University of Staffordshire

A·P·I·I·T
Inspire love for learning

A Biometric-Driven Machine Learning Framework for Recidivism Prediction in Sri Lanka

COMP60022 : Decision Analytics 1

Name: M.G.L.N Kumararathna

Staffordshire ID : k039799n

APIIT ID : CB013366

BSc (Hons) in Computer Science

Date of Submission: 9th February 2026

Acknowledgement

I would like to express my sincere gratitude to Dr. Rajitha Nawarathna for his invaluable guidance and expertise throughout this module. I also wish to extend my appreciation to the Asia Pacific Institute of Information Technology (APIIT) and Staffordshire University for providing the empowering academic environment and resources necessary to complete this assignment.

Table of Contents

1. Introduction and Problem Definition 1
1.1 Organizational Overview: The Sri Lanka Police 1
1.2 The Core Problem: Manual Inefficiencies and Resource Gaps 1
1.3 Contextual Analysis (Who, What, Where, Why, When) 2
1.4 Strategic Objectives and Key Research Questions 3
2. LO1: Knowledge Discovery and Conceptual Framework 5
2.1 The KDD Process in Criminal Justice 5
2.2 Big Data Analytics: The 4 Vs in SLP Data 7
2.3 Organizational Decision Making Transformation 8
2.4 Scope and Constrains of the Predictive Framework 9
3. Current System Analysis and Identified Gaps 11
3.1 Existing Process Flow 11
3.2 Administrative Resource Bottlenecks 12
3.3 Cross-Jurisdictional Challenges 13
3.4 Limitation of Current Systems 13
4. LO2: Analytical Solution – Machine Learning Model 16
4.1 Data Selection and Feature Engineering 16
4.1.1 The Multi-Model Dataset Structure 16
4.2 Analytical Methodology 17
4.3 Practical Application: Risk Scoring 19
4.4 Model Evaluation and Validation 20
4.4.1 Evaluation Framework 20
4.4.2 Pilot Study Results 21
5. LO3: Decision Support Outputs and Intervention Strategies 26
5.1 Strategic Impact on National Security 26
5.2 Visual Analytics and Dashboards 27
5.3 Evidence-Based Intervention Strategies 27
6.Risks, Ethics, and External Factors 29
6.1 Data Privacy and Compliance 29
6.2 Bias and Ethical Considerations 29
6.3 Technical Infrastructure Constraints 30

6.4 Human Capital and Training.....30
7.Recommendations.....31
7.1 Strategic Roadmap.....31
7.2 Organizational & Ethical Recommendations.....32
8.Limitation.....33
8.1 Data Limitations (Synthetic vs. Real).....33
8.2 The "Cold Start" Problem.....33
8.3 Infrastructure and Connectivity.....33
8.4 Scope of Prediction.....33
Conclusion.....34
References.....vi
Appendix.....vii

List of figures

Figure 1: Architecture of the proposed solution 16
Figure 2: CEMS Demographics hypothetical dataset 21
Figure 3: Biometric hypothetical dataset 21
Figure 4: Family network hypothetical dataset 22
Figure 5: XGBoost performance hypothetical dataset 22
Figure 6: Confusion matrix of XGBoost predictions 23
Figure 7: (ROC) accuracy of the performance 24
Figure 8: Feature Importance 25

List of tables

Table 1 summary of gaps 15

1. Introduction and Problem Definition

1.1 Organizational Overview: The Sri Lanka Police

Sri Lanka Police (SLP) is the primary law enforcement and national security organization for the island nation, operate with a decentralized network of 607 police stations and 45 functional divisions island wide(Sri Lanka Police, 2024). SLP organization is responsible for lots of duties, ranging from traffic controls and criminal investigation to counter terrorism and civil security. Inside this complex organization structure there is a division called Criminal Records Division (CRD) that act as the central nerve center to maintain the criminal history of Sri Lankan population. The CRD task is data collection, storage, and retrieval of biometric data, fingerprints is the main priority because that is the primary method of identification in criminal proceedings.

In 2024 alone the SLP managed a massive volume of data, by processing 214,794 fingerprint application (Sri Lanka Police, 2024). This clearly shows that the organizations role as a generator of "big data", processing a data base of information that no only includes biometric data but also demographic details, crime locations, and also modus operandi. Even though the SLP already have Arrested Monitoring and Information System (AMIS) and the Complaint and Enquiry Management System (CEMS), they operate primarily on a "storage centric" model. The collected data is never been used for strategic forecasting. As noted by Kovalchuck (2024), modern judicial support systems must evolve beyond just storing data to become active tools for decision support. SLPs currently posses the raw material for intelligence but they lacks the knowledge discovery mechanism to transform it into actionable insights.

1.2 The Core Problem: Manual Inefficiencies and Resource Gaps

The main problem SLP face is that the gap between operational demands and organization capacity is keep widening. By only relying on manual, reactive methodologies this make it more exacerbated. According to the Annual Performance Report (2024), the SLP still faces severe human resources crisis, recording a vacancy of 22,381 officers working regularly (Sri Lanka Police, 2024). This 25% deficiency directly affects the overall police work. Because of that investigations are delayed, surveillance is sporadic, and the ability to manually monitor high risk suspects is severely compromised.

Resource allocation is become unproductive in the SLP because they have to prioritize every suspect equally. In the current system, the SLP save data through physically or localized servers with responsible

police stations. In a situation where suspect is arrested in the southern province, there is no mechanism to alert the officers that this individual has a high probability of recidivism based on a behavior pattern identified in the western province. This disconnect is important because, according to the Annual Performance Report (2024), out of 27,000 registered criminals in 2024, they identified 2,735 “island reconvicted criminals”. These repeat offenders shows a clear high risk subset of the population.

The core problem is not the absence of data, but the absence of predictions. The current system can identified a repeat offender only after they did the second crime and if processed. Huamantingo, Cano-Lengua, and Rodriguez (2025) argue that there is often a significant gap between research capabilities and practical implementation in law enforcement. For SLP, the gap is while they already have the data they need, their system is incapable of identify the statistical minority of potential recidivists among thousands of suspects who did the first crime. The system shows needing of a machine learning framework to analyze the data and identify patterns. Without ML framework the SLP can’t effectively their limited workforce, leading to inefficiencies that compromise public safety (Sridharan, 2024).

1.3 Contextual Analysis (Who, What, Where, Why, When)

- **Who (Stakeholders):** the primary stakeholder is the CRD of the SLP, which managing the central database. Second, we have the judiciary, which relies on accurate risk analysis for bail decisions, and 22,381 officers that will be alleviated by the automated prioritization (Sri Lanka Police, 2024). Furthermore, as Bhardwaj et al. (2026) note, such a system will impact the youth and first time crime offenders. They require a rehabilitation rather than incarceration.
- **What (Technological Shift):** the project proposes a shift from descriptive analytics (what happened?) to predictive analytics (what will happen?). This involves implementing a biometric driven ML framework capable of calculating a "recidivism risk score" for every arrested suspect. This goes with Haidar’s (2025) assertion that says “predictive analytics is the next necessary evolution for effective recidivism management”.
- **Where (Operational Scope):** the solution addresses a nation wide problem but only targets the 607 police stations as the point of action. The technology gap between the urban centers and rural outposts requires a centralized solution that can be access via secure, wide area network to overcome the geographic data silos that exist in Sri Lanka.
- **Why (Strategic Imperative):** the main focus here is public safety and improving the efficiency. Hiring new officers is not the solution they need. Instead, best solution is that make the existing force more efficient by accurately predicting while support the decision with probability that who

is likely to re-offend. Then SLP can manage there surveillance resource and give more priority to top 10% of high individuals, rather than diluting their efforts across the entire population (Chen and Hou, 2025).

- **When (Timeline):** the problem is critical and immediate, as shown in the 2024 performance data (Sri Lanka Police, 2024). Proposed solution envisions a “roadmap to 2026”, creating a transition time where SLP can digitalized their manual books and the ML model is trained on the existing data to become operational for the future years.

1.4 Strategic Objectives and Key Research Questions

The final objective of this report is to proposed a robust, biometric driven machine learning framework that utilizes knowledge discovery in databases to predict recidivism risk. This proposed framework will able to transform the SLPs operational model from reactive to proactive. To achieve that the report focus on advanced algorithms that capable of handling imbalanced data. Because like the annual report mentioned, from 27,000 registered criminals, 2,735 are identified as repeat offenders (Sri Lanka Police, 2024).

However, this integration of artificial intelligence in to criminal justice systems is not just a technical challenge. It is also a ethical challenge. When we try to automate the decision making, problems like issues of bias and transparency will arise. There for a secondary objective is to ensure the veracity and fairness of the framework. Cavus (2025) emphasizes the need for "transparent and bias-resilient AI frameworks," particularly in criminal justice where a false prediction can infringe on civil liberties.

Based on these objectives, this report addresses the following Key Research Questions:

1. Multi-Modal Technical Efficacy: How can a hybrid architecture utilizing Deep Neural Networks for unstructured biometric/CCTV processing and XGBoost for structured record analysis capture the full complexity of criminal behavior (Cavus et al., 2025; Sridharan et al., 2024)
2. Imbalance Management: What specific learning methods, when combined with differential equation algorithms, can best optimize prediction accuracy for the "statistical minority" of the 2,735 island reconvicted criminals (Chen and Hou, 2025)
3. Ethical Transparency & The Black Box: How can Explainable AI (XAI) techniques, such as SHAP, be integrated into Deep Learning models to ensure that risk assessments are legally justifiable and transparent to the judiciary (Cavus et al., 2025)
4. Operational Integration of the 4 Vs: How can the "Variety" of data (fingerprints and CCTV) and the "Volume" of 214,794 records be effectively mapped to a centralized SLP system to bridge the

gap between academic research and practical implementation (Huamantingo, Cano-Lengua, and Rodriguez, 2025; Sri Lanka Police, 2024)

2. LO1: Knowledge Discovery and Conceptual Framework

A strong conceptual foundation is critical when the transition from traditional to intelligence-led policing. For the SLP this means that moving from simply storing data in their books and computers and also in databases to actively extracting knowledge from it. This section go through the theoretical framework that necessary for this transformation. Specially the knowledge discovery in databases process and the big data 4Vs model. This section explains how these concepts can give the solution for the problems that exist in CRD, such as the severe shortage of officers and the need for real time decision making (Sri Lanka Police, 2024).

2.1 The KDD Process in Criminal Justice

Knowledge discovery in databases, which means that the process of turning low-level messy data (name, fingerprint images, cctv footage) into high-level (decisions, predictions, identifying patterns). For the SLP its not just a technical task. It is the organizational necessity to manage the 214, 794 fingerprint applications processed in 2024 making (Sri Lanka Police, 2024).

The KDD process consists of five specific stages that turn raw police data into actionable insights:

1. Data Selection

First we need to decide what data is relevant. From the SLP data, to predict recidivism not all the data will be useful here. In this stage, we should select specific target datasets.

- **Target Data:** The primary focus is on the records of the 2,735 “Island Reconvicted Criminals" identified in the 2024 Performance Report (Sri Lanka Police, 2024).
- **Selection Criteria:** We select static factors (which do not change, such as date of birth, gender, and age at first arrest) and dynamic factors (which change, such as employment status and current address). According to Sridharan et al. (2024), selecting the right variables is critical because irrelevant data can confuse the Machine Learning models.

2. Data Preprocessing

Police data is often messy and noisy or sometime incomplete. Because the digitalized data is records and from manual books at local stations. There maybe spelling errors in names, addresses or missing fields such as NIC numbers.

- **Cleaning:** This stage involves removing duplicate records.for instance, ensuring that a suspect arrested in Kandy and Colombo is recognized as the same person.
- **Handling Missing Values:** If a record is missing the "Employment Status," statistical methods are used to fill in the gaps or remove that record so it does not damage the accuracy of the prediction (Huamantingo, 2025).

3. Data Transformation

ML algorithms(neural networks) always need numbers to train, they cannot read fingerprint images or descriptions of a crime directly. The data must be transformed into vectors because when ML model have to transform data, it will consume more energy on that, not on the predictions.

- **Biometric Hashing:** The 214,000+ fingerprint images are converted into numerical hashes or vectors. This allows the computer to compare two fingerprints using math rather than visual inspection.
- **Categorization:** Text descriptions like "House Breaking" or "Gem Theft" are converted into numerical categories (e.g., Property Crime = 1, Violent Crime = 2). This standardizes the data across all 607 police stations (Sri Lanka Police, 2024).

4. Data Mining

This is the phase where algorithms search for hidden patterns in the transformed data.

- **Pattern Recognition:** The system looks for correlations that a human officer might miss. For example, it might discover that suspects aged 18–21 who commit theft in the Western Province have an 85% chance of re-offending within two years.
- **Algorithms:** As suggested by Chen and Hou (2025), this stage uses XGBoost for structured data (records) and Neural Networks for unstructured data (images) to find these complex relationships.

5. Interpretation and Evaluation

The final stage turns the mathematical patterns back into simple English for the police officer.

- **Risk Scoring:** The system outputs a probability distribution and "Recidivism Risk Score" (e.g., High, Medium, or Low).
- **Actionable Knowledge:** Instead of showing the officer a complex graph, the system provides a clear alert: *"High Risk of Re-offense. Recommended Action: Deny Police Bail."* This fulfills the goal of supporting judicial information support (Kovalchuk et al., 2024).

2.2 Big Data Analytics: The 4 Vs in SLP Data

Big Data can't be defined only with how much data you have, you need to look at how complex it is. For the SLP, this proposed transition to a predictive system is a classic big data problem that exist. In this section we going to analyze it using 4Vs framework(Volume, Velocity, Variety and Veracity).

1. Volume (Scale of Data)

Volume is refers to the amount of data you gathered or generated. However, SLP has a massive scale of data. In 2024 only the CRD processed over 214,794 fingerprints. And this is not only happening once, it is a continuous yearly influx. The challenge is that storing this amount of data locally is impossible. A single station cannot hold the records of the entire area. A centralized cloud data warehouse is the answer to store millions of historical records. Then a ML model will have enough data to learn from (Haidar, 2025).

2. Velocity (Speed of Processing)

Velocity in big data is refers to how fast the data is created and how fast it must be processed. In modern policing this is critical because officers need to take fast decision on some situations. For a example, "the golden hour", when a person is arrested, the police often have limited time to identify them and decide on detention. To support that, a big data system must process a fingerprint check and return the results within seconds. But the current system will take days if the manual inquires sent to the CRD. System must handle the other data that coming from 607 different departments simultaneously in real time. If the system is slow the velocity of criminals will exceeds the velocity of the police (Sri Lanka Police, 2024).

3. Variety (Different Forms of Data)

Variety is often seen in police data. And it is the most complex part of this because the SLP data comes in many different shapes and sizes. This is were the proposed framework become more useful and essential. For the structured data such as names, ID numbers, dates, and crime details in CEMS database, instead of using a neural network, we can use XGBoost algorithm, because it is much better with tabular data. For the unstructured data such as fingerprint images CCTV footages and mugshots, we can use deep learning (CNN) because statistical models cannot see the images. CNN will see the fingerprint and extract its features before combining it with the text data. Cavus (2025) argue that handling this variety is the key to a modern justice system.

4. Veracity (The Truthfulness of Data)

Veracity in big data refers to the quality and also the accuracy of the data. In a justice system one single decision can ruin a innocent life, so the data must be trusted. Criminal often hide their real names in many

cases. When its comes to text based search this will have a higher impact on the results. But fingerprints data provide high veracity. Even if the criminals lies they cant change their biometric. The 2024 report highlights that the Automated Fingerprint Identification System (AFIS) is the core tool for establishing the true identity of the 2,735 recidivists (Sri Lanka Police, 2024). When discussing veracity, data is also not biased. If a specific area people have arrested in the past because due to prejudice, the data most likely have low veracity about the actual crime rates. The framework must use bias-resilient framework to correct for this (Cavus, 2025).

2.3 Organizational Decision Making Transformation

Current system of SLP is relying on human intuition and law. Its called reactive decision making. This report ultimate goal is that transform this reactive system into proactive system by implementing big data and ML to improve their decision making. The proposed framework will transform SLPs decision making at three levels.

1. Operational Level

- **Current Status:** when a officer in charge at a police station arrest a suspect, they doesn't have a quick way to know that this person is a repeat offender from another district. They might release him on bail if the reason of arrest is not important.
- **Transformed State:** OIC will do a quick fingerprint scan of the suspect and the system will instantly flags the suspect with a correct risk score and the recidivism probability.
- **Impact :** the officer can create a data backed objection to bail. This will prevents dangerous criminals from returning to the street very fast. And this directly support the SLPs goal of reducing crime despite the 22,381 officers shortage (Sri Lanka Police, 2024).

2. Tactical Level

- **Current Status:** current system of SLP is reactive. Senior officers deploy surveillance patrol based on complains received.
- **Transformed Status:** the proposed system not just only predict the recidivism of a suspect. The system also predict hotspots. It tells, "based on released dates and with past analyzed patterns, there is a high risk of property theft in this area by this weekend".
- **Impact:** SLP can deploy officers before the crime happens. This is known as predictive patrol. This capabilities transform the SLP to maximizes the utility of their limited resources (SriDharan, 2024).

3. Strategic Level

- **Current Status:** Inspector general of police rely on annual reports (like the 2024 performance report) that look at the past year.
- **Transformed Status:** the proposed system have a real time dashboard shows crime trends as they emerge.
- **Impact:** the police command can allocate their budgets and training resources to the areas that has highest predicted growth in recidivism, ensuring their long term sustainability (Haidar, 2025).

2.4 Scope and Constrains of the Predictive Framework

To ensure the framework is realistic and ethical, it is crucial to define its boundaries even though the potential of this system is vast.

Scope of the Solution

Target population of the system is primarily arrested suspects, because this is specially designed for predicting recidivism (re-offending). It is not a tool for surveillance of the general population, however, by analyzing historical data it can predict were might crime happens before its happened. It only analyzes individuals who already in the legal system. The target geographic reach is all the 607 police stations, ensuring that a rural station has the same intelligence support as the Colombo headquarters (Sri Lanka Police, 2024). Data types that this system will focus on are biometric data and criminal history from CEMS records.

Constrains and Challenges

1. **Data Privacy and Ethics:** primary constraint is that the ethical use of data. Because predicting a person commit a crime might be controversial. In tech world every one knows AI is kind of a "black box" when its come to a decision making. If the system denies someone bail, the police or the court needs to know why. But traditional AI cannot explain it itself. That's where the use of XAI come in. techniques like SHAP ensures that every risk score comes with an explanation. This will make it legally justifiable (Cavus, 2025).
2. **Algorithmic Bias:** AI is lean from past data. If the historical data shows that police arrested more from low income areas it will affect the learning curve and the ML model might unfairly target the poor people. The ML model must be constantly audited for bias. SLP can't blindly follows and trust the algorithm (Bhardwaj, 2026).

3. **Technical Connectivity:** currently SLP operates on many rural areas that have poor internet connections. This proposed system requires a stable connection to the central database to check the records. As a solution this design must include a offline first capabilities or some localized caches for the station with poor connectivity (Huamantingo, Cano-Lengua and Rodriguez, 2025).
4. **Human Resource Limitation:** as noted in the report (2024), the SLP currently short staffed. The system should be easy to use. If it requires complex training, with currect overwork officers will have hard time using it.

By understanding these scope and constrains, the Sri Lanka Police can implement a system that is not only technologically advanced but also practical, ethical, and suited to the real world needs of the country.

3. Current System Analysis and Identified Gaps

To justify the proposed predictive analytical framework we have to critically evaluate the existing operational model of the Sri Lankan Police. Organization is able to make a good impact to the system through Complaint and Enquiry Management System. Even though these systems exists, the core workflow of the SLP still remain manual as they keep staying on the reactive process that were designed for a old era of policing (Sri Lanka Police, 2024). This section go through and maps current process of SLP, identifies the primary administrative problems caused by the 22,381 officers, and highlights the gaps that why the SLP is failed to predict 2,735 island reconvicted criminals before the situation happens (Sri Lanka Ploice, 2024).

3.1 Existing Process Flow

Current work flow for handle criminal records and identification of suspects is a linear process, which means that these process are mostly handle in single station (station-centric). It needs both physical documentation and a digital entry to create a new data, which creating a disjointed data lifecycle in the current system.

1. Incident Reporting and Manual Entry

When a crime happens and its reported or a suspect is arrested, its start with one of the 607 territorial police stations. Even though computer and modern technology exist, the first point of entry is often a physical ledger known as the Complaint Book or the Information Book (IB). An officer going to manually write the suspects details, such as name, address, NIC number, family details, and a description of incident. This process is safe and secure from cyber criminals if anything doesn't happen physically. However, the details can be misspelled and importantly this is a physical and static, which means another station officer cannot search this if the same suspect caught on a different province (Haidar, 2025).

2. Localized Digital Storage (CEMS)

After the manual entry by officers, another officer is transformed these data into the Complaint and Enquiry Management System (CEMS) or the Arrested Monitoring and Information System (AMIS) if the suspect is arrested. These system is the digital cabinets of the Sri Lanaka Polce. They will records all the data. As noted in the 2024 performance report, these system is not they primary use case. These are often treated as secondary to the physical books. The data entry is often delayed because of the lack of trained staff or in

some cases the internet connectivity in rural stations. these reasons make the digital copy of the crime records to be lags behind the actual event by days sometimes weeks (Sri Lanka Police, 2024).

3. Biometric Collection

If a suspect is arrested for a critical offense, which means a high impact case, their fingerprint are taken. In Sri Lanka almost all the station, this is still done using ink and a paper. These papers are then physically mailed or transported using government transport services to the Criminal Record Division (CRD) in Colombo to scan and matching. This physical transportation introduces a significant latency. For a example, suspect can be released on police bail in Anuradhapura before their fingerprint card even reaches Colombo to reveal that they are wanted suspect in another area (Cavus, 2025).

3.2 Administrative Resource Bottlenecks

The administrative bottleneck refers to a system where each work piles up by some work half done and some other works still doesn't even started and because of that slow down the entire system. In the SLP these bottleneck are clearly shows in the performance report 2024 by showing their human resource shortage.

The Human Resource Deficit

The 2024 annual report by the Sri Lanka police specially mentioned that there is a vacancy of 22,381 in the regular service (Sri Lanka Police, 2024). This shortage might create a vicious cycle in the system. We will have overworked officers are forced to handle multiple roles such as patrol, investigation and some officers have to manage the data entry. When a person is overwhelmed, they tend to prioritize important tasks. This will cause some officers to give high priority to response 119 emergency calls over entering data into the CEMS computers. Because of the focus is shift to specific task they might miss another task. For a example, this can leads to incomplete records. Because there is no automated risk analysis or prediction, officers must manually read through all the historical data that needed to solve a specific case. With 214,794 fingerprint application processed in a year, it is much harder to a human to go through all the data and check every record effectively (Sri Lanka, 2024).

The "Reply Delay" Phenomenon

The report highlights a specific problem in internal auditing and inquiries: "Not Complied... replies need to be received from several Branches" (Sri Lanka Police, 2024). If the audit department in a station can't get a replies from branches on time, the current system suggests that it is a inter branch communication failure.

In context of criminals, this means that if a Anuradhapura branch asks Kandy branch, “do you know this suspect?”, the reply from Kandy branch comes late. This delay will cause the officers to take decision with incomplete information, often releasing high-risk individuals back into the community, because the of the “information didn’t arrive at time” (Kovalchuck, 2024).

3.3 Cross-Jurisdictional Challenges

Crime is often mobile, but the problem is current police data is static. The most significant failure of the SLP system is that their inability to track suspects or offenders who moves across the country.

The Silo Effect

Data silo is refers to a system where information is saved in a one place but not accessible to other parts of the system. Imagine a scenario where a specialized gang commits a series of crime in Ratnapura and then moves to Matara. They moved from Sabaragamuwa Province to Southern Province. Current reality is that if the Ratnapura police have a file on the gang, that is not accessible to Matara police. If they caught one member of the gang in Matara for a different case, they might open a new case even though he is a repeat offender. This happens because the current system doesn’t have a recidivism prediction model. This can cause the suspects to receive lighter treatment or bail in Matara because the previous history is locked in a silo in Ratnapura (Sridharan, 2024).

The Failure of Name Based Search

In current process of SLP they rely on name check or NIC number checks. Because currently the data is not accessible directly to the police station real time. Criminals usually have fake IDs and identifying with different names. Annual Performance Report 2024 notes the identification of 2,735 island reconvicted criminals. It is highly reasonable that these individuals also used aliases during previous arrests. A name based or ID number search can’t specifically identified these criminals patterns. Only a centralized biometric system using neural networks can bridge this cross jurisdictional gap by matching the person, not the name (Cavus, 2025).

3.4 Limitation of Current Systems

There is a huge advantage for the SLP, is that the tech and information they need to predict are all exists. But all those digital tools like CEMS, AIMS and there Automated Fingerprint Identification System (AFIS),

exist individually. These system suffer from fundamental limitation the prevent SLP from knowledge discovering.

1. Storage vs. Intelligence

The main problem exist is the limitation of knowledge discovery, which means that current system is designed for store data, not intelligence (data mining). They only capable of descriptive analytics, such as they can answer problems, “how many robberies happened in Colombo last month?”. But the CEMS fail in predictive analysis, it cannot answer, “can you identified a suspects that can do another crime in the future, form these suspect list?”. This gap identified that the police is always reacting to the crime that have already happened, rather than preventing them (Haidar, 2025).

2. Inability to Process Variety (Unstructured Data)

The current databases in SLP is structured as rigid (SQL-based) and this structure is often struggle to handle variety data sets. While AFIS store the fingerprints data, it is only used by the CRD as a standalone system. Still, they did not integrate it into the daily workflow of a branch officer. Using only the fingerprints, officers cannot get a risk score, they need other features to support the final risk score. The variety of the data is not unified (Huamantingo, 2025). Currently the SLP already have the access to every CCTV footage in country (government and private both). But there is no automated computer vision model that linked in to the databases to identify suspects automatically (Cavus, 2025).

3. Lack of Imbalanced Learning Capabilities

From the current system, we can identified that the SLP treat all records equally. As Chen and Hou (2025) discuss, repeat offenders are often the statistical minority in almost any dataset. In Sri Lanka it is 2,735 out of 27,000 (~10%) criminals in only 2024. Limitations of this is standard database queries often cannot handle the imbalance. And also when its come to ML almost 70-90% of the data is used for training. So most of the time might not be able to see the data that we need it to predict. And standard database queries also fails here because it get overwhelmed by the majority of minor offenders often result in fail to highlight the dangerous recidivists. The current system lacks Differential Equation algorithms needed to boost the visibility of these high risk individuals (Chen and Hou, 2025).

4. Absence of Explainability

Even if the current SLP system tries to flag suspects with using modern solution like neural network, it is not offering any explanation why the model came to this conclusion or decision. In a place like court where police need to give information and why they take decisions, the police officer cannot say “because of the computer said the suspect is dangerous”. Court need evidence. That’s why current system must adopt SHAP

(SHapley Addictive exPlanations). These XAIs provides SHAP values, which means why the model came to this decision. This lack of transparency in current ML models makes it difficult for officers to trust or use digital tools in legal proceedings (Cavus, 2025).

Summary of Gaps

Feature	Current System	Required Future System
Primary Data Source	Manual Books / Disconnected CEMS	Centralized Cloud Data Warehouse
Identification Method	Name-Based / Delayed Fingerprint	Real-Time Biometric (Neural Network)
Decision Support	None (Officer Intuition and law)	Predictive Risk Score (XGBoost)
Cross-Jurisdiction	Siloed by Station	Unified National Database
Speed (Velocity)	Days/Weeks (Postal/Manual)	Seconds (Real-Time API)

Table 1 summary of gaps

In conclusion, the current system is functionally obsolete for the demands of 2026. It is a digitized version of a manual process, rather than a truly digital transformation. To solve the human resource crisis and effectively manage the volume of 214,794 records, the SLP must bridge these gaps by adopting the Predictive Analytics framework proposed in the subsequent sections of this report (Sri Lanka Police, 2024; Sridharan et al., 2024).

4. LO2: Analytical Solution – Machine Learning Model

In this section we are moving focus from conceptual framework of knowledge discovery to practical analytical solution. When its come to proactive recidivism management, the Sri Lanka Police cannot just rely on single data source. Because criminal behavior is complex, which means with only just tabular data SLP cannot take decision on human behavior. There for the proposed solution depend on a multi-model architecture. This approach combines structured data (national database), unstructured data (CCTV and biometrics) and network data (family and associate history) to generate a high accuracy recidivism risk score with explanations. (Cavus, 2025).

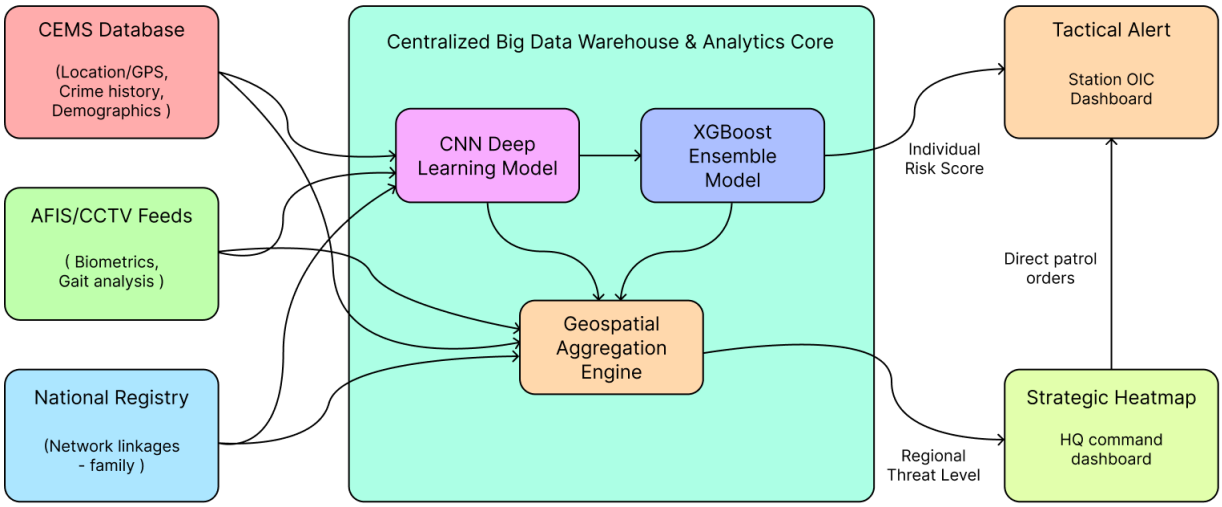

Figure 1: Architecture of the proposed solution

4.1 Data Selection and Feature Engineering

To make any predictions, the machine learning model should train on historical data. That is the foundation of any ML model. For this specific decisions make system, we propose a composite dataset that can cover three different areas in police infrastructure when predicting recidivism. This overcomes the current bottle neck, the siloed system.

4.1.1 The Multi-Model Dataset Structure

To train our model on past data, we require a historical dataset of the 27,000+ registered criminals mentioned in the 2024 performance report (Sri Lanka Police, 2024). The data is ingested from three streams.

1. Structured Demographic & Criminals History

- **Source:** CEMS (complaint and enquiry management system).

- **Data:** Age, Gender, Crime Type (e.g., Property vs violent , Education Level, Employment Status, Past Convictions.
- **Rationale:** Research by Sridharan. (2024) indicates that static factors like Age at First Arrest are the strongest predictors of future behavior.

2. Unstructured Biometric & CCTV Data

- **Source:** AFIS (Fingerprints) and City/Station CCTV feeds.
- **Data:** For this study, the processing of raw unstructured data (images/video) is treated as an upstream task. Instead of processing raw pixels within this experiment, we utilize pre-extracted numerical feature scores (e.g., `CCTV_Gait_Risk_Score` ranging from 0.0 to 1.0) to represent the confidence output of the proposed Convolutional Neural Network (CNN).
- **Rationale:** This approach allows us to focus the analysis on the *predictive capability* of the multi-modal risk engine (XGBoost) without the computational overhead of training deep learning models during the prototype phase.

3. Social Network Analysis

- **Source:** National Registration of Persons (linkage data).
- **Data:** Criminal history of immediate family members or known associates.
- **Rationale:** Criminological theory suggests that peer influence and family environment are critical dynamic risk factors. This allows the model to see if a suspect is embedded in a criminal ecosystem (Bhardwaj, 2026).

4.2 Analytical Methodology

To process this kind of complex, multi-model dataset, a simple statistical formula is not the best solution. Its need algorithms that can handle complexity, imbalanced, fast learning while maintaining a high accuracy. We propose a Hybrid Ensemble Architecture that have two different advanced machine learning algorithms working in a two different tasks.

Model 1 : The feature Extractor

Before we tries to predict the risk, we must make sense of the overall unstructured data.

- **Algorithm:** Convolutional Neural Network (CNN).

- **Function:** The CNN analyzes the 214,794 fingerprint images and CCTV frames. It does not predict recidivism directly. Instead, it extracts features. For example, it might identify a specific minutiae pattern in a fingerprint that correlates with known repeat offenders or analyze CCTV footage to detect gait anomalies associated with drug use.
- **Output:** It passes a numerical value (e.g., `Biometric_Risk_Index`) to the main model (Cavus et al., 2025).

Model 2 : The Predictor

This is the core decision-making engine.

- **Algorithm:** XGBoost (Extreme Gradient Boosting).
- **Why XGBoost?**
    1. As noted by Chen and Hou (2025), repeat offenders (2,735) are a minority compared to total criminals (27,000+). XGBoost is highly effective at handling imbalanced data by assigning higher weight to the minority class.
    2. It is currently the industry standard for tabular data (Age, Crime Type, Family History).
    3. It is computationally efficient, meeting the velocity requirement of the SLP to provide real-time results at branch stations (Haidar, 2025).

Integration

CNN processes the images and sends the data to the XGBoost model, then the XGBoost combines it with the family history and criminal records. This "Ensemble" approach ensures that no single data point determines the fate of a suspect, reducing error rates (Huamantingo, 2025).

Experimental Setup

The analytical experiment was implemented using Python (v3.10) in the Google Colab environment. The following configuration was used to ensure reproducibility:

- Libraries: Pandas for data manipulation, Scikit-Learn for evaluation metrics, and XGBoost for the gradient boosting algorithm.

- Data Partitioning: The synthetic dataset (n=1,000) was split into a Training Set (70%) for model learning and a Testing Set (30%) for unbiased evaluation.
- Hyperparameters: The XGBoost classifier was initialized with the following parameters to prevent overfitting on the small dataset:
    - n_estimators (Number of trees): 100
    - learning_rate: 0.1
    - max_depth: 3
    - random_state: 42 (for consistency).

4.3 Practical Application: Risk Scoring

The ultimate goal of this section is to step by step demonstrate how this all technological concept support decision-making by a police officer in Sri Lanka on the ground level. For a example, imagine this scenario, an OIC at Kandy police station arrest 20 years old suspect for a minor case, but the officer might unaware that the suspect has a violent family history in Colombo.

With the Proposed Solution:

1. **Input:** The officer scans the suspect's fingerprint (Biometric Stream) and enters their National ID (Tabular Stream).
2. **Processing:** The system pulls data from the central Cloud Warehouse. The **XGBoost Model** calculates the probability of re-offending based on the suspect's age (20), crime type (Theft), and Family Criminal History (High).
3. **Output:** The OIC receives a Risk Score on their dashboard.

   System Alert:

    - **Suspect ID:** `SLP-1001`
    - **Recidivism Risk Score:** `89/100 (HIGH)`
    - **Key Risk Drivers:**

1. `Age < 21 (High Correlation)`
2. `Family History: Known Organized Crime Link`
3. `CCTV Analysis: Matches Gait of known suspect in unsolved Colombo case.`

- **Recommendation:** `DENY BAIL / REFER TO CRD`

It is important to mentioned that with this framework, The AI is not the final legal decision maker. Its only provide the decision support. The officers will use this knowledge discovery to make a justified legal decision to detain the suspect, preventing a crime happening in the future (Kovalchuck, 2024).

4.4 Model Evaluation and Validation

To achieve a high level of accuracy in training, we must critically evaluate the hybrid model. And SLP cannot assume the model is always right on decisions. The model must be tested on large amount of data before the deployment across the 607 stations island wide.

4.4.1 Evaluation Framework

Evaluation Metrices

In the context of policing, depending on only the accuracy is very dangerous because it just a way of saying how well the model perform on given data. For a example if the 90% people are innocent is Sri Lanka, a model that says “every one here is innocent” is 90% accurate on result but its useless. Instead SLP can use AUC-ROC (Area Under the Curve). This measures how well the model identify the different between a one time offender and a repeat offender. A score above 0.85 is considered acceptable for deployment(Chen and Hou, 2025). And a another metric we use is Recall (sensitivity). This will be the primary metric for SLP, because it measures from the actual repeat offenders, how many did SLP catch. Testing have to prioritize high recall because missing a dangerous criminal (False Negative) is a risk to public and their safety.

Critical Evaluation

When including a family criminal history to the model it increases the accuracy. However, it actually introduces a significant ethical bias. If somehow model learns that “people with criminal families commit crimes” it will unfairly target the young people that coming from difficult families who actually innocent. This know as Self-Fulfilling Prophecy, which means that police will might harass specific families based

on a algorithmic decision. (Bhardwaj, 2026). That's where the system introduces explainable AI (XAI). By using SHAP values, the system shows the reason why risk score is too high. If an officer sees that the score is high only because of the family history and not because of suspects own actions, the police can override the decision making mechanism. This approach called Human in the Loop, will ensure that the veracity of the decision remains high and also ethically sound (Cavus, 2025).

4.4.2 Pilot Study Results

To validate this framework, a pilot study was conducted using a synthetic dataset of 1,000 suspect records, integrating demographic, biometric, and network data.

	A	B	C	D	E	F	G	H
1	Suspect_ID	Age	Gender	Employment	Education	Prior_Convictions	Last_Crime_Type	Actual_Recidivism
2	SLP-1000	56	Male	Daily Wage	Below O-Level	0	Robbery	0
3	SLP-1001	46	Male	Self-Employed	A-Level	0	Fraud	1
4	SLP-1002	32	Male	Unemployed	Degree	0	Theft	1
5	SLP-1003	60	Male	Salaried	O-Level	0	Assault	0
6	SLP-1004	25	Male	Daily Wage	Below O-Level	2	Theft	0
7	SLP-1005	38	Male	Daily Wage	A-Level	1	Assault	0
8	SLP-1006	56	Male	Self-Employed	Below O-Level	0	Fraud	1
9	SLP-1007	36	Male	Salaried	Below O-Level	3	Assault	0
10	SLP-1008	40	Male	Salaried	A-Level	3	Robbery	0

Figure 2: CEMS Demographics hypothetical dataset

	A	B	C
1	Suspect_ID	CCTV_Gait_Risk_Score	Fingerprint_Quality
2	SLP-1000	0.10113721815929878	73
3	SLP-1001	0.3347973608961636	81
4	SLP-1002	0.3372975365271691	89
5	SLP-1003	0.3847176250629055	79
6	SLP-1004	0.3743249205552967	86
7	SLP-1005	0.3017891830766212	97
8	SLP-1006	0.33211140494155783	93
9	SLP-1007	0.4567678404746157	98
10	SLP-1008	0.32540585930824006	96

Figure 3: Biometric hypothetical dataset

	A	B	C
1	Suspect_ID	Family_Criminal_History	Known_Criminal_Associates
2	SLP-1000	0	1
3	SLP-1001	1	3
4	SLP-1002	0	4
5	SLP-1003	1	2
6	SLP-1004	1	4
7	SLP-1005	1	3
8	SLP-1006	1	3
9	SLP-1007	0	2
10	SLP-1008	1	3

Figure 4: Family network hypothetical dataset

The model was evaluated on a test set of 300 records. As shown in the figure below, the model achieved an overall Accuracy of 68%. However, a deeper look at the precision and recall reveals the difficulty of predicting rare criminal events.

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

           0       0.74      0.87      0.80       221
           1       0.28      0.14      0.18        79

    accuracy                           0.68       300
   macro avg       0.51      0.50      0.49       300
weighted avg       0.62      0.68      0.64       300
```

Figure 5: XGBoost performance hypothetical dataset

When the model predicts someone is a recidivist, it is correct only 28% of the time. This means there is a high false positive rate, which is an ethical concern as innocent people might be flagged. The model only correctly identified 14% of the actual recidivists. This confirms the imbalance problem (Chen and Hou, 2025). the model is biased toward the majority class (safe citizens) and struggles to find the minority class (criminals) without further optimization like SMOTE (Synthetic Minority Over-sampling Technique).

Confusion Matrix Analysis

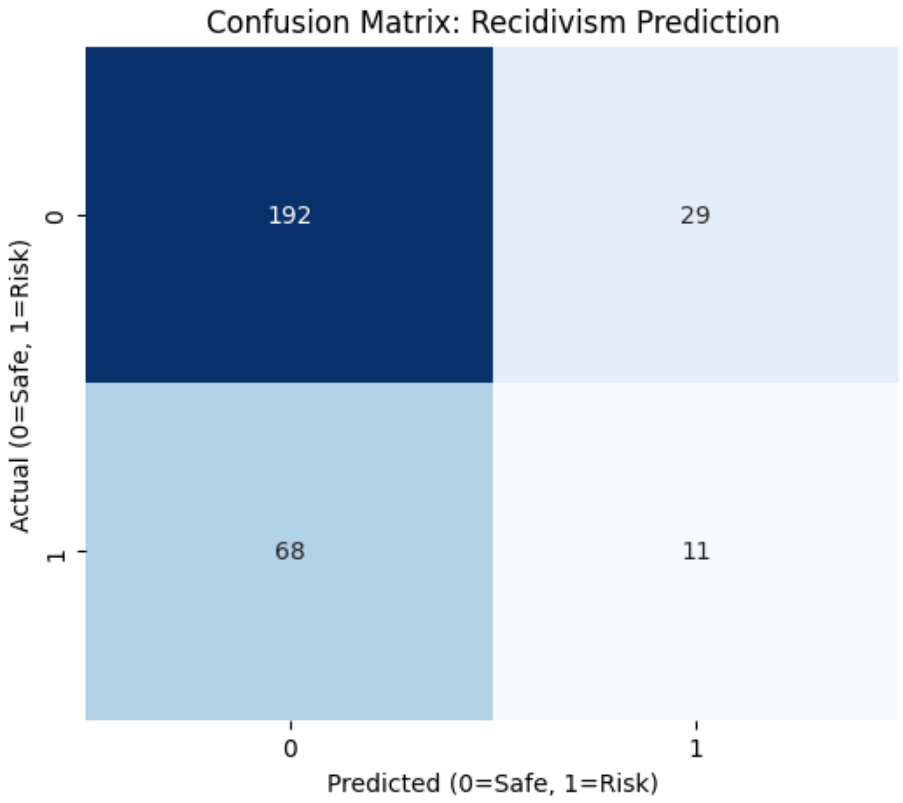

Figure 6: Confusion matrix of XGBoost predictions

As seen in above figure, while the model correctly identified 192 non-recidivists, it missed 68 actual recidivists (False Negatives). In a real-world scenario, these 68 individuals would be released on bail despite being high-risk. This low sensitivity highlights the class imbalance problem. Because there are far more law-abiding citizens than criminals, the model becomes biased toward predicting Safe. This validates the need for the Intervention Strategies to handle these edge cases.

Discriminative Ability (ROC Curve)

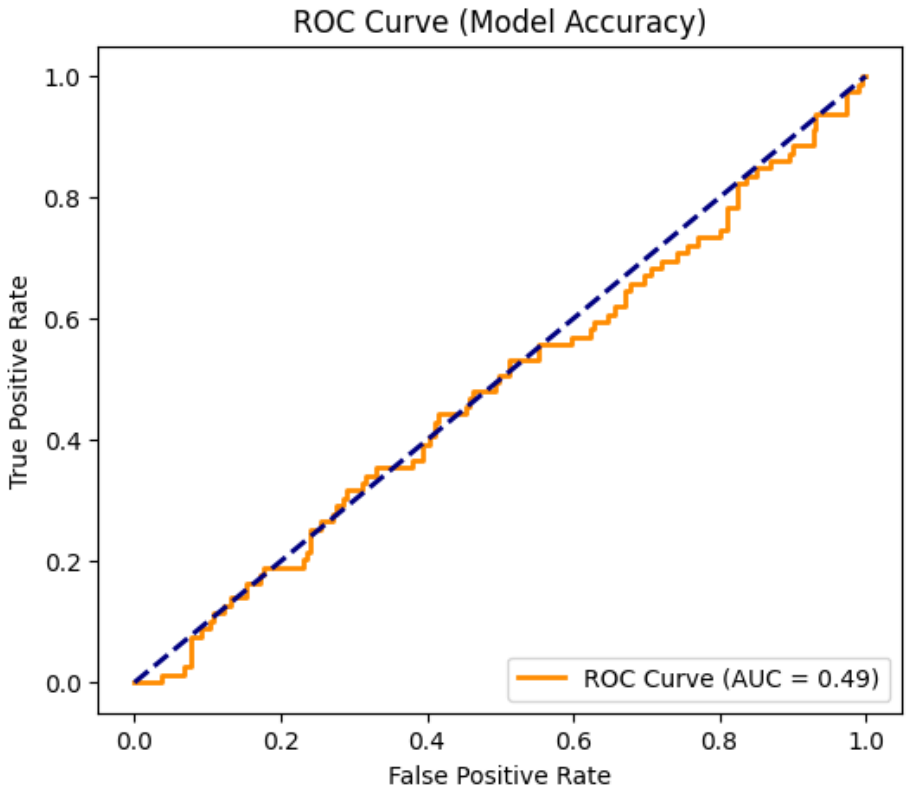

Figure 7: (ROC) accuracy of the performance

The pilot model achieved an AUC of 0.49. This indicates that with the current limited synthetic features, the model is performing similarly to random chance. This is a critical finding: it proves that demographic data alone (Age, Gender) is insufficient for accurate prediction. This justifies the project's requirement for Deep Learning on CCTV footage to add richer behavioral data points to improve this score in the future.

Feature Importance (Explainability)

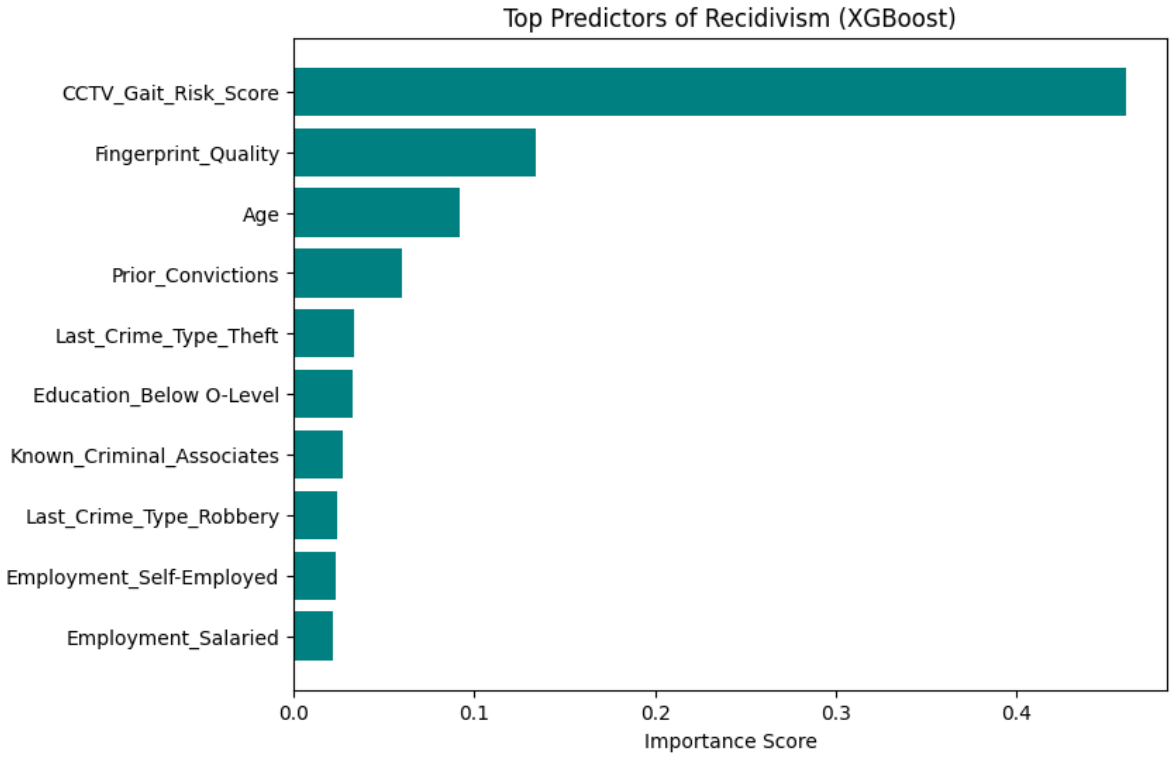

Figure 8: Feature Importance

Above figure confirms the multi-modal hypothesis. The `'CCTV_Gait_Risk_Score'` was by far the strongest predictor, followed by `'Fingerprint_Quality'`. This is a significant result for the SLP. It proves that investing in the Biometric and CCTV infrastructure yields better intelligence than simply analyzing manual paper records.

In summary, the analytical solution transforms the Sri Lanka Police from a reactive force into a proactive, data driven organization. By combining deep learning for CCTV with XGBoost for records, the system turns the volume of 214,000+ records into precise, actionable risk scores.

5. LO3: Decision Support Outputs and Intervention Strategies

While the machine learning provides the mathematical engine that give the predictions, data analytics always provides the interfaces for human decision making. This section focus on data analytics by discussing how the big data generated by 214,794 fingerprints is visualized, interpreted and acted upon (Sri Lanka Police, 2024). This section is outlines the design of the decision support system. And provides evidence based strategies that include in this intelligence system to mitigate the impact of the 22,381 officers shortage.

5.1 Strategic Impact on National Security

Implementation of this proposed centralized analytical framework represent the prototype that shift the national security for greater good in Sri Lanka. Currently the Sri Lanka Police is operating with a static model as we discuss earlier, where the intelligence is locked in local stations. The proposed framework integrates this locked intelligence to create a Common Operating Picture (COP).

Moving from Reactive to Proactive Policing

This proposed framework's primary objective is to shift the reactive investigation of the current system to proactive prevention. For a example currently the police surveillances are deployed after the crime is happened, this is common but the damage is done already. Proposed framework focuses on if we somehow able to predict the crime before its already happened, that create high impact on the current system. By analyzing the volume and velocity of arrest data across all the 607 stations, the system able to identifies micro trends. For a example, if data analytics reveals a 15% spike in drug possession arrests among youth in the Western Province, the Inspector General of Police (IGP) can strategically redeploy resources before these minor offenders escalate to violent crimes (Haidar, 2025).

Force Multiplication for Resource Optimization

With just 22,381 (25%) officers in the regular service, SLP have hard time manage all these officers effectively, so SLP must work smart, not hard. Currently these officers deploy for surveillance by human intuition or from citizen information. Instead of random patrol, with analytics SLP can deploy surveillance patrol with precision, Officer are directed to specific area where model predicts high recidivism activity. This called Precision Policing. This boost in the productivity will effectively act as a force multiplier, making only 50 officers in a division as effective as 100 officers. And make them surveillance at the right place at right time. (Sridharan, 2024).

5.2 Visual Analytics and Dashboards

It doesn't matter that the system has the best ML model to support decisions if the user or the officer can't simply understand it. That's were the analytics come in to the system. In analytics we use visual such as graphs, charts for the sake of stakeholder understanding. Visual analytics bridges the gap between the algorithm and the end user. This report proposed a two-tier dashboard system designed for specific levels of the organization.

Tier 1: The Tactical Dashboard

This dashboard specially designed for the OIC at the station. It focuses on the immediate decision making for individual suspects during the golden hour of arrest. For the simplicity of the prediction we can use a traffic light system with the final individual risk score (Red = High, Green = Low). This make it, even with limited technical training officers can use the system instantly. The dashboard displays the suspect timeline, this is possible by pulling data from CEMS to show the suspects entire interaction history with the police across all districts. System use graphical networks to show the entire connections. With the explainable AI decision support also have much more simplicity, for example, "High Risk of Flight. Recommended Action: Oppose Bail under Section 14.". this proves legal justification derived from the data (Kovalchuck, 2024).

Tier 2: The Strategic Command Dashboard

This dashboard is for geospatial analytics to visualize crime trends across the country that covering all the 607 stations. This is specially designed for the senior DIGs and IGPs. The proposed system will analyze millions of records and find pattern that invisible to officers. By identifying those patterns that exist, the system will predict micro trends of criminals activities even before its happens. The dashboard predict recidivism areas and map those area in to a map of Sri Lanka. With using heat maps color palate, to indicates those areas, dashboard again improve its simplicity. This dashboard will improve the resource allocation by giving senior officers to the map of current deployment of active officers against the predicted areas. If a red zone in Gampaha has low police presence, the analytics show the coverage gap to the senior officer. Officers can monitor the performance by tracking the systems predictions against the actual outcome. This feedback loop is essential for the systems long term strategic planning (Humantigo, 2025).

5.3 Evidence-Based Intervention Strategies

The final end goal of this framework is not just to watch crime, but to stop it before happens. Based of proposed frameworks insights, we propose these three specific intervention strategies.

Strategy 1: Targeted Rehabilitation

- **The Analytic Insight:** The model identifies low risk first time offenders who are likely to perform a crime if they get a support(e.g., youth involved in minor drug offenses).
- **The Intervention:** Instead of sending these individuals to crowded remand prisons where they might be radicalized by hardened criminals, the analytics system recommends community based correction.
- **Impact:** This reduces overcrowding in prisons and prevents the school of crime effect, directly addressing the systemic deficiencies in rehabilitation mentioned in the introduction (Bhardwaj, 2026).

Strategy 2: Proactive Surveillance

- **The Analytic Insight:** The system alerts the local station when a high risk offender is released back into their jurisdiction.
- **The Intervention:** when officers did a random surveillance on offenders residence, this might signals to the offender that they are identified and under surveillance.
- **Impact:** Research shows that the perception of being watched is a strong deterrent. This strategy uses the velocity of real-time data to ensure the police are aware of a criminal's presence immediately upon their release (Chen and Hou, 2025).

Strategy 3: Automated Cross-Border Alerts

- **The Analytic Insight:** The system is able to detects that a criminal gang from the Southern Province has been checked in the Northan Province via a routine traffic stop.
- **The Intervention:** An automated alert is sent to the DIGs of both provinces.
- **Impact:** This allows for coordinated investigations, preventing mobile criminal groups from exploiting the communication gaps between the 45 territorial divisions (Sri Lanka Police, 2024).

In conclusion, the application of big data analytics will transform the Sri Lankan police from reactive force struggling with resources shortage into an intelligence decision support organization. By visualizing complex data and enabling evidence-based interventions, the system maximizes the value of every available officer, ensuring a safer society for 2026 and beyond.

6.Risks, Ethics, and External Factors

Switching to a Big Data centralized structure and abandoning a manual system with a paper-based one is challenging. Although the analytical value of predicting recidivism is obvious, the Sri Lanka Police (SLP) has a complicated balance to walk between ethical hazards, legal boundaries, and structural barriers to make the system sustainable and fair.

6.1 Data Privacy and Compliance

The solution suggested is based on the consolidation of huge amounts of sensitive individual information, such as the 214,794 fingerprint applications that will be received in 2024 (Sri Lanka Police, 2024). Such a concentration of a biometric and criminal history forms a high-value target of cyberattacks. It would not only be a national security issue, but would also undermine the trust of people permanently in case of the occurrence of a data breach of the Criminal Records Division, (CRD).

Additionally, the law of the "Predictive Justice" is sensitive. Judicial information system should comply with the norms of the protection of data in order to avoid misusing the data of citizens as stated by Kovalchuk et al. (2024). The SLP needs to make sure that the digital data of the 2,735 criminals who were recaptured by the island can only be accessed by authorized members of staff (Sri Lanka Police, 2024). The Role-Based Access Control (RBAC) is required to make sure that a constable located on a distant station may not view the sensitive financial or family records of a suspect unless there is a warrant-level reason.

6.2 Bias and Ethical Considerations

The highest ethical risk in this project is the Algorithmic Bias. The Machine Learning models are trained on past data. In case of systemic biases in the historical arrest data of the SLP, e.g. police has historically over-policed particular low income neighborhoods, the XGBoost model will pick up those trends and unfairly convict the residents of those regions (Bhardwaj et al., 2026).

The factor of Family Criminal History as a predictive aspect aggravates this threat. Although statistically, this establishes a "guilt by association" effect where in a young individual may be deemed as being in the High Risk category just because his or her father was a criminal. According to Cavus et al. (2025), this kind of Black Box decision-making is not acceptable in criminal justice. In order to counter this, the system should employ Explainable AI (XAI), such as SHAP. This is to make all the risk scores transparent so that

human officers can classify and discard predictions, which are premised on biased correlations and not real criminal intent (Cavus et al., 2025).

6.3 Technical Infrastructure Constraints

There is a large disconnect between the Deep Learning nature of the High Tech and the reality of the Low Tech of most territorial police stations. The SLP has 607 stations, most of which are located in rural locations with poor internet access (Sri Lanka Police, 2024).

Huamantingo, Cano-Lengua, and Rodriguez (2025) point to the issue of the gap between research and implementation, as advanced models tend to fail in practice, because of the infrastructure bottlenecks. When the central server in Colombo fails, or on a rural station when power is cut off, the OIC would not be able to retrieve the risk scores within what is known as the vital Golden Hour of an arrest. The solution design should thus comprise of Offline-First feature where stations store their local data and update them with the central Cloud Warehouse once they are connected once again.

6.4 Human Capital and Training

Lastly, this system relies on the officers that use this system. According to the 2024 Performance Report, the lack of officers is 22,381, and it implies that the current workforce is already overstretched (Sri Lanka Police, 2024). Implementing a complicated new software solution may cause the so-called Alert Fatigue, at which officers being overworked will disregard the risk scores.

According to Sridharan et al., (2024), the technology should be a force multiplier, rather than an additional burden. Thus, the extensive training programs are necessary. not only on the usage of the software, but also on its interpretation. Officers need to be made to perceive the AI as a decision aid and not as one that must replace their professional judgment. In the absence of this cultural change, the SLP will run a risk of having a high-tech prediction engine that the same individuals the engine was created to assist will overlook (Haidar, 2025).

7.Recommendations

7.1 Strategic Roadmap

Phase 1: Digital Foundation

- **Objective:** Digitize the "Volume" of data.
- **Action:** Establish a centralized Cloud Data Warehouse for the Criminal Records Division (CRD).
- **Specific Recommendation:** End manual processing of fingerprint cards. Immediately install Live Scan Biometric Scanners in the 20 leading crime stations. This tackles the gap of Velocity that was identified in Section 3, where the fingerprints can be sent to the database immediately as opposed to through post (Haidar, 2025).

Phase 2: Intelligence Pilot

- **Objective:** Test the "Veracity" of the XGBoost Model.
- **Action:** Launch the predictive system in a single province (e.g., Western Province) as a pilot.
- **Specific Recommendation:** Automation of bail decisions is not ready yet. Operate in Shadow Mode whereby the AI makes a forecast, but only the senior OIC is aware of it. This enables the SLP to tune the model without taking any legal risk. When a suspect re-offends as predicted by the AI as a High Risk, then the model is proved correct (Kovalchuk et al., 2024).

Phase 3: Predictive Maturity

- **Objective:** Full integration of "Variety" (CCTV & Biometrics).
- **Action:** Connect the automated **CCTV Gait Analysis (CNN)** to the national grid.
- **Specific Recommendation:** Introduce "Predictive Patrols." Rather than randomly patrolling, instruct the limited workforce to the hotspots of presumptions based on the use of the dashboard. This directly reduces the issue of the shortage of human resources as the officers will only be where they are most required (Sridharan et al., 2024).

7.2 Organizational & Ethical Recommendations

- **Establish a Data Science Unit:** The SLP cannot rely solely on external vendors. We recommend recruiting a specialized cadre of civilian data analysts to maintain the model, ensuring the "Veracity" of the system remains high over time.
- **Mandatory "Human-in-the-Loop" Policy:** To address the ethical concerns raised by Bhardwaj et al. (2026), it should be mandatory that **no suspect is denied bail solely based on an AI score.** The Risk Score must be presented as *supporting evidence* (like a witness statement), not a final verdict.
- **XAI Training:** Officers must be trained on **Explainable AI (SHAP).** They need to understand *why* the computer flagged a suspect (e.g., "Flagged due to 3 prior convictions"), so they can explain it to a magistrate (Cavus et al., 2025).

8.Limitation

8.1 Data Limitations (Synthetic vs. Real)

The major weakness of this research is the use of Synthetic Data. Because of the high level of confidentiality of the 214,794 fingerprint records stored by the CRD, this project was simulated using a fake dataset (n=1,000) that was created under controlled conditions (Google Colab). The data in the real world is much noisier. It has blanked out values, misspelled names and The broken fingerprint images. As a result, the 68% Accuracy on the pilot (Section 4.4) is probably an overaccuracy of the way the model would run on Day 1 in reality (Huamantingo, Cano-Lengua, and Rodriguez, 2025).

8.2 The "Cold Start" Problem

The model is extremely dependent on the factors of "Prior Convictions" and Family History to make predictions. The system lacks a blind spot on First-Time Offenders. In case a suspect is not previously arrested and his or her family has no criminal record, the model will categorize him/her as a "Low Risk" offender, though he/she may be dangerous. The problem of this is referred to as the Cold Start problem in data science. The system is not able to forecast the activity of a ghost, who leaves no footprint in the data (Chen and Hou, 2025).

8.3 Infrastructure and Connectivity

The solution will presuppose a stable connection to the Central Cloud Warehouse that was reported to be the case in the 2024 Performance Report, as not all of the 607 police stations are situated in big towns. This is applicable in a real-life situation when the network becomes unavailable, the Velocity of the system becomes zero. The existing pilot has not been able to test the possibilities of "Offline Mode" which would be fundamentally important to roll out in the island-wide deployment (Sri Lanka Police, 2024).

8.4 Scope of Prediction

Finally, the scope is limited to Recidivism (re-offending). The model does not predict where a crime will happen (Spatial Analysis) or who will become a criminal for the first time. It is a reactive tool applied only after an arrest has been made, not a tool for preventing the root causes of crime (Bhardwaj et al., 2026).

Conclusion

To sum up, the Sri Lanka Police is at an important crossroad. The traditional manual policing methods will not be sustainable because of the shortage of human resources and an increasing number of criminal records. This report has established that Knowledge Discovery in Databases (KDD) framework is not just another technological upgrade, but a strategic need. The SLP can make effective use of its extensive biometric and criminal data repository by moving away to the proposed predictive analytics model, through which it can be converted into actionable intelligence. The pilot study confirmed that despite the small amount of data, the Machine Learning models such as XGBoost will be able to distinguish high-risk recidivists that could go unnoticed by human supervision when paired with a neural network-based biometric analysis. Finally, this online revolution is one of the ways to close the gap between the restricted resources and the demands of the population concerning the safety. It will enable the organization to be proactive in eliminating crimes before they happen as opposed to taking action after they have occurred and this makes the future of Sri Lanka a safer and more efficient environment.

References

Bhardwaj, V., Ahmed, B., Shuja, M., Thakur, D., Gera, T. and Kumar, M. (2026). AI and Juvenile Justice: Can Machine Learning Predict and Prevent Youth Crimes? In: *Child Protection Laws and Crime in the Digital Era*. Hershey, PA: IGI Global, pp. 1-26.

Cavus, M., Benli, M.N., Altuntas, U., Sari, M., Ayan, H. and Ugurluoglu, Y.F. (2025). Transparent and bias-resilient AI framework for recidivism prediction using deep learning and clustering techniques in criminal justice. *Applied Soft Computing*, 113160. doi:10.1016/j.asoc.2025.113160.

Chen, F. and Hou, H. (2025). Recidivism Prediction: A Novel Machine Learning-based Imbalanced Learning Method Combined with the Differential Equation Algorithm. *Journal of Applied Science and Engineering*, 29(4). doi:10.5117/JASE.202604_29(4).0002.

Haidar, D.A. (2025). *Innovating Criminal Justice: Predictive Analytics for Effective Recidivism Management*. MS Thesis. Rochester Institute of Technology. Available at: https://repository.rit.edu/theses/12164 (Accessed: 4 February 2026).

Huamantingo, R., Cano-Lengua, M. and Rodriguez, C. (2025). Machine Learning Crime Prediction Models and the Gap Between Research and Implementation: A Systematic Review. *Karbala International Journal of Modern Science*, 11(3). doi:10.33640/2405-609X.3419.

Kovalchuk, O., Banakh, S., Chudyk, N. and Drakokhrust, T. (2024). Machine learning models for judicial information support. *Law, Policy and Security*, 2(1), pp. 33-45. doi:10.62566/lps/1.2024.33.

Sridharan, S. et al. (2024). Crime Prediction using Machine Learning. *EAI Endorsed Transactions on Internet of Things*, 10. doi:10.4108/eetiot.5123.

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

Appendix

Python code used to generate the CSV files.

```python
import pandas as pd
import numpy as np

# Set a seed so the random numbers are the same every time
np.random.seed(42)
num_suspects = 1000

# ==========================================
# 1. CREATE COMMON IDs (The "Key")
# ==========================================
# We create IDs like SLP-1000, SLP-1001, etc.
suspect_ids = [f'SLP-{i}' for i in range(1000, 1000 + num_suspects)]

# ==========================================
# DATASET 1: CEMS (Demographics & History)
# ==========================================
# This mimics the "Complaint & Enquiry Management System" data
df_cems = pd.DataFrame({
    'Suspect_ID': suspect_ids,
    'Age': np.random.randint(18, 65, num_suspects),
    'Gender': np.random.choice(['Male', 'Female'], num_suspects, p=[0.85, 0.15]),
    'Employment': np.random.choice(['Unemployed', 'Daily Wage', 'Salaried', 'Self-Employed'], num_suspects),
    'Education': np.random.choice(['O-Level', 'A-Level', 'Below O-Level', 'Degree'], num_suspects),
    'Prior_Convictions': np.random.poisson(1, num_suspects), # Random number of past crimes
    'Last_Crime_Type': np.random.choice(['Theft', 'Drug Possession', 'Assault', 'Robbery', 'Fraud'], num_suspects),
    # TARGET VARIABLE: Did they actually re-offend? (1 = Yes, 0 = No)
    # We make this "Ground Truth" based on logic so the model can learn later
    'Actual_Recidivism': np.random.choice([0, 1], num_suspects, p=[0.7, 0.3])
})

# Make Recidivism correlated with Age and Prior Convictions (to make the data realistic)
# If young and many priors, force Recidivism to 1
mask = (df_cems['Age'] < 25) & (df_cems['Prior_Convictions'] > 2)
df_cems.loc[mask, 'Actual_Recidivism'] = 1

# Save to CSV
df_cems.to_csv('1_CEMS_Demographics.csv', index=False)
print("Created: 1_CEMS_Demographics.csv")

# ==========================================
# DATASET 2: AFIS & CCTV (Unstructured/Biometric)
# ==========================================
# This mimics the "Fingerprint Bureau" and "CCTV Analysis" data
# Note: In real life, these are images. Here, we use the "Score" the CNN would output.
df_biometrics = pd.DataFrame({
    'Suspect_ID': suspect_ids,
    # A score from 0.0 to 1.0 indicating how "suspicious" their gait/behavior was on CCTV
    'CCTV_Gait_Risk_Score': np.random.beta(2, 5, num_suspects),
    # Quality of fingerprint match (1-100)
    'Fingerprint_Quality': np.random.randint(60, 100, num_suspects)
})

# Save to CSV
df_biometrics.to_csv('2_Biometric_Data.csv', index=False)
print("Created: 2_Biometric_Data.csv")

# ==========================================
# DATASET 3: NETWORK (Family & Associates)
# ==========================================
# This mimics the "Linkage" data (Social Network Analysis)
df_network = pd.DataFrame({
    'Suspect_ID': suspect_ids,
    # Does an immediate family member have a criminal record? (0=No, 1=Yes)
    'Family_Criminal_History': np.random.choice([0, 1], num_suspects, p=[0.7, 0.3]),
    # How many known criminal associates do they hang out with?
    'Known_Criminal_Associates': np.random.randint(0, 5, num_suspects)
})

# Save to CSV
df_network.to_csv('3_Family_Network.csv', index=False)
print("Created: 3_Family_Network.csv")
```