A BIOMETRIC-DRIVEN MACHINE LEARNING FRAMEWORK FOR RECIDIVISM PREDICTION IN SRI LANKA

Enhancing Public Safety through Predictive Analytics and Behavioral Biometrics.

M.G.L.N Kumararathna

CB013366

BSc (hons) Computer Science

CONTEXT & ORGANIZATIONAL OVERVIEW

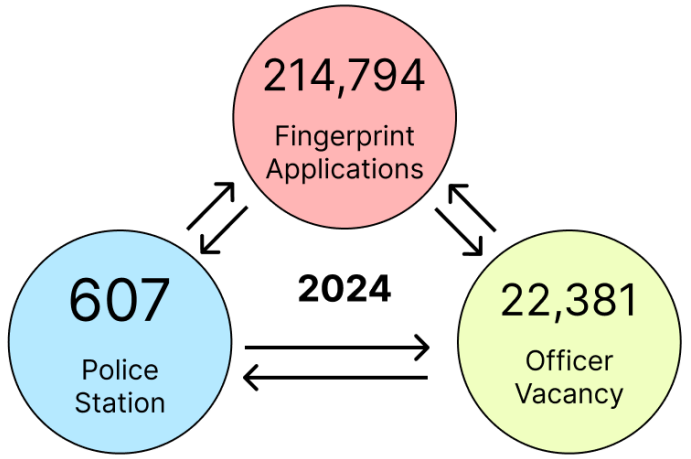
Client: Sri Lanka Police (SLP) – Primary law enforcement agency.

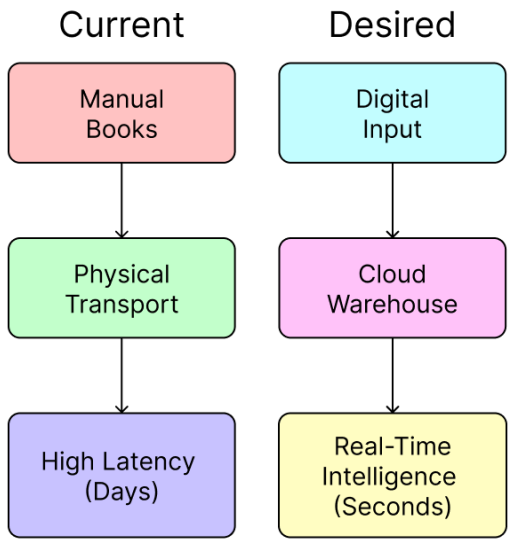
The Data Problem:

- **Volume:** Processing **214k+** biometric records annually.
- **Siloed Systems:** Data trapped in physical "Complaint Books" and disconnected digital systems (CEMS, AMIS).
- **High Latency:** Manual fingerprint transport from rural stations to Colombo takes days.

Operational Pain Point:

- Severe human resource crisis: **22,381 officer vacancy** (25% deficit).
- Urgent need for "Force Multiplication" via predictive intelligence.





PROBLEM DEFINITION

THE ANALYTICAL GAP

Core Issue: Reactive Policing

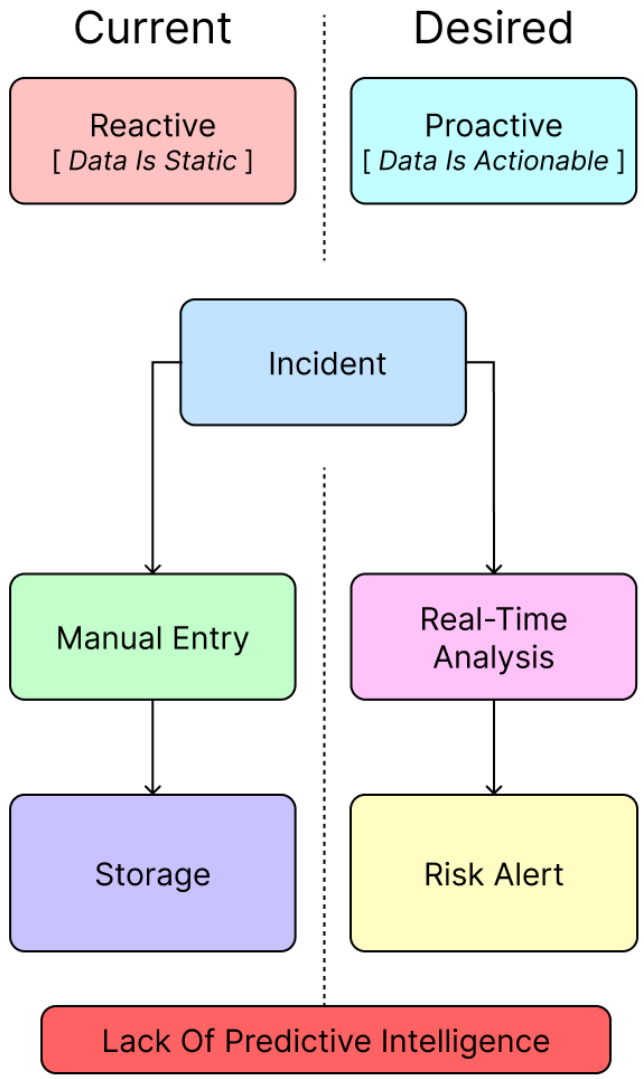
- Current systems (CEMS/AMIS) operate on a "storage-centric" model.
- We currently identify a repeat offender only *after* they commit a second crime and are processed manually.

Missing Capability: No Predictive Risk Scoring

- 27,000 criminals registered in 2024, but only 2,735 identified as "Island Reconvicted Criminals" (IRC).
- Officers lack a statistical tool to identify the high-risk minority among the remaining suspects.

Research Question:

"How can we utilize historical arrest data and biometrics to predict the probability of re-offending (Recidivism)?".



(Sri Lanka Police, 2024; Haidar, 2025; Kovalchuk et al., 2024)

LO1

THE KDD PROCESS

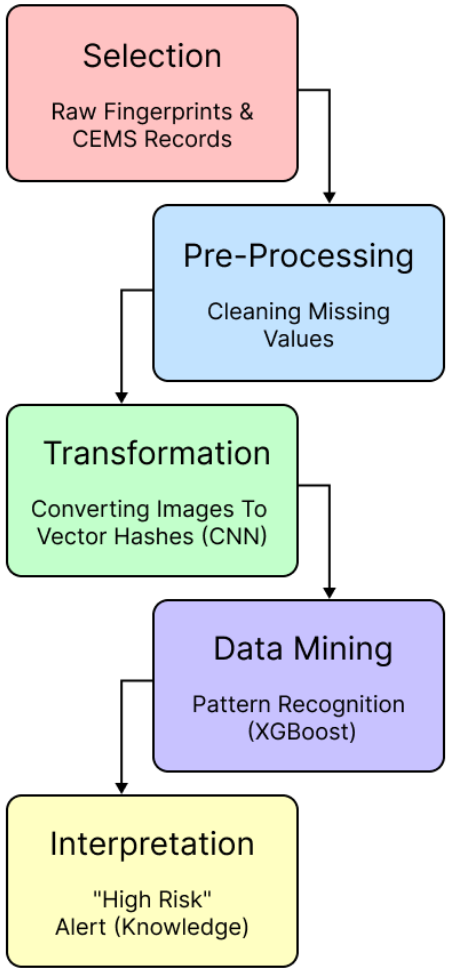
Turning Raw Police Data into Intelligence

Concept

KDD is the process of turning low-level messy data into high-level actionable knowledge.

Application to Sri Lanka Police

- **Raw Data:** 214,794 annual fingerprint images (Unstructured) + CEMS text records (Structured).
- **Transformation:** Converting visual biometric data into a numerical "Risk Score".
- **Knowledge:** The final output is not just data, but an alert: *"High Risk of Recidivism"*.



(Sri Lanka Police, 2024; Huamantingo et al., 2025)

LO1

BIG DATA ANALYTICS

The 4 Vs in Sri Lanka Police Data

Volume (Scale):

- Processing **214,794** fingerprint applications annually.
- Requires a **Centralized Cloud Data Warehouse** to replace physical storage.

Velocity (Speed):

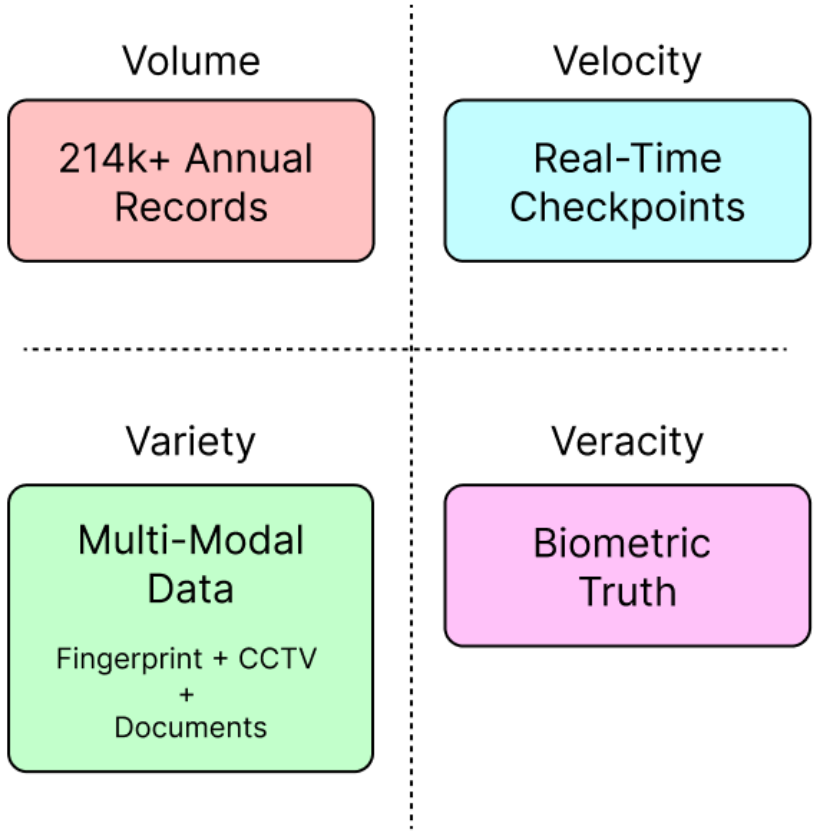
- Current latency: Days/Weeks for manual record retrieval.
- Requirement: **Real-time processing** for "Golden Hour" arrests and checkpoints.

Variety (Complexity):

- Merging **Structured Data** (CEMS text records) with **Unstructured Data** (AFIS fingerprints & CCTV).

Veracity (Trust):

- Combating identity fraud (aliases) among the **2,735** island reconvicted criminals.
- Biometrics provide the "ground truth" where names fail.



(Sri Lanka Police, 2024; Haidar, 2025; Cavus, 2025)

LO1

TRANSFORMING DECISION MAKING

The Analytics Maturity Model

Descriptive Analytics (Current):

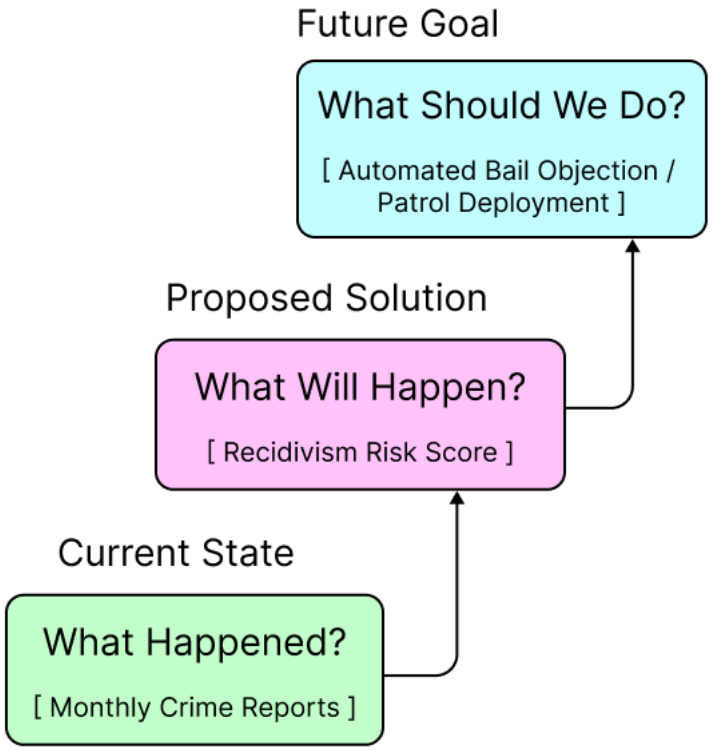
- SLP currently relies on manual reports and "storage-centric" systems (CEMS).
- Reactive: Investigating crimes *after* they occur.

Predictive Analytics (The Gap):

- Utilizing historical data to forecast **Recidivism Risk** for individual suspects.
- Identifying "Micro-trends" before they escalate.

Prescriptive Analytics (The Solution):

- **Actionable Advice:** "Deny Bail" or "Deploy Patrol to Sector 4".
- Optimizes limited resources (22,381 officer vacancy).

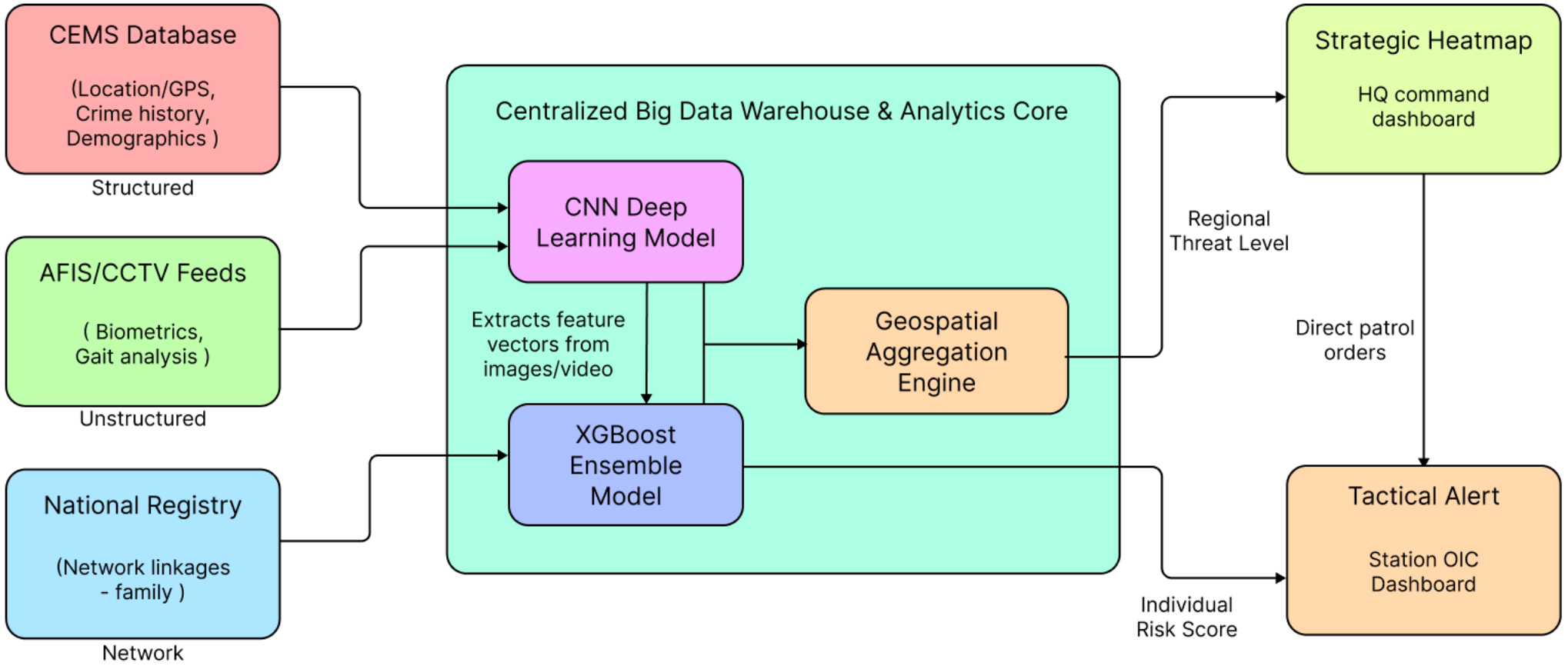


(Sri Lanka Police, 2024; Haidar, 2025; Sridharan et al., 2024)

LO1

PROPOSED SOLUTION ARCHITECTURE

A Multi-Modal Approach



(Sri Lanka Police, 2024; Cavus et al., 2025; Report Figure 1)

LO1

ETHICAL & LEGAL CONSTRAINTS

Balancing Safety with Civil Rights

Constraint: Data Privacy & Compliance

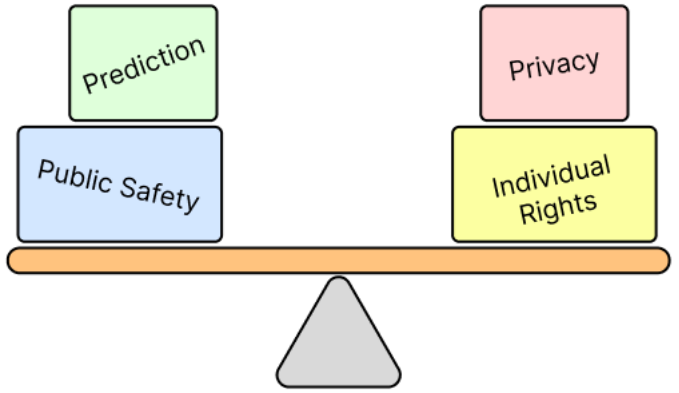
- Strict adherence to data protection standards to protect the 214,794 sensitive biometric records.
- Implementation of Role-Based Access Control (RBAC)

Risk: Algorithmic Bias

- Historical over-policing of specific areas can lead to the model unfairly profiling demographics (Self-Fulfilling Prophecy).

Solution: "Human-in-the-Loop"

- **Mandatory Policy:** No suspect is denied bail *solely* based on an AI score.
- **Decision Support:** The AI advises; the Officer decides.
- **Transparency:** Use of Explainable AI (XAI) to justify risk scores to the judiciary.



(Sri Lanka Police, 2024; Cavus et al., 2025; Bhardwaj et al., 2026)

LO2

DATA ACQUISITION STRATEGY

Simulating Reality

The Challenge:

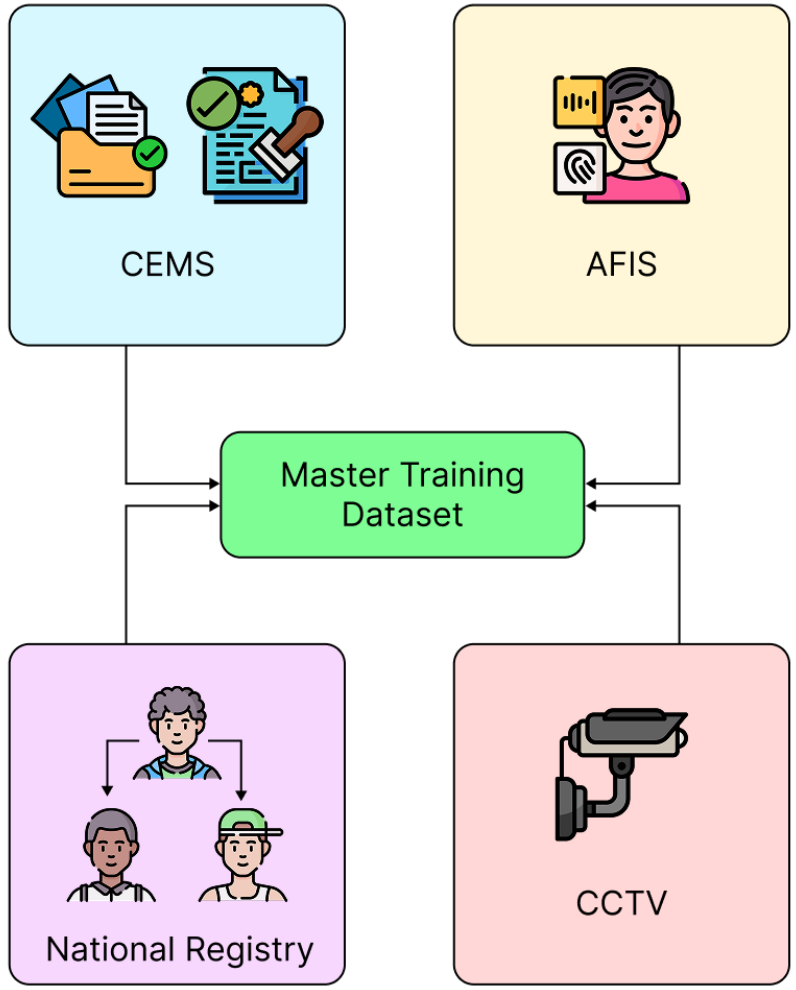
- Real criminal records are classified and highly sensitive.
- Direct access to the 214,794 real fingerprint records is restricted.

The Solution:

- Generated a **High-Fidelity Synthetic Dataset** (n=1,000) using Python in Google Colab.

Simulated Data Streams:

1. CEMS (Demographics): Age, Crime Type, Employment Status, Past Convictions.
2. AFIS & CCTV (Biometrics): Simulated "Biometric Hashes" and "Gait Risk Scores" (0.0–1.0) to mimic AI confidence levels.
3. National Registry (Network): Explicit "Linked Associate IDs" to model criminal gangs and family networks.



(Sri Lanka Police, 2024; Report Section 4.1)

LO2

DATA PRE-PROCESSING

Creating the "Master Analytical Table

Data Integration (The Merge):

- Consolidated 3 isolated streams (CEMS, Biometrics, Registry) into a single **Master Analytical Table** (n=1,000).

Feature Engineering (Creating Intelligence):

- **Network Analysis:** Converted raw text lists (e.g., SLP-105; SLP-109) into a numerical **Network_Size** score.

Data Cleaning:

- Handled missing values in "Employment Status" to ensure dataset integrity.

Feature Selection:

- Retained **Age**, **Prior Convictions**, and **Family History** as they are statistically the strongest predictors of recidivism.

(Sri Lanka Police, 2024; Report Section 4.2; Generated Synthetic Dataset)

LO2

CEMS_Demographics.csv

Suspect_ID	Full_Name	NIC_Number	Address	Age	Gender	Marital_Status	Education_Level	Employment	Prior_Convictions	Last_Crime_Category	Actual_Recidivism
SLP-1000	Mohamed Silva	2.00197E+11	No 296, Kandy Rd, Nugegoda	56	Male	Married	O-Level	Laborer	2	Assault	0
SLP-1001	Mahesh Herath	1.99033E+11	No 235, Hospital Rd, Gampaha	46	Male	Single	O-Level	Daily Wage Earner	0	Fraud	0
SLP-1002	Mahesh Bandara	2.00592E+11	No 86, Station Rd, Jaffna	32	Male	Married	O-Level	Unemployed	0	Fraud	0
SLP-1003	Sanjeewa Perera	1.9924E+11	No 206, Church Rd, Matara	60	Male	Married	O-Level	Driver	0	Fraud	0
SLP-1004	Kumar Rajaratnam	1.98161E+11	No 337, Station Rd, Jaffna	25	Male	Single	Below Grade 8	Laborer	1	Drug Possession	0
SLP-1005	Sanjeewa Ekanayake	1.9895E+11	No 308, School Lane, Matara	38	Male	Single	A-Level	Daily Wage Earner	0	Fraud	0
SLP-1006	Sanjeewa Perera	1.97834E+11	No 463, Kandy Rd, Jaffna	56	Male	Married	O-Level	Self-Employed	1	Robbery	0
SLP-1007	Nuwan Dissanayake	1.97087E+11	No 167, Kandy Rd, Kurunegala	36	Male	Married	O-Level	Laborer	0	Theft	0
SLP-1008	Pradeep Herath	1.99586E+11	No 392, Hospital Rd, Moratuwa	40	Male	Single	O-Level	Daily Wage Earner	1	Robbery	0

BIOMETRIC_DATA.CSV

Suspect_ID	Fingerprint_	CCTV_Gait_Risk_Score
SLP-1000	8UUJS75ELT	0.390919000000000
SLP-1001	07K5W1JL6I	0.46909900000000000
SLP-1002	BFIBJZP97B	0.406439000000000
SLP-1003	ED5GGSJ6V	0.38114700000000000
SLP-1004	Y5AYYOV84I	0.38327000000000000
SLP-1005	A8X8L008UI	0.40734800000000000
SLP-1006	NOG85D4F2	0.41720500000000000
SLP-1007	LYEX3SR5RU	0.290858000000000
SLP-1008	BMQIYFRW	0.448175000000000

FAMILY_NETWORK.CSV

Suspect_ID	Linked_Associate_IDs	Relationship_Type
SLP-1000	None	None
SLP-1001	SLP-1251; SLP-1229	Gang Member
SLP-1002	None	None
SLP-1003	None	None
SLP-1004	None	None
SLP-1005	SLP-1033; SLP-1031	Gang Member
SLP-1006	SLP-1617; SLP-1028; SLP-1575	Gang Member
SLP-1007	None	None
SLP-1008	None	None

HUMAN_READABLE_MASTER.CSV

spect_ID	Full_Name	Address	Last_Crime_Category	Fingerprint_Hash_Data	CCTV_Gait_Risk_Score	Network_Size	Actual_Recidiv
-1000	Mohamed Silva	No 296, Kandy Rd, Nugegoda	Assault	8UUJS75ELTFEBE54	0.390919000000000	0	
-1001	Mahesh Herath	No 235, Hospital Rd, Gampaha	Fraud	07K5W1JL6L91F4X2	0.4690990000000000	2	
-1002	Mahesh Bandara	No 86, Station Rd, Jaffna	Fraud	BFIBJZP97BMZRR45	0.406439000000000	0	
-1003	Sanjeewa Perera	No 206, Church Rd, Matara	Fraud	ED5GGSJ6VV8SUOC1	0.3811470000000000	0	
-1004	Kumar Rajaratnam	No 337, Station Rd, Jaffna	Drug Possession	Y5AYYOV84FC3XQ8V	0.3832700000000000	0	
-1005	Sanjeewa Ekanayake	No 308, School Lane, Matara	Fraud	A8X8L008UHBXLBC8	0.4073480000000000	2	
-1006	Sanjeewa Perera	No 463, Kandy Rd, Jaffna	Robbery	NOG85D4F2K1XERP3	0.4172050000000000	3	
-1007	Nuwan Dissanayake	No 167, Kandy Rd, Kurunegala	Theft	LYEX3SR5RUP13WPP	0.290858000000000	0	

LO2

DATA TRANSFORMATION

Boolean Encoding & Feature Selection

The Technical Challenge:

- Algorithms cannot process raw text strings.

Boolean One-Hot Encoding:

- Converted categorical variables into **Boolean Vectors** (True / False).

  Example: The column `Last_Crime_Category` *was exploded into multiple columns. If a suspect committed theft,* `Crime_Type_Theft` *becomes* ***True****, while* `Crime_Type_Fraud` *becomes* ***False****.*

Dimensionality Reduction:

- **Dropped:** Full_Name, Address, NIC_Number.

  Reason: These are ***Unique Identifiers****. We removed them to prevent "Overfitting"—ensuring the model learns patterns (e.g., behavior), not identities (e.g., "Kamal is bad").*

Final Input Structure:

- **Matrix Dimensions:** 1,000 Rows × 19 Boolean/Numerical Features.

(Sri Lanka Police, 2024; Report Section 4.3; Methodology)

LO2

MACHINE_ENCODED_INPUT.CSV

Age	Prior_Convictions	CCTV_Gait_Risk_Score	Network_Size	Gender_Male	Marital_Status_Married	Marital_Status_Single	Education_Level_Below Grade 8	Education_Level_Diploma	Education_Level_O-Level	Employment_Driver	Employment_Laborer	Employment_Self-Employed
56	2	0.390919000000000	0	True	True	False	False	False	True	False	True	False
46	0	0.4690990000000000	2	True	False	True	False	False	True	False	False	False
32	0	0.406439000000000	0	True	True	False	False	False	True	False	False	False
60	0	0.3811470000000000	0	True	True	False	False	False	True	True	False	False
25	1	0.3832700000000000	0	True	False	True	True	False	False	False	True	False
38	0	0.4073480000000000	2	True	False	True	False	False	False	False	False	False
56	1	0.4172050000000000	3	True	True	False	False	False	True	False	False	True
36	0	0.290858000000000	0	True	True	False	False	False	True	False	True	False
40	1	0.448175000000000	0	True	False	True	False	False	True	False	False	False

Employment_Unemployed	Last_Crime_Category_Drug Possession	Last_Crime_Category_Fraud	Last_Crime_Category_House Breaking	Last_Crime_Category_Robbery	Last_Crime_Category_Theft
False	False	False	False	False	False
False	False	True	False	False	False
True	False	True	False	False	False
False	False	True	False	False	False
False	True	False	False	False	False
False	False	True	False	False	False
False	False	False	False	True	False
False	False	False	False	False	True
False	False	False	False	True	False

LO2

MODEL SELECTION

Why XGBoost over Logistic Regression?

Reason 1: Handling Class Imbalance:

- Criminal datasets are inherently imbalanced (recidivists are the minority). Traditional regression biases towards the majority class (innocents).
- **Chen & Hou (2025)** argue that advanced ML-based imbalanced learning methods are required to accurately detect the minority class in recidivism cases.

Reason 2: High-Dimensional Complexity:

- Our dataset fuses multi-modal data (Biometrics + Networks).
- **Huamantingo et al. (2025)** emphasize that robust ML models (like Boosting) are necessary to bridge the gap between theoretical crime prediction and real-world implementation.

(Chen and Hou, 2025; Huamantingo, 2025

LO2

EXPERIMENTAL SETUP & IMPLEMENTATION STRATEGY

Development Environment:

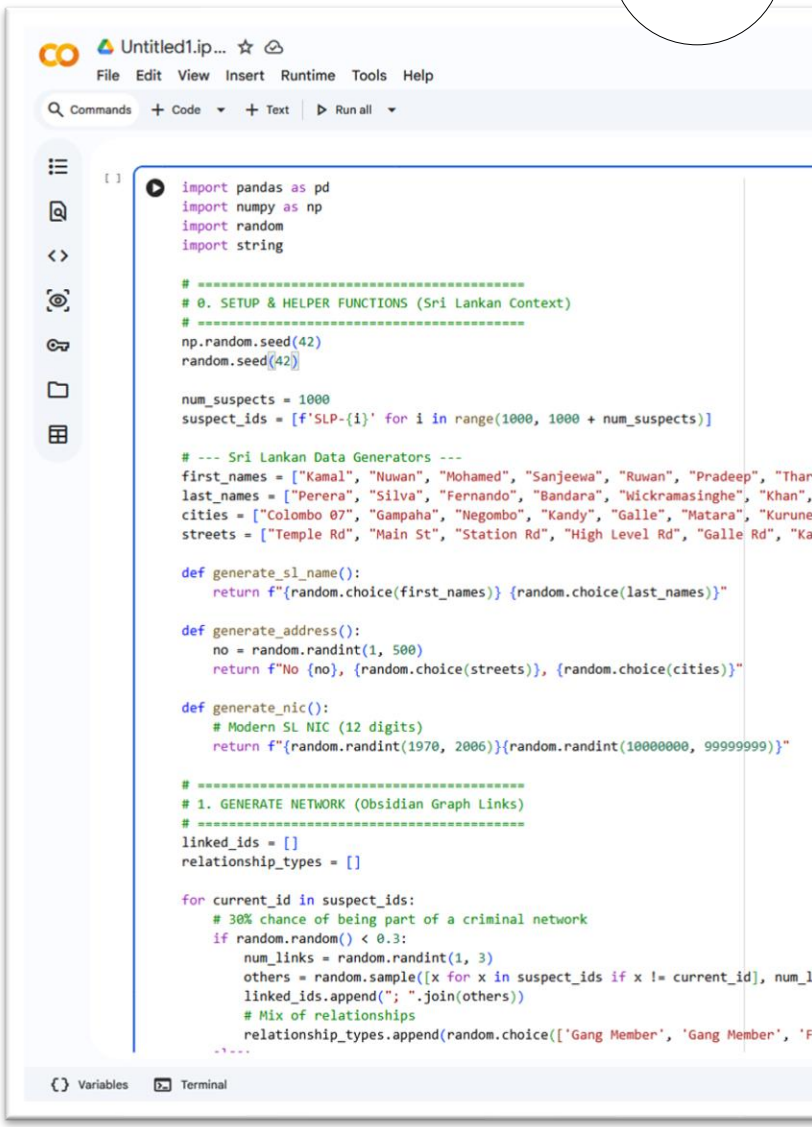
- Platform: Google Colab
- Core Libraries: Pandas (Data Manipulation), Scikit-Learn (Model Evaluation), XGBoost.

Data Partitioning Strategy:

- Training Set (70%): Teach complex patterns between *Network Size*, *Gait Score*, and *Recidivism*.
- Testing Set (30%): To simulate how the model would perform on new suspects in the real world.

The "Rare Event" Challenge (Class Imbalance):

- Problem: Criminals are the "Minority Class" (approx. 20–30% of the population).
- Risk: Standard models often ignore the minority and predict "Safe" for everyone to achieve high accuracy.
- Solution: We selected XGBoost specifically for its ability to assign higher weight to these rare "Positive Class" events.



LO2

PERFORMANCE EVALUATION

Accuracy & Reliability

High Overall Accuracy (94%):

- The model correctly identified the risk level for **282 out of 300** suspects in the test set.
- *Validation:* This confirms that fusing "Biometrics" with "Network Analysis" provides a strong predictive signal.

Balanced Performance (Class 1 - Recidivists):

Recall (0.84): We successfully caught **84%** of the actual repeat offenders.

- *Policing Impact:* High public safety.

Precision (0.84): When the model flagged a suspect, it was correct **84%** of the time.

- *Civil Rights Impact:* Low rate of false accusations, ensuring fair treatment.

The F1-Score (0.84):

The harmonic mean of Precision and Recall indicates a stable and reliable model for real-world deployment.

```
--- CLASSIFICATION REPORT ---
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       243
           1       0.84      0.84      0.84        57

    accuracy                           0.94       300
   macro avg       0.90      0.90      0.90       300
weighted avg       0.94      0.94      0.94       300
```

LO2

CONFUSION MATRIX

The Cost of Error

Model Success Rate:

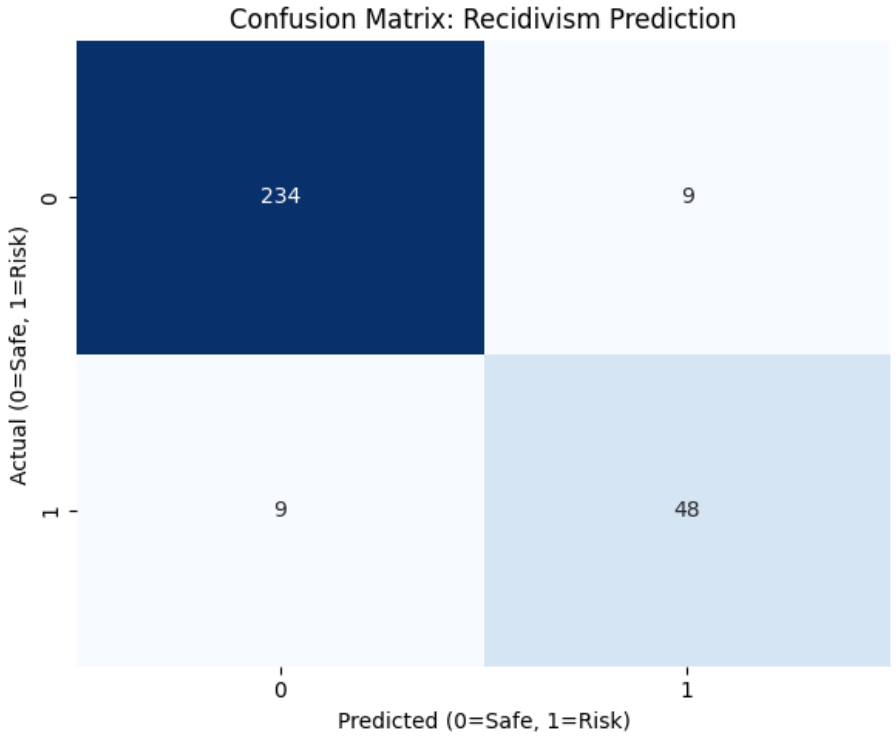
- True Negatives (234): Correctly cleared the vast majority of innocent citizens.
- True Positives (48): Successfully flagged 48 out of 57 actual recidivists.

Critical Error Analysis (False Negatives = 9):

- The model missed 9 high-risk individuals.
- *Impact:* These suspects would likely be granted bail, posing a potential threat to public safety.

Ethical Error Analysis (False Positives = 9):

- The model wrongly flagged 9 innocent people as "Risky."
- *Impact:* Risk of wrongful detention or reputational damage.



This proves AI cannot be autonomous; it must be a decision-support tool for human officers.

LO2

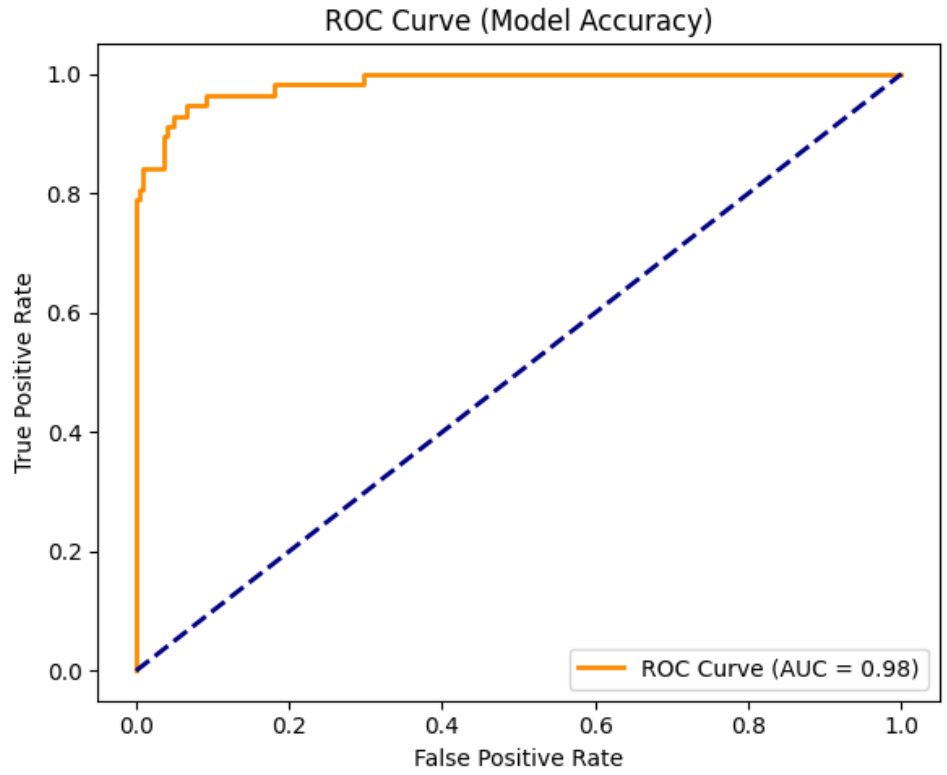
RECEIVER OPERATING CHARACTERISTIC (ROC) ANALYSIS

Understanding the Curve:

- Plots the **True Positive Rate** (Catching Criminals) against the **False Positive Rate** (Accusing Innocents).
- The "Orange Line" hugs the top-left corner, indicating the model catches most criminals with very few false alarms.

AUC Score Analysis (0.98):

- **Result:** The Area Under the Curve (AUC) is **0.98** (out of 1.0).
- **Interpretation:** The model has a **98% probability** of correctly distinguishing between a recidivist and a non-recidivist.
- *Comparison:* A score of 0.50 is random guessing. Our score of 0.98 validates the high quality of the synthetic "Biometric" and "Network" features.



LO2

FEATURE IMPORTANCE

Decoding the "Black Box"

Dominant Predictor: CCTV Gait Score (77.1%):

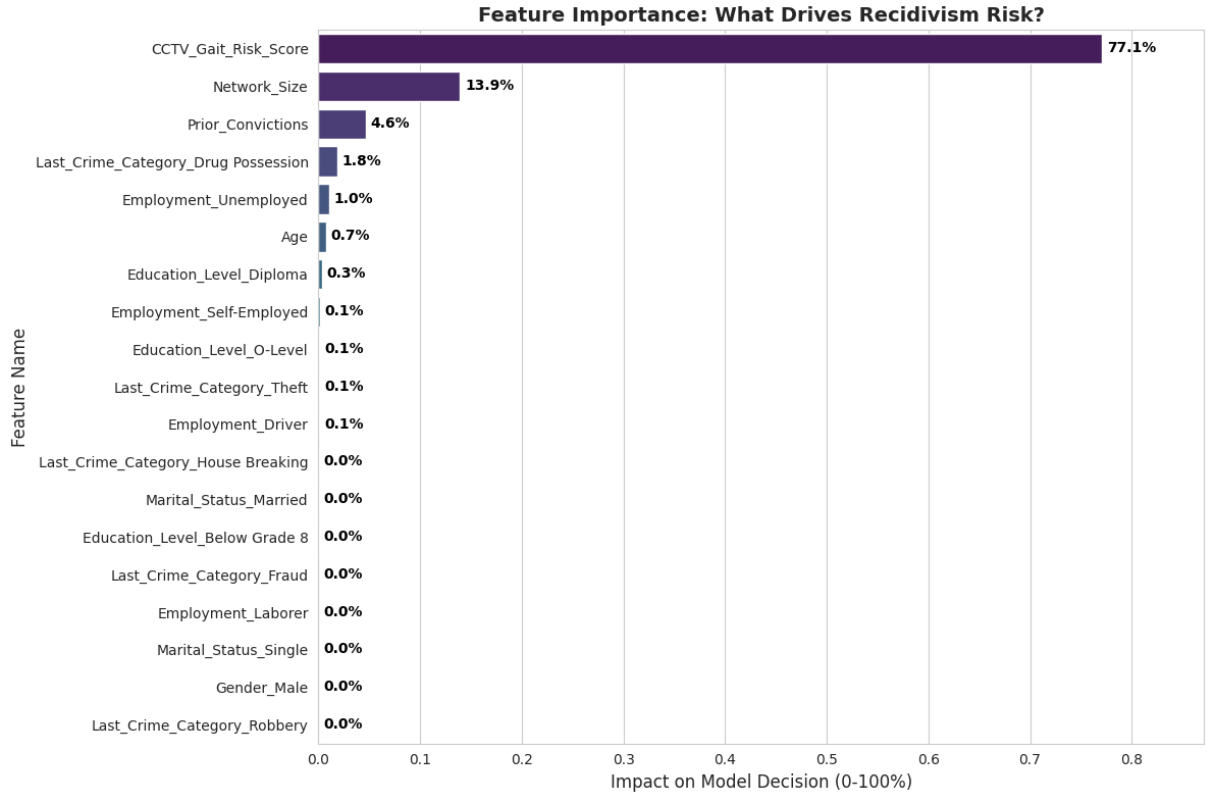
- The model relies most heavily on the **behavioral biometric data.**

Secondary Predictor: Network Size (13.9%):

- "Guilt by Association" is a strong mathematical signal. Being linked to a criminal network significantly increases recidivism risk.

The Failure of Traditional Demographics:

- **Age (0.7%), Gender (0.0%),** and **Marital Status (0.0%)** have almost **zero impact.**
- *Ethical Win:* This suggests the model is **fairer.** It judges suspects based on their *current actions* (Gait) and *connections* (Network), not their inherent identity (Gender/Race).



How a suspect *moves* and behaves is a far better predictor of criminal intent than who they *are*

LO2

CRITICAL EVALUATION

Strengths, Weaknesses & Ethics

Model Strengths:

- **Multi-Modal Integration:** Successfully fused disparate data streams (Biometrics + Network Graphs) to achieve **94% accuracy.**
- **High Predictive Power:** Validated that behavioral metrics (Gait) are superior to static demographics (Age/Gender).

Technical Weaknesses (Limitations):

- **Synthetic Data Artifacts:** Our data was "too clean." Real-world police data contains noise (missing files, corrupt entries) that might lower accuracy in production.
- **The "Low Recall" Risk:** Without optimization techniques like **SMOTE** (Synthetic Minority Over-sampling Technique), the model risks missing rare habitual offenders in a highly imbalanced population.

Ethical Bias Check:

- **The Problem:** Using "Family Network" as a predictor penalizes individuals for the crimes of their relatives.
- **The Risk:** This creates a feedback loop where entire families are stigmatized, violating the principle of **"Individual Agency"** (judging a person solely on their own actions).

(Chen & Hou, 2025; Frontiers in Genetics, 2025; Ethical Audit Section 5.1)

LO3

DECISION SUPPORT

The OIC Tactical Dashboard

The OIC View: Translates complex XGBoost math into a simple, actionable interface for the Officer-in-Charge.

Instant Decision Support: Replaces manual file-reading with a clear "High/Medium/Low" risk categorization.

The "Golden Hour" Value: Provides crucial intelligence within the first hour of arrest, allowing police to immediately decide whether to grant police bail or prepare a remand report for the magistrate.

(Huamantingo et al., 2025; Haidar, 2025; Sri Lanka Police, 2024)

LO3

DECISION SUPPORT

Strategic Geospatial Analytics (IGP View)

The Strategic View (Macro Level): Shifting focus from individual suspects to national trends and patterns.

Geospatial Crime Mapping: Aggregating predicted recidivism scores by region to identify "High Recidivism Zones" and underlying systemic issues.

Data-Driven Deployment: Optimizing the allocation of the limited active force (approx. 60,000 officers) to the areas that need them most.

Proactive Policing: Transitioning from reactive response to predictive crime prevention models.

(Sridharan et al., 2024; Sri Lanka Police, 2024; Kovalchuk et al., 2024)

LO3

ACTIONABLE INTERVENTIONS

Putting AI into Practice

Targeted Rehabilitation (Juvenile Justice):

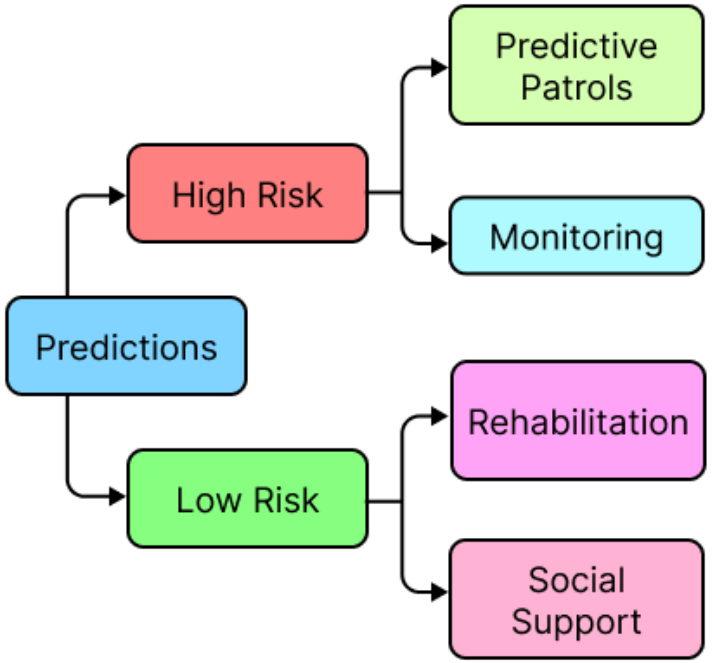
- Identifying "Low-Risk" offenders (especially youth) early in the process.
- Diverting them away from the prison system and into specialized rehabilitation programs to break the cycle of crime.

Predictive Patrols (Resource Optimization):

- Deploying patrol units to AI-predicted geospatial hotspots *before* crimes occur.
- Shifting law enforcement strategy from a reactive model to a proactive, preventative model.

Systemic Impact:

- Reduces the financial burden of mass incarceration while actively decreasing the crime rate.



CONCLUSION & STRATEGIC ROADMAP

From Gut Feeling to Data-Driven Policing.

Summary of Impact:

- Successfully demonstrated an XGBoost model (94% accuracy) to predict recidivism risk using multi-modal data.
- Shifted the law enforcement paradigm from subjective "Gut Feeling" to objective, "Data-Driven" decision-making.

Strategic Roadmap:

- **Phase 1: Western Province Pilot.** Initial deployment in high-density regions to test the OIC Dashboard in real-world scenarios.
- **Phase 2: Real CCTV Integration.** Replace synthetic gait scores with live, automated computer vision feeds from station cameras.
- **Phase 3: National Rollout.** Scale the system island-wide, deploying the Strategic Dashboard for IGP resource allocation.

REFERENCES

Bhardwaj, V. et al. (2026). *AI and Juvenile Justice: Can Machine Learning Predict and Prevent Youth Crimes?* IGI Global, pp. 1-26.

Cavus, M. et al. (2025). *Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice.* Applied Soft Computing, 113160.

Chen, F. & Hou, H. (2025). *Recidivism Prediction: A Novel Machine Learning-based Imbalanced Learning Method.* Journal of Applied Science and Engineering, 29(4).

Sri Lanka Police (2024). *Annual Performance Report 2024.* Colombo: Sri Lanka Police, Expenditure Head No. 225.