National Forensic DNA Database Management System For Criminal Investigations

NCD Arambawela[1#], PPNV Kumara[1], CP Waduge[1] and A Manamperi[2]

[1] Department of IT, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka
[2] Molecular Medicine Unit, Faculty of Medicine, University of Kelaniya, Sri Lanka
[#]ncdarambawela92@gmail.com

Abstract *— A DNA profile is a genetic fingerprint which is unique to each person and it is used as a powerful evidence to identify individuals in criminology. A DNA database holds DNA profiles taken from individuals and crime scenes and helps in quick identification of criminals that leads to improve the efficiency of the judiciary system of a country. The current process of using DNA evidence in court cases in Sri Lanka is more time consuming since the widely used electrophoresis gel image analysing technique requires expert scientists and often has human errors because of the anomalies that can occur in the faint gel images. Although there are novel advanced technologies and equipment based on capillary electrophoresis, the enormous cost of implementing and licensing them is unaffordable to third world developing countries like Sri Lanka. Therefore, this research paper proposes a fully automated and cost effective methodology to re-engineer the current manual process by integrating recent advancements in the computer science field such as image processing and machine learning techniques to spare human being from voluminous and tedious image analysing and to provide accuracy and rapid speed without human errors. The computerized system will get the scanned electrophoresis gel image as the user input and it will enhance the quality of the gel image by using image processing techniques. After that it will generate the DNA profile and save it as a number pattern to the database for speedy retrieval. The system provides matches linking an individual to a crime scene or a crime scene to another crime scene. Furthermore, the system facilitates family relationship identification based on machine learning techniques. It ensures the security by integrating fingerprint authentication. Establishing a National Forensic DNA Database Management System in Sri Lanka will benefit in identifying criminals and excluding the innocent efficiently. More importantly, it will support to prevent criminals from having many opportunities to commit mass murders, rapes and robberies and to identify potential victims.*

Keywords— **DNA profiling, DNA database, Electrophoresis gel images, Image processing and Machine learning**

I. INTRODUCTION

DNA typing is an accurate scientific method that uses to identify individuals from the differences in their DNA (Deoxyribose Nucleic Acid) structure. It produces a pattern which is termed as a DNA profile. An individual's DNA profile is a genetic fingerprint which is unique to each person with the exception of identical twins that have the same DNA content. But the probability of having the same DNA profile is less than one in a billion (Great Britain et al., 2015). Therefore, DNA profiles are used as a powerful evidence to identify individuals in forensic science.

Every cell in a person's body contains the same DNA content. Therefore, DNA profiling can be performed by using a very little amount of biological specimens using blood, sperms, muscle, salvia, bones, teeth or even sweat.

The nucleus is the commanding center of a cell and it houses the DNA that codes genetic information responsible for all cellular functions. Several DNA molecules that comprise genes are called as chromosomes. In a human cell, there are a total of 46 chromosomes; 23 chromosomes are inherited from the mother and the other 23 chromosomes are inherited from the father. There are repeating units of the same type of DNA sequence in some chromosomal regions and the number of those repeating units in individuals can vary. Hence those repeated DNA sequences which are known as Short Tandem Repeat (STR) or STR markers are used in human identification in criminology. Tetra-nucleotide repeats are mostly used in DNA typing. DNA profiling is done by counting the number of times each repeating unit occurs within a specific area on the chromosome.

Electrophoresis gel image analysis is the most widely used technique that is used for DNA fingerprinting. Electrophoresis is an electrochemical separation method that makes DNA molecules to migrate through a specific substrate such as polyacrylamide gel under the influence of an electrical current. The output of this electrophoresis process is called as an electrophoresis gel image and that scanned image is analysed by the expert scientists for DNA profiling.

In forensic science, DNA typing is used for two main purposes; to identify individuals from biological samples and to determine familial relationships. Since a DNA profile has recognized as a powerful evidence in forensic casework, the concept of national forensic DNA database was introduced to the field of criminology. Hence national DNA databases have been established in several countries such as United Kingdom, Netherlands, Australia, Germany, Finland, Norway, etc. and many other countries are developing DNA database systems. Researches highlight that DNA databanks of criminal offenders benefit in quick identification of criminals and help to reduce the crime rates of a country.

Although the existing legal framework in Sri Lanka provides the legal acceptance for DNA evidence in court cases, still a national DNA database is not available. In Sri Lanka DNA profiling was used for the first time in a criminal case when six family members were murdered in Hokandara. Aftermath over 1400 criminal cases and over 3000 cases of disputed parentage have had the advantage of DNA evidence ("Genetech," n.d.). Some significant incidents such as Sarath Ambepitiya murder case, Royal Park murder case and Seya Sadewmi's murder case can be defined as landmark cases in terms of using DNA typing as expert evidence.

The current process of using DNA profiles as an evidence in court cases in Sri Lanka is more time consuming and not cost effective. It requires expert scientists to match profiles. Sometimes DNA samples that are taken from crime scenes are full of dust, sault, etc. The electrophoresis gel images of those samples contain anomalies, hence it is very difficult to analyse those faint images. Security issues also have arisen regarding the current process. Nowadays, most countries are adopted to a novel DNA analysing technique which is called as capillary electrophoresis. The huge cost for purchasing, implementing, maintaining and annual licensing of the equipment based on capillary electrophoresis method is unaffordable to third world developing countries like Sri Lanka.

Therefore, this research paper suggests to deploy recent advancements in computer science field to re-engineer the existing electrophoresis gel image analysing process. Image processing techniques can be used to eliminate the anomalies that can occur in electrophoresis gel images. All the DNA profiles of individuals and crime scenes are stored as a number pattern along with the personal information in the database for further manipulation. Thereby, it will save the time taken for investigations. Machine learning techniques will be integrated to the system to enhance the efficiency of the family relationship matching process. Web technology is combined with the proposed solution to speed up the existing process. It will gain the advantage of using biometric authentication technique to ensure the security of the sensitive data that handles in the system.

The aim of this research is to enhance the efficiency, accuracy and security of the current DNA analysing procedure in forensic casework in Sri Lanka by re-engineering it through integrating appropriate newest technologies introduced in the computer science field.

II. LITERATURE REVIEW

The literature review emphasizes an appraisal of the similar existing systems. The objective of the literature review is to analyse the focused problems, adopted technologies, proposed solutions and research gaps in previous similar researches. Furthermore, it focuses on evaluating the applicability of previous studies into the problem domain of this research.

A research highlights a four steps algorithm that addresses the bottleneck for further development and reproducibility issues of manual or semiautomatic DNA profiling processes. The four steps include in this algorithm are automatic thresholding, shifting and filtering, detecting and annotating gel bands and data processing. Automatic thresholding is engaged to equalize the grayscale levels of electrophoresis gel image background without affecting the size of DNA bands. The purpose of the second step is to shift the minimum level of the gel image to zero and to remove as much noise as possible using top hat filter. An indirect method consists of top-hat filtering and bottom-hat filtering has been used in the third step to improve the quality of gel images. In the data processing step object detection technique has been used to detect all the DNA bands and to compute quantitative information (Kaabouch et al., 2007).

A method of detection of DNA fingerprint to identify family relationship through the use of image processing based on medical knowledge is presented in literature. In this method, when a DNA testing performer inputs a DNA fingerprint to a commercial program, the DNA gel image is improved by image enhancement as the first step. For the next step the software converts the DNA image to a binary code, reduces small noisy spots and increases the quantity and the size of noisy spots of the size of the binary image to be 45*26 pixels. In the third step it is correlated by template matching to identify the same DNA positions in terms of mother, father and child. As the final step a complete relationship or no relationship is verified in terms of 10 positions of similarity (Kiattisin and Leelasantitham, 2008).

Various image processing techniques had been used in a three steps algorithm to analyse DNA gel images. Firstly an enhanced fuzzy c-means algorithm is designed for image segmentation to extract the helpful information from the DNA gel images and exclude the unnecessary background that includes blurred noise. The next step named as lane detection uses Gaussian function to estimate and detect the location of each lane on the gel images. For the final step renewing lost bands and eliminating repetitive bands are applied in order to ascertain each band more accurately (Lee et al., 2011).

Literature emphasises image analysis techniques and pattern recognition techniques can be used to obtain quantitative and qualitative information from gel images. The background of most gel images varies because of the presence of noise and some bands of gel images are not aligned horizontally because of non-uniform migration called as ‘smile’ on gel. The goal of image enhancement is to obtain accurate quantification from distorted gel images. Top-hat transform technique has been used for background removal. The clustering analysis addresses the issue of measuring similarity and dissimilarity between two samples based on the distance between them. The K-means algorithm, ISODATA algorithm and vector quantization are used as clustering methods based on discriminant analysis (Ye et al., 1999).

The “GELect tool” is an appropriate software for DNA diagnosis from 1D electrophoresis gel images. The workflow of GELect comprises three main procedures; lane segmentation, DNA band extraction and band genotyping. This tool efficiently segments lanes from the gel image by detecting curved lanes automatically and it constructs a band model by performing band registration against a reference band. GELect tool has used the band classification technique to perform genotyping from DNA gel images (Intarapanich et al., 2015).

A research done by Maxwell and William provides an overview of the applicability of machine learning approaches for analysing genome sequencing data sets. It presents challenges and considerations in the application of supervised, unsupervised and semi-supervised machine learning methods as well as generative and discriminative modelling approaches. Although the generative modelling gives more compelling benefits than discriminative modelling, discriminative modelling achieves more performance than generative modelling (Libbrecht and Noble, 2015).

Previously discussed researches emphasise, the image quality of electrophoresis gel images plays a significant role in automatic analysis of DNA signatures. Existing researches highlight the filtering techniques still require considerable human intervention and pre-assumptions to define appropriate values for thresholding due to the diversity and quality of images. Moreover, those researches underlines that some techniques are not suitable for DNA electrophoresis gel images because they alter the size of the DNA bands.

Most of the existing algorithms consider the data structure of DNA sequences as tree, string and graph. Some researches highlight that it can limit the efficiency of retrieval and the speed of processing.

Some existing systems have used only the image processing techniques according to the addressed problem domains. Some systems have used signal processing techniques or pattern recognition techniques combine with image processing techniques to achieve better performance in DNA sequence analysis. Some researches emphasise signal processing techniques has received a great attention in analysing numeric sequences compared to other technologies such as pattern recognition. Very few researches have considered of applying supervised machine learning techniques such as Artificial Neural Networks in analysis of DNA sequences.

Even though previously discussed DNA image analysing systems are dealing with sensitive data, those systems did not consider about applying robust security mechanisms to those systems since they are not combining with national DNA databases and crime analysis.

This research focuses on addressing the above identified research gaps to deploy a more accurate, efficient and secured IT solution in forensic science in order to speed up the traditional DNA profiling process in Sri Lanka.

III. METHODOLOGY AND EXPERIMENTAL DESIGN

A. *Data Gathering*

Qualitative and quantitative data that required for carrying out the research were gathered through interviews and document reviews. An interview was conducted with the Head of Department of the Molecular Medicine Unit of the Faculty of Medicine, University of Kelaniya to acquire the medical knowledge that needs for the project. Another interview was done with the Technical officer of the DNA laboratory of the Molecular Medicine Unit of University of Kelaniya to gather information of the electrophoresis gel image analysing process. An interview was conducted with the Deputy Government Analyst of the Government Analyst Department to gather data about the existing procedure of analysing DNA sequences for court cases. Some DNA electrophoresis gel images, reports and excel sheets of

DNA profiles were reviewed during the data gathering process.

B. *Data Analysis*

The data which was gathered during the data collection process was analysed in this phase to define the problem and to identify the limitations with the existing process. Suggestions of the users to improve the current process could be identified through the collected data. Consuming much time and cost, low security, limited analysing capabilities due to unclear images are some of the drawbacks with the manual procedure which could be identified through data analysis. Moreover, the data analysis phase highlighted that the users are also willing to adopt to an automated system to speed up the DNA analysing process. And also this phase emphasized that re-engineering the existing process for better functionality is more important to enhance the efficiency of crime investigations in Sri Lanka.

C. *Approach*

Users of this system are authorized police officers, jailers and laboratory staff. Inputs for the system are scanned DNA electrophoresis gel images and the personal details of offenders or suspects. Outputs from the system are basically a report that indicates the matching probability of the profiles, an email of the report and a SMS alert. The system receives inputs and executes user requests to generate DNA reports and display the output through an email.

D. *Technology adopted*

The developed system consists with a web application and a standalone application. The web application has been developed using Bootstrap UI design framework and programming languages such as HTML, CSS, JavaScript and PHP. The standalone application was designed using Swing framework. Java programming language with object oriented programming concepts was used to programme the functions of the standalone application. Image enhancement process was designed through image processing filters and functions in MatLab. An Artificial Neural Network that is known as a supervised machine learning model which identifies the patterns based on a previously trained dataset, has been integrated in order to develop the family relationship matching process. A MySql database has designed to store data. Desktops, Scanners and fingerprint scanners are some of the hardware technologies deployed in the developed system.

E. *Design*

The overall architecture of the system can be defined based on three main layers; client layer, application layer and database and server layer. Architectural perspective of the developed system is shown in figure 1.

1) Client Layer: The client layer provides the access to the users of the system. There are two main user levels (authorized police/jail officers and authorized DNA laboratory staff) in this system and access levels are varies from one user level to the other. Therefore the authorized users will be predefined.

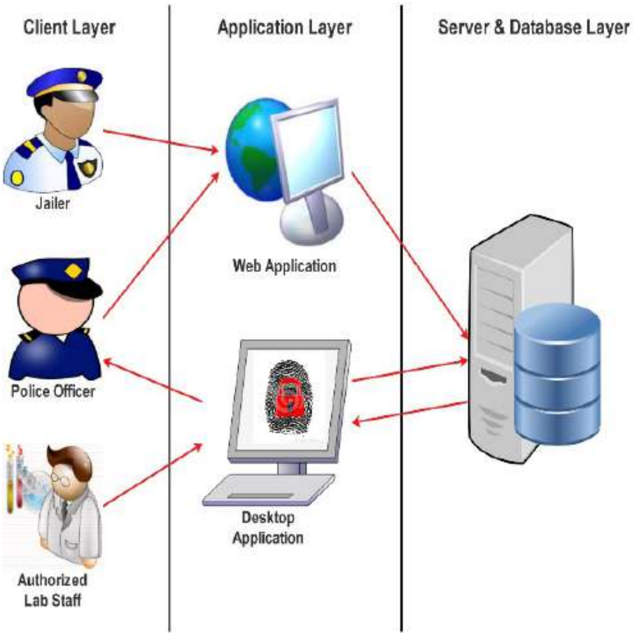

Figure 1. System Architecture

2) Application Layer: The application layer consists with a web application and a standalone application. The web application is designed for the jailers and police officers. It can be further described through the below discussed modules.

i. Login module

This module verifies the username and password that are entered by the users and gives the access to the system for the predefined authorized users. It has given an option to recover username and password if the username or password is forgotten.

ii. Insert suspects module

In this module police officers can enter the personal details of the suspects whose biological samples have been submitted for DNA profiling.

iii. Insert offenders module

Jailers can insert the personal details of the offenders whose biological samples have been submitted for DNA profiling.

iv. Update suspects module

When the suspects are termed as offenders, relevant details of them such as imprisonment date, number of years of imprisonment, etc. can be inserted to the system through this module.

v. Update offenders

This module facilitates jailers to update the information of offenders when necessary.

vi. Delete offenders

When the offenders die, the details and DNA profiles of them can be deleted through this module.

The desktop application allows authorized DNA laboratory staff to perform the functions through below described modules.

i. Login module

This module allows authorized users to access the system by providing the fingerprint scan.

ii. One-to-one profile matching module

In this module user can browse and insert the scanned electrophoresis gel image. The “image enhancement” option has been provided to improve the quality of the image. Then the DNA profile can be generated by clicking the “generate profile” option. After that generated profile will be saved to the database and then it goes for a search against the database to find a full match or a family match. A report will be generated indicating the DNA profile, matching probability and other relevant details. The workflow of this module is shown in figure 2.

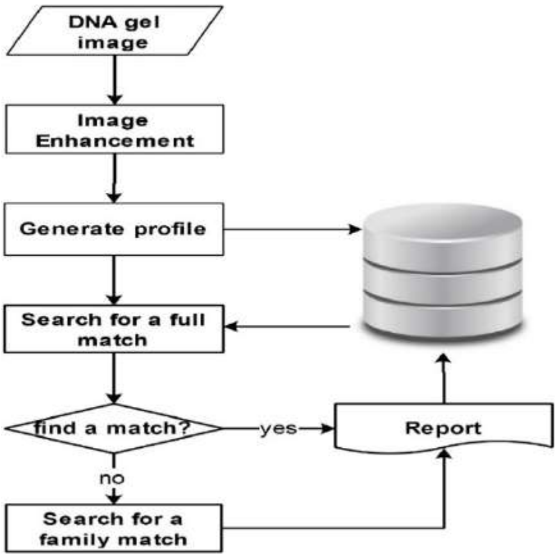

Figure 2. One-to-one matching workflow

iii. Family relationship matching module

Scanned electrophoresis gel image will be browsed and inserted to enhance the quality of the image. Then the DNA profiles of mother, father and child will be generated respectively. Those profiles will be saved to the database and find whether there is a relationship in terms of father, mother and child. After that a report will be generated. Figure 3 shows the workflow of this module.

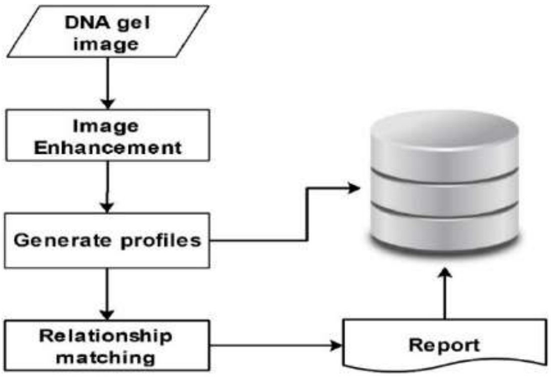

Figure 3. Family relationship matching

iv. Email sending module

The generated reports will be send to the relevant personals through emails in this module.

v. SMS sending module

This module allows to send SMS alerts automatically to inform about the emails.

3) Server and Database Layer: This layer is responsible for managing the server and the database of the system. The server will provide the web connectivity to the web application and the database will store the data that are entered to the system by two applications and the stored data will be provided to the users when necessary.

IV. RESULTS AND DISCUSSION

The above identified filtering techniques such as top-hat filtering, bottom-hat filtering, Fourier transformation, etc. that described in previous researches could not applied directly to the addressed research domain. Therefore, a combination of image processing filters, functions and algorithms were used for the image enhancement process of the system to improve the quality of electrophoresis gel images without altering the size of the DNA bands.

The speed of retrieval and manipulation of data could be upgraded by representing the DNA profiles in number patterns rather than symbolizing them as tree, string or graph.

Moreover, using machine learning techniques along with image processing techniques in DNA profile matching provides a better combination to analyse DNA sequences. It gives more accurate and efficient results than other combinations which are described in the previous studies that discussed in the literature review section.

More importantly, security breaches of standalone application could be eradicated successfully by using fingerprint authentication mechanism rather than using traditional username and password mechanism.

Implementation of the web application that allows police officers and jailers to insert personal details from different places helps to spare laboratory staff from tedious data entering. The distribution of works benefits in enhancing the efficiency of the system. Email sending facility and SMS gateway supports to reduce the paper work.

More efficient and accurate electrophoresis gel image analysing program for DNA profiling and a secured solution for establishing a national forensic DNA database could be provided through this research by addressing the research gaps found in the previously proposed solutions in literature.

V. CONCLUSION AND FURTHER WORK

In this research work, the author could develop an efficient and secured National Forensic DNA Database Management System by integrating novel advancements in computer science field in order to speed up the traditional electrophoresis gel image analysing process. It will help in quick and accurate identification of criminals that leads to speed up the crime investigation process which plays a major role in an efficient judiciary system of a country. The newest system will support not only to prevent criminals from having many opportunities to commit crimes rapidly, but also to identify potential victims of serial murders, rapes and robberies.

In the near future, the database will populate with DNA profiles day by day. Researches has estimated that the stored DNA profiles of the DNA databases are increased by about 75% in each year (Great Britain et al., 2015). This rapidly increasing data may provide bottlenecks for speedy retrieval and manipulation of data. Therefore, the author infers to integrate Big Data theories and Data Warehousing technologies to handle the large amount of data for further work.

DNA samples that are taken form crime scenes such as gang rapes provides a DNA profile which is called as a mixed DNA profile. There is no proper automatic technique to analyse those mixed DNA profiles. Finding a novel accurate and efficient computerized solution for mixed DNA profile analysing will be an interesting future direction that will help leads to a great upheaval in forensic science and computer science fields.

REFERENCES

Abeykoon A., Yapa, R.D., Sooriyapathirana S., (2013) An Automated System for Gel image Analysis Using Image Processing Technologies and Signal Processing Technologies.

Bailey D.G., Christie B.C., (1994) Processing of DNA and protein electrophoresis gels by image analysis, in: Proceedings of the Second New Zealand Conference on Image and Vision Computing, Palmerston North. pp. 1–2.

Genetech [WWW Document], n.d. . Genetech. URL http://www.genetechsrilanka.com (accessed 7.1.17).

Great Britain, Home Office, National DNA Database (Great Britain), Strategy Board, 2015. National DNA Database Strategy Board annual report 2014/15. Stationery Office, London.

Herisson J., Payen G., Gherbi R., (2007) A 3D pattern matching algorithm for DNA sequences. Bioinformatics 23, 680–686. doi:10.1093/bioinformatics/btl669

Intarapanich A., Kaewkamnerd S., Shaw P.J., (2015) Automatic DNA diagnosis for 1D gel electrophoresis images using bio-image processing technique. BMC Genomics 16, S15.

Kaabouch N., Schultz R.R., Milavetz B., (2007) An analysis system for DNA gel electrophoresis images based on automatic thresholding an enhancement, in: Electro/Information Technology, 2007 IEEE International Conference on. IEEE, pp. 26–31.

Kazhiyur-Mannar R., Smiraglia D.J., Plass C., (2006) Contour area filtering of two-dimensional electrophoresis images. Med. Image Anal. 10, 353–365.

Kiattisin S., Leelasantitham A., (2008) A Detection of DNA Fingerprint Using Image Processing Based on Medical Knowledge. UTCC Eng. Res. Pap.

Koprowski R., Wróbel Z., Korzyńska A., (2013) Automatic analysis of 2D polyacrylamide gels in the diagnosis of DNA polymorphisms. Biomed. Eng. Online 12, 68.

Lee J.D., Huang C.H., Wang N.W., (2011) Automatic DNA sequencing for electrophoresis gels using image processing algorithms. J. Biomed. Sci. Eng. 04, 523–528. doi:10.4236/jbise.2011.48067.

Leung M.K.K., Delong A., Alipanahi B., (2016) Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. Proc. IEEE 104, 176–197. doi:10.1109/JPROC.2015.2494198.

Libbrecht M.W., Noble W.S., (2015) Machine learning applications in genetics and genomics. Nat. Rev. Genet. 16, 321–332. doi:10.1038/nrg3920.

Liew A.W.C., Yan H., Yang M., (2005) Pattern recognition techniques for the emerging field of bioinformatics: A review. Pattern Recognit. 38, 2055–2073. doi:10.1016/j.patcog.2005.02.019.

Moreira B., Sousa A., Mendonça A.M., (2013) Automatic Lane Segmentation in TLC Images Using the Continuous Wavelet Transform. Comput. Math. Methods Med. 2013, 1–19. doi:10.1155/2013/218415.

Ye X., Suen C.Y., Cheriet M., (1999) A recent development in image analysis of electrophoresis gels, in: Vision Interface. pp. 19–21.

ACKNOWLEDGEMET

Author would like to acknowledge the supervisor Mr. P.P.N.V. Kumara and co-supervisor Ms. C.P. Waduge for their motivation and inspiration that triggered throughout the research. The author is highly indebted to Prof. A. Manamperi for the guidance as well as the massive support regarding the research in innumerable ways. Author would also like to thank the technical officer and all the staff of DNA laboratory of Faculty of Medicine, University of Kelaniya, whose steadfast support of this research was deeply appreciated.