PROJECT REPORT)

PAPER WITH PLAGIARISM REPORT

Solving Cold Case Using Historical Case Files

Kanaga Shanmugam P
Computer Science and Engineering
Rajalakshmi Engineering College
Chennai, India
210701103@rajalakshmi.edu.in

Karan Balaji R.S
Computer Science and Engineering
Rajalakshmi Engineering
Chennai, India
210701105@rajalakshmi.edu.in

***Abstract*—The unsolved cases in the records of police are often referred to as Cold Cases. The statistics of the crimes and attempt to crimes happening in India are eye opening. Close to 60,00,000 crime cases are registered in India every year and majority of these cases go unsolved and fall under the category of Cold Cases due to a number of factors like lack of evidence, lack of witness and lack of information. By using machine learning algorithms, the gap between the cases that are solved and the cases which are unsolved can be bridged. Taking in the important features of case files into consideration, the solved cases can be fed as an input to a model which is trained by clustering algorithms like K Means, DBSCAN etc. When the unsolved cases are fed to the trained model, the case files of the cluster can be suggested. These case files can then be used as references to solve the cases. The use case of this model can prove to be extremely beneficial in the context of solving age old cases where the records of those cases might be misplaced to lost or the evidences pertaining to those cases might not have been preserved for very long. While the project's scope is not restricted to solving crimes, if extended to the area of cybersecurity can have boundless applications as the threat level in the domain of cybersecurity is very high and if extrapolated, can help in protecting various cybercrimes from happening.**

***Keywords*—Cold Cases, Clustering, DBSCAN and K Means**

I. INTRODUCTION

Crime is considered an action which constitutes an offence and unlawful, which can be punished by law. There are different types of crimes which are punishable by law. Some of them, not limited to, are, (i) Organized crimes which is defined as the illegal activities carried out by group of people in a systematic and well-planned way in order to gain maximum profit out of a vulnerable situation, (ii) Cyber-crimes are the most serious and threatening crime which has an impact over a wide geographical area. It is the type of crime that happens through the digital medium and can range from a simple email phishing activity to gaining unauthorized access to large and sensitive government databases, (iii) Arson is a type of crime where damage is caused to a collateral property with a specific intention which can include personal vengeance or business rivalry, (iv) Assault is a type of crime which includes physical harm or threat of such an harm as an act of revenge or intimidation, (v) Burglary which can also be termed as a felony is the act of unauthorized entry into a restricted place with the intention of theft, (vi) Fraud involves manipulation or false representation of information for personal or financial gain, (vii) Domestic crime which accounts to a large section of crime is an act where physical or mental harm is caused by a partner to another in a relationship, and many more. With increasing number of crimes day by day, it has become more and more difficult to solve these crimes.

Crime across various countries are measured using crime rate which denotes the number of crimes reported to the law enforcement agencies per 1,00,000 persons of a population. According to the recent study in 2024 which includes data till the month of April, Venezuela stands first in highest crime rate having a crime index of 82.1 most of which includes Homicide, Kidnapping and Drug trafficking. Second on the list comes Papua New Guinea with a crime index of 80.4 while India has a crime index of 46.7. But when a closer look is taken at India's historical crime rate, the statistics ring a big alarm as the trend of number of crimes have increased exponentially in the recent years. According to the National Crime Records Bureau, starting with over 14 lakh crime incidences having a crime rate of 180.8 in the year 1988, the country stands with a staggering 58 lakh crime incidences having an alarming crime rate of 422.3 in the year 2022. The major contributors of these cases have been from the metropolitan cities with more the 25% cases being registered in these cities where majority of the population resides. Out of these crimes which include Murder, Kidnapping and abduction, Crime against Women and Children and many more, there is one sector of crime whose statistics instills fear and poses a severe threat to the public. Having more than 70,000 cases registered in a single year, Cyber-crimes trends have been on the rise having a 24.4% increase from the previous years. Out of these crimes, 64.8% were registered under the motive of fraud and 5.5% under the motive of extortion. A general study on Cyber-crimes worldwide has shown that Data breaches costs an average of $4.35 million for businesses in the year 2022. Accounting close to 400 million ransomware attacks in 2022, cyber-crimes have grown exponentially with a significant jump of 358% cyber crime cases in 2020 compared to 2019. The post pandemic era has witnessed the rise of cyber-crimes so much so that the governments of certain countries have allocated significant amount of their law enforcement forces only to the department of cyber-crime.

While the number of crimes are on the rise, the number of crimes which are unsolved accounts to a large and threatening number. While solving active cases have been challenging, it has been even more difficult to solve the unsolved cases dating back to 20 years. These cases are called Cold Cases (ie) the cases which remain unsolved for a longer period of time and remain a persistent challenge for the law enforcement officers. These cases not only pose as a great challenge to the realm of law enforcement but also affects the levels of trust on the law enforcement officers by the public. While the conventional methods of solving a crime may be effective in many cases, it may not be effective when it comes to solving cold cases due to the lack of evidence and resources. Cases like these generally go unsolved due to the lack of evidence or lack of witnesses. Moreover it can also be due to the negligence of the law enforcement officers or tendency of them to zero in on someone who is not the suspect known as tunnel vision. In India, close to 21 lakh cases are unsolved.

In order to solve these crimes, we have to make use of the technological advancements in the field of Machine Learning and Deep Learning. By exploiting the power of Machine Learning algorithms, the law enforcement agencies can potentially unlock now leads and re-examine old evidences and retouch upon the earlier overlooked information. The exponential increase in the digital data has made internet the house of information and using ML techniques, patterns and unnoticed relations can be established.

With the passage of time, the crime rate worldwide is only going to increase which calls for new measures and aid from the recent technological advancements. This paper explores the power of Machine Learning and its impact on the field of law enforcement. By analyzing disconnected data points which are often overlooked by humans, the Machine Learning algorithms suggests already solved case files which might help the case to take a new direction. The ensemble model of classical clustering algorithms like K Means, DBSCAN, Hierarchical clustering algorithms have been used to increase the accuracy of the case files suggested.

K Means is a clustering algorithm is an unsupervised Machine Learning algorithm that is used to cluster similar data points according to a specific parameter. This algorithm takes the value of ‘k’ which is a pre-determined value which denotes the number of clusters. K Means is a centroid model which uses clustering of data points using the centroid value. The variance of the data points which are relatively closer to the centroid is calculated and based on the variance, the points clustered together. This process is repeated until the best possible group of clusters are formed and the number of outliers are decreased.

DBSCAN which stands for Density Based Spatial Clustering of Application with Noise, is another type of unsupervised clustering algorithm which is an improvement to the K Means clustering algorithm. In contrast to the centroid model used in K Means, DBSCAN uses density based clustering which determines whether the data points belong to a cluster or not using the density of the region. The DBSCAN algorithm uses two parameters minPts and eps.minPts is the parameter which is the minimum number of points required to form a cluster. eps is the parameter which is the distance measure that is used to locate the points in the neighborhood of any point.

Hierarchical Clustering is a type of clustering algorithm which considers each datapoint as a separate cluster and later combines them into larger clusters. There are two types of clustering namely Agglomerative Clustering and Divisive Clustering. Agglomerative Clustering is a type of clustering algorithm which follows a bottom up approach where initially, each data point is considered as a single entity and eventually, bigger clusters are formed by combining the clusters which are close to each other. Divisive clustering algorithm follows the top down approach where initially the data points are considered to be as a large cluster and as the distance between the data points decrease, the clusters are sub divided into individual components.

Ensemble learning is a Machine Learning strategy which uses the output of multiple models in order to generate a more accurate result. This is done by using a voting classifier which uses one of the two voting techniques namely hard voting and soft voting to choose the best model. Hard voting is a voting technique where the output class is based on the highest majority of votes whereas Soft voting technique determines the output class by taking the average probabilities of classes. While the above-mentioned clustering algorithms have their own advantages and benefits, the lack some key features which can be rectified by using the ensemble learning model. While K Means uses the centroid model for clustering, it is sensitive to initial conditions as the number of clusters needs to be predefined which may not be ideal in cases involving large number of datasets having large number of features to cluster with. While DBSCAN which uses density-based clustering, can form clusters or arbitrary size and does not need the number of cluster before-hand, it poses its own challenges that needs to be rectified. It is sensitive to the choice of Eps and MinPts and the cost of computation is high when the number of data points is large. Meanwhile Hierarchical clustering provides rich information and insight to the dataset but the inability to handle scalable datasets and time complexity makes the algorithm computationally expensive. An ensemble model of the above-mentioned algorithms will be robust to different data distributions, with enhanced cluster separation and outlier detection.

Using an ensemble model in this project improves pattern recognition as cold cases generally involve complex and multi-dimensional datasets which may be challenging to detect by using a single clustering algorithm. Understanding these patterns and relationships is crucial for the law enforcement officers in order to prioritize their investigation. By incorporating hierarchical clustering into the ensemble model, prioritization of the cases can be done as similar cases files will be clustered together. This may help them to lead the investigation in a certain direction.

II. LITERATURE SURVEY

This paper [1] discusses the importance and benefits of using ensemble learning methods in crime prediction and the advantages it holds over the conventional machine learning models when used as a single classifier. Due to the dynamic nature of crimes, it is difficult to find the right configuration for the datasets to feed as an input to the ensemble model. The paper proposes a model called assemble-stacking based crime prediction method (SBCPM) which applies SVM model to achieve configurations that are specific to the domain. This model achieves 99% classification accuracy on training data.

One of the main types of crime, organized crimes have been studied in this paper [2] by identifying organized crimes through social media analysis. By analysing language cues in social media as indicators, this paper classifies crime type and location. The system generates Organized crime concepts to alert analysts of potential criminal activity by grouping information sources. Analysts can also investigate organized crime concepts through a prototype software system featuring social media scanning and map-based visualization. The system is illustrated using human trafficking and modern slavery.

This paper [3]describes a Crime Monitoring System (CMS) designed to detect crimes in real-time using camera surveillance. It aims to overcome human limitations like slow reactions by combining CCTV cameras with deep learning techniques. The system operates in three stages: detecting weapons, violence, and recognizing faces. It achieves high accuracy rates in each stage: over 80% for weapons, 95% for violence, and 97% for faces. Real-world testing shows the system effectively detects crime and alerts authorities promptly, improving overall security and safety measures.

The study conducted in the area of cross domain learning on crime prediction [4] gives an insight about the data insufficiency problem faced in small cities. As the number of researches on crime prediction are on the rise, the urban data has become more accessible to the researchers. But this paper discusses on how to overcome data insufficiency in small cities with respect to Canada. By using ensemble learning as it generalizes new data and the classified data is compared against the baseline models.

This review paper [5] gives a take on Artificial Intelligence being used in the field of crime prediction. By intensively analysing various criteria, the models are evaluated. Intensive research is carried out after reviewing 120 research papers that were published between 2008 and 2021. The research has concluded that crimes and spatial are the most applied categories in analysing crimes. The various ML models used across the 120 research papers were noted and supervised learning models were found to be the major contributors with 31% while a combination of supervised and unsupervised learning models contributed with 22% and unsupervised learning models alone contributed with 10%.

This study [6] draws attention towards the importance of crime prediction and forecasting to enhance urban safety which are considered as hotspots of crime. As an improvement to the existing studies which lack accuracy on learning models, this study uses various machine learning algorithms like SVM, XGBoost, KNN and ARIMA model to better fit the crime data. The findings suggest a moderate increase in Chicago's overall crime rate while Los Angeles experiences a decline. The study concludes that these predictive models can aid law enforcement in directing patrols and developing effective strategies to combat crime.

Aimed at aiding the Police Department with proper crime forecasting, this study [7] concentrates on machine learning algorithms for crime forecasting. By using Folium for data visualization, the year wise trends of crimes were discovered. The machine learning algorithms used were Random Forest, K-Nearest-Neighbours, AdaBoost and Neural network out of which Neural Network provided promising results when tested with Chicago Police Department's records with an accuracy of 90.77%.

This review paper [8] focuses on the growing interest among researchers in using machine learning and deep learning techniques to predict crime by analysing over 150 articles in the process. It examines the diverse algorithms employed and datasets utilized for crime prediction, exploring emerging trends and factors influencing criminal behaviour. By providing a overview of the research in crime prediction, it serves as a valuable resource for both academics and law enforcement agencies.

This paper [9] compares machine learning algorithms for crime prediction using historical data of public property crime from a coastal city in China between 2015 and 2018. It finds that LSTM model outperforms other algorithms like KNN, SVM, Random Forest and that incorporating environmental factors improves prediction accuracy compared to the model which only uses historical crime alone. Thus the paper concludes that crime prediction techniques should use both environmental factors and historical crime to maximize the accuracy.

This study [10] specifically focuses on classification of Crime category in the United States of America by using the data collected from socio-economic data from the US Census and crime data from FBI UCR. By using supervised classification algorithms like Naïve Bayes and Decision Tree, the study has concluded that the Decision Tree algorithm outperforms Naïve Bayes by having an accuracy of 83.9519% whereas the latter has an accuracy of 70.8124% in predicting crime of different states of the country.

This paper [11] deals with a unique type of crime called economic crime, which generally takes a lot of time for the law enforcement officers to solve. This paper develops an algorithm that detects fictitious enterprise using a classification algorithm called the Support Vector Machine. This model proved to be efficient as it resulted with a 99.7% accuracy in the testing data which consisted of the economic activities of 1100 companies in Ukraine out of which 355 were defined as fictious.

Various data mining algorithms and ensemble learning techniques applied in crime data analysis and prediction have been discussed in this paper [12]. It deals with the significance of crime forecasting to reduce criminal activities based on historical data. With the increasing rate of crime cases, accurate crime prediction becomes crucial. Data mining methods helps in finding out patterns. The study aims to analyse and discuss the effectiveness of different methods applied in crime prediction, which can be used as a foundation in solving further crimes.

Data Mining is one of the best practices that can be used to find out patterns and relationships within the dataset. This paper deals with [13] analysing each data mining technique extensively by finding out the pros and cons of each technique. This technique is mainly used in Crime Detection as the patterns which generally go unnoticed can be detected can be found out by applying Data Mining techniques. This survey is intended to serve as a state-of-the-art crime detection guide.

Criminology is a field which identifies crime characteristics. This study [14] uses data mining techniques to identify crime characteristics by using Decision Tree(J48). Decision Tree algorithm is considered to be most efficient among all the machine learning algorithms as experimental results have proved that Decision Tree(J48) holds an accuracy of 94.25% which can be considered as a safe score to be relied on.

As crime rates continue to rise, it poses a severe challenge to the law enforcement officers. To address this issue, this paper [15] proposes extracting data from crime records and applying data mining techniques, including classification and regression algorithms, to predict future crime trends. The law enforcement agencies can allocate their resources effectively by using this system since the model is trained on historical data and using this, future trends can be found out. The system also suggests visualizing predicted outcomes using clustering algorithms like K-means, providing a user-friendly interface for understanding and interpreting the data.

This study [16] looks at how people connect in social networks and predicts behaviour using fuzzy systems. It uses colors to show different levels of possible criminal behaviour based on factors like background and habits. By studying these connections, it helps spot unusual behaviour and adjusts the network to keep things safe, using fuzzy logic methods.

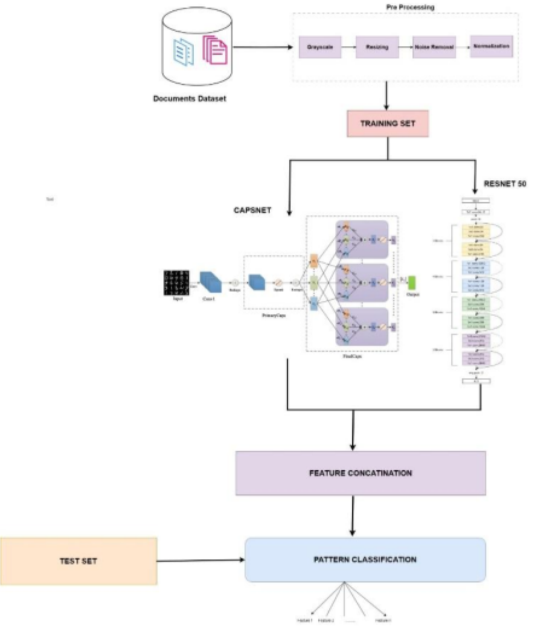
This research [17] explores using data mining and machine learning to predict violent crime patterns. It compares crime data from a dataset with actual statistics from Mississippi. Three algorithms are tested: Linear Regression, Additive Regression, and Decision Stump. The study finds that Linear Regression performs the best. It emphasizes the importance of using data mining in law enforcement for tasks like identifying crime hotspots and understanding trends. Despite challenges, the study highlights the value of these methods in improving public safety.

The changing nature of crime is making traditional approaches to crime ineffective. This paper [18] discusses on a growing pattern which is the combined use of computer vision and machine learning technology to deal with these limitations. It makes use of machine learning algorithms that use historical crime data to predict crimes in the future. In public places, computer vision algorithms are used simultaneously for anomaly detection and real-time surveillance. The capacity of these integrated techniques to improve law enforcement operations and reduce the negative effects of criminal activity on society is examined.

This paper [19] reveals the importance of predicting crimes based on various factors like weather, geographic location, literacy rate of the location and so on. These features creates a base for these crimes. Moreover, the paper suggests some us to use Hotspots based on geographic location to predict and stop it. The paper proposes a K-Means for clustering the data and creating the hotspot. This paper also proves the vitality of these kind of applications in police department. Plotting of 2D Hotspot in location of crimes are based on historic crime data.

This research [20] assumes that traditional algorithms use concepts which are stationary and expect them to be stationary. Using such algorithms in real world forecasting where the concepts and scenarios would change could cause a real problem as the machine is not future proof and it is susceptible to errors. Further, the paper deals with the predicting vulnerable victims of crimes occurred in large cities. It mentions that a significant number of types of victims are changed based on police countering.

III. PROPOSED MODEL



A. PRE-PROCESSING:

Digitalizing Tamil and Hindi text from FIR (First Information Report) comes with a challenge of ignoring background noises as the paper's age can be decades. Before feeding the image into the models it has to be cleaned and uniformly resized. To reduce the number of dimensions of the data the image is grey scaled using binarization for separation of foreground and background which further helps in removal of image noise and text classification. The background is changed to black and the text and noise is converted to white. The colors values of 0 to 255 is normalized to 0 to 1. The ununiform images of the dataset is being resized into 64x64 size in order to reduce the training and working time of the system implemented along with that the zoom, height shift and other shifts are applied to make the character centered. The morphological operation of images are applied into the dataset in order to convert the text and background more clear and less vulnerable to errors. These helps in identifying shapes of different letters and strokes and erosion operations are done to diminish the sizes of boundaries. The data is then processed using Vott in order to convert it into JSON to feed it into CAPSNET and RESNET 50 the image is resized into 9x9 after Vott. Each image is given its own JSON file and then their coordinates are stored accordingly. The JSON can be further used in formation of classes for CAPSNET.

B. Training Set

In the domain of computer vision, there are two important techniques which greatly help in the application for the problem statement discussed here: CASPNET for handwritten regional text recognition and ResNet50 for image classification in crime scenes. CASPNET, an advanced convolutional neural network, specializes in deciphering handwritten text in various languages and styles, aiding in document digitization and analysis. Meanwhile, ResNet50, a deep learning model, excels in recognizing objects and patterns within crime scene images, aiding in forensic investigations. By deploying CASPNET, researchers can gather valuable information from handwritten documents, enhancing archival and investigative processes. On the other hand, ResNet50 plays a crucial role in identifying crucial elements within crime scenes, aiding law enforcement agencies in identifying suspects and reconstructing events accurately. Together, these technologies offer great tools for enhancing efficiency and accuracy in document analysis and crime scene investigation, contributing to the advancement of forensic science and law enforcement practices.

CAPSNET:

Image Annotation and Segmentation:

Data from various case files that were documented in the regional language where the crime was reported are collected together. The languages can be in English, Tamil or Hindi.

Initially, images are segmented individual characters. This is carried out using the Vott tool, an open-source software designed for image annotation. Vott provides the annotation process where users can define bounding boxes around objects of interest, in this case, characters within the case files. These boxes bounded by the users provide a visual representation of the segmentation process, outlining the boundaries of each character on the image. After processing the input image, Vott generates a JSON file for each annotated image, containing detailed coordinates of the bounding boxes, including various characteristics like width, height, and the coordinates of the upper-left corner. This enables precise identification and separation of characters for further analysis. Vott is capable of saving segmented character images, further helping the segmentation process. Following segmentation, all isolated characters were standardized to a size of 9 by 9 pixels to prepare them for input into the developed model. This step proved crucial in preprocessing case files written in regional language, enabling segmentation into lines, words, and individual characters for subsequent analysis.

CapsNet Model:

CapsNet is a type of neural network that uses a group of neurons to represent different parts of an object, like an object's characteristic or a specific part of it. Two convolutional layers are combined with a fully connected layer called RecCaps . The first layer converts the character image into blocks of activity. The second layer acts like a primary capsule and turns single output neurons into vectors with 8 dimensions. Then, RecCaps is used to capture the spatial relationships between all the local features from the primary capsule, and fed all these features into a higher dimensional capsule with 16 dimensions. The network's first layer works like a typical convolution layer in a CNN, using the ReLU activation function. In further layers, a special "squashing" activation function to shrink short vectors to zero and longer ones to a number close to 1 is used. This helps represent the probability of certain features being present. To run the network, segmented characters is fed from the case files into it. The first layer, extracts lower-level features from the characters. Then, the PrimaryCaps layer applies convolutional operations to get a 3-D matrix. This matrix has dimensions of 18 * 16 * 128, which is then split into 16 capsules, each with dimensions 18 * 16 * 8. We use a dynamic routing algorithm to connect primary capsules with advanced capsules, helping the model understand how different features relate to each other. This improves the model's ability to recognize and interpret handwritten characters. The algorithm calculates coupling coefficients for each iteration, which are used to compute input vectors for parent capsules and output vectors for capsules in the next layer. Finally, the result is gathered by the agreement between all capsules

Implementation:

The developed CapsNet model, created using the Keras Python library is implemented on computers having minimum specifications RAM and graphic support. The CapsNet model has a nested structure with 800 hidden units, and tested against 399 different classes. Adam optimizer is used for model training, which is a popular model in deep learning. The model used here is a hybrid one, which includes convolution layers, dense layers, and hidden units. ReLU and sigmoid activation functions is used, along with a loss function called binary cross-entropy.

RESNET50:

ResNet50 is a deep residual network designed to address the problem of network degradation. It introduces cross-layer connections to construct residual blocks, which learn the difference between input and output. This approach helps protect information integrity and simplifies the learning process. ResNet's structure accelerates network training by enabling faster loss reduction. The ResNet50 model consists of an initial independent convolutional layer, followed by pooling and four distinct convolutional residual modules. Each residual block comprises multiple convolutional layers and cross-layer connections, concluding with a pooling layer. The pool5 and fc1 features are extracted from the ResNet50 network for experimentation.

The Faster R-CNN object detection algorithm is utilized for semantic annotation of images, followed by content-based image retrieval. This method involves using ResNet50 for feature extraction, where the ResNet50 model is enhanced with a multi-scale pooling technique at the ROI-Pooling layer to improve feature representation. Specifically, the ROI-Pooling layer, responsible for mapping candidate regions back to the original image, undergoes modification by replacing single-scale pooling with multi-scale pooling. This alteration involves employing multiple pooling panes for maximum pooling at various scales (8x8, 4x4, 2x2, and 1x1) on the feature maps generated by the last convolutional layer, resulting in 85-dimensional feature vectors. This enhancement aims to overcome limitations in feature representation encountered with single-scale pooling. The improved multi-scale features are then utilized for object classification, contributing to enhanced accuracy in the classification process. Additionally, the semantic information obtained from Faster R-CNN object detection is integrated with the content-based image retrieval process, leading to improved retrieval rates and accuracy. This integrated approach involves utilizing the semantic information derived from object detection for image filtering, thereby reducing the search space for subsequent content-based retrieval based on depth features.

C. Pattern Classification

After digitalizing the case files, some features are extracted from the case files based on various parameters. Some of the many possible features are listed below:

Case Identifier: Each case will be represented by a unique identifier. Though this may not be a contributing factor towards the feature concatenation, this feature will contribute towards removing duplicate files from the dataset as combining case files from various jurisdiction can result in the dataset having duplicate case files and by using this feature, duplicate files will not be considered for further process.

Address: This field contains the location details of the crime containing the house number, street name, city, and country where crime has reported. Further, each of the above-mentioned attribute is separated for ease of classification for the model and the attributes present as text are converted into an integer for standardization .

Date of Crime: This feature contains the date and time of the crime. This date is in the format of Date, Month, Year followed by the time when the crime took place. Later each of the individual aspects are separated and concatenated along with the place where the crime took place to uniquely identify the crime.

Binary Classification: Certain features contain only binary classification of true or false. Features like whether the crime is Domestic Crime, whether an arrest was made or not. These classifications contribute to the overall classification to the type and severity of the crime.

Evidence Description: This features contains the details of evidences that are collected from the crime spot during the investigation process like DNA, Fingerprints, Weapon description.

Case Status: Case status contains the current status of the crime (ie) whether it is Active, Closed, Under Process or if there are developments since the initial report.

Based on the various features extracted, patterns are identified and classified together.

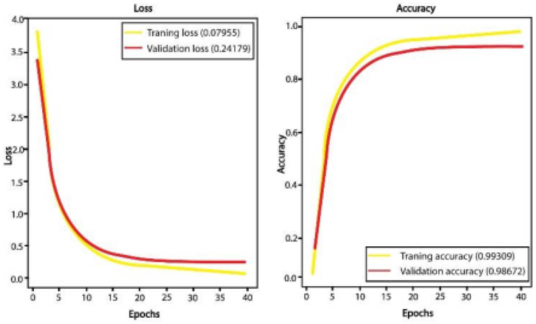
The test set, which is an active case file, is tested against the pattern classified cases files. When a new case file is given to the model, it is compared with the historical crime files and pattern matching is done with those files. Based on the output from the pattern classification algorithms, the historical case files which are closely related to the test case file are grouped together and are given as suggestions.

IV. RESULT

A. STATISTICAL ANALYSIS:

The analysis shows that strokes, edges, and curves are identified in caspnet to identify the specific regional language and also preprocessing removed all the unnecessary noises. The analysis on text recognition in Tamil gives around 90.12 +or - 2.24% and Hindi gives around 95.32 + or - 4.34%. The resnet50 used to identify crime scenes gives an sensitivity analytical score of 88.43 + or - 2.23% .This is a huge score in crime scene image classification. ResNet-50 exhibited the highest accuracy of 95.39% in digitizing handwritten text, surpassing alternative methods such as HWT, CSO, and BBO. ResNet-50's effectiveness is attributed to its utilization of residual modules and convolution layers, enabling superior feature extraction and training efficiency.

Characters	precision	recall	f1-score	support
ka	0.95	0.96	0.98	300
cha	0.95	0.97	0.97	300
ta	0.96	0.95	0.95	300
pa	0.97	0.95	0.93	300
ya	0.93	0.94	0.94	300
ra	0.94	0.91	0.91	300
va	0.91	0.93	0.95	300
na	0.95	0.97	0.97	300
gna	0.98	0.95	0.96	300
zha	0.97	0.98	0.95	300
Accuracy			0.951	3000
Macro Average	0.95	0.95	0.95	3000
Weighted Average	0.95	0.95	0.95	3000



(a) Loss curve

(b) Accuracy curve

The graph is plotted for Accuracy and Loss. The graph is plotted against the number of Epochs. The Loss Curve results in having a training loss of 0.07955 and a validation loss of 0.24179. The accuracy curve increases exponentially having a training accuracy of 0.951 and validation accuracy of 0.98

V.CONCLUSION

In conclusion, this paper has concentrated two important areas of research: handwritten regional language text classification using CapsNet and crime scene image classification using ResNet50 which provides a foundation for prediction of crime using historical crime data. In the domain of handwritten regional language text classification, the implementation of CapsNet proved to be promising. By taking advantage of the unique architecture of CapsNet, the model showed excellent performance in accurately categorizing handwritten text of various regional languages. This achievement helps greatly in digitalizing old case files which can help to solve cases which remain unsolved in that period. On the other hand, in the domain of crime scene image classification, the utilization of ResNet50 showcased remarkable capabilities in accurately identifying and classifying objects and scenes within crime scene images. This can be of great help in forensic investigation. The robustness and efficiency of ResNet50 make it a valuable tool for law enforcement agencies and forensic experts in analyzing and interpreting visual evidence, ultimately aiding in criminal investigations and ensuring justice.

Moreover, this research emphasizes the importance of using state-of-the-art deep learning techniques in addressing complex real-world problems. However, it's important to acknowledge the limitations and areas for future exploration. While CapsNet and ResNet50 show promising results, further research is needed to enhance their performance, scalability, and applicability across different datasets and scenarios. Additionally, the ethical implications of deploying such technology in sensitive domains like law enforcement warrant careful consideration as there might be a potential data leak issue and if done, it might have a serious impact.

In summary, this study lays a solid foundation for future researches aimed at advancing the fields of handwritten text classification and crime scene image analysis. This lays a solid foundation on which crime prediction is done. By incorporating regional language handwritten recognition, the scope and application of the project is increased multifold.

REFERENCES

[1] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," *IEEE Access*, vol. 9, pp. 67488–67500, 2021, doi: 10.1109/ACCESS.2021.3075140.

[2] S. Andrews, B. Brewster, and T. Day, "Organised crime and social media: a system for detecting, corroborating and visualising weak signals of organised crime online," *Secur Inform*, vol. 7, no. 1, Dec. 2018, doi: 10.1186/s13388-018-0032-8.

[3] M. M. Mukto *et al*., "Design of a real-time crime monitoring system using deep learning techniques," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, doi: 10.1016/j.iswa.2023.200311.

[4] F. K. Bappee, A. Soares, L. M. Petry, and S. Matwin, "Examining the impact of cross-domain learning on crime prediction," *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00489-9.

[5] F. Dakalbab, M. Abu Talib, O. Abu Waraga, A. Bou Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," *Social Sciences and Humanities Open*, vol. 6, no. 1. Elsevier Ltd, Jan. 01, 2022. doi: 10.1016/j.ssaho.2022.100342.

[6] W. Safat, S. Asghar, and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," *IEEE Access*, vol. 9, pp. 70080–70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

[7] A. Tamir, E. Watson, B. Willett, Q. Hasan, and J.-S. Yuan, "Crime Prediction and Forecasting using Machine Learning Algorithms," 2021. [Online]. Available: https://www.researchgate.net/publication/355872171

[8] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," *IEEE Access*, vol. 11, pp. 60153–60170, 2023, doi: 10.1109/ACCESS.2023.3286344.

[9] X. Zhang, L. Liu, L. Xiao, and J. Ji, "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*, vol. 8, pp. 181302–181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[10] R. Iqbal *et al*., "An Experimental Study of Classification Algorithms for Crime Prediction." [Online]. Available: www.indjst.org

[11] A. Krysovatyy, H. Lipyanina-Goncharenko, S. Sachenko, and O. Desyatnyuk, "Economic Crime Detection Using Support Vector Machine Classification."

[12] A. Almaw and K. Kadam, "Survey Paper on Crime Prediction using Ensemble Approach." [Online]. Available: http://www.ijpam.eu

[13] S. Qayyum and H. Shareef Dar, "A Survey of Data Mining Techniques for Crime Detection," 2018.

[14] E. Ahishakiye, D. Taremwa, E. O. Omulo, and I. Niyonzima, "Crime Prediction Using Decision Tree (J48) Classification Algorithm," 2017. [Online]. Available: www.ijcit.com188

[15] V. Pande Student, C. Engg, V. Samant Student, B. E. Computer Engg, and S. Nair Asst Professor, "Crime Detection using Data Mining." [Online]. Available: http://www.ijert.org

[16] S. Gupta and S. Kumar, "Crime Detection and Prevention using Social Network Analysis," 2015.

[17] L. McClendon and N. Meghanathan, "Using Machine Learning Algorithms to Analyze Crime Data," *Machine Learning and Applications: An International Journal*, vol. 2, no. 1, pp. 1–12, Mar. 2015, doi: 10.5121/mlaij.2015.2101.

[18] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1. Springer, Dec. 01, 2021. doi: 10.1186/s42492-021-00075-z.

[19] G. Hajela, M. Chawla, and A. Rasool, "A Clustering Based Hotspot Identification Approach for Crime Prediction," in *Procedia Computer Science*, Elsevier B.V., 2020, pp. 1462–1470. doi: 10.1016/j.procs.2020.03.357.

[20] A. J. De Souza, A. P. Borges, H. M. Gomes, J. P. Barddal, and F. Enembreck, "Applying ensemble-based online learning techniques on crime forecasting," in *ICEIS 2015 - 17th International Conference on Enterprise Information Systems, Proceedings*, SciTePress, 2015, pp. 17–24. doi: 10.5220/0005335700170024.

kags

ORIGINALITY REPORT

26%	5%	8%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to University of Illinois at Urbana-Champaign Student Paper	18%
2	Aditi Moudgil, Saravjeet Singh, Vinay Gautam, Shalli Rani, Syed Hassan Ahmed. "Handwritten Devanagari Manuscript Characters Recognition using CapsNet", International Journal of Cognitive Computing in Engineering, 2023 Publication	2%
3	Dongyuan Li, Xiaojun Bai. "Criminal Investigation Image Retrieval Based on Deep Learning", 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), 2020 Publication	1%
4	www.medrxiv.org Internet Source	1%
5	ijcsmc.com Internet Source	<1%
	sci-hub.se	

6	Internet Source	<1 %
7	Fatima Dakalbab, Manar Abu Talib, Omnia Abu Waraga, Ali Bou Nassif, Sohail Abbas, Qassim Nasir. "Artificial intelligence & crime prediction: A systematic literature review", Social Sciences & Humanities Open, 2022 Publication	<1 %
8	link.springer.com Internet Source	<1 %
9	www.researchgate.net Internet Source	<1 %
10	Submitted to Capitol College Student Paper	<1 %
11	Guofu Zhai, Zhigang Sun, Guotao Wang, Pengfei Li, Qi Liang, Min Zhang. "Instance-based transfer learning method for locating loose particles inside aerospace equipment", Measurement, 2023 Publication	<1 %
12	www.aiirjournal.com Internet Source	<1 %
13	repository.tudelft.nl Internet Source	<1 %
14	5wwwww.easychair.org Internet Source	<1 %

15	file.techscience.com Internet Source	<1 %
16	www.cse.griet.ac.in Internet Source	<1 %
17	"Intelligent Systems and Applications", Springer Science and Business Media LLC, 2019 Publication	<1 %
18	Asit Kumar Das, Priyanka Das. "Graph based ensemble classification for crime report prediction", Applied Soft Computing, 2022 Publication	<1 %
19	Oghenevovwero Zion Apene, Nachamada Vachaku Blamah, Gilbert Imuetinyan Osaze Aimufua. "Advancements in Crime Prevention and Detection: From Traditional Approaches to Artificial Intelligence Solutions", European Journal of Applied Science, Engineering and Technology, 2024 Publication	<1 %
20	doaj.org Internet Source	<1 %
21	polynoe.lib.uniwa.gr Internet Source	<1 %
22	scitepress.org Internet Source	<1 %

23	www.jcdronline.org Internet Source	<1 %
24	Junxiang Yin. "Crime Prediction Methods Based on Machine Learning: A Survey", Computers, Materials & Continua, 2023 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		