Section B – Information Technology

Invited Paper

Towards Intelligent Crime Investigation

M.A.P. Chamikara[1], Y.P.R. D. Yapa[1], S.R. Kodituwakku[1], R.D.Nawarathna[1]
[1]*University of Peradeniya*
[*]*Corresponding Author: salukak@pdn.ac.lk*

Abstract—The tendency of the grave crimes shows that the security agencies have to shoulder the burden of the criminals in larger numbers than the past because considerable number of criminals already have been confined in many prisons and some were released after the punishment while many criminals have not even been caught. This perpetual trend has made the process of maintaining, investigating crimes very tedious and manual crime recording and investigation system has resulted inefficient crime analysis. Therefore, the retrieval an analysis of crime data has become a huge burden for the crime investigators. This paper presents an efficient method of utilizing data mining techniques along with a robust data mining framework (SL-CIDSS: Sri Lanka Crime Investigation Decision Support System) for intelligent crime investigation. The solution consists of an affluent set of data mining tools such as social network analysis, offender profiling, entity extraction, association rule mining, modus operandi analysis, etc. which provides a systematic way of crime investigation utilizing a proper controlling mechanism of the grave crimes. Security of data is ensured with Apache Shiro security framework for sensitive information and resources being accessed by a VPN (Virtually Private Network) implemented by Sri Lanka Telecom (SLT). The data mining tools have been tested for validity and accuracy by testing their results against the knowledge provided by a domain expert from Sri Lanka police department.

Keywords— Crime Analysis, Data Mining, Data Integration, Decision Support Systems, Intelligent Led Policing, Law Enforcement

I. INTRODUCTION

In Sri Lanka, it is observed that there is a growing discourse about crime wave in the society. This is evident by the large number of crimes recorded each year by the Sri Lanka police department [1].

The Department of Police divides crimes took place in Sri Lanka in to two categories: Grave crimes and Minor offences. Crime complaints are complained either by police officer or a person anonymously or with his/her identity. If the complaint is about a grave crime the complaint will be recorded in the grave crime record book. Currently Sri Lanka police defines 26 grave crime types. Namely, Abduction, Kidnapping, Arson & Mischief, Fraud of Mischief causing a damage greater than 25000 rupees, Burglary, Grievous Hurt, Hurt by Sharp Weapon, Homicide, attempted or committed homicide, Rape, statutory Rape, Unlawful assembly and riots, Robbery ,Unnatural Offence, Extortion, cheating by Trust, Theft, Counterfeiting and Forging Currency, Offense against State, Child Cruelty, Child sexual abuse, Human trafficking, Offensive Weapon act, Use of Automatic or Repeater Guns, Using Dangerous drugs, Obstruction to duty. The complaints which are not related to these grave crimes and which seem to be minor are categorized under minor offences. After a particular grave crime is recorded, as shown in Fig.1, a new investigation is initiated. The crime place will then be protected and it will be visited by the corresponding police personal for further investigations.

This manual process is conducted for all the crimes recorded each year. As Fig.1 depicts, after the crime place is protected, the evidences will start to be collected. Crime case properties will be taken into custody and the police will publish a gazette called PGIII (Police Gazette III) mentioning about the missing properties and other information about criminals and crimes so that all the police stations in Sri Lanka can be aware of them. This will help them to recover the properties of crime scenes and capture the criminals with the help of other police stations. If any suspect is identified he/she will be taken into custody and there will be identification parades in which the police will call for special evidences such as DNA, Finger Print, Analyst data, etc.

After a particular GCR is recorded, the details of that particular GCR will be reported to the Magistrate. This will initiate a court process as depicted in Fig.2. The details of the crime will be reported to the corresponding court. Charge sheets will be issued with the advices from the attorney-general. Evidences will be processed along with the information of specialist evidences. The court decision will be sentenced by the judge and it will be recorded under six categories, namely B1: False Information Reported, B2: Intentional False Accusation, C1: A Complaint solved with an accused being punished, C2: Suspect being freed with accusations, C3: No Accused, C4: other. After an accused being sentenced, a police IRC list will be published.

These two processes involve a huge amount of data recording, data retrieving, and the investigation is extremely difficult under the conditions of large stacks of data files.

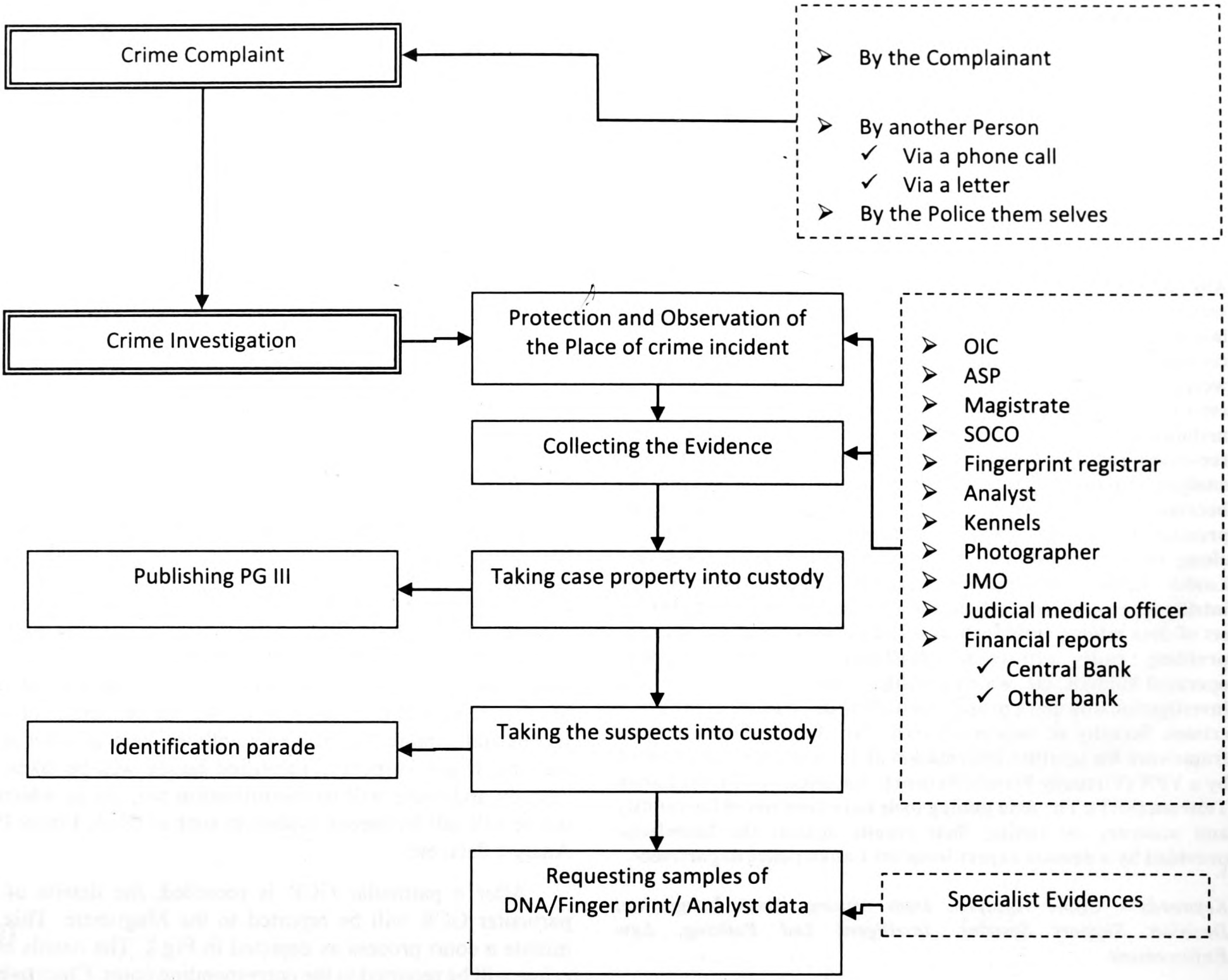

Fig.1. Flow of events in the manual crime investigation system of Sri Lanka Police Department

The technology development and the high growth of community have resulted this high magnitude of crimes, most of the time with bizarre patterns due to solicitation of the cognitive skills in implementing crimes. Crime investigators without an extensive training as data analysts will not be able to analyze these crimes in easy manner due to the availability of this high crime magnitude. Therefore, computing techniques can be used to analyze these data quickly and efficiently since computers can process thousands of records within seconds and they are not error prone like humans. Development of automated system by incorporating data mining and other techniques for decision making can expedite the crime investigation process [2].

II. RELATED WORK

Literature provides enough evidence for systems and candidature methods implemented for crime data investigation and integration. These methods and systems provide important information not only about automated crime analysis, but also about the details related to legislation and crime investigation procedures of different domains.

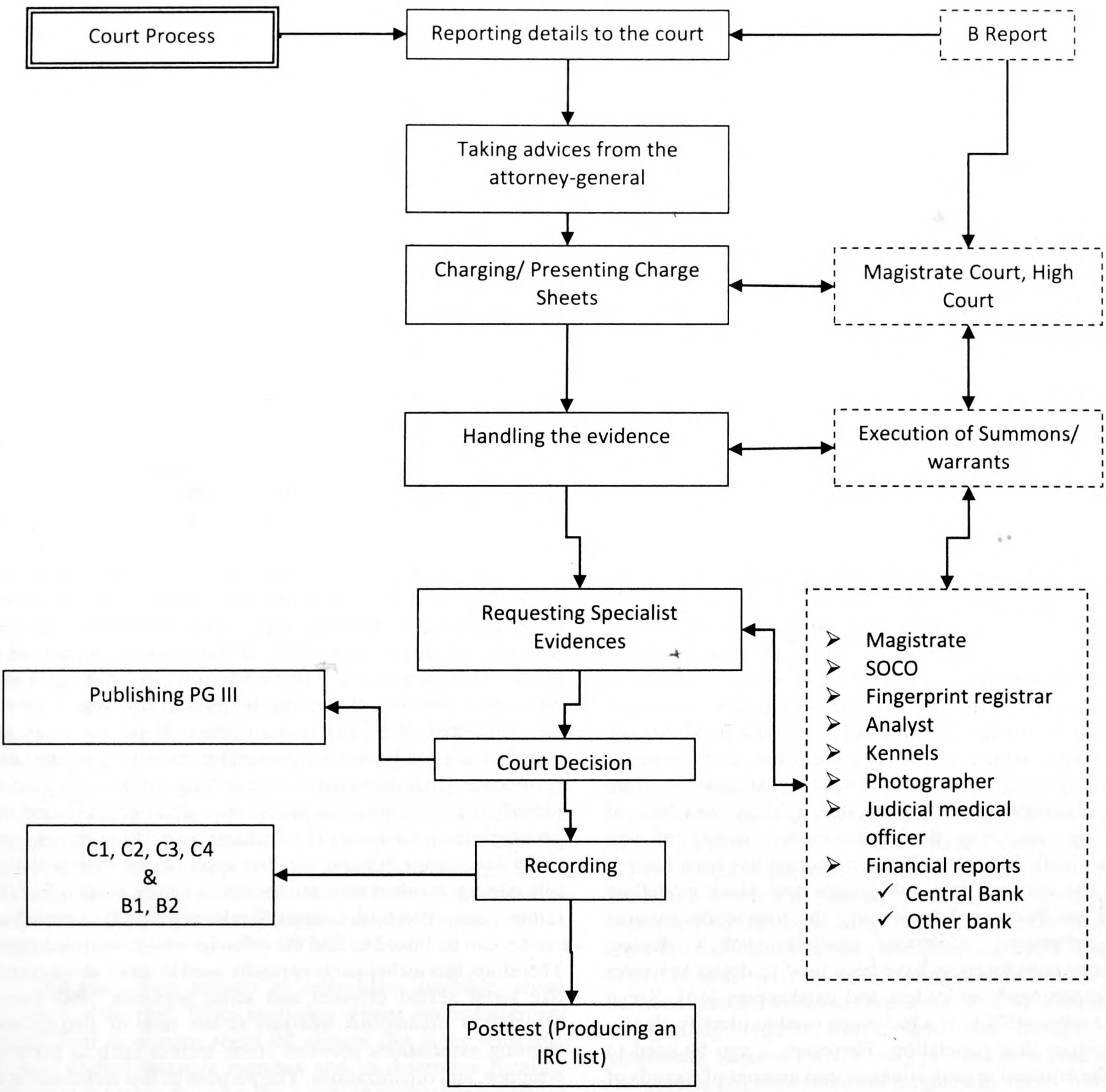

Fig.2. Architecture of the manual court details processing system of Department of Police, Sri Lanka

A. Existing Crime Investigation Systems.

COPLINK is a comprehensive crime investigation and analysis system which has incorporated a collection of data mining tools to support the investigation process Tucson police department, USA. COPLINK provides one easy-to-use interface that integrates different data sources such as incident records, mug shots and gang information, and allows diverse police departments to share data easily raising the concept of implementing a thorough centralized database in information retrieval [3]. Uniform Crime Reporting: National Incident-Based Reporting System [4] is one other automated information system which has aimed at providing reporting standards. H. Chen has introduced the ways of using machine learning methods for information retrieval in information retrieval [5] which can be effectively used for crime data retrieval.

TAS (Timeline Analysis System), AICAMS project, FALCON (Future Alert Contact Network) and CCHRS (Consolidated Criminal History Reporting System) are some other systems which serve as information management or intelligence analysis tools for law enforcement. Each of these systems has its own drawback. One of the main drawback is the

unavailability of a complete knowledge base [6]. Although there are several live crime information systems available, they have been custom made for legislative authorities in different countries and those systems are not accessible to outside of those respective authorities. Crime Reports online system [7] is one such automated system. Calhoun et al. (2008) have developed a web-based crime analysis toolkit designed especially for Virginia law enforcement agencies, in United States, called WebCAT 2.2 with improved data sharing capabilities compared to their previous versions [8]. They have introduced data sharing, analysis, mapping, and querying capabilities which are not available in the other crime mapping and analysis software programs, such as Rigeel, CrimeStat III and Dragnet where crime mapping refers to mapping, visualization and analysis of crime incidents with the help of geographical maps. Most of these systems are expensive [8], not easily customizable to different domains of legislation and do not provide a complete knowledgebase.

B. Existing Crime Investigation and Analysis Techniques

Clustering crimes, finding links between crimes, profiling offenders and criminal network detection are some of the common areas of data mining applied in crime analysis [9]. Association analysis, classification and prediction, cluster analysis, and outlier analysis are some of the traditional data mining techniques which can be used to identify patterns in structured data. New data mining techniques assists in identifying patterns in both structured and unstructured data [2]. The technique concept-space approach has been used in COPLINK project to extract criminal relationships from incident summaries and have created a likely networks of suspects by measuring the co-occurrence weight of two criminals. Single link hierarchical clustering has been used to partition the network into sub groups and block-modelling approach has been used to identify the interaction patterns between subgroups. Centrality measures such as degree, betweenness and closeness have been used to detect key roles in each group, such as leaders and gatekeepers [10]. Social network analysis (SNA) is a technique used to identify closely related cliques in a population. Therefore, it can be used to explore the criminal organizations in vast amount of records of offenders in databases. SNA identifies substructures which has a high local density but separated from the rest of the data points which are considered as subgroups [2]. The k-core method is one of the most common methods used in SNA [9]. Furthermore, affinity propagation and Bayesian networks can provide promising results in identifying relationships between entities and structure of the network [11], [12] . Offender profiling is a methodology which is used in profiling unknown criminals or offenders. There are several other synonyms such as criminal profiling, criminal personality profiling, criminological profiling, behavioral profiling and criminal investigation analysis which are used to refer the same concept. The purpose of offender profiling is to identify the socio-demographic characteristics of an offender based on information available at the crime scene/ scenes [13]. Entity extraction is a technique used to identify the specific patterns in text, image, or audio data. Neural networks can be used efficiently in entity extraction. Entity extraction mainly helps in identifying behavioral patterns of serial offenders [2]. COPLINK project has used a method called named-entity extraction which is a modified version of AI entity extractor system. It uses three steps to identify the names of persons, locations, and organizations in a document. Step one is to identify the noun phrases according to linguistic rules. Second step is to calculate a set of feature scores for each phrase based on pattern matching the lexical lookup. Step three uses a feed forward/back propagation neural network to predict the most likely entity type for each phrase [2].

Clustering techniques are applied to group crimes or offenders in to classes with similar characteristics. For example, this technique can be used to identify the criminals or gangs who do the crimes in similar fashion, who has common interests and same person with multiple false identities. Clustering techniques are more effective in crime association detection and prediction. In unsupervised learning method, similar data items grouped into clusters without knowing their class membership [2]. Complete-link algorithm, single-link algorithm, k-means algorithm, self-organizing maps and affinity propagation are some examples for cluster techniques. Association rule mining discovers items in databases which have frequent occurrences and present them as rules [2].Since, this method is often used in market business analysis to find which products are bought with what other products, it can also be used to find which crimes are conducted with what other crimes. Here, the rules are mainly evaluated by the two probability measures, support and confidence [14]. Association rule mining can also be used to identify the environmental factors that affect crimes using the geographical references [15]. Incident association mining and entity association mining are two applications of association rule mining. Incident association mining can be used to find the crimes committed by the same offender and then the unresolved crimes can be linked to find the offender who committed them. Therefore, this technique is normally used to solve serial crimes like serial sexual offenses and serial homicide [16]. Entity association mining/link analysis is the task of finding and charting associations between crime entities such as persons, weapons, and organizations. The purpose of this technique is to find out how crime entities that appear to be unrelated at the surface are actually linked to each other [16]. Attribution can be used to link crimes to offenders. If two offences in different places involve same specific type, they may be readily attributed to the same offender [9].There are three types of link analysis approaches. They are Heuristic-based, Statistical-based and Template based [16]. Sequential pattern mining is also a similar technique to association rule mining. This discovers frequently occurring items from a set of transactions occurred at different times [2]. Deviation detection detects data that deviates significantly from the rest of the data which is analyzed. This is also called outlier detection. This is used in fraud detection [2]. In classification the data points will be assigned to a set of predefined classes of data by identifying a set of common properties among them. This technique is often used to predict crime trends. Classification needs a reasonably

complete set of training and testing data because high degree of missing data would limit the prediction accuracy [2]. This is a supervised learning method [16]. Methods: Bayesian models, decision trees, artificial neural networks, support vector machines. Applications of classification: Fraud detection, computer and network intrusion detection, bank failure prediction, image categorization [16]. Classification techniques: Bayesian models, decision trees, artificial neural networks, support vector machines. String comparator techniques are used to detect the similarity between the records. It compares the database record pairs and determines the similarity among them. This concept can be used to avoid deceptive records of information of the offenders because information of offenders such as name, address and etc. might be deceptive and therefore the crime database might contain multiple records of the same offender and making the process of identification of their true identity hard [2]. There are advantages and disadvantages in each of these methods. Scientists have also proposed some hybrid methods of the above methods to tally with different classification tasks.

III. FRAMEWORK

This paper proposes a framework which has been tested with a data set of around 250000 of crime records (GCR Crime record of the 5 years from 2010-2015) fed by the police officers of Sri Lanka police department. The results prove that this system improves the efficiency and the reliability of crime recording and analysis conducted by the police personals. The system also facilitates online crime mapping, recording, analyzing and viewing , pattern detection, hotspot detection, models operandi association, link analysis, social network analysis, etc. which ease the process of the existing system allowing any police station to work with the system online without having to physically present in the regional police station.

There is a vast amount of crime data available, about present and the past. When analyzing crimes law enforcement agents need to concern about the present fact along with past records to find possible matches and to determine possible actions to be taken to solve the cases. In manual approach of solving crimes taking into account of all related information is not practical due the capacity of the human calculation and analytical abilities. Hence identifying criminals who are using multiple identities, possible resources accessible to criminals and possible potential threats encountered may not be done as the demanded. Moreover, changing nature in the criminal behavior required dynamic models to the data analysis process. Static models in the manual approach fall behind the par to this concern. Furthermore, accessibility and mobility of data are in a low rate too. A solution to these problems has been proposed with the introduction of a novel scalable data mining framework named as SL-CIDSS (Sri Lanka Crime Investigation Decision Support System) for intelligent led policing.

Fig. 3. represents the layered architecture of the proposed framework which provides the room for extendibility and expandability. The layered architecture facilitates to add new entities which reflect the database relations, new services and new visualization API (Application Programming Interface) in demand in an independent manner.

A. Model

SL-CIDSS has been implemented as a web based system which runs in the servers located at the Headquaters of Sri Lanka police department located at Colombo so that all the police stations located all around the country can access the system though the VPN (virtually private network) which is facilitated by Sri Lanka Telecom, network service. SL-CIDSS is a platform independent. Therefore it provides the capability of being installed into any platform. The layered architecture of SL-CIDSS provides a high scalability of new tools being installed so that the systems completeness can be increased.

B. SL-CIDSS Architecture

As Depicted in Fig.3, SL-CIDSS has been implemented using Spring MVC (version – 4.1.5) framework [17]. AngularJS [18] was used in the view to have a MVC (Model-View- Controller) capability over the view. Since, SL-CIDSS core is written in Java programming language, any Java enabled web server can be used to run the system. Currently, Apache Tomcat 7.0 Web Server is used in running the service. As SL-CIDSS is a web based system, when a user wants to access system, he/she will have to use a web browser. A user can access one of the visualization tools including any report, any map or any CRUD (Create Read Update Delete) GUI, using an HTTP request. A particular tool will correspond to an ng-model which is a directive in AngularJS which binds form elements to a property on the scope using NgModel Controller. NgModel Controller will allow sending and receiving JSON data through HTTP requests and responses. Spring MVC RESTful [17] web services have been used to invoke the SL-CIDSS analysis tools though http request which were sent though NgModel controllers. Jackson JSON library [17] has been used to convert the objects returned from the handlers to JSON format. Map layers are rendered through the OpenLayers Library [19]. SL-CIDSS is composed of around 250 CRUD elements excluding the elements such as Crime Map, Crime Clock, Crime Hotspot Tool, Crime Network Visualization Tool, Geographic Profiling Tool, Pattern Plotter and Reports. Each of these elements communicate through JSON based GET/POST requests which makes SL-CIDSS a scalable system with independent control for each layer.

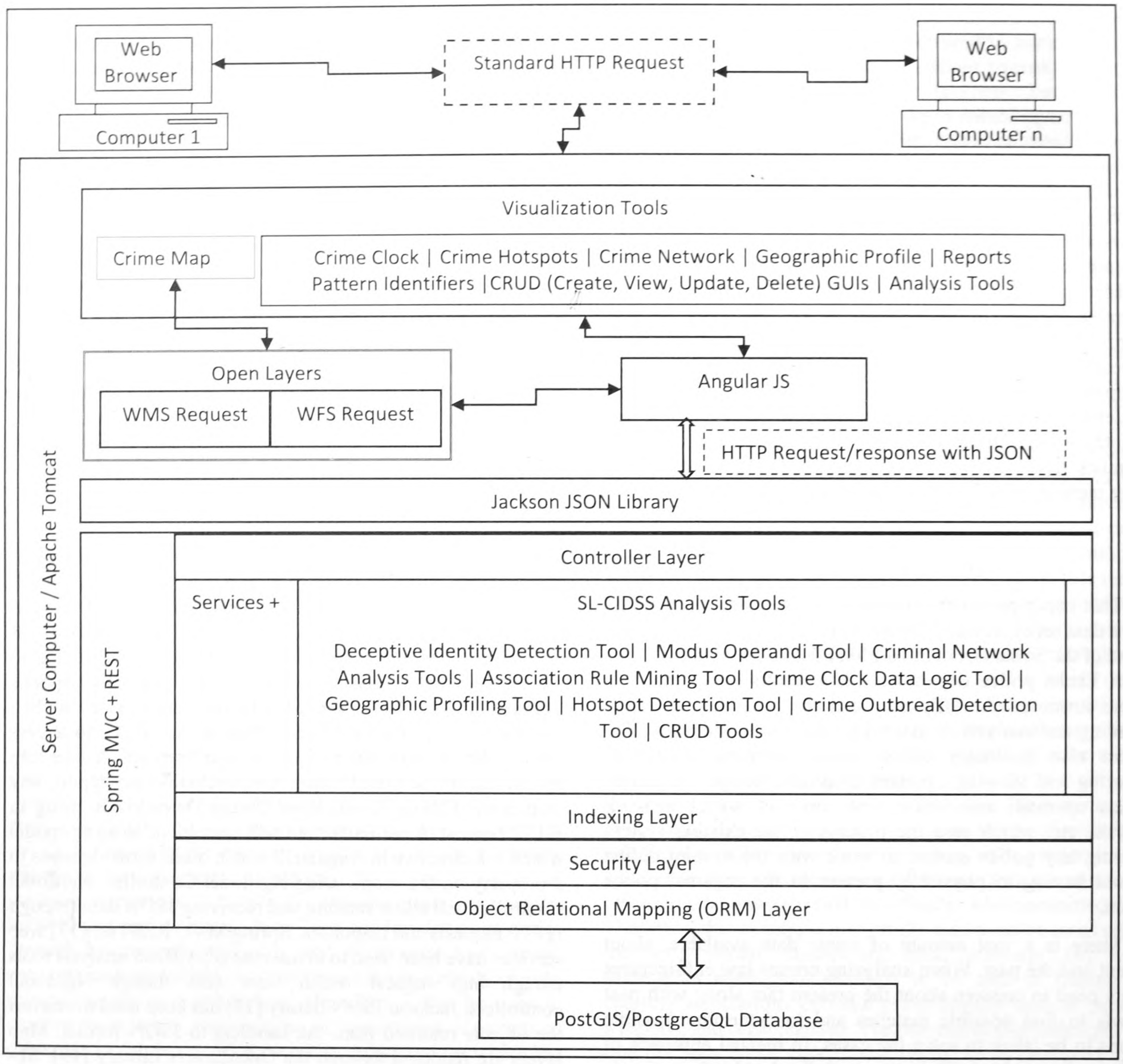

Fig.3. The underlying architecture of SL-CIDSS

C. Information Access and Date Retrival

The graphical user interfaces and the visualization tools shown in Fig.4 and Fig.5 are self-explanatory as the target user group does not have a high computer literacy to cope with complex system scenarios. As shown in Fig.1 and Fig.2, Sri Lanka police have to store many entity sets of information. Therefore, SL-CIDSS has also facilitated this problem by providing a GIS based database using PostgreSQL which stores data including GIS information of all the entity sets. Currently the database SL-CIDSS contains 71 entity sets. Object Relation Mapping has been done through Hibernate 4.4.2 [20].

Since, SL-CIDSS is a web based system, it faces problems such as network bandwidth in information retrieval. Since, the system is accessed through a VPN which has a maximum bandwidth of 128 Kbps for the police stations island wide while only for the main police stations of the 41 divisions there is a

reserved bandwidth of 512 kbps. Although this is the main scenario, a connection of 128 Kbps will normally utilize around 50 Kbps in busy hours. Therefore the system page sizes should be very low. But, since SL-CIDSS purely runs though AJAX based JSON http requests, the bandwidth problem does not affect the systems speed in a drastic manner. The system carries around 52000 crimes each year. Therefore, the amount of data available in the database is very high. When a search query is run in the system, it is subjected to many delays such as I/O delays and network communication delays. Data indexing using Apache Lucene framework [21] has provided a solution for this by letting a search query of searching for a word in a collection of tuples around 10 million restricting to an access time of 30 milliseconds.

D. Information Security

SL-CIDSS is incorporated with a security layer as shown in Fig.3. Which encrypts sensitive data using Apache Shiro security framework [22]. Information such as accused name, age, address are very sensitive information which should not be exposed to outside. The passwords have been encoded with bcrypt password hash algorithm which comes with Spring Security [23]. All the entities of all the entity sets have been incorporated with fields to store IP address of the machine used in inserting the record, record inserted user's ID, record inserted time, record inserted date, record updated user's ID, record updated time , record updated date, IP address of the machine used in updating the record, authority value to a particular tuple so that it will be validate before the record being updated by a particular user. As the information holds high value in the means of law enforcement, when a user deletes a particular record, the status of that record will be changed to 'deleted' with a help of a flag field. Then the record will not be displayed. But, the record will not be removed from the database. All the activities done in the system will be logged in textual format. An automated backup will be generated from the database once per day and saved in two networked partitions located at the Crime Record Division of Sri Lanka police department. Once a user is logged in to SL-CIDSS, another user with the same account credentials will not be allowed to login to the system from anywhere. Inactive sessions will automatically be invalidated after 5 minutes.

IV. Results and Discussion

As SL-CIDSS is composed of much functionality, to minimize the complexity of accessing them, the user interfaces have been categorized into 3 sections, namely Add/Edit/View Tab, Search Tab and Analysis/Reports Tab. Users are provided a composite view doing all the CRUD operations in one window as depicted in Fig. 6. It provides the user a simplified outcome of the whole structure. The navigation through the system screen and the menus have been made easy by providing a sequence for the screens which goes along with the existing crime investigation (See Fig.1) system and the court processing system (See Fig.2). Fig. 4 and Fig.5 depicts the flow of screens of SL-CIDSS which literally inherits the flow of events in Fig.1. and Fig.2. This makes it easy for the police personals to work with SL-CIDSS in an easy manner. Fig.4 depicts the flow of screens of Analysis/reports tab. Fig. 7 shows the suspect interface which holds all the categories of information related to a particular suspect, such as, Suspect general information, suspect photos, suspect biological information, suspect last known address information, suspect relative information, suspect classification code information, suspect trade information, system also generates a dynamic suspect modus operandi by linking all of his/her previously convicted crime scenes. Fig.8 (Crime Comparator) and Fig.9 (Crime Clock) depicts two of the analysis tools which has been incorporated into SL-CIDSS. The system is currently composed of 10 analysis tools in which the flow of the analysis tools is depicted in Fig.4.

Since the framework has been incorporated with indexing which allows the data retrieval very fast this has made the processes of criminal profiling, link analyzing and predicting, and Modus Operandi analysis very efficient in correspondent to a very large data set of crime information. Further, web based implementation provides the mobility and accessibility of the data to law enforcement personnel at required time. However, the accuracy of the results depends on the algorithms used and the implementation of them. Finding the most suitable model can be still situational due to the rational, dynamic criminal behavior and dynamic nature in criminal organizations [11].

V. Conclusion

A novel data mining framework which tallies the GCR crime maintenance structure has been introduced. The decision support tools of the system have directly influenced the crime solving rate due to its fast data retrieval capability which is utilized with data indexing. SL-CIDSS now makes the process of crime reporting to the main police station free from physically being present. The framework has utilized an open space for more data mining tools such as entity mining, string comparator techniques, image processing techniques, etc. to be incorporated in an easy manner without having to worry about the internal architecture.

Acknowledgment

This work was funded by the National Research Council (NRC) of Sri Lanka [Grant number: 11-071].

References

[1] Information Technology Division Sri Lanka Police. (1998) Sri Lanka Police. [Online]. http://www.police.lk/index.php/component/content/article/190

[2] H., Chung, W., et al. Chen, "Crime Data Mining: A general framework and some examples.," *Computer*, vol. 37, no. 0018-9162, pp. 50-56, April 2014.

[3] Jenny Schroeder, Roslin V.Hauck Linda Ridgeway, Homa atabakhsh, Harsh Gupta, Chris Boarman, Kevin Rasmussen, Andy w.clements Hsinchun Chen, "COPLINK Connect: information and knowledge management for law enforcement," *Decision Support Systems*, vol. 34, pp. 21-285, 2002.

[4] United States Department of Justice, *Uniform Crime Reporting: National Incident-Based Reporting System, Data Collection Guidelines.*, 1998, vol. 1.

[5] H. Chen, "Machine learning for information retrieval: neural," *Journal of the American Society for Information Science*, vol. 46, no. 3, pp. 194-216, 1995.

[6] Jenny Schroeder,Roslin V. Hauck, Linda Ridgeway, Homa Atabakhsh, Harsh Gupta, Chris Boarman, Kevin Rasmussen, Andy W. Clements Hsinchun Chen, "COPLINK Connect: information and knowledge management for law enforcement," *Decision Support Systems*, no. 34, pp. 271-285, 2002.

[7] (2009) Crime Reports. [Online]. https://www.crimereports.com/

[8] Calhoun, C.C., Stobbart,C.E., Thomas, D.M., Villarrubia, J.A., Brown, D.E., "Improving Crime Data Sharing and Analysis Tools for a Web-Based Crime Analysis Toolkit: WebCAT 2.2," in *Proceedings of the 2008 IEEE Systems and Information,Engineering Design Symposium, University of Virginia*, Charlottesville, 2008, pp. 40-45.

[9] Brian Ewwart Giles Oatley, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.*, 2011, vol. 1.

[10] (2011) About COPLINK project. [Online]. http://ai.bpa.arizona.edu/research/coplink/

[11] Brendan J. Frey and Delbert Dueck, "Clustering by Passing Messages Between Data Points," *SCIENCE*, vol. 315, pp. 972-976, Feb. 2007.

[12] C. J. Rhodes and E. M. J. Keefe, "Social Network Topology: A Bayesian Approach," *The Journal of Operational Research Society*, vol. 58, no. 12, pp. 1605-1611, Dec. 2007.

[13] (2011) Offender Profiling. [Online]. http://www.liv.ac.uk/psychology/ccir/op.html

[14] R., Imielinkski, J., and Swami, A. Agrawal, "Mining Association rule between sets of items in large databases," in *Proceedings of the ACM SIGMOD International Conference of Management of Data*, New York, 1993, pp. 207-216.

[15] K. and Han, J. Koperski, "Discovery of spatial association rules in geographic information databases," in *Proceeding of the 4th International Symposium on Spatial Databases*, 1995, pp. 47-67.

[16] Hsinchun Chen, *Intelligence and security informatics for international security.*, 2006, vol. 10.

[17] Craig Walls, *Spring in Action.*: Manning Publications, 2011, vol. 3.

[18] Brad Green and Shyam Seshadri, *AngularJS.*: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2013, vol. 1.

[19] Dev Team OpenLayers. (2006) OpenLayers. [Online]. http://openlayers.org/

[20] Red Hat. (2015, Jan.) Hibernate. [Online]. http://hibernate.org/

[21] The Apache Software Foundation. (2012) The Apache Software Foundation. [Online]. https://lucene.apache.org/

[22] The Apache Software Foundation. (2008) The Apache Software Foundation. [Online]. http://shiro.apache.org/

[23] Pivotal Software. (2015) Spring. [Online]. http://docs.spring.io/autorepo/docs/spring-security/3.2.4.RELEASE/apidocs/org/springframework/security/crypto/bcrypt/BCrypt.html

[24] Jenny Schroeder, Roslin V. Hauck, Linda Ridgeway, Homa Atabakhsh, Harsh Gupta, Chris Boarman, Kevin Rasmussen, Andy W. Clements Hsinchun Chen, "COPLINK Connect: information and knowledge management for law enforcement," no. 34, pp. 271-285, 2002.

APPENDIX A. FIGURES.

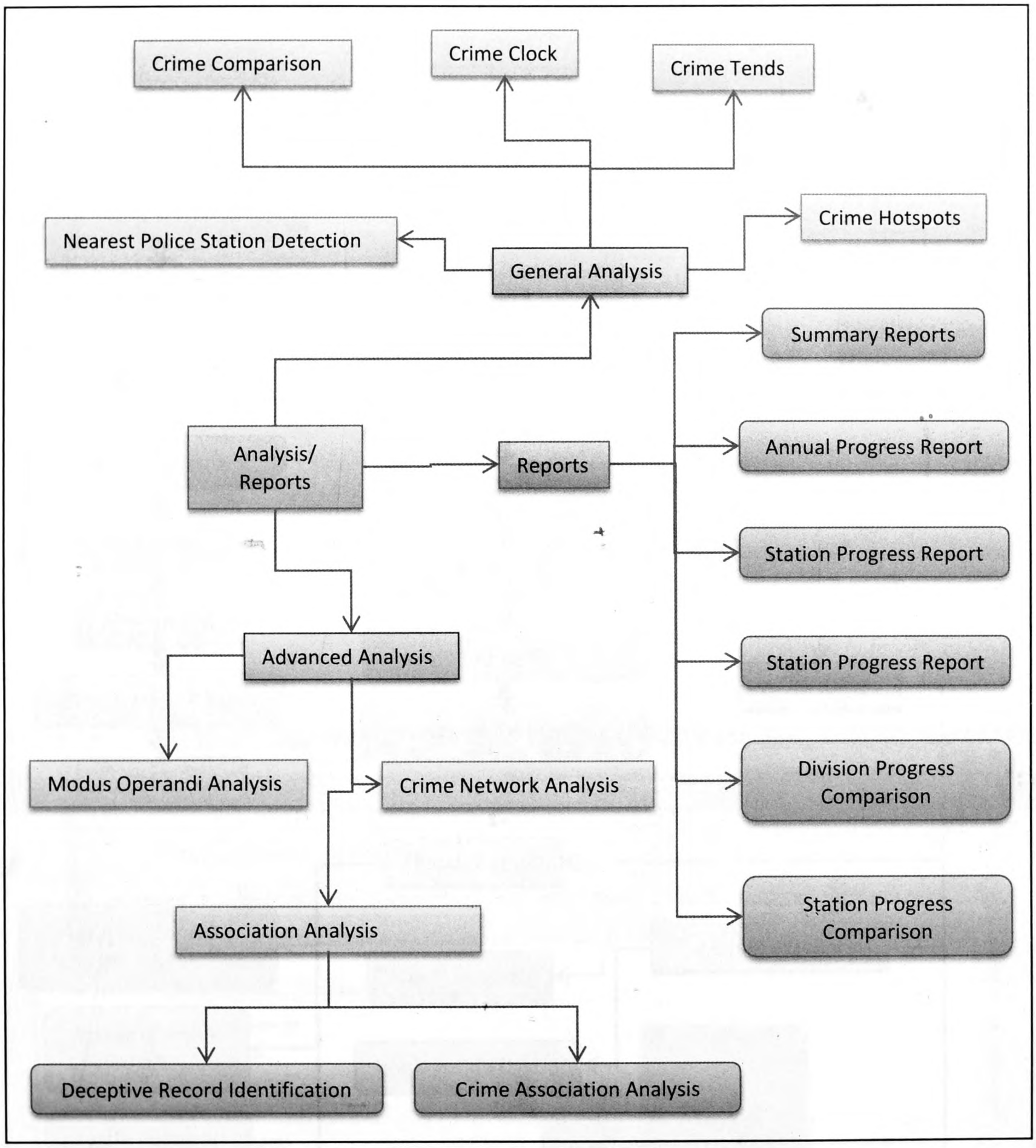

Fig.4. Flow of screens of Analysis /Reports Tab

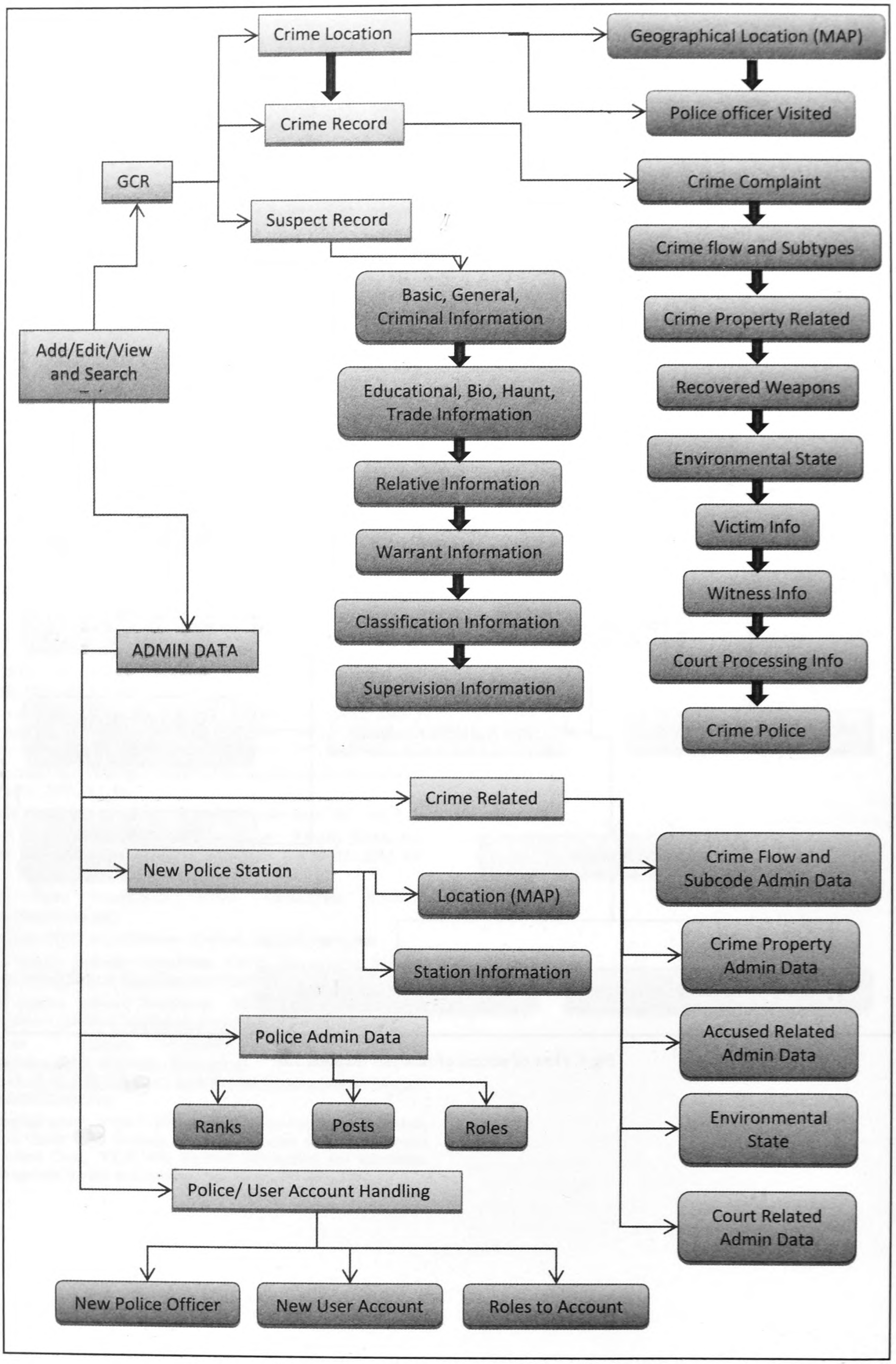

Fig.5. Flow of screens of Add/Edit/View and Search Tab

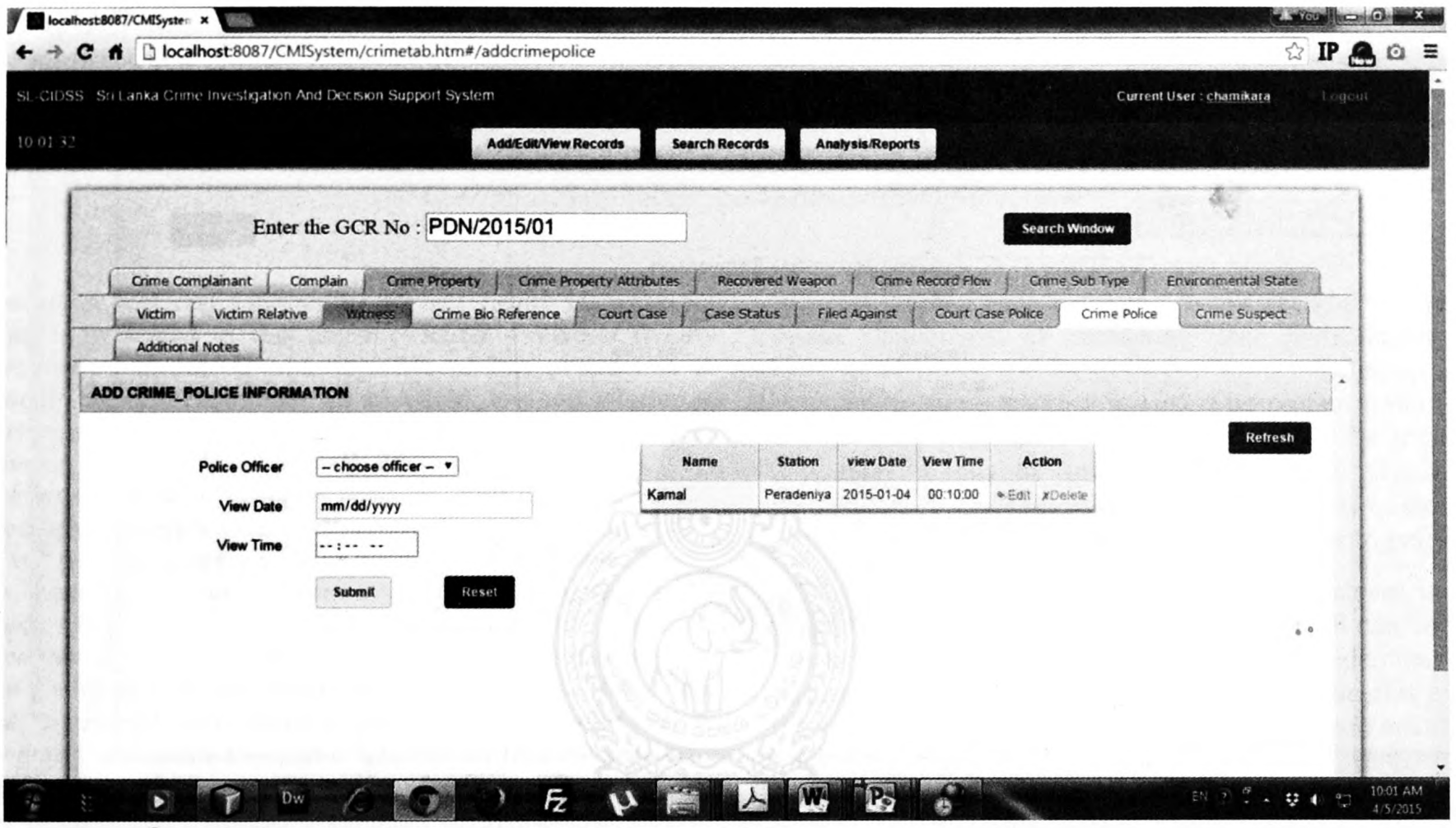

Fig.6. A composite view of GCR window of SL-CIDSS

Fig.7. Suspect view

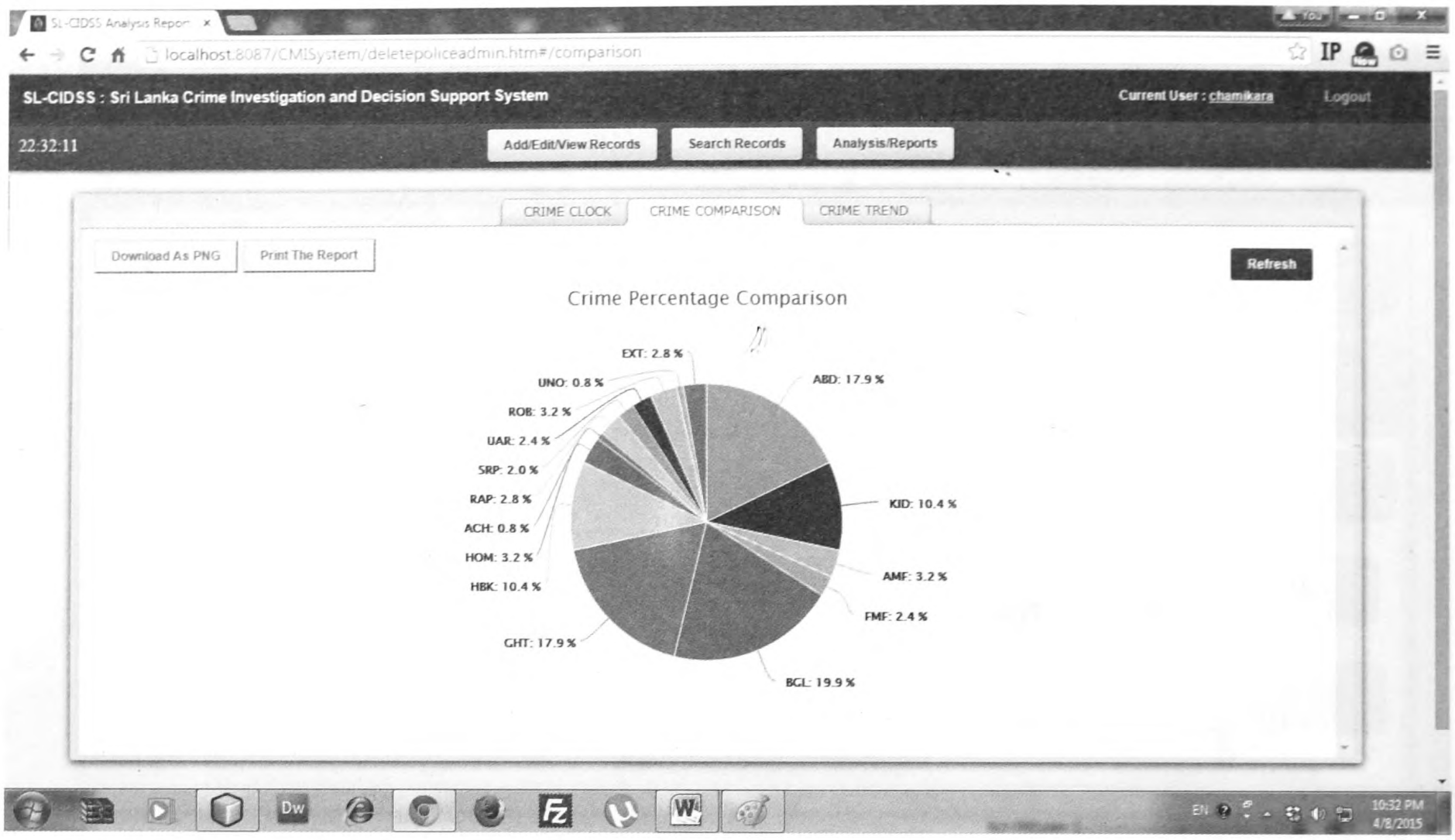

Fig.8. Crime Comparison Tool

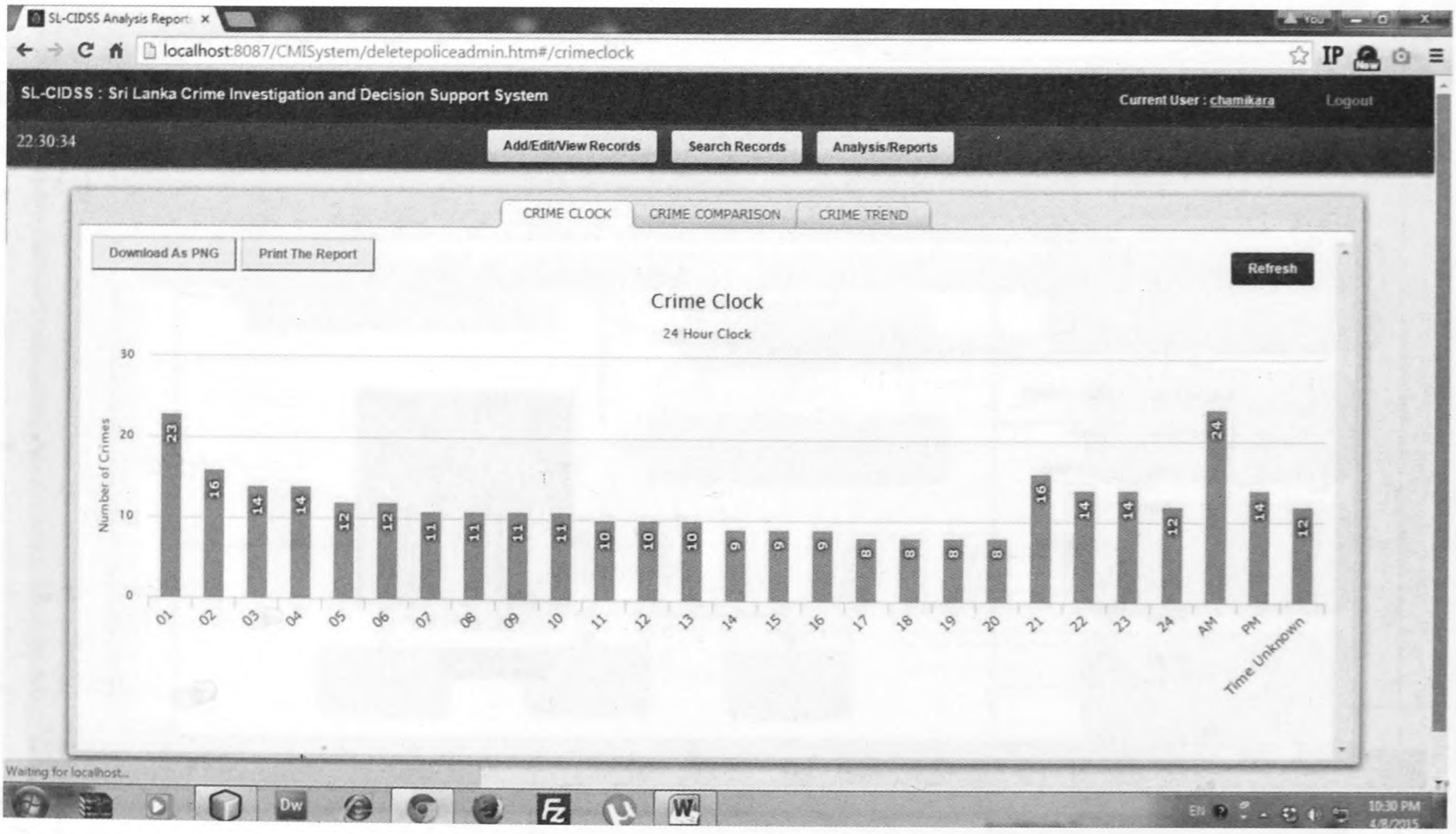

Fig.9. Crime Clock Tool