Strathmore UNIVERSITY

SU+ @ Strathmore
University Library

Electronic Theses and Dissertations

2023

Sexual and Gender-Based Violence reporting and pattern analysis tool: a case of Nairobi County.

Makau, Agnes Nzembi
School of Computing and Engineering Sciences
Strathmore University

Recommended Citation

Makau, A. N. (2023). *Sexual and Gender-Based Violence reporting and pattern analysis tool: A case of Nairobi County* [Strathmore University]. http://hdl.handle.net/11071/13521

Follow this and additional works at: http://hdl.handle.net/11071/13521

This work is availed for free and open access by Strathmore University Library.
It has been accepted for digital distribution by an authorized administrator of SU+ @Strathmore University.
For more information, please contact library@strathmore.edu

Sexual and Gender-Based Violence Reporting and Pattern Analysis Tool: A Case of Nairobi County

By

Makau Agnes Nzembi

147514

Master of Science in Information Technology

2023

Sexual and Gender-Based Violence Reporting and Pattern Analysis Tool: A Case of Nairobi County

By

Makau Agnes Nzembi

147514

Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Information Technology at Strathmore University

School of Computing & Engineering Sciences

Strathmore University

Nairobi, Kenya

June, 2023

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Declaration and Approval

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

© No part of this thesis may be reproduced without the permission of the author and Strathmore University

Student's Name: Makau Agnes Nzembi

Sign: _________________________ Date: _________________________

Approval

The thesis of Makau Agnes Nzembi was reviewed and approved for examination by the following:

Dr. Allan Omondi,
Lecturer, School of Computing & Engineering Sciences,
Strathmore University

Dr. Julius Butime,
Dean, School of Computing & Engineering Sciences,
Strathmore University

Dr. Bernard Shibwabo,
Director of Graduate Studies,
Strathmore University

Abstract

Globally, sexual gender-based violence (SGBV) has affected individuals irrespective of age, gender, and social status. This harmful act includes sexual, mental, physical, and economic harm to both the private and economic sectors. SGBV has devastating consequences that can last a lifetime for survivors and even lead to death, hence the need for immediate medical attention. In the Kenyan context, SGBV continues to be a severe problem for society despite government efforts to address it through legislative and/or policy frameworks The inability of health, non-governmental organizations, and the government, as well as duty bearers, to combat SGBV has been impeded by a lack of up-to-date data. Governmental limitations in terms of planning to ensure that services are put in place to prevent SGBV and provide assistance to victims have attracted a rise in the number of affected victims. In this study, data was collected using questionnaires, and the collected data was used to find out the challenges in the reporting of SGBV cases and data management processes by the duty-bearers involved. The data was analysed using Microsoft Excel as a correlational tool, and the results were used to guide the researcher in the design of the mobile and web application. The study adopted an applied research design, and Agile methodology was used to develop the application. In order to understand the trends and patterns of SGBV, a KNCHR database with SGBV and feature-related information was first compiled and prepared to be used for ML modelling purposes, making use of the Poisson Regression Algorithm because it is best suited for identifying patterns and relationships between different factors in this case; gender, location, month, and year, and how these variables may affect the frequency of SGBV. Secondly, an android-based mobile application was developed to assist victims in reporting the violations in a timely manner, assist hospitals and police stations in data management by allowing them to update the details of the case with necessary information on the application, seek duty bearers' attention and assist in the mapping of the reported cases for purposes of visualizing the number and geo-location of these cases, ensuring transparency throughout the ecosystem. The patterns depicted from the ML model were integrated with the web application for duty bearers to view the patterns of previously reported violations as well as a reports module on the web application displaying analysis in form of charts of the reported cases through the mobile app including the gender, age, type of violation and location. These analyses will aid in planning ahead for better resource optimization and comprehension of the magnitude of SGBV in order to mitigate future incidents.

Table of Contents

VT OMNES VNVM SINT

VT OMNES VNVM SINT

VT OMNES VNVM SINT

List of Figures

VT OMNES VNVM SINT

List of Tables

List of Equations

List of Abbreviations

ANN	Artificial Neural Networks
ARIMA	Autoregressive Integrated Moving Average
ATM	Automated Teller Machine
BERT	Bidirectional Encoder Representation from Transformers
CNN	Convolutional Neural Networks
CNTK	Cognitive Toolkit
DPP	Director of Public Prosecutions
ERD	Entity Relationship Diagram
ESRC	Economic and Social Research Council
FSI	Forensic Science International
GAN	Generative Adversarial Network
GBV	Gender Based Violence
GBVRC	Gender Based Violence and Recovery Centres
GPS	Global Positioning System
HMM	Hidden Markov Models
IPV	Intimate Partner Violence
JVM	Java Virtual Machine
KNCHR	Kenya National Commission on Human Rights
KNN	K Nearest Neighbour
LDA	Latent Dirichlet Allocation
LRM	Local Repair Most
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
NGEC	National Gender and Equality Commission
NC	Nearest Centroid
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NMS	Nairobi Metropolitan Services
NPS	National Police Service
OOAD	Object Oriented Analysis and Design
ODPP	Office of the Director of Public Prosecutions
PEP	Post Exposure Prophylaxis
PHR	Physicians for Human Rights
PLWD	Persons Living with Disabilities
POS	Part of Speech
PRC	Post Rape Care

VT OMNES UNVM SINT

RAINN	Rape, Abuse, Incest National Network
RF	Random Forest
ROC	Receiver Operating Characteristics
RNN	Recurrent Neural Networks

Acknowledgements

I thank God for His strength and grace throughout this entire journey toward my master’s degree, because with Him, anything is possible.

Special thank you to my supervisor, Dr. Allan Omondi, for his advice and insights that helped shape the thesis.

I also appreciate the other lecturers who interacted with my work and shared their feedback, which helped enrich the study.

I would also like to express my gratitude to my encouraging and supportive family and friends. This doesn’t go unnoticed.

Finally, I would like to thank all of the researchers and writers who contributed to my research. Your efforts in the field do not go unrecognized.

VT OMNES VNVM SINT

Dedication

This study is in honour of my late mother, Mary Muthoki Makau, who passed away just as I was beginning my Master of Science in Information Technology program. Initially, I considered giving up on the degree altogether, but then I kept in mind that she always wanted the best for me and that I should always be outstanding, so I persisted.

Chapter 1: Introduction

1.1 Background to the Study

Sexual and gender-based violence (SGBV) is defined by the World Health Organization (WHO) as any coercive sexual act, effort to obtain a sexual act, or act against a person's sexuality committed by anyone, regardless of that person's relationship to the victim, in any situation. Instances include denial of access to resources or services, physical, psychological, or emotional harm, and sexual assault. Coercion and threats of violence are both forms of violence. SGBV is a grave violation of numerous human rights, causing harm to women, men, girls, and boys ("Sexual and Gender Based Violence", n.d).

The Sexual Offenses Act of Kenya makes rape, attempted rape, compelled or induced indecent activities, and defilement all illegal and prohibited. According to Kenya's national violence against children study, one out of every three girls and one out of every five boys has experienced at least one incident of sexual assault by the age of 18. According to the Abuse Against Children report, by the age of 18, 73% of boys and 66% of girls have experienced physical abuse. According to the most recent National Demographic and Health Survey, 47% of women aged 15 to 49 have witnessed physical or sexual violence. These findings paint a picture of Kenyans experiencing violence on a daily basis. Kenya National Commission on Human Rights (2017) found that during the 2017 electioneering period—which included political campaigns—25.17% of all human rights violations were sexual and gender-based violations recorded from victims' personal statements that included political party primaries, campaigns, the election, repeat elections, and post-election scenarios.

Nairobi Metropolitan Services (2021) reported 5,589 rape survivors, 52 of whom were infected with HIV-positive and 104 became pregnant. In 2021, the Health Chief Administrative Secretary stated that the country had recorded at least 5,000 rape cases since March 2021, with many involving girls under the age of 18.

According to the World Health Organization (WHO), nearly one-third of women will experience sexual or physical abuse from an intimate partner during their lifetime ("Violence against Women", 2021). Wado (2021), indicated that in Kenya, about 41% of women have reported experiencing sexual and physical violence from their marriage partners in their lifetime. He also indicated that there is a lack of up-to-date information on SGBV, but time-bound data usually indicates how often this type of violence is happening. Human Rights Watch, in their report "I Had Nowhere to Go," highlighted that during the Covid- 19 pandemic, there was an increased rise in cases of SGBV due to a less than adequate or late response by the Kenyan government. Sexual and other forms of violence against women and girls may have increased because Kenya's response to Covid-19 did not include services to prevent gender-based violence and aid survivors. The Kenyan government has let victims down by not providing them with adequate mental health care, medical care and protection services, financial aid, or a thorough investigation and prosecution of the crimes committed against them ("Kenya," 2021).

With considerable concern, it is noticed that policy texts have advanced the promise to dealing with SGBV instances more promptly than in actual practice. However, there are substantial gaps between real legislation, the experiences of sexual and gender-based violence victims resulting from the implementation process, and a great deal of operational misunderstanding at all levels.

1.2 Problem Statement

Sexual and gender-based violence continues to be a severe problem for the society despite government efforts to address it through legislative and/or policy framework (National Gender and Equality Commission, 2017). The inability of health, non-governmental organizations, governmental as well as duty bearers, to combat SGBV has been impeded by a lack of up-to-date data. Governmental limitation in terms of planning to ensure that services are put in place to prevent SGBV and provide assistance to victims has attracted a rise in the number of affected survivors. Lack of timely, quality and comprehensive mental health care, protection services, financial assistance, medical attention, proper investigation and prosecution of their cases from the duty bearers and related stakeholders to survivors have increased due to lack of timely data and integration of this information to the affected parties (Nnoko-Mewanu, 2021).

1.3 Objectives

1.3.1 General Objectives

The general objective of this study was to develop a mobile application for the timely reporting of SGBV violations, assist in the data management of the reported violations by integrating hospitals and police stations services, and use an ML algorithm to come up with an SGBV model that will aid in identifying patterns and relationships between different factors that may affect their frequency, which will help improve policies by the duty bearers.

1.3.2 Specific Objectives

i. To identify the challenges in the current methods of data collection for SGBV.
ii. To review the current techniques in place for reporting sexual and gender-based violence cases.
iii. To develop a mobile application for the timely reporting of sexual and gender-based violence cases.
iv. To develop a machine learning model for pattern analysis of sexual and gender-based violence cases.
v. To test and validate the proposed tool.

1.4 Research Questions

i. What are the challenges faced in reporting sexual and gender-based violence cases?
ii. What are the techniques used to report sexual and gender-based violence cases?
iii. How can the mobile application ensure the quality of evidence collected?
iv. How can we develop a model to perform pattern analysis of sexual and gender-based violence cases?
v. How can the functionality of the proposed tool be tested?

1.5 Justification

Survivors of SGBV face difficulties when it comes to receiving the services they need after a violation has occurred owing primary by lack of up-to-date data by the duty bearers, as well as lack of fundamental understanding or awareness of crime reactions, with evidence lost at this point due to tampering by the victim or perpetrator. According to international forensic best practices, the proposed model will ensure that SGBV violations are properly reported within the 72 hours recommended period and proper data management by the duty bearers which is critical in terms of ensuring that services are in place to provide the necessary assistance to the victims and aid in preventing more violations from occurring. The National Police Service (NPS) will benefit from this proposed solution as it will aid in data management of complaints received from survivors as well as have a database indicating cases that have been finalized and those that are pending. The data collected will help the Office of the Director of Public Prosecutions (ODPP) better understand the challenges of case prosecution and how to improve it. The proposed solution and data collected will also benefit the Ministry of Health by revealing trends and patterns in reported cases as well as improvements that can be made to assist SGBV survivors. This will also help to increase the number of convictions, giving victims better and timely access to justice. Furthermore, it will assist the government in allocating resources of all forms that are required for dealing with these situations before they occur.

1.6 Scope

The research focused on the development of a mobile app that victims will use to report sexual and gender-based violence cases, more so within the 72-hour period as per the international forensics' threshold to ensure evidence is preserved and maintains its validity. Proper data management is ensured as the proposed solution will aid in integrating the victim reports, updates by the law enforcement and medical officers in regards to the violation reported. All these leads up to one common goal which is the accountability process.

With the centralization of the data collected, it is easier for courts to prosecute the perpetrators, which will eventually lead to justice for the victims. The aspect of prosecution is outside the scope of the research, as it wasn't achievable within the timelines of this project.

1.7 Limitations

The anticipated limitation of the study was the time limitations that were there because of the many concurrent activities being carried out by the researcher, the study swas only limited to what could be efficiently achieved within the available time.

VT OMNES VNVM SINT

Chapter 2: Literature Review

2.1 Introduction

There has been a significant decline in the number of convictions for sexual and gender-based violence cases, primarily due to the absence of a proper data management mechanism ensuring the centralization of all data pertaining to sexual and gender-based violence cases and the lack of fundamental knowledge of crime scene reactions where evidence is lost due to tampering by either the victim or suspect (Forensic Science International, 2019). As a result, the duty bearers are unable to render necessary and timely assistance to the survivors.

2.2 Theoretical Framework

2.2.1 Information Theory

According to Martignon (2015), information theory provides methods for distinguishing between genuine information and noise. Shannon (1948) states that the fundamental challenge of communication is duplicating a message conveyed from one point to another, either accurately or approximately. The noisy-channel coding theorem establishes in information theory that no matter how contaminated with noise a communication channel is, it is possible to send information almost error-free. Communication noise is any factor that interferes with the clarity of messages sent and received. Physiological, psychological, cultural, semantic, and physical noise are all types of noise in communication (Kobiruzzaman, 2019).

Table 2. 1 Types of Communication Noise (Kobizurraman, 2019)

Physical Noise	Physiological Noise	Psychological Noise	Semantic Noise	Cultural Noise
Communication Disturbance created by Environmental Factors.	The physical condition of the communicator, which is caused by physical illness.	Communication obstacles caused by psychological aspects of the communicator.	Communication barrier created from confusion over the meaning of words.	Communication barrier from the wrong explanation of another person’s behaviours.
Raining Sounds, Thunderstorms, Outside Building Sounds.	Headaches, Deafness and Blindness, Talking Fast or Slow.	Beliefs, Attitudes, Behaviours, Sensitive Issues i.e., Religious, Ethnic and Politics.	Syntactical Barriers, Mispronunciations, Grammatical Errors or Wrong Sentences.	Wrong interpretation of messages conveyed from nonverbal communication.

The use of the Claude Shannon model, which was designed for effective communication between the sender and receiver, is illustrated below in Figure 2.1. This will ensure survivors' reports are captured, stored, and analysed accurately and appropriately.

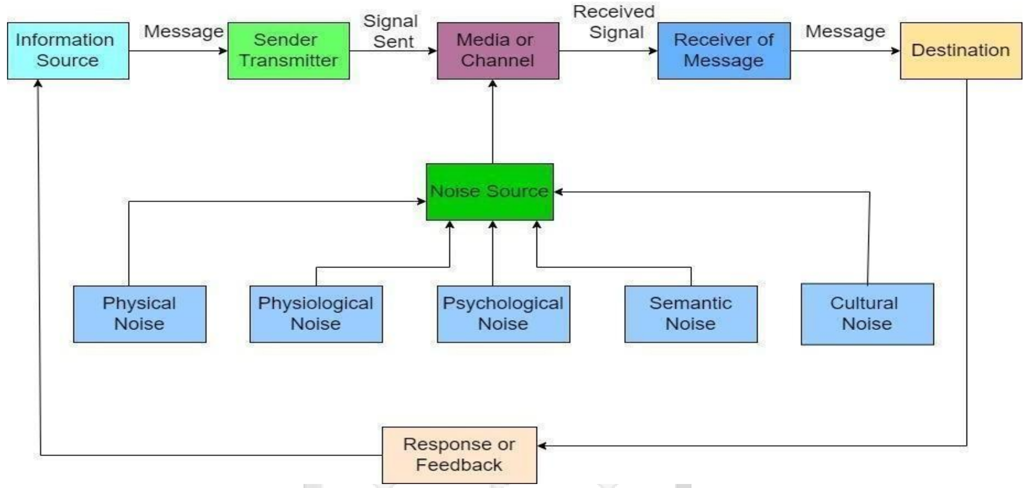

Figure 2. 1 Claude Elwood Shannon ’s Model of Communication (Drew, 2022)

2.2.2 Review of ML Models Instrumental in Crime Detection, Prediction and Analysis

Prieto et al. (2021) created a tool to help law enforcement officers choose the most appropriate recidivism prevention techniques. In order to develop models that reliably predicted the recidivism risk of perpetrators of gender-based violence, they employed Machine Learning methods like Decision Trees, Logistic Regression, K-Nearest Neighbors, and Nearest Centroid. The proposed ML method was put to the test on a database containing over 40,000 reports of gender-based violence in order to see if it performed better than the current risk assessment algorithm based on classical statistical techniques. To help authorities make the switch from the old model to the new one based on ML, they proposed a hybrid model that uses both ML and statistical prediction techniques. If the VioGen model is not accurate in practice, the police may have been less effective when it was used to issue fairly good risk predictions. NC was found to be the most effective ML model for solving the aforementioned risk prediction issues.

Machine-learning regression algorithms have been used to predict crime (Ajagbe et al., 2020). When it comes to handling the randomness of test samples, linear regression has been found to be more effective than decision trees. More importantly, the accuracy of machine learning in crime prediction is valuable to understand for its potential to reduce crime rates (McClendon & Meghanathan, 2015). When predicting criminal activity, random forest as a regressor produces superior results compared to other regression algorithms, such as linear regression (Kadar et al., 2016).

According to Dakalbab (2022), the majority of the methods for predicting crimes were developed for generic crimes. In the process of crime prediction, various models were used and tested to determine which one was the most effective in relation to the dataset that was provided. An investigation into the use of ML to forecast generic criminal behavior was carried out by Kim et al. (2019). On the other hand, specific methods have been developed to address particular types or categories of criminal activity. For instance, Srivastava et al. (2008) modeled the sequence of operations involved in credit card transactions by employing a hidden Markov model (HMM).

Bello et al (2020) state that the safety, dignity, health, and autonomy of those who are victims of gender-based violence (GBV) are all compromised, despite the taboo nature of the subject matter. Even though there have been a lot of studies done on the subject, nobody really knows how much of an impact the media has on this kind of violence. To extrapolate the effect that the news will have on GBV, the tools of machine learning are utilized here. It is possible to recover the topic information associated with each article by feeding neural networks with news. The study found a link between public awareness and gender-based violence (GBV) news, as well as the effect of media coverage of GBV cases and the intrinsic thematic connection between GBV news stories. Due to the adaptability of the neural model employed, we can apply our methodology to other media sources or topics.

VT OMNES VNVM SINT

According to Abdulkareem & Karan (2022), a prediction study was conducted in Los Angeles, California. The main objective of this research is to foresee future instances of intimate partner violence (IPV) by identifying the underlying causes and relevant contextual factors; this will allow local authorities to take preventative measures. An online survey administered to Los Angeles shelters for the homeless and youth centres provided the data for this analysis. Questions about one's social life, romantic connections, and other topics are included in the survey. More than a hundred different factors are represented in the columns of the data sets. To reduce the size of the dataset and make it more manageable during training, feature shrinking is performed by using ML methods such as P-value and Lasso to figure out which features were necessary for the final dataset. As a result of its superior ability to foresee anomalies in data patterns, ANN has been incorporated into the final prediction process.

2.3 Empirical Framework

Bello et al. (2020) created a machine-learning tool that estimates GBV rates from news articles in Spain using natural language processing (NLP). Web scraping was used during the research to extract data from the region's websites. 784,259 data points were collected between 2005 and 2020. Each article used was labeled with a set of tags, making topic classification easier. Information was extracted from the data source to assist in determining the relationship between the subject of the article and its relationship to GBV. The stack protocol was used by the researchers, which consists of two neural network models: first, they determined the general subject classification by analyzing the content of each text, and then they applied a binary gender-based violence classification, also known as the GBV probability, which is a rate probability where zero indicates the absence of GBV.

Mishra & Kulkarni (2022) conducted a study to demonstrate the use of ML to attempt to accurately predict types of violence such as physical or sexual, mental, and verbal experienced by adolescent girls in public spaces, homes, and the educational institutions, ultimately concluding that the Winnow algorithm had a sufficient accuracy rate (97%) to assist institutions in identifying victims of violence and providing assistance. Another research group developed a scalable, supervised-learning-based, automated sexual violence report tracking model with 80.4% accuracy and 83.3% recall for identifying reports of sexual violence (Hassan et al., 2020).

Petering et al. (2018) developed an IPV perpetration triage tool that could be designed and applied in the field to identify young individuals at high risk of perpetrating violence using several supervised ML algorithms. They investigated supervised machine learning tools such as SVMs, LRMs, CNNs, RFs, and deep SVMs. These models were trained and tested with K-fold (5-fold) cross-validation, and their performance was measured with metrics like precision, recall, F1-score, ROC curve, and accuracy. When compared to other classifiers, the F1 score of 0.58 achieved by the SVM classifier (using an RBF kernel) was the highest possible.

Rodriguez et al. (2020) used ML techniques to model and predict the occurrence of GBV over a six-month period. They accomplished this by analysing 30 features from a Spanish national database to determine which had the greatest impact on the occurrence of GBV. GBV forecasting is studied and compared using four prediction algorithms (Linear Regression, Gaussian Process, Random Forest, and Support Vector Machines) so that governments can improve policy planning on this issue and thus optimize and maximize strategies. SVM and LR outperformed in terms of short-term predictions, but their errors increased over time. The results of the tests showed that using Random Forest as the predictive algorithm, it can estimate the number of GBV complaints filed in court over a six-month period in Spain with an accuracy (Root Median Squared Error) of 0.1686 complaints per 10,000 people.

Crowdsourcing apps can provide valuable insight into SGBV patterns and mechanisms, as well as its spread in a specific region. Stop it! Report it! The app was created to allow Kenyan users, either victims or witnesses, to pin incidents of violence on an interactive map and, if desired, add a description of the incident. The information was then compiled into a comprehensive database of GBV hotspots in Kenya, which was not only made public but also shared with government officials, public transportation operators, and civil society organizations in Kenya (Dickins & Mwaura, 2020).

According to Roback and Legler (2021), Poisson Regression includes predicting a response using one or more explanatory variables, which may include counted response variables. It is usually used to calculate counts. D, Osgood (2020) advocated using Poisson Regression models to analyse aggregate crime rates. They are best suited for modeling offense counts since they are based on error distribution assumptions that are consistent with the nature of event counts.

Flowers, R, et al. (1980) indicated that Weber (1970) evaluated the validity of the assumption that accidents follow a Poisson distribution. Weber categorized 148,000 California Driver Record Study participants into 2880 groups based on gender, marital status, age, place of residence, conviction history, and accident history. Accident distributions were discovered for the 193 groups with at least 100 participants. Weber discovered that the Poisson distribution hypothesis was satisfactory at the 0.05 level for 86.5% of the groups. The purpose of the analyses is to identify the variables that influence the probability of a future accident, which are incorporated into the Poisson regression model.

2.4 Architectures

According to Twin (2022), "data mining" is the practice of sifting through massive datasets in search of patterns and relationships. In Figure 2.2, we observe the stages of a data mining process being implemented.

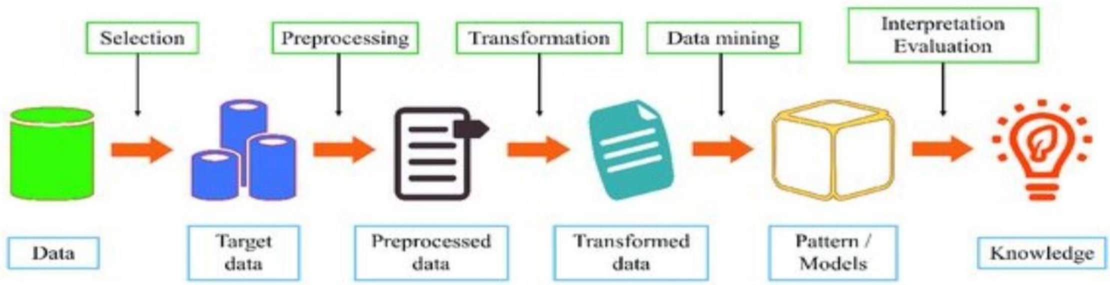

Figure 2. 2 The steps for data mining process (Yang et al., 2020)

2.4.1 Convolutional Neural Networks (CNNs)

According to Zamri et al., (2021), a classification approach or architecture for crime prediction was through the convolutional neural network (CNN), which accepts images as input data and assigns weights and biases. CNN works best for image prediction, classification, and recognition.

Figure 2.3 depicts CNN's basic structure, which includes a pooling layer, a convolutional layer, and a fully connected layer.

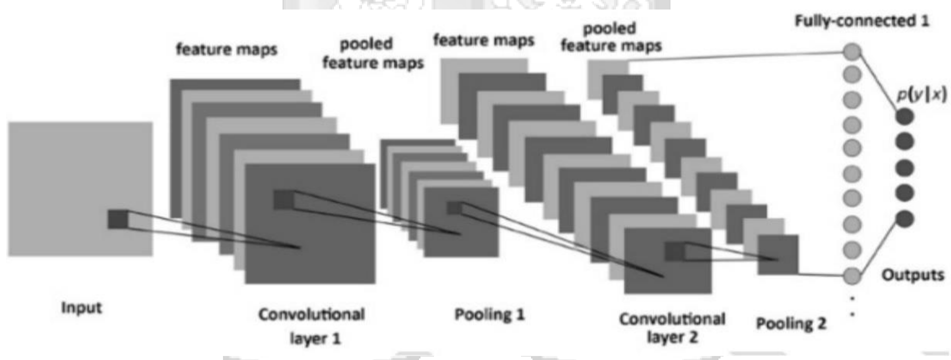

Figure 2. 3 Convolutional Neural Network Architecture (Zamri et al., 2021)

Xia et al. (2016) used a multitask CNN that's pre-trained to recognize the front as well as both sides of the face when detecting face obstruction in ATM proximity. In this study, new face datasets for criminal identification around ATMs were made. These datasets included images of people wearing hats, sunglasses, or face masks. Rasanayagam et al. (2018), using a feature-based method, on the other hand, identified and extracted feature characteristics from input images, such as the mouth, nose, and eyes. Furthermore, the Caffe model framework was combined with CNN to improve classification accuracy. The developed CNN can also detect the presence of blood, a knife, and different types of guns.

Dinama et al. (2019) used ReLu as an activation layer because the convolutional layer consists of a linear filter and a non-activation layer capable of detecting the presence of humans. This is also used to predict the potential running direction of a person after committing a crime, such as pick-pocketing. CNN is prone to overfitting because the network may be overly vulnerable during data training. This could result in low recognition accuracy during the testing phase, despite the fact that the training data achieved high accuracy. To overcome this limitation, Kim et al. (2018) used a dropout method to randomly disconnect the connection by inserting a 50% dropout probability between the first and second fully connected layers.

2.4.2 Recurrent Neural Networks (RNNs)

RNN is a type of artificial neural network which uses sequential data or time series data. Figure 2.4 depicts a recurrent neural network (RNN). According to Zamri et al. (2021), CNN is one of the best image classifiers, but not for non-sequential data or the temporal domain. Han et al. (2019) therefore utilized RNN to analyse the temporal domain in the middle of sequential data. In this instance, training and testing data were processed using a fully connected CNN layer, followed by RNN as video feature extraction in order to generate a new training set for RNN in order to improve the accuracy of the CNN when working in conjunction with the RNN.

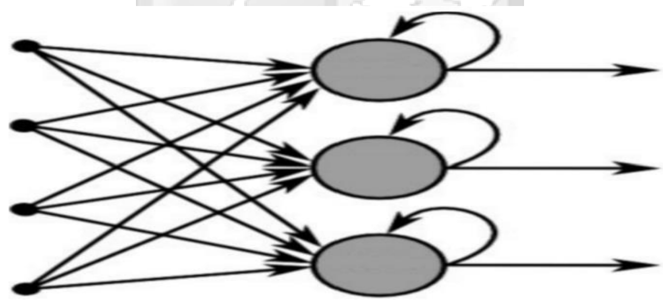

Figure 2. 4 Recurrent Neural Network Architecture (Zamri et al., 2021)

Another type of artificial RNN architecture is Long Short-Term Memory (LSTM), which analyses the entire data sequence, including statistical data, images, and videos, as well as classifying, processing, and predicting based on time series. There is also a feedback connection.

Choi et al. (2019) proposed a two-stack LSTM structure for encrypting each passer-by motion information as a character vector and predicting the direction chosen by the passer-by. This can be used to monitor a criminal's movements in a crowd and foresee his or her next steps after a crime has been committed. By adding more layers of LSTM, each with its own set of memory cells, the model can now make more precise predictions (Zamri et al. 2021).

The introduction of a conditional gate was necessary because LSTM uses a memory block connected via layers. This LSTM architecture was used to predict crimes such as robbery, burglary, or violence based on data sorted by timestamp. The LSTM was also used to predict how many cases would occur in the next day or even the next month (Zamri et al. 2021).

2.4.3 Generative Adversarial Networks with Bidirectional Encoder Representations from Transformers

BERT is a transfer learning technique in which a model is initially pre-trained on general tasks and then refined on the final target tasks. BERT is a model that is pre-trained on a corpus of raw texts and then fine-tuned on annotated target data. BERT's foundation is the Transformer, which is an attention-based mechanism that learns contextual relationships between words, sub-words, and word fragments in a text.

Croce et al., (2020) define the first step as performing some pre-processing steps on tweets to remove redundant content that does not contribute significantly to model performance. Following that, normalize some content to aid the model's learning. During the training process, GAN-BERT architecture is used with unlabeled data in a generative adversarial setting. GAN-BERT is appropriate for small datasets when we are unable to collect more data due to high costs and time constraints. GAN-architecture BERT's is depicted in Figure 2.5.

SS-GAN (Semi-Supervised Generative Adversarial Networks) are composed of two networks: a discriminator D for classifying input datasets and a generator G for distinguishing fake data in an adversarial manner. Both the generator and the discriminator in GAN-BERT are multi-Layer Perceptron networks (MLP). The real data, which includes two categories (labeled and unlabeled), will be passed to BERT in order to receive text embeddings.

For instance, the training set is labeled data because its tweets have already been labeled 1 for violence, and 0 for no violence). The test set and validation are unlabeled data because their data is unknown. In the study by Croce et al. (2020), because there are only two categories (0 and 1), we label violent tweets as labeled data; otherwise, data (including non-violent tweets) is unlabeled. Meanwhile, based on real text embeddings, noises or fake tweets will be generated (vectors).

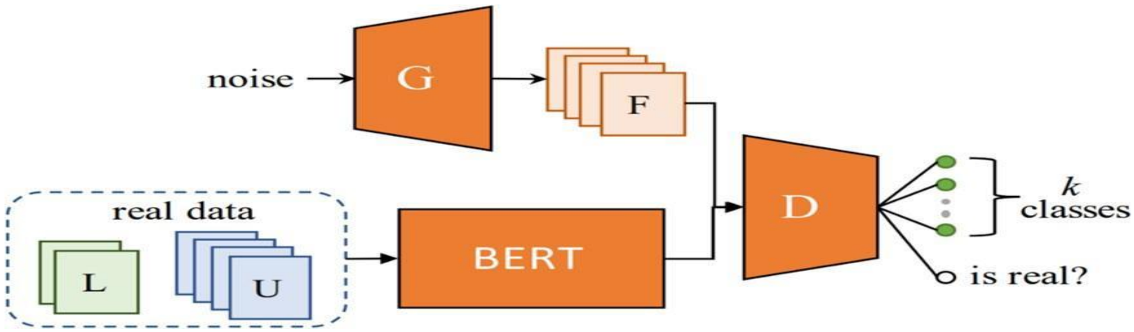

Figure 2. 5 Generative Adversarial Networks with Bidirectional Encoder Representations from Transformers (GAN-BERT) Architecture (Croce et al., 2020)

Based on a random distribution, G will generate a set of fake sentence pairs F. The discriminator D will use the fake pairs, unlabeled U, and labeled L vector representations computed by BERT as input for the training process.

2.5 Algorithms

Lynch (2022) describes an algorithm as a series of well-defined steps or rules that must be followed in order to achieve a specific result. The defined steps to follow on the input to get the intended result would thus be an algorithm.

2.5.1 Random Forest (RF) Algorithm

Pyzhov et al. (2017) define Random Forest as a classification method that builds a set of decision trees to define the class that is the mode of the classes produced by individual trees. The basic idea is that each tree constructs nodes that represent a test of specific attributes of objects, which are then split into branches based on the presence of this attribute. This method is one of the most accurate algorithms, and it works well with both large datasets and a large number of variables. However, it overfits training datasets, the results are difficult to interpret by humans, and its results cannot be implemented in other contexts. Berk et al. (2016) determined that 20% of people who are released following an arraignment for domestic violence are arrested for a new domestic violence offense within two years using Random Forest algorithms. Figure 2.6 depicts how a random forest determined its final prediction based on the votes of each individual tree.

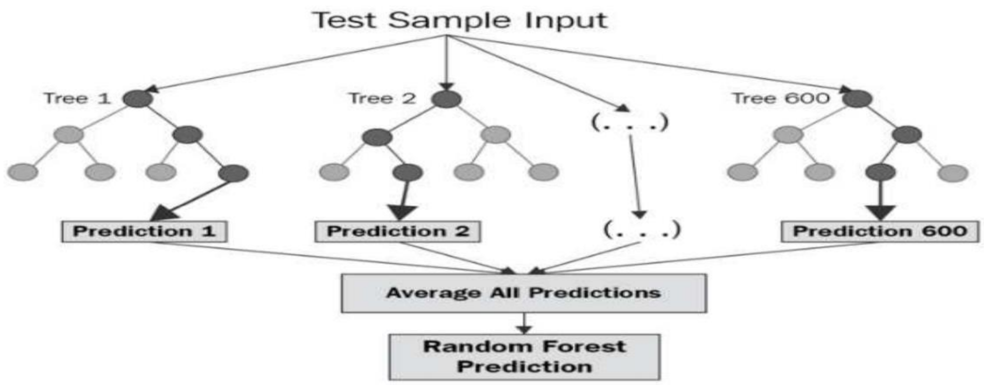

Figure 2. 6 Random Forest Algorithm (Asor et al., 2022)

2.5.2 Support Vector Machines (SVM) Algorithm

SVM is a supervised machine learning algorithm that can perform classification as well as regression. These dual learning algorithms perform data processing solely by calculating dot-products between variable arrays, an operation that can be efficiently performed by a kernel function. SVM learning seeks a hyperplane that divides the examples given this function (margin). Due to the max-margin criterion is used in the optimization process, SVMs are known to be resistant to overfitting and maintain good generalization performance. Given their convex optimization formulation, SVMs, unlike other solutions, are guaranteed to converge to a global optimum (Dhiraj, 2017).

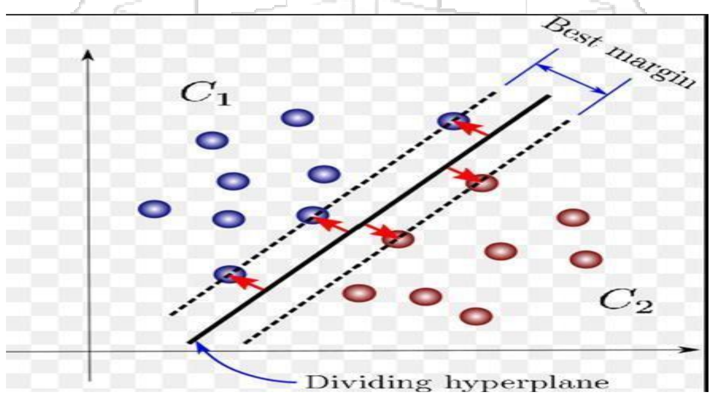

Figure 2. 7 Support Vector Machine Algorithm (Dhiraj, 2017)

2.5.3 Latent Dirichlet Allocation (LDA) Algorithm

According to Manzoor et al. (2021), topic modeling is utilized to deduce hidden meanings and semantic structures from massive amounts of text. It is understood that the document may contain keywords such as "violence," "gender," and topics related to gender violence. The model examines the collection of text documents by examining the statistics of the words contained in each one in order to determine the hidden topic of each text document and the balance of multiple topics within it. There are many well-implemented versions of this algorithm that are known to be faster and provide better topic segregation. LDA is an unsupervised machine learning model that stores words in a bag for each text. As depicted in Figure 2.8, the LDA hypothesizes that each text document is a set of topics containing a set of words.

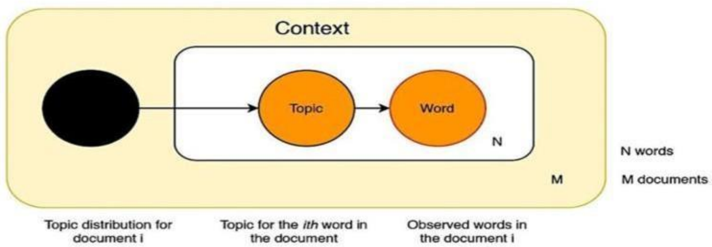

Figure 2. 8 Basic workflow of LDA topic modeling (Manzoor et al., 2021)

A series of experiments are carried out in order to analyse GBV content in the newspapers selected. Initially, LDA-based topic modeling is used to analyse media-reported topics using a dataset of major newspapers. By extending existing keywords, we create a gender- based violence keywords dictionary (Purohit et al. 2015). The trained ML model is then integrated into a web portal to analyse GBV-related web content. To conclude the case studies, a comprehensive analysis of this dataset was performed using a variety of visualization techniques to generate various reports (Manzoor et al. 2021).

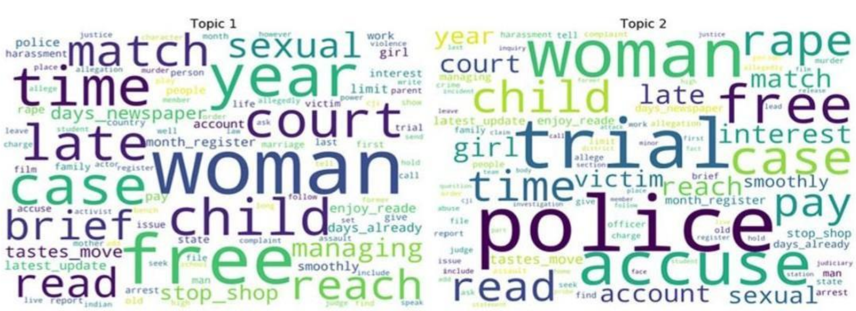

Figure 2. 9 Dominant GBV topics from LDA analysis (Manzoor et al., 2021)

2.5.4 Poisson Regression Model Algorithm

According to Roback and Legler (2021), Poisson Regression includes predicting a response using one or more explanatory variables, which may include counted response variables. It is usually used to calculate counts. D. Osgood (2020) advocated using Poisson Regression models to analyse aggregate crime rates. They are best suited for modeling offense counts since they are based on error distribution assumptions that are consistent with the nature of event counts.

Flowers, R, et al. (1980) indicated that Weber (1970) evaluated the validity of the assumption that accidents follow a Poisson distribution. Weber categorized 148,000 California Driver Record Study participants into 2880 groups based on gender, marital status, age, place of residence, conviction history, and accident history. Accident distributions were discovered for the 193 groups with at least 100 participants. Weber discovered that the Poisson distribution hypothesis was satisfactory at the 0.05 level for 86.5% of the groups. The purpose of the analyses is to identify the variables that influence the probability of a future accident, which are incorporated into the Poisson regression model.

2.6 Frameworks

Machine learning frameworks are tools or libraries that enable developers to easily build ML models or applications without having to understand the underlying algorithms. It offers a more complete machine learning development pipeline (Hari, 2022). Some of these frameworks include; TensorFlow, Keras, PyTorch, Theano, Deeplearning4j, Scikit-learn, and Caffe.

2.6.1 TensorFlow

Venkatesh (2021) defines tensors as data structures used by machine learning systems. It is a container for numerical data and the information can be stored which will be used within the system. TensorFlow's robust and adaptable ecosystem of tools, libraries, and community resources enables researchers and developers to rapidly build and deploy machine learning (ML)-powered applications. The system is sufficiently versatile to be applicable in a variety of other domains. TensorFlow offers stable C++ and Python APIs, and backward compatibility with other languages (Jsmisa, 2022).

TensorFlow comes with a plethora of tools. For example, neural network activation functions can do all of the hard work of statistics. TensorFlow is capable of performing neural networks, which

VT OMNES VNVM SINT

is what deep learning is defined as. It can also handle more ordinary issues such as regression (Rowe & Johnson, 2020).

2.6.2 Keras

Keras is a high-level deep learning API that can be used to build neural networks on top of existing machine learning frameworks like Microsoft Cognitive Toolkit (CNTK), Google TensorFlow, and others. The end goal is to have a universal API that facilitates communication between them all (Rowe, 2019). Keras is relatively simple to learn as it provides a high-level python frontend with the option of multiple backends for neural network computation. Keras is also deeply integrated with TensorFlow, allowing you to easily create customized workflows (Simplilearn, 2022).

2.6.3 Deeplearning4j

According to Goyal (2022), deeplearning4J (DL4J) is a Deep Learning library that runs on the Java Virtual Machine (Java Virtual Machine). Therefore, it works with any language that runs on JVM including Scala, Clojure, and Kotlin. DL4J uses C, C++, Java, Scala, and Cuda for its underlying computations. Both Apache Spark and Hadoop are utilized by the platform, which expedites the training of models and enables the incorporation of AI into enterprise environments for use on dispersed CPUs and GPUs. There are numerous neural network types that it can support, including LSTM, CNN, and RNN.

2.6.4 Natural Language Toolkit (NLTK)

Python developers have had a lot of success creating applications that interact with human language data using the NLTK framework. It provides users with easy access to more than fifty corpora and lexical resources, including WordNet, as well as a set of text processing libraries for classification, parsing, tokenization, tagging, stemming, and semantic reasoning (Gudivada & Arbabifard, 2018). NLTK is widely used in education and research because it has resulted in numerous breakthroughs in text analysis. It has a large number of pre-trained models and corpora, which makes it very easy to analyze things. It is an excellent library when a specific combination of algorithms is required. The learning curve is steep and, most of the time, slow, and frequently does not match the demands of real-world production usage (Bak, 2019).

2.7 Current tools used for reporting SGBV

2.7.1 Use of Autoregressive Integrated Moving Average (ARIMA) Model

SGBV is a global public health issue that disproportionately affects victims. Sexual and gender-based violence has both immediate and long-term effects on health. Using autoregressive integrated moving average (ARIMA) models, the goal is to predict the number of SGBV cases reported to hospitals in Lusaka. The ARIMA model can assist hospital administrators and physicians in dealing with victims of sexual and gender-based violence (Lubeya et al., 2020).

2.7.2 MediCapt App

The vast majority of sexual violence crimes are never prosecuted mainly due to a lack of evidence. Doctors, nurses, and other professionals who assess survivors frequently acquire insufficient information that cannot be used as evidence in court for the following reasons: medical clinics frequently run out of required forms, and police frequently fail to obtain the evidence required to support investigations from clinics. Most health and police facilities retain private medical files on desks or floors due to lack of secure storage, putting them at danger of damage, manipulation, loss, or theft. To address these concerns, Physicians for Human Rights (PHR) developed the "MediCapt" smartphone app that allows healthcare practitioners to utilize it to gather medical evidence, photograph survivors' injuries, and securely communicate the data to lawyers, police, and judges involved in sexual violence prosecutions (MediCapt, 2019).

2.7.3 SV Case Study App

An application that aims to create a database of information that will be analyzed to inform advocacy, enhance, coordinate and allow policy formulation that will allow for intervention. The dignity of the survivors is key and to have it restored requires that we increase awareness, remove barriers to reporting and at the same time ensure access to justice while providing comprehensive care and support to the victims and or survivors. In essence, the major purpose of this application is to help reduce the frequency of sexual assault in Kenya and improve efforts to provide critical services to survivors of sexual violence through effective and ethical data collecting and analysis.

2.7.4 Use of POLICARE in Kenya

The POLICARE program was developed by the Kenyan National Police Service as an integrated response to sexual and gender-based violence (SGBV). Its goal is to serve as a "ONE STOP CENTER" for victims, providing accessibility to a numerous service under one roof, including those of the police, forensics investigators, health care providers, psychologists, the attorney's office, the magistrate on call, the medical-legal system, the gender experts, and the corrections staff. The general goal of POLICARE is to increase the capacity of the NPS to prevent and respond to SGBV situations. The current SGBV protection environment in Kenya is both challenging and promising, which should be carefully examined while developing countermeasures.

2.8 Research Gaps

As the studies in this chapter show, SGBV is a threat to everyone around the world, no matter their gender, age, social or economic status, or where they live. Several types of research have been done to try to find a solution. However, policies and governance have moved forward more than the way things are actually done. The legal and policy frameworks, as well as the identified measures and procedures, are still not being put into place in a consistent or strong way. This is because of factors such as a lack of human and financial resources at the duty bearer level, such as service provider skill levels, gaps in systems, and tools that make it hard to provide and get effective SGBV prevention and response services (World Bank, 2017). Ajema et al. (2011) say that the lack of clear rules about the roles of survivors, communities, healthcare workers, and police officers in handling forensic evidence is another reason why sex offenders are charged late or not at all. By coming forward and reporting, survivors of SGBV can get the medical, mental health, and legal help they need to lessen the effects of the violence on their health, and the people who hurt them can be held responsible. Also, when SGBV is reported in a formal way to medical staff, lawyers, or community leaders, it is possible to get a good idea of how common the violence is. This enables the right resources to be put toward efforts to stop SGBV and give survivors the care they need. Technologies to combat SGBV are existent, but they lack the ability to combine up-to-date data collected from the victims with an SGBV model that will aid in identifying patterns and relationships between different factors or variables that may affect their frequency that will help understand the magnitude of the incidents in order to plan ahead for better optimization of resources needed by the duty bearers to better assist the victims.

2.9 Conceptual framework

The conceptual framework is a rationalization of the problem at hand and a proposed way in which the researcher intends to solve it. The conceptual framework for the research is depicted in Figure 2.10. A user downloads the application. The mobile app will be used as a channel for reporting violations. The obtained datasets will be stored in a database, and they will be modeled using a machine learning algorithm to study patterns of previously reported cases based on specific predictor variables such as gender, location, months, and years. These patterns will aid duty-bearers in the formulation of policies to mitigate the prevalence of SGBV. After a violation is reported, feedback is sent to the victim to guide them on the actions they need to take next. An analytics interface will be used by the duty-bearers to view a summary of statistics that will be beneficial to them and the victims.

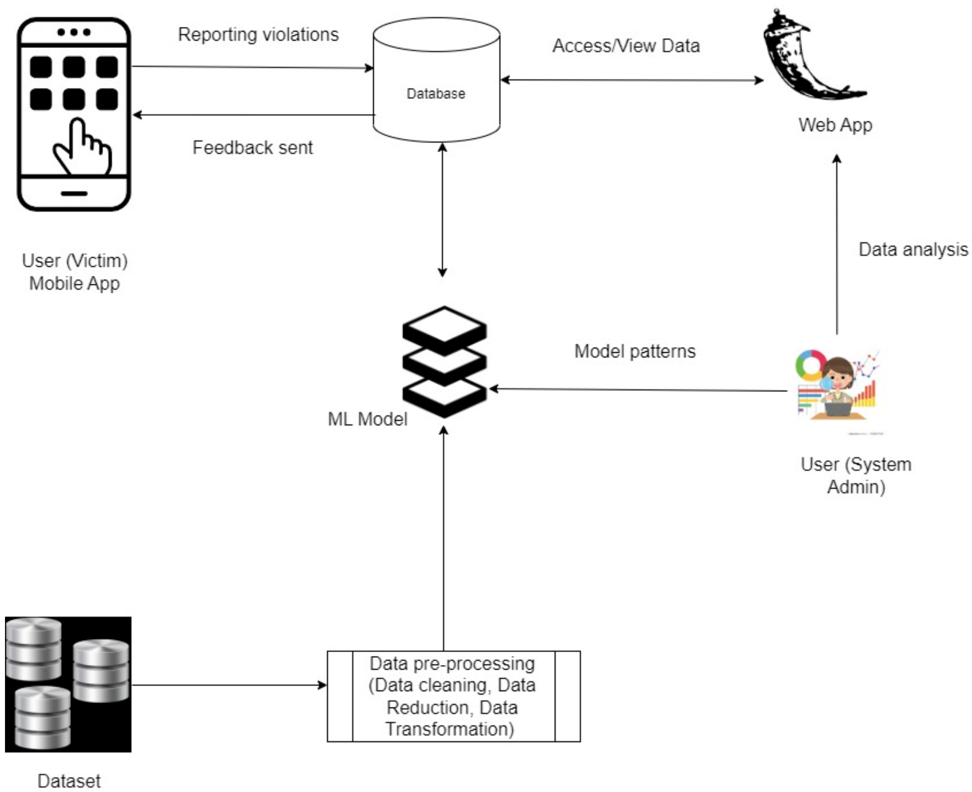

Figure 2. 10 Conceptual Framework of the proposed solution

Chapter 3: Research Methodology

3.1 Introduction

A research methodology provides credibility to the study and generates scientifically sound outcomes. It provides a detailed plan for keeping the researcher on track and making the process as easy, effective, and manageable as possible. The methodology of a researcher allows the reader to understand the approach and procedures utilized to attain the conclusions.

3.2 Research Design and Philosophy

The goal of this study was to design, develop, and test a prototype that can be used to ensure timely reporting of sexual and gender-based violence cases, a recommender engine to create awareness to the survivors on what evidence to preserve and what needs to be done, as well as a ML model that will be used to tailor analysis that will be shared with duty bearers. The selection of a research method is a crucial step in the design of a study because it determines how relevant data will be collected (Jilcha, 2019).

A research philosophy is a set of rules for conducting studies that are based on presumptions about reality and the importance of the data they yield (Collis &Hussey, 2014). The research made use of Pragmatism research philosophy because the research combined elements of qualitative and quantitative research approaches. This takes advantage of the benefits of both approaches while accounting for their respective limitations. This study used an applied research approach in which the proposed model solved the highlighted problem.

3.3 Population and Sampling

3.3.1 Target Population and Sampling

A population is the total number of persons who share a characteristic (Momoh, 2022). Nairobi County was considered as the target population for this study. It is estimated that there are 5,119,000 people who reside in Nairobi County. A sample represents the complete population. To eliminate bias in the questionnaires, random sampling procedures were utilized. This entails selecting a subset of participants to examine from a larger group known as the population. To represent the larger group accurately this research used the Yamane’s formula illustrated in equation 3.1 to get the sample size.

$$n = \frac{N}{1 + N(e^2)}$$

Equation 3. 1 Yamane's Formula

Where: n = Sample size, e = margin of error, N = Population of Nairobi

The level of confidence that was used is 95%, the N = 5,119,000, and margin of error is 5% thus the sample size is 40 participants.

The sample size of 40 participants was further broken down to include the four target groups i.e., the victims, medical officers, police officers, and community-based organizations. The stratified random sampling method was used, the random sample will made use of the stratums listed above and equation 3.2 to come up with a representation of the population.

$$sample\ size\ of\ the\ strata = \frac{size\ of\ entire\ sample}{population\ size} * layer\ size$$

Equation 3. 2 Stratified sampling formula

Table 2. 2 Representation of stratum

Category	**% Of participants**	**Number of people in sample**
Victims	35%	40/5,119,000*(0.35*5,119,000) =14
Medical Officers	25%	40/5,119,000*(0.25*5,119,000) =10
Police Officers	25%	40/5,119,000*(0.25*5,119,000) =10
Community based organizations	15%	40/5,119,000*(0.15*5,119,000) =6

The importance of collecting data from the target participants was to gather more information on whether the victims report violations and if they seek medical assistance from the victim's perspective as well as from medical officers and police officers. In many instances the victims first contact is usually the community-based organizations therefore it was crucial to gather information from them.

3.4 Data Collection Methods

Several data collection instruments were used in the study. Several instruments were utilized to collect data for the investigation. Instruments for data collecting are essential to the research procedure because they provide the analytical foundation for the search for answers to a particular research problem (Moyo, 2017). The data collection instruments should be comprehensive in scope, reliable and credible as they inform the data collected and analysis process of the data in order to inform the necessary duty bearers or stakeholders. The following data collection tools were used in this study:

3.4.1 Questionnaires

The questionnaire was aligned with the study questions to ensure that it sought meaningful information that improved the research's conclusion. Participants from several target groups, including medical officers, police officers, survivors, and community-based organizations, were given questionnaires to complete. The questionnaires allowed for an understanding of SGBV cases through the subjective experiences and perspectives of the respondents, and these helped answer the research questions. The researcher wanted to understand the victim's ages, gender, types of SGBV violations, frequency of SGBV violations at police stations and hospitals, responses of victims in seeking assistance, challenges in reporting SGBV violations, techniques used to report SGBV violations, and data management and storage of SGBV information. The results obtained were used to guide the researcher in the design of the mobile and web application.

3.4.2 Prototyping

Prototyping entails disseminating the developed application modules to the general public and facilitating brainstorming sessions in order to obtain critical feedback, particularly concerning the satisfaction and inclusivity of desired functionalities, as well as generating new ideas. This is useful in ensuring that the victims' challenges and needs are taken into account in the proposed solution.

3.4.3 Document Review

As part of this research, the following documents were reviewed: Searching the library and reviewing information from numerous sources to have a better knowledge of the difficulties in reporting incidences of sexual and gender-based violence. Documents and literature on SGBV awareness and reporting, as well as relevant studies, frameworks, and models were reviewed. This information provided insight into the study that led to a better solution.

3.5 Data Analysis Methods

The study's goal was to ensure the collection of accurate data. Based on the data collected, proper analysis was done using correlational analysis tools such as Microsoft Excel to explain the relationships between sets of data. The data was presented using graphs and charts that were used to diagrammatically represent the findings of the data.

3.6 Research Quality and Reliability

Reliability refers to the consistency and stability of the used measuring tool over time. Changes in the population and sample, as well as variations in the measurement device, make it unlikely that the same results will be achieved repeatedly. The accuracy of the measurement technique determines the validity of the study's conclusions. As a result, it was critical to guarantee that the measurement tools utilized were accurate (Surucu & Maslakci, 2020).

The study's quality is ensured by its reliability and validity. The proposed model guarantees that the qualitative and quantitative findings of the data collected was used to assess how well the study met the research objectives and collection of accurate data. This was made possible by utilizing proper measurement methods and employing the proper methodology, as described above.

3.7 System Development Methodology

The agile software development methodology is centred on a short iterative software release cycle (Edeki, 2015). Continuous planning, learning, and improvement are used in the agile development methodology approach. This strategy was used for this study because it can combine iteration to improve ongoing feedback during the development process, which helps to refine and successfully accomplish the study objectives. The Agile technique helps with the creation of the suggested solution because it promotes feedback from all parties concerned, ensuring the solution's accuracy.

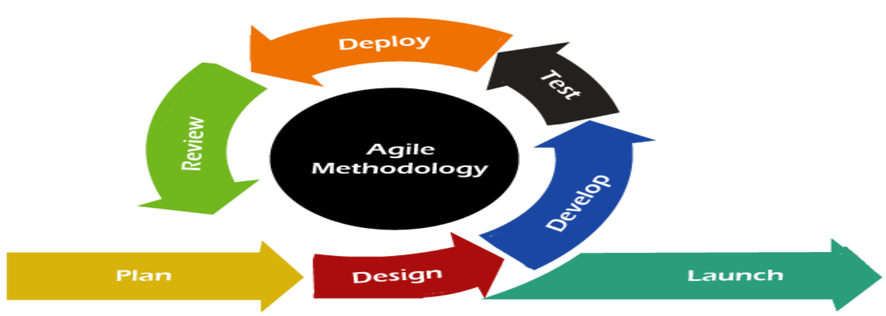

Figure 3. 1Agile Software Development Methodology (javaTpoint)

3.7.1 Planning Phase

Outlining the objectives and establishing what is needed for the proposed solution to be developed was conducted at this phase. These requirements were categorized as user requirements, operational requirements and system requirements which were documented in a requirements document. These are what was used to design the solution and to validate the correctness of the design. The information was gathered through questionnaires distributed to the target participants.

3.7.2 Design Phase

The study used Object Oriented Analysis and Design (OOAD). This ensured that the requirements are properly captured. The system was designed using Unified Modeling Language (UML) diagrams because they provide an understandable model that reduces the system's complexity these include:

i. Entity Relationship Diagrams (ERD) was used in the study to depict entities, attributes, and relationships during database design.
ii. Use-case diagrams were used to represent the interaction of the actors and system processes.
iii. In the study, class diagrams were used to build and visualize object-oriented systems.
iv. Object interactions were depicted in time sequence using sequence and collaboration diagrams.
v. System Sequence Diagram (SSD) helps track how the functions and the use case functions are performed inside the system. Also, to model the software in concern with how the system interacts with the events.

These diagrams were required to demonstrate the process, information flow, and relationships among the entities in the proposed solution.

3.7.3 Development Phase

During this phase, an application prototype was developed i.e the system was developed into modules that were then integrated in the subsequent phases. Each module was subjected to testing of the functionality to find out whether the objectives and requirements were met.

3.7.4 Testing Phase

Since the proposed solution is based on agile methodology, adopting continuous functional testing was employed. The testing that was carried out include: functional and compatibility testing. The different components or units were tested incrementally as each component was developed to test its functionality.

3.7.5 Deployment Phase

This phase involves the deployment of the solution into the production environment for use. After the deployment, it was important to conduct a system review in order to find out whether there are any discrepancies that will arise that may require fixes or upgrades which are then addressed.

3.7.6 Review Phase

Since the proposed solution is based on agile methodology, adopting a continuous reviewing process was employed. Any changes based on the system review ensured that the solution is per the requirements and objectives.

3.8 Utilization and Dissemination of Research Results

The study's findings will be shared with stakeholders or decision-makers to help improve the handling of SGBV cases. The findings will be shared with the following individuals:

i. National Police Service (NPS) to empower police officers at stations to ensure proper data documentation and management of complaints received from survivors;
ii. Office of the Director of Public Prosecutions (ODPP) to use the data collected to allocate adequate personnel with specialized competencies in sexual violence prosecution to ensure perpetrator conviction.
iii. The Ministry of Health can use the data gathered to better understand the trends and patterns associated with SGBV, as well as the reforms that can be implemented to ensure better handling of the situation.

The findings are also useful to future researchers interested in improving SGBV case management. Publications and academic articles will be used to disseminate the findings.

3.9 Ethical Considerations\Issues

During the data collection procedure, ethical challenges such as protecting data confidentiality in terms of privacy, permission, privacy, discrimination, and stigmatization of SGBV survivors may occur. Data collected was used exclusively for this study, guaranteeing its confidentiality, anonymity and privacy. All participants in the questionnaires had provided their consent to participate voluntarily. To ensure that all ethical considerations are addressed during the study, the researcher:

i. Created a trust agreement. It was used between the researcher and the survivors. The two parties gave informed and explicit consent to the study's requirements.
ii. Guaranteed that participants are taking part voluntarily and are fully aware of the benefits and risks, adhere to the informed consent rules.
iii. Uphold individual rights by respecting individuals’ confidentiality and privacy.
iv. Used the third ethical principle. According to the Economic and Social Research Council (ESRC), "the confidentiality of the information supplied by research subjects, as well as the anonymity of respondents, must be respected" (Smill, 2003). Anonymity was encouraged in cases where confidentiality will be limited.

Chapter 4: System Analysis, Design and Architecture

4.1 Introduction

This chapter talks about the system's analysis, design, and architecture, as shown in Figure 2.11. The system made use of Object-Oriented Analysis and Design and Unified Modeling Language (UML) diagrams because they provide an understandable model that reduces the system's complexity.

4.1.1 Age Group

The researcher wanted to know which age groups experience SGBV victimization and its impact more than others. The researcher presented a query to law enforcement agents to determine this. According to the responders, the age group most affected by the cases they have in their records is young people under 18 years old. In an increasing number of defilement cases, the perpetrators are known to the family, including close relatives and/or neighbors. Figure 4.1 depicts the age categories based on the reactions of law enforcement.

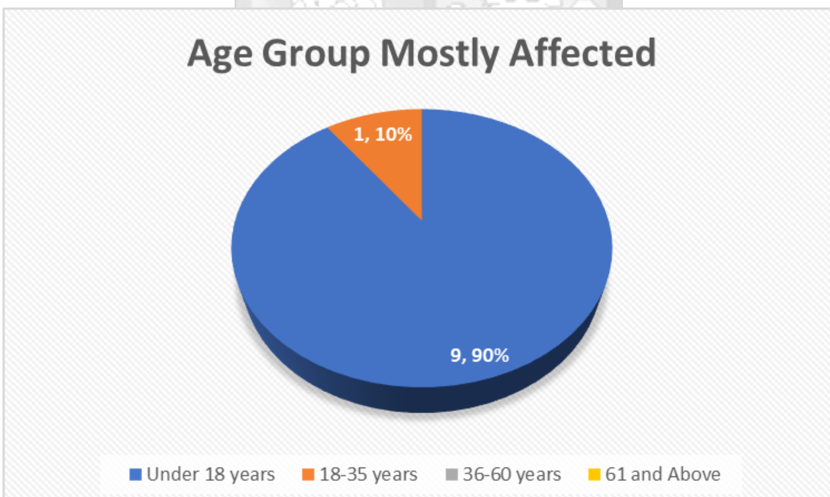

Figure 4. 1 Responses from Law Enforcement Officers on Age Groups Affected

Figure 4.2 shows that the researcher looked at the age groups of the respondents to find out which age groups are more likely to be victims of SGBV.

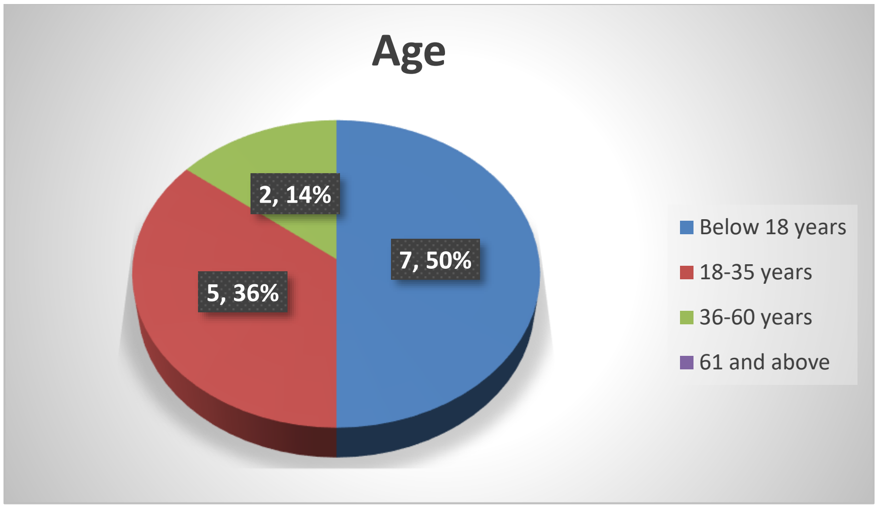

Figure 4. 2 Victim's Age Groups

4.1.2 Gender

The participants' gender was used as a data point to find out which gender was most affected by SGBV. Figure 4.3 shows that nine of the fourteen respondents were women and five were men, but this doesn't show how many incidents there were. When it comes to sexual violence, male victims do not speak up because they believe that men must always be strong. The police officers echoed this sentiment and added that males should receive more consideration when it comes to awareness and sensitization forums.

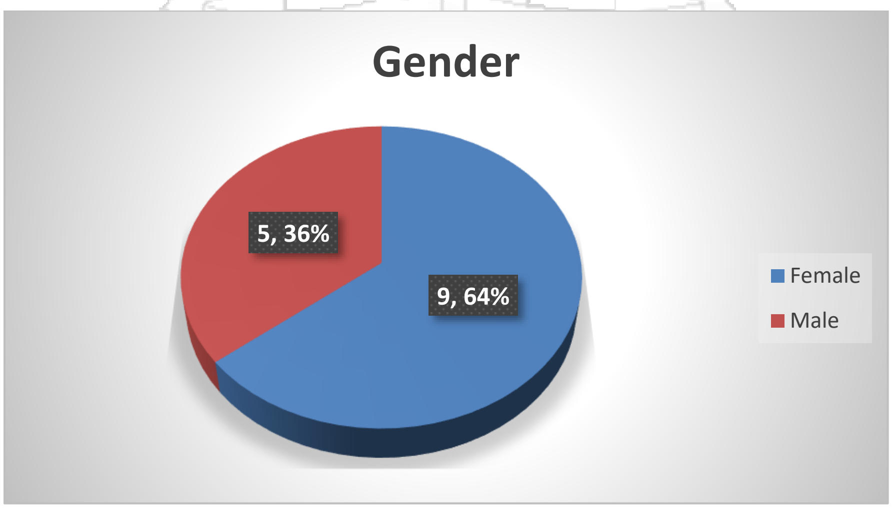

Figure 4. 3 Gender

4.1.3 Types of SGBV

The researcher hoped to learn more about the patterns of sexual and gender-based violence by looking at the answers to the questionnaire. The distribution of questionnaires to several target groups, including victims, community-based organizations or social justice centers, medical practitioners, and law enforcement officials, made this possible.

As shown in Figures 4.4, 4.5, and 4.6, all respondents, including medical professionals, law enforcement officers, and community-based groups, said that defilement, followed by rape and sexual assault, was the most common form of sexual and gender-based violence. Figure 4.7 shows the types of SGBV that have been reported by victims, based on information gathered from them.

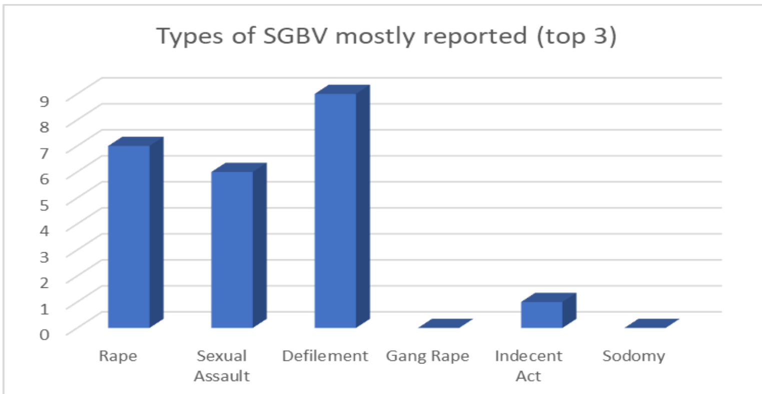

Figure 4. 4 Responses from Law Enforcement Officers

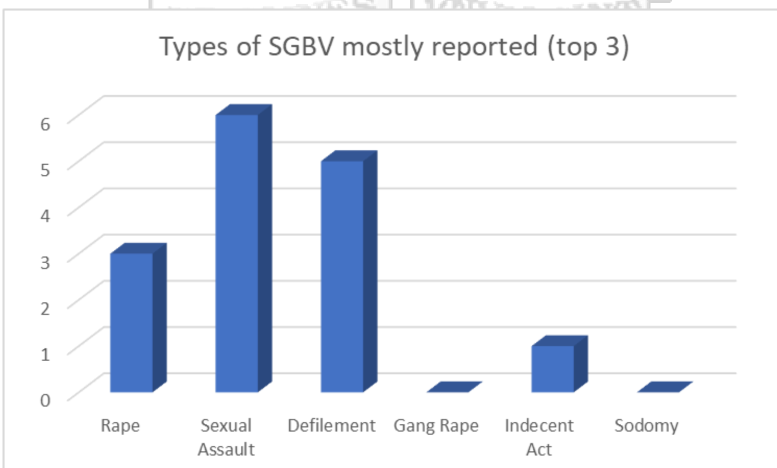

Figure 4. 5 Responses from Medical Practitioners

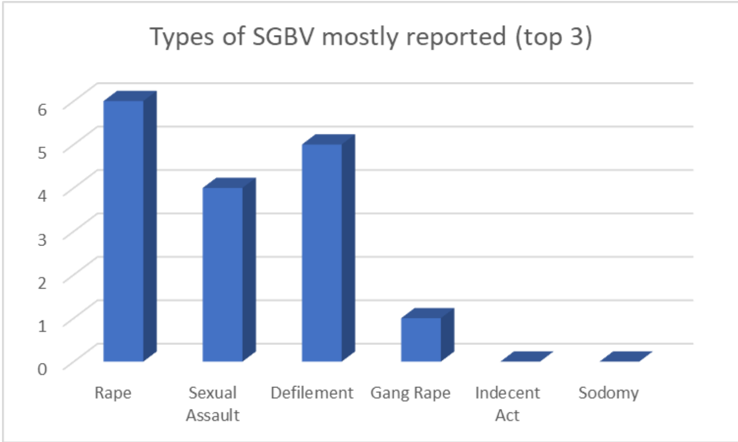

Figure 4. 6 Responses from Community Based Organization

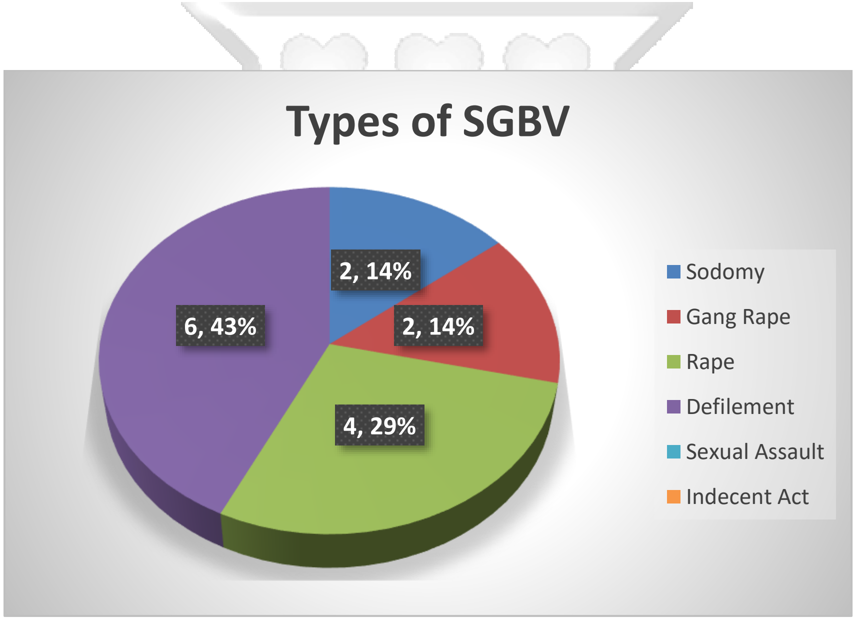

Figure 4. 7 Responses from the victims

4.1.4 Frequency of SGBV cases

The researcher questioned the respondents, the law enforcement officers, and medical officers, on how frequently they receive SGBV cases at their centres to better understand the frequency of cases that are usually reported and those that seek medical assistance.

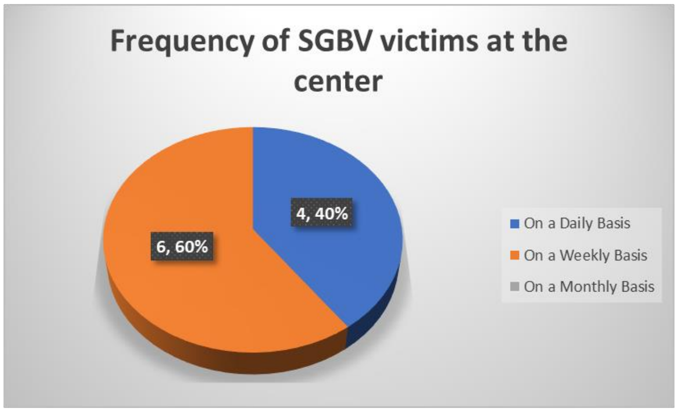

Figure 4. 8 Responses from Law Enforcement Officers

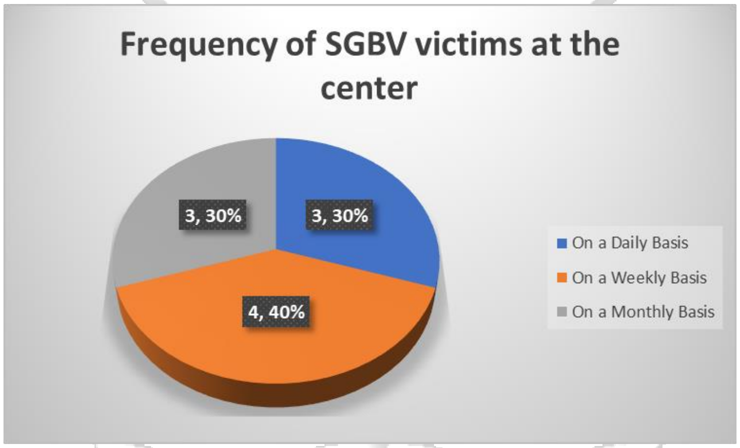

Figure 4. 9 Responses from Medical Officers

According to the law enforcement officials who responded, 40% of the cases occurred on a daily basis, and 60% occurred on a weekly basis. Whereas medical practitioner responses revealed that 40% of cases are on a daily basis, 30% are on a weekly basis, and 30% are on a monthly basis. This demonstrated that there are instances where victims report violations to police stations but do not seek medical attention, and vice versa. In addition, the general public's understanding of what is expected of them when a violation occurs has improved. However, Figure 4.10 demonstrates that SGBV is very rampant, as eight (8) out of the fourteen (14) victims have happened in the last one-year period.

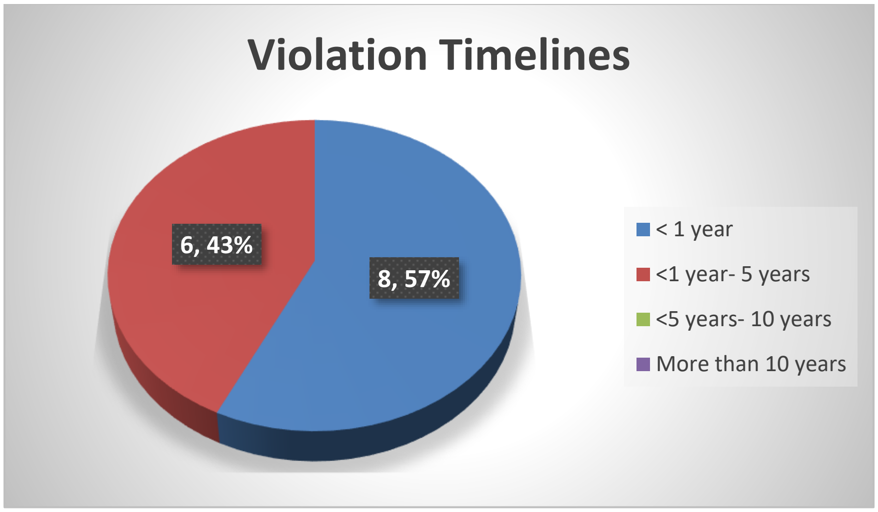

Figure 4. 10 Responses from Victims

4.1.5 Responses to seeking assistance

The researcher wanted to know if victims know that they should report crimes and get medical help, and if they usually follow the advice to get medical help within 72 hours, which gives doctors time to do a medical exam, collect evidence, and keep it so that it can be used to bring criminals to justice in the future. Officers say that victims know they need to report the crime right away, and if they come to report after a month, they are treated as an afterthought and turned away. They also stated that victims under the age of 18 were delayed in reporting from four days to one week after the assault because they feared repercussions from the perpetrators because the majority of them were close relatives. According to medical practitioners, victims are now informed of what to do next after a violation occurs, so medical treatment is mostly sought within the 72-hour recommended time frame. It is worth noting that some victims do not report the violation to the police station due to intimidation by the perpetrators, stigma, fear, or other alternative dispute resolution processes. The figures below show a summary of the victims' responses when it comes to reporting the violation to the police station and seeking medical care.

Figure 4. 11 Responses from Law Enforcement Officers on reporting violation

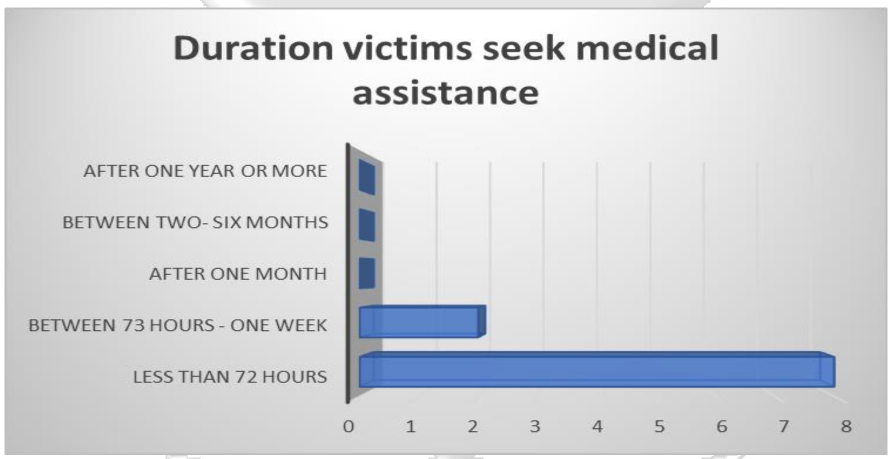

Figure 4. 12 Responses from Medical Officers on medical assistance

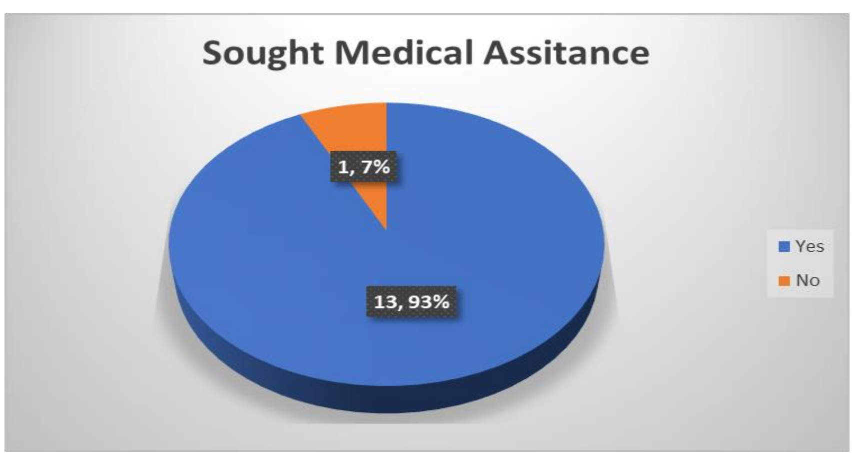

Figure 4. 13 Responses from victims on medical assistance

Figure 4. 14 Responses from victims on reporting violation

4.1.6 Data Management and Storage

To answer Section 1.3, the researcher tried to figure out what was hard about the way data is collected and managed right now. All of the police officers and medical officers who answered the survey said that data collection, management, and storage were all done manually. The information is written down in "occurrence books," which are kept for 10 years and then thrown away. Monthly reports are generated and sent to the sub-county headquarters. In other stations, the records are stored in the records office, which sends them every week to the division office. Also, in crime registers and exhibit rooms. Therefore, retrieving and analyzing the information collected is a challenge, leading to difficulty in informing policies.

4.1.7 Other Findings

In the past few years, community-based organizations have become a big help when it comes to dealing with problems at the local level. For the purpose of this study, the researcher collected data from sample social justice centers in Nairobi to understand SGBV at the community level. The researcher wanted to find out what makes people more likely to commit SGBV in the sampled communities. Some of the things they said were poverty, traditions and cultures, not knowing about SGBV, and holidays when schools are closed and children are more likely to get married young, be defiled, or be sexually assaulted or harassed by relatives or neighbors. The researcher also asked how the communities protect themselves from SGBV and how they respond to incidents. Their answers were similar and included holding forums to raise awareness, helping victims by following up on their cases at the police station and in court, and having good links with the police stations in the area. This will help victims get justice, and some communities have gender desks where victims can report violations. Based on how the communities felt about SGBV, it was important to find out if the problem has gotten better, worse, or stayed the same in the last year. All of the people who answered the question said that the situation has definitely gotten better over the years, mostly because of community programs to raise awareness and hold perpetrators accountable. There are still challenges or barriers to reporting to the police station and seeking medical assistance for the victims, which explains the under-reporting of violations based on the responses from the sampled CBOs, including financial constraints, stigma, and fear of what the community will perceive them as, a lack of knowledge of what action they should take, threats and intimidation by the perpetrators, and delayed justice for victims. In most cases, the CBOs are the first point of contact with the victims, and they tend to help them at the community level. The CBOs have services that are available at all times to SGBV victims in the community, and they include psychosocial support for the victims, sensitization forums, help in following up on cases at police stations and in court, and partnerships with NGOs that offer health services for the victims.

4.2 System Analysis

Nwakanma, et al., (2018) state that system analysis entails detecting problems, acquiring and evaluating facts, and applying the knowledge to offer possible system improvements. The system analysis will identify system users and provide an overview of the performance of the system. System analysis is concerned with defining the system requirements, analysing the system hardware, analysing the system software, identifying the research requirements, and defining the functional and non-functional system requirements.

4.2.1 Functional Requirements

The functional requirements outline the system's behaviour, function, and input or output, all of which should be in line with the study's objectives and goals. The functional requirements are as follows:

i. The user should be able to register and log in to the system.
ii. The system should allow the user (victim) to report a violation.

iii. The system should send feedback to the victim to guide them on their next steps based on the type of violation.
iv. The system should accurately identify the patterns and relationships of variables that influence the frequency of SGBV.
v. The system should allow the user (medical officer, police officer and pro-bono lawyers) to view and update the necessary information to the reported violation.
vi. The system should generate key reports to share with duty-bearers.

4.2.2 Non-Functional Requirements

These requirements define the general attributes of a system. The non-functional requirements are as follows:

i. Usability: The prototype should be simple to use so that the user can perform the necessary tasks with minimal or no training. Since it deals with a sensitive topic, the system must provide accurate and reliable information to its users.
ii. Security: The system should ensure all the data collected is only accessible to authorized users and ensure confidentiality.
iii. Availability: The system should be accessible at all times.
iv. Reliability: The system should provide users with accurate information.
v. Scalability: The system should be able to grow so that new features can be added.
vi. Accuracy: The output from the system should be accurate.

VT OMNES VNVM SINT

4.3 System Architecture

The system architecture shows how the different components of the system will work together to do what they are supposed to do. The system architecture will comprise the server side and the client side. The data is stored and retrieved on the server side, whereas the client side is where the user accesses the application to report a violation and access their information. Figure 4.15 illustrates the components and how they are connected to each other to achieve the reporting and feedback engine and model the number of occurrences. The components used in the proposed system are users, a mobile application with a GPS module enabled, a web application, a database, and an ML model for the duty bearers to help inform policy. Before they can log in, the user must first register with the system. The user sends a message to report a violation through the mobile application, and the texts are stored in the database. When the user is filling in the information, it picks up their exact geo-location and lists the five nearest police stations and hospitals that they can go to. The system generates a unique number for the user and shares it with them. If the victim chose a specific hospital or police station, the system sends an alert to the station or hospital. The hospital user is able to view the cases under their centers and can update the necessary information, including the PRC details, for further medical examination. They then call the victim to the hospital for diagnosis. Thereafter, sends the report to the nearest police station, and the victim and police station receives an alert. The police user can update the report with the necessary information, including the OB number, and P3 details, and start an investigation. The pro bono lawyers and judicial representatives will have administrator access to view all the updated information. The ML model will be used to model the number of cases with respect to specific variables, including gender, month, year, and location, which will help analyze the collected data, and aid duty-bearers inform policy decisions on SGBV prevention and response services. Through the use of the firebase database, the feedback engine will send a response to the victim outlining the next steps to take based on the type of violation. The information is stored in the database, and based on the ML model data analysis is performed to depict the trends and patterns that will help understand the magnitude of the incidents in order to plan ahead for better optimization of resources needed by the duty bearers to better assist the victims.

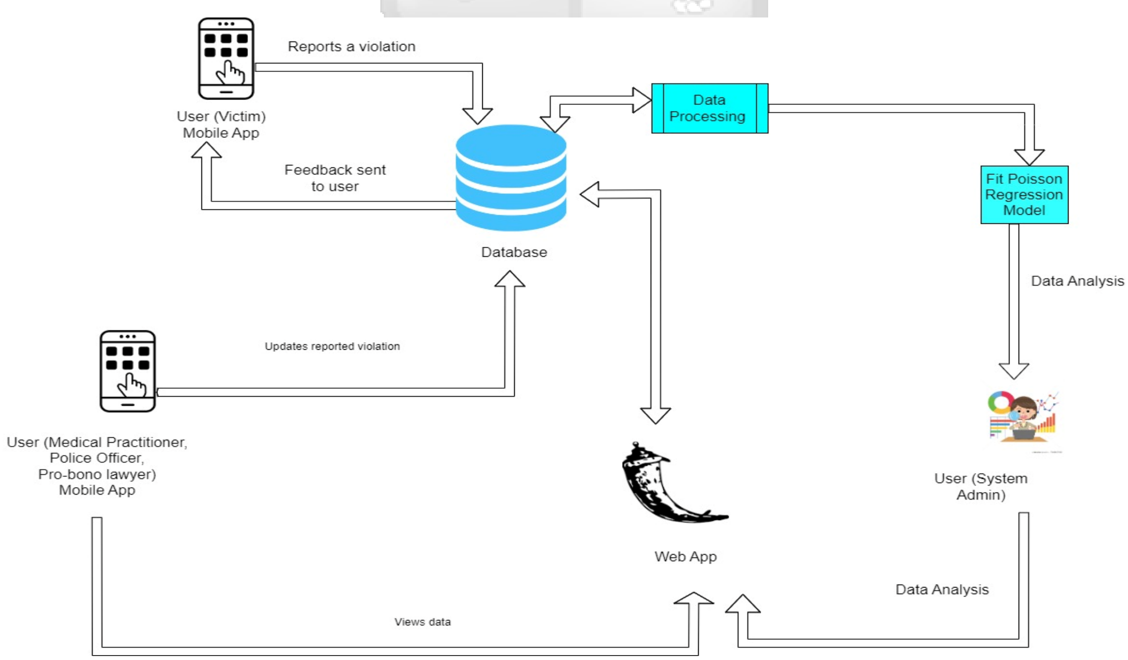

Figure 4. 15 System Architecture

4.4 System Design

System design is the process of defining and designing a system's architecture, interfaces, and data to meet certain needs (Schaffer, 2022). Unified Modeling Language (UML) diagrams will be used to represent the application's various interactions. Diagrams used include use case diagrams, sequence diagrams, data flow diagrams, entity relationship diagrams, and class diagrams.

4.4.1 Use Case Diagram

The use case diagram describes the various actors and their interactions with system processes. Figure 4.16 illustrates the behavior of the system. The description tables are documented in the Appendix A section.

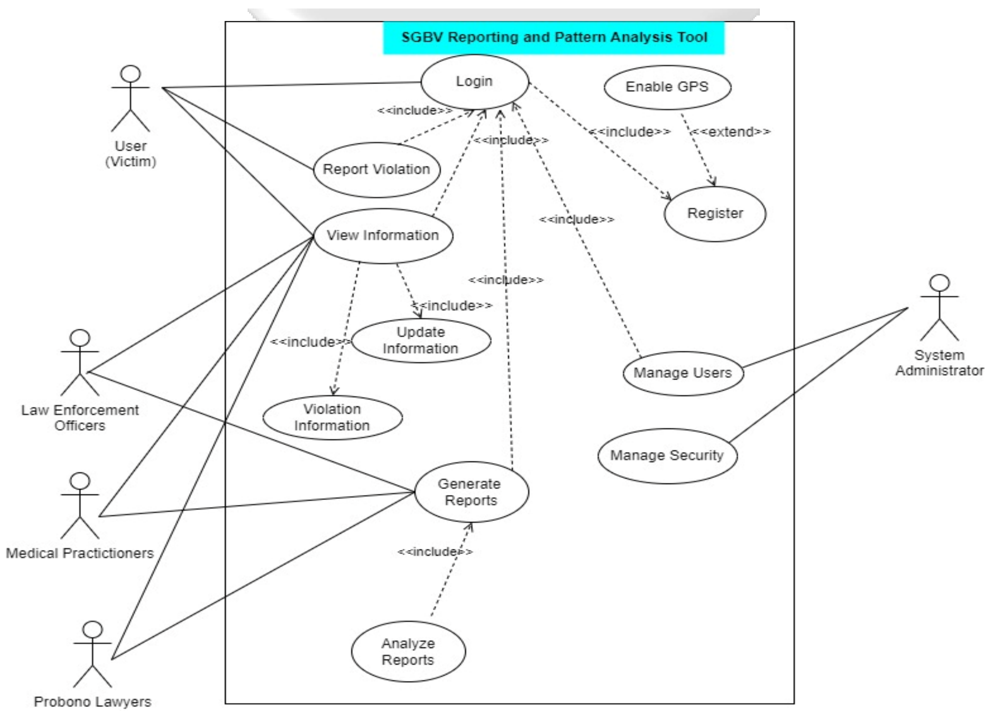

Figure 4. 16 Use Case Diagram

4.4.2 Sequence Diagrams

A sequence diagram, like the one shown in Figure 4.17, will show how users interact with the system. These interactions are important as they allow one to understand how the application works in the different sections. There are different actors, including the victim, medical officers, police officers, pro-bono lawyers, and system administrators, who interact with different modules or sections of the system.

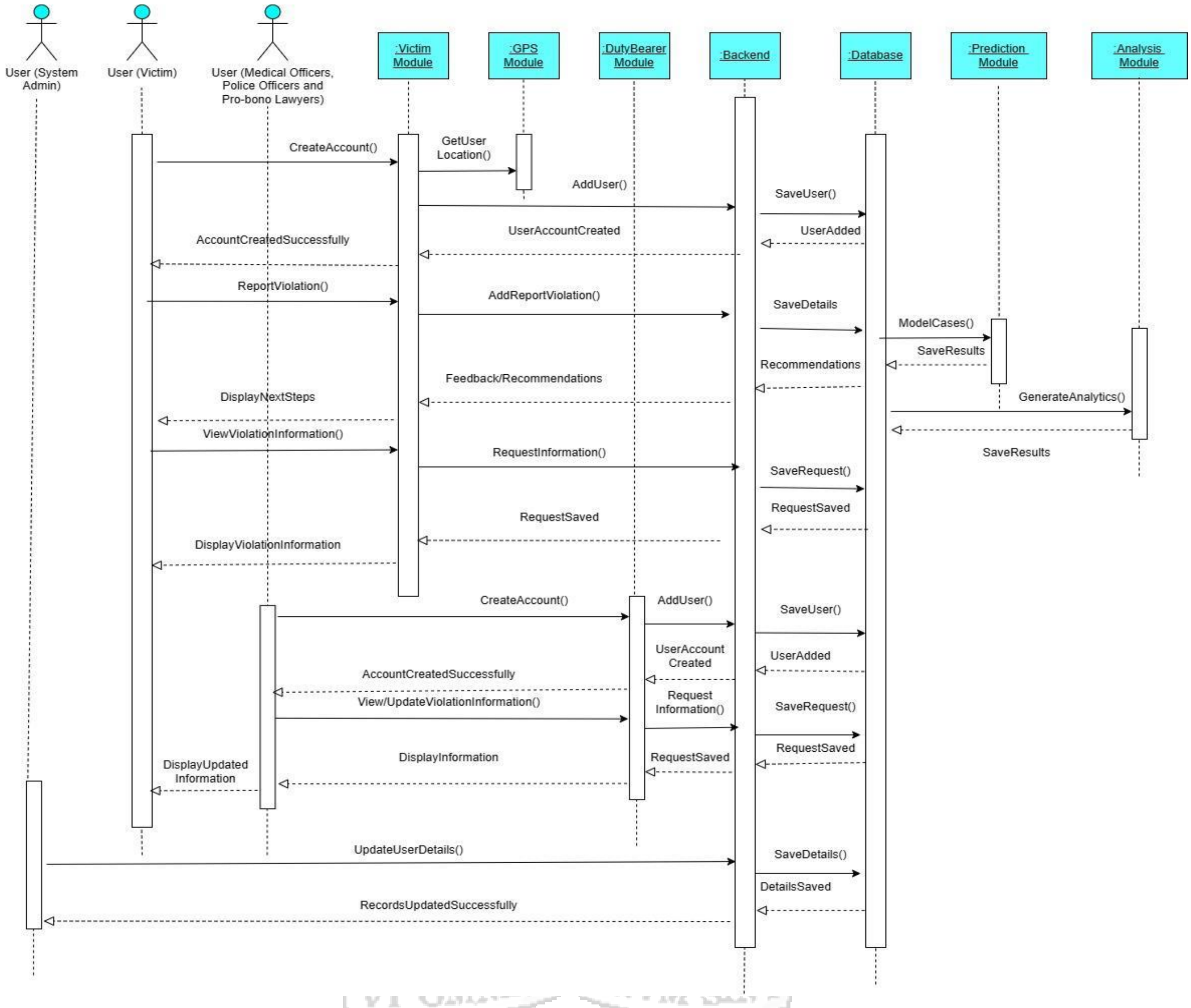

Figure 4. 17 Sequence Diagram

4.4.3 Entity Relationship Diagrams

The entity relationship diagram illustrates the database conceptualization of the system, as shown in Figure 4.18, which illustrates the relationship between database tables. These included the victim's information and the different users who interacted with the information.

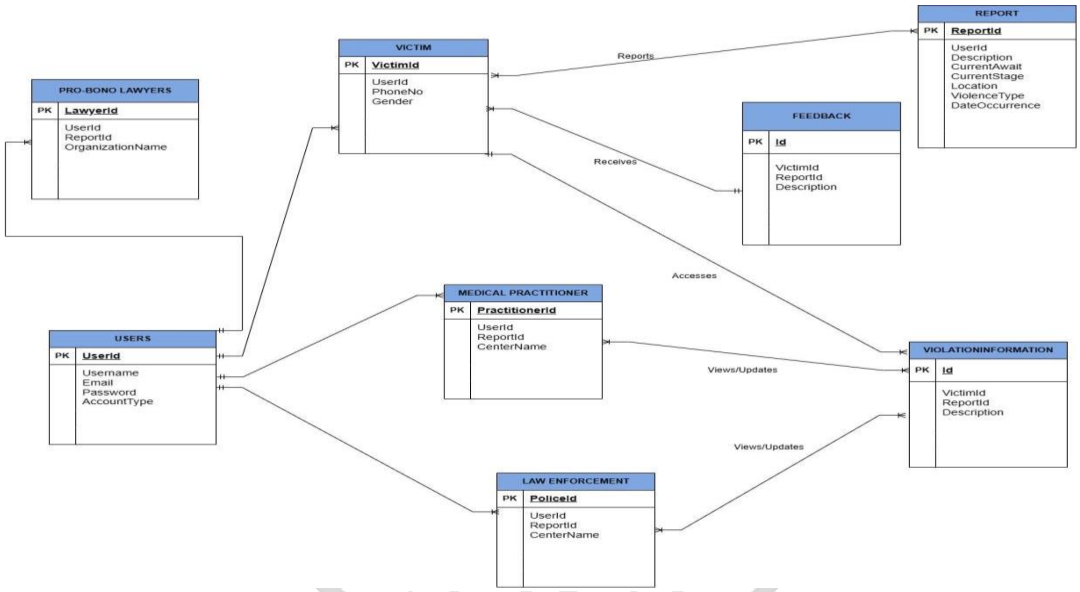

Figure 4. 18 Entity Relationship Diagram

4.4.4 Class Diagram

In this study, class diagrams will be used to show how the system is put together. Class diagrams, as shown in Figure 4.19, are used to model the objects that make up the system, display the relationships between the objects, and define what these objects do and the services they provide.

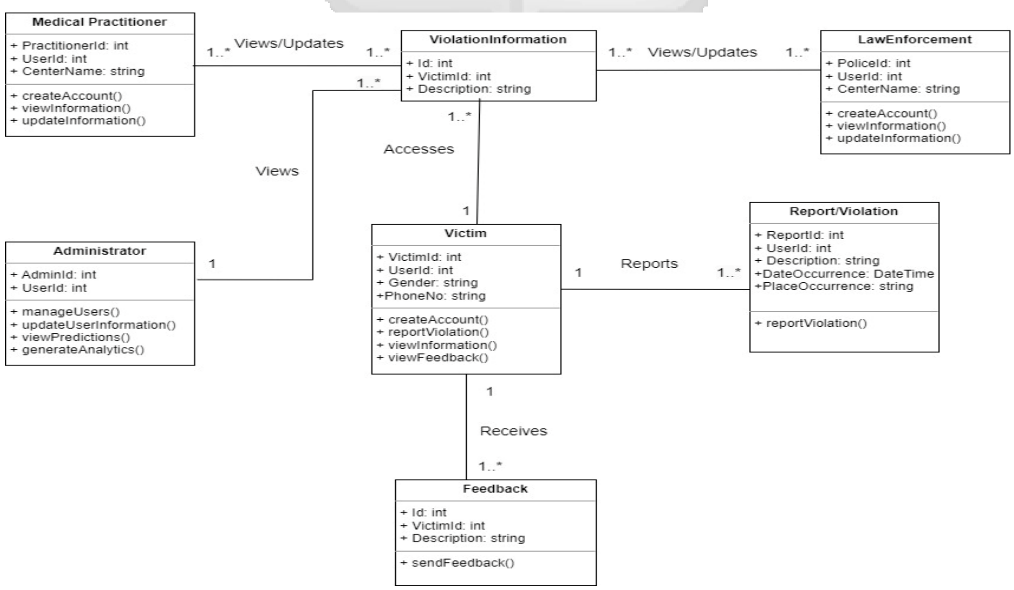

Figure 4. 19 Class Diagram

4.4.5 Data Flow Diagram

A data flow diagram displays the flow of information through a system or process. It includes data inputs and outputs, database systems, and other sub-processes that illustrate how data flows through the system.

4.4.5.1 Context Diagram

Also referred as a DFD Level 0 Diagram. It represents the complete business procedure as a single procedure. (process 0). The purpose of this diagram is to show the system as a single high-level process and how it interacts with the external entities, as seen in Figure 4.20.

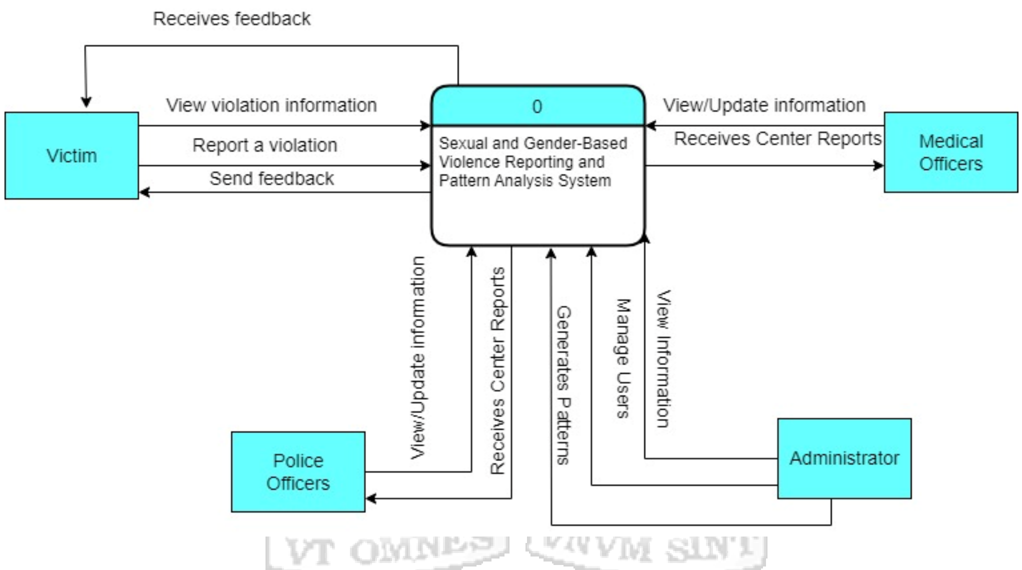

Figure 4. 20 Context Diagram

4.4.5.2 Level 1 Data Flow Diagram

Level 1 DFD decomposes the context diagram into several processes. This level focuses on the primary functions of the system by decomposing the level 0 DFD's high-level processes into subprocesses.

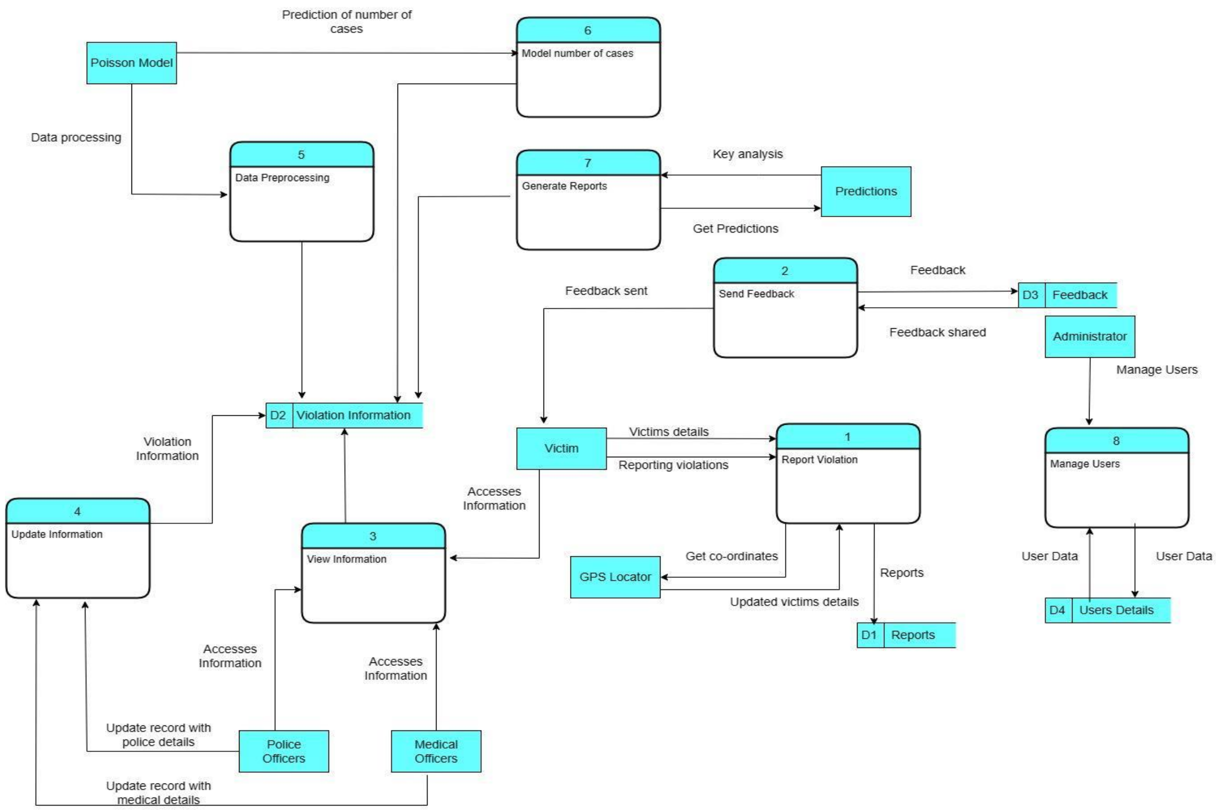

Figure 4. 21 Level 1 Data Flow Diagram

4.5 Wireframes

Wireframe diagrams are used to develop a visual representation of the design of the solution. It takes into account users' and developers' needs and requirements to come up with the designs, as it is critical to ensuring that the application is created with the goals and objectives in mind. This is achievable using wireframes because they set expectations about how features will be implemented by depicting how they will work; however, they are not always completely correct because the features can be altered during the development period.

4.5.1 Account Types Wireframe

Figure 4.22 depicts the application’s different account types in wireframe, which allow the user to register as a police officer, hospital user, anonymous victim, or pro bono lawyer.

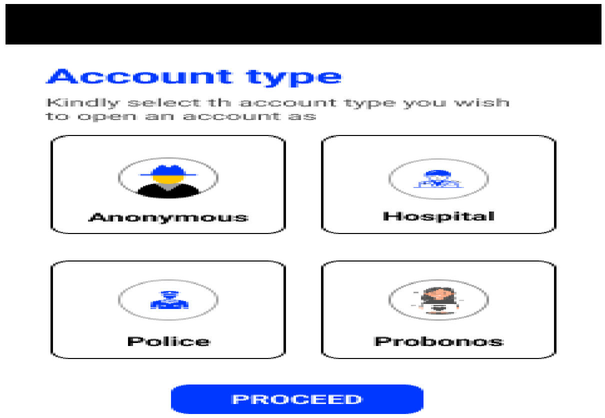

Figure 4. 22 Account Types Wireframe

4.5.2 Register Wireframe

Figure 4.23 depicts wireframes for signing up or registering as a new user. The user accounts include those of a victim, a medical practitioner, police representatives, and pro bono lawyers. It contains a form that requires one to provide the requested information for the successful creation of user accounts.

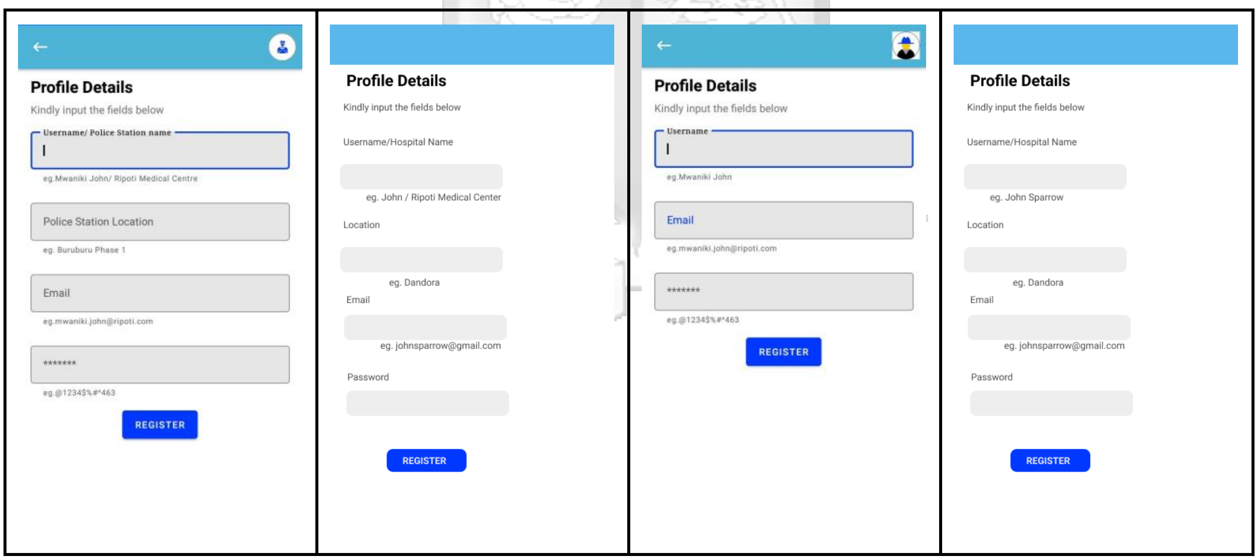

Figure 4. 23 Registration Wireframe

4.5.3 Login Wireframe

Figure 4.24 below depicts a login wireframe that contains a form requiring a user to make use of valid credentials, including the username and password. After entering the credentials, the user clicks the Sign in button.

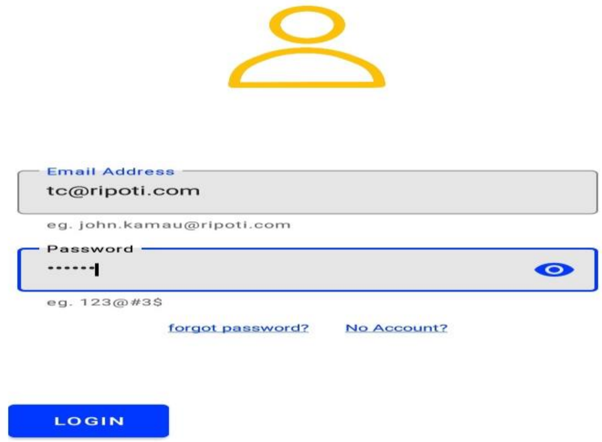

Figure 4. 24 Login Wireframe

4.5.4 Reporting Wireframe

Figure 4.25 below depicts the reporting of violations wireframe. On successful login, the victim will be able to report a violation and follow up on the status of the case from the application.

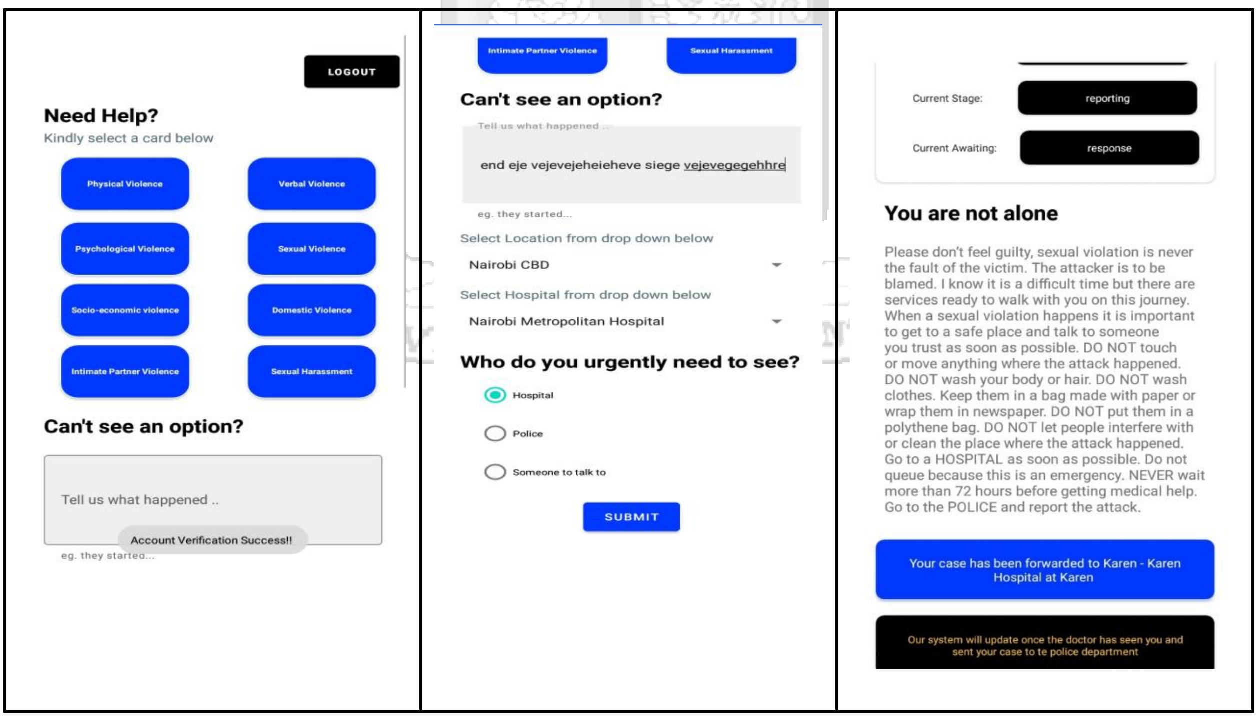

Figure 4. 25 Reporting Wireframe

Chapter 5: System Implementation and Testing

5.1 Introduction

System implementation designs and creates a structure that adheres to the architectural design and specifications established during the system analysis phase and meets the system's and stakeholders' requirements (Snoderly, J. and Faisandier, A., 2022). The purpose of system testing is to examine the full workings of the system against the requirements and objectives before delivery of the final product.

5.2 System Implementation Environment

The sexual and gender-based violence pattern analysis system was designed and developed using the Agile software development methodology as mentioned in Chapter 3.

5.2.1 System Software Environment

The model was developed using the Poisson Regression algorithm for demonstrating pattern analysis of sexual and gender-based violence cases reported. The advantage of using the Poisson Regression model is that it has a minimum value of 0 therefore it will not predict negative values leading to accurate prediction. The Android Studio 2022.1.1 for 64-bit Windows OS was used for the mobile application, and the model was designed in Python using the Jupyter Notebook IDE.

5.2.2 System Hardware Environment

The system was successfully designed and implemented utilizing a Hewlett Packard (HP) Intel Core i7 laptop with 16GB RAM and a 1TB Solid State Drive. (SSD).

5.2.3 Cloud Environment

The prototype made use of the Firebase Database because it is a real time cloud-hosted NoSQL database. The advantage of using Firebase is it has analytics and it is efficient in relaying asynchronous data.

5.2.4 Network Environment

Wireless Connection was required for the network environment.

5.2.5 Mobile Application Prototype

The mobile app was built on the Android platform, and its design was made to work well with all versions of Android, from version 4.0 to the latest version, 12.0. The Firebase database was used for database query language because of its real-time capabilities. The mobile application was developed using the Kotlin programming language.

5.2.6 Web Application Prototype

The JavaScript programming language was used to develop the web application. The JavaScript framework was used in the web application because it is easy to learn and understand and has the ability to run on multiple platforms directly. The Firebase database was used to develop the database of the system since it is open source and offers real-time data storage and access.

5.3 Implementation Details

5.3.1 Mobile Application Modules

5.3.1.1 Registration and Login Module

The system implementation incorporated different types of user accounts profiles including law enforcement officers, medical practitioners, pro-bono lawyers and victims. Therefore, a user was required to register their personal details with respect to the account type. On successful registration of an account the user was able to login and access the application. Figure 5.1 shows the different user accounts available; Figure 5.2 depicts the user registration module, whereas Figure 5.3 depicts the application login module.

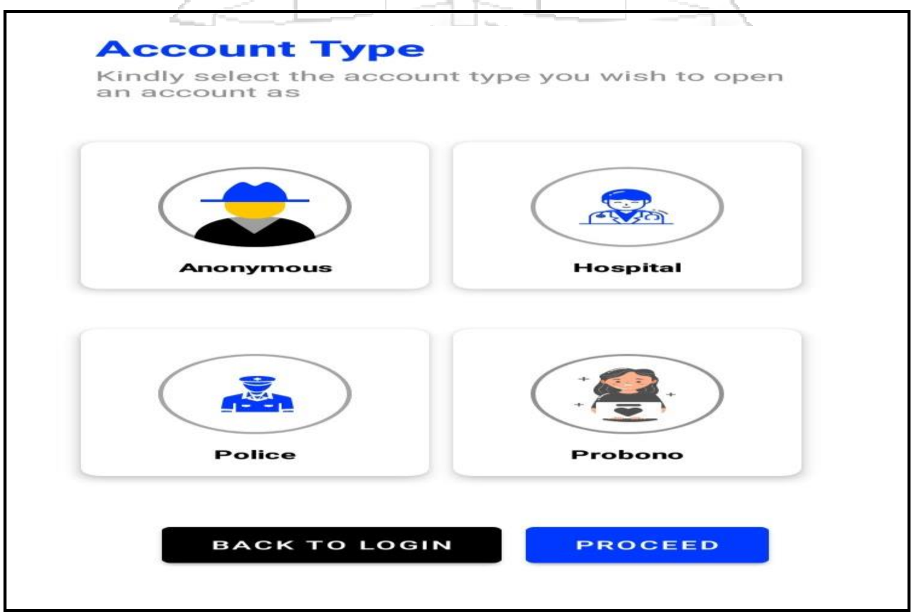

Figure 5. 1 Account Type

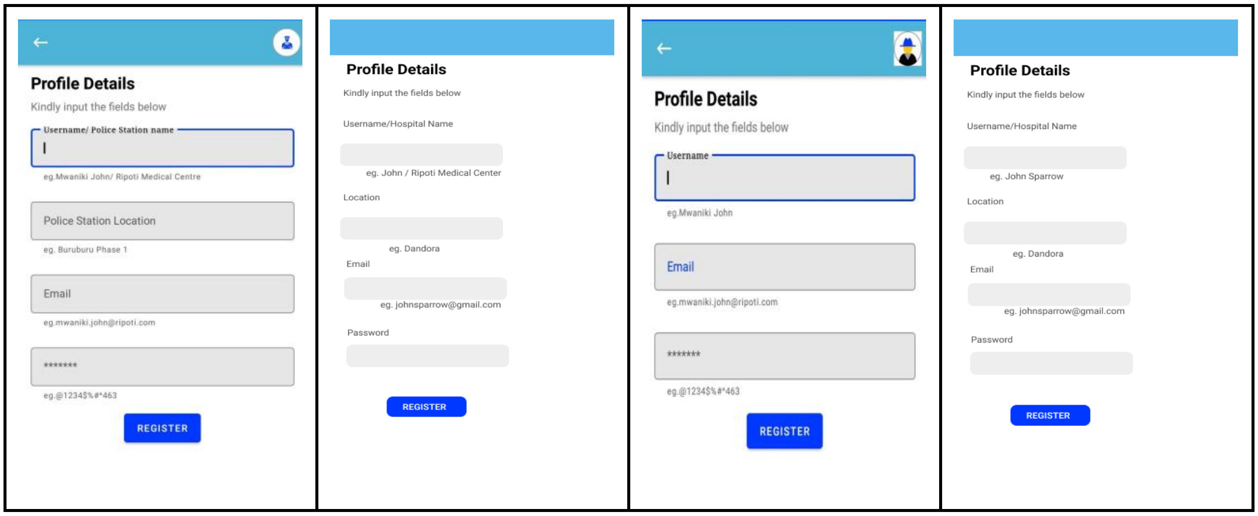

Figure 5. 2 Registration of different account users

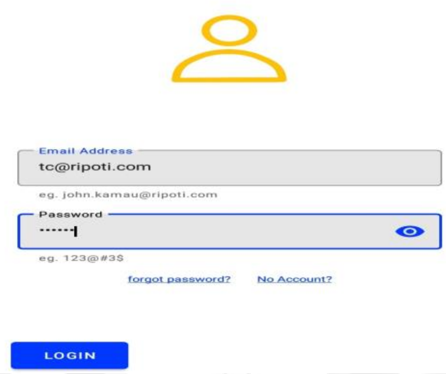

Figure 5. 3 Login Screen

5.3.1.2 Sexual and Gender-Based Violence Victim Reporting Module

After registering and logging into the app, a victim can report a violation by choosing the category that best fits the situation. The app then lets the victim choose where they are and shows them a map with the top 5 closest hospitals they can go to for a check-up and the top 5 closest police stations they can go to report the crime. The app also lets the victim choose who they need to see right away, whether it's a police officer, a doctor, or a counsellor. When the victim reports a violation, the system creates a unique ID and sends an alert to the police or hospital, depending on what the victim chose. Also, the system updates the user with feedback on what they need to do and displays a banner showing the stages the report is in and updates when it moves to other stages.

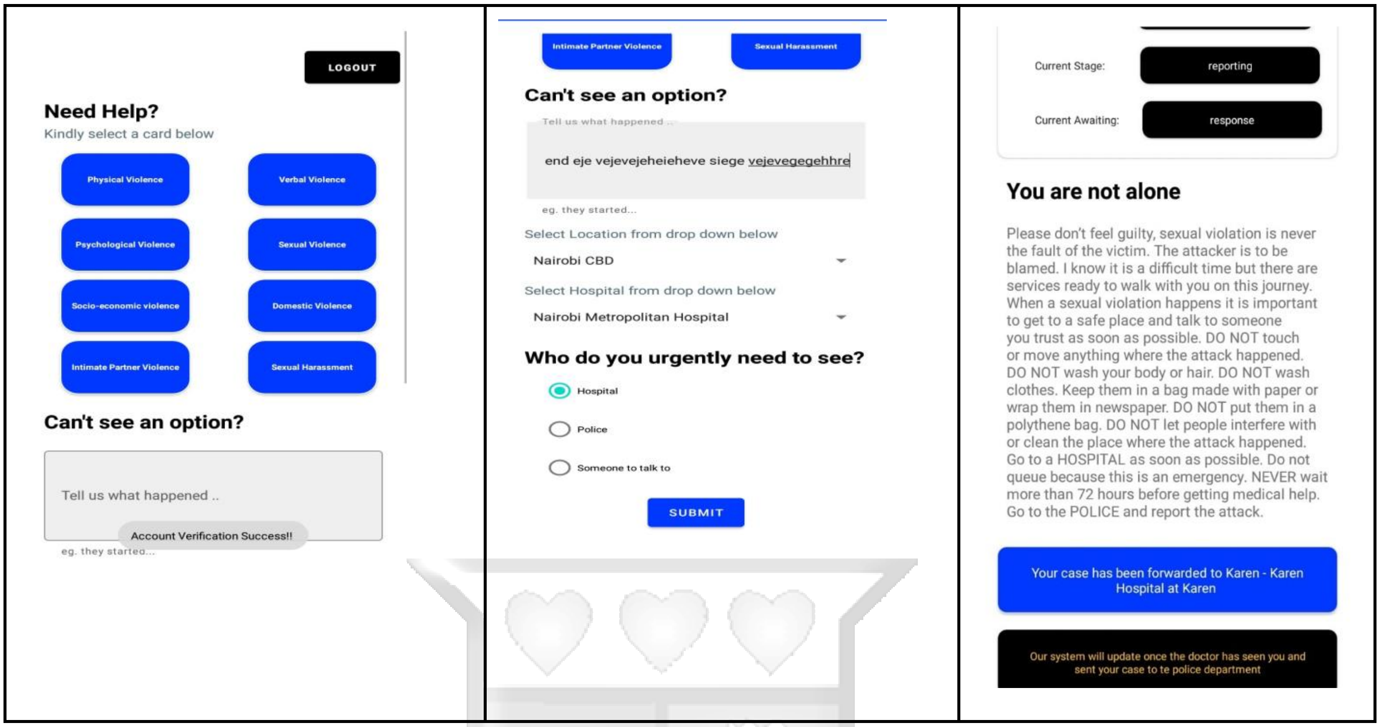

Figure 5. 4 Victim Reporting Module

Figure 5.5 below depicts a snippet of the system interacting with the user by sending notifications to update them of every stage of the reported violation and what to expect next.

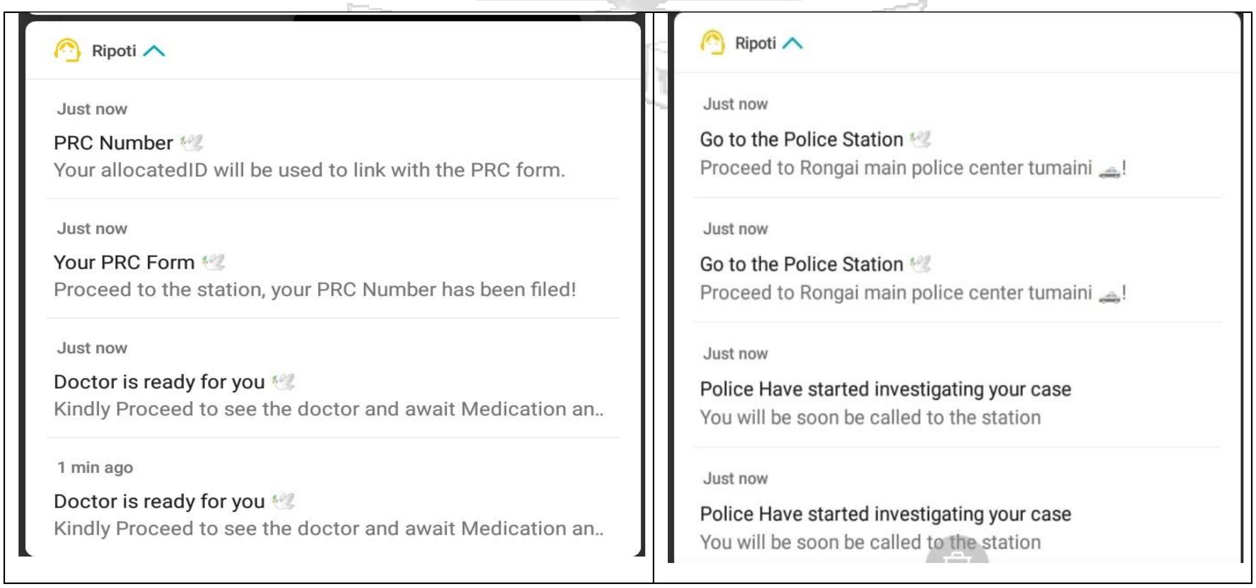

Figure 5. 5 Notifications/Alerts

5.3.1.3 Sexual and Gender-Based Violence Hospital User Module

When the victim sends the report to a certain hospital, the hospital gets a notification that a new report has been sent. The hospital user can sign up and log in successfully. Once that is done, the hospital user can see how many cases are at their centres at any given time. The hospital user can click on the process case button to access the violation details, then call the patient to come to the centre for examination and update the PRC form on the app. On completion of the medical examination, the user can click on the send to police button to forward the case to the nearest police station for follow-up.

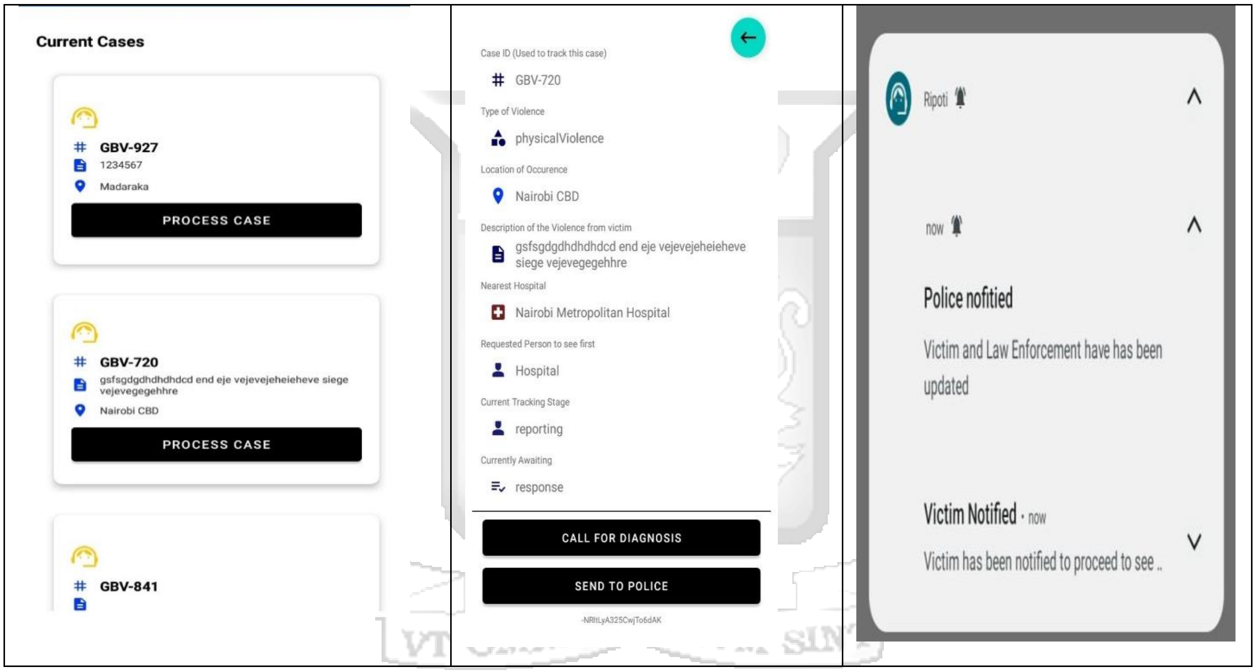

Figure 5. 6 Hospital User Module

5.3.1.4 Sexual and Gender-Based Violence Police Station User Module

If the victim chose to go to the hospital first after the violation, the hospital user calls the victim to get a diagnosis and then updates the application with the necessary medical information. Thereafter, the hospital user sends the report to the nearest police station, and alerts are sent to the victim, police station, and hospital user, updating them on the progress. When a police officer registers and logs in to the app as a police user, they can see how many cases have been reported to their centre, how many are waiting to be processed, and how things are going at any given time. The police officer can view the details of the case and start the investigation.

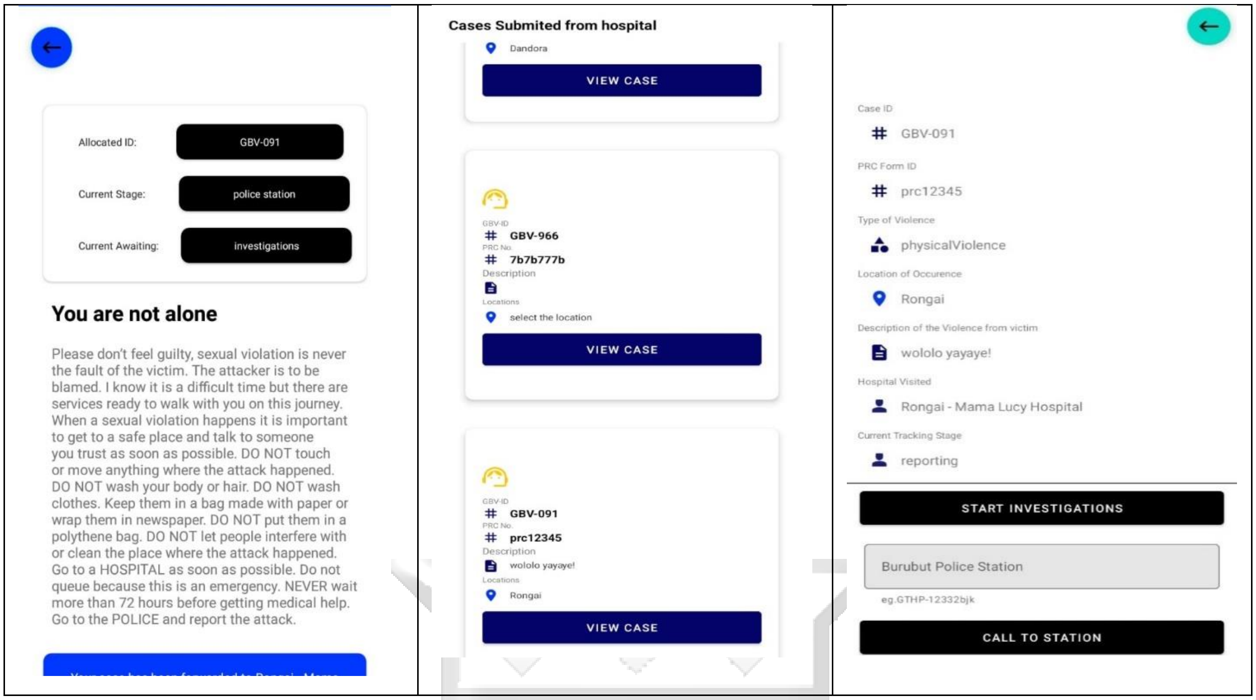

Figure 5. 7 Police Officer User Module

5.3.2 Firebase Database

The prototype made use of the Firebase Database because it is a real-time, cloud-hosted NoSQL database. The benefit of using Firebase is that it can analyse data and send asynchronous data quickly. Figure 5.8 shows a user account record that was saved successfully and a report of a violation.

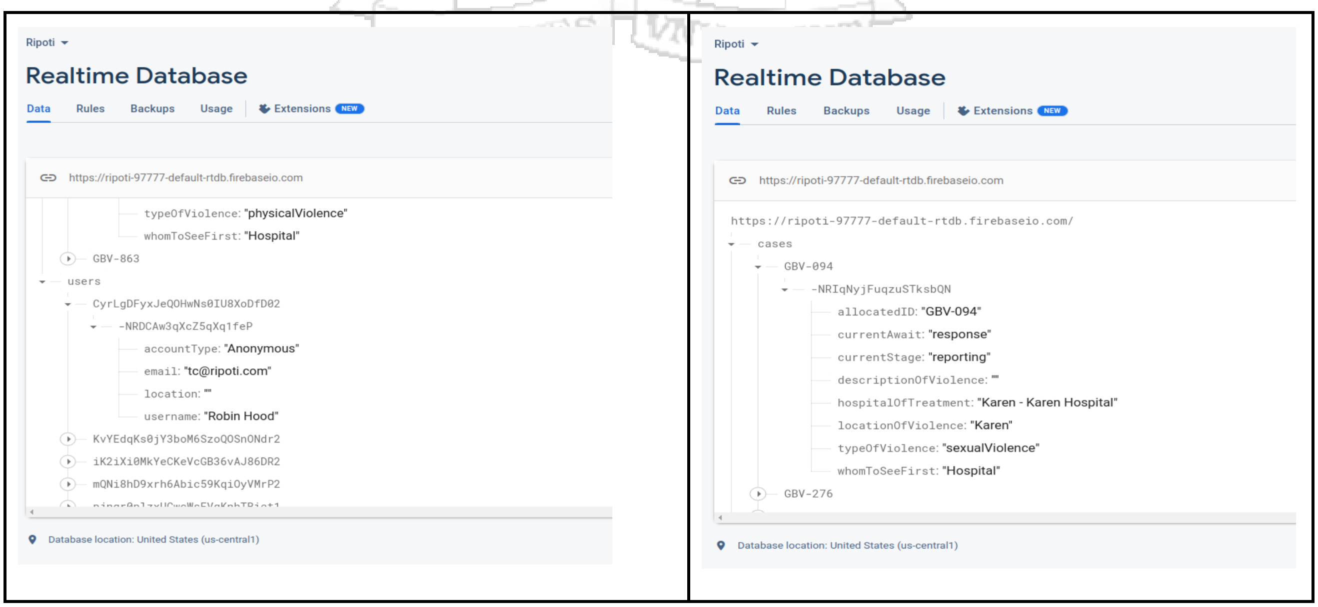

Figure 5. 8 Database Records

5.3.3 Web Application Modules

5.3.3.1 Locations Module

When a user logs in successfully, the system shows them the different modules they can use. Figure 5.9 depicts the locations module displaying a summary of cases reported with respect to their locations in the form of a map. The locations module is accessible to different user types and allows them to view only a summary of the locations.

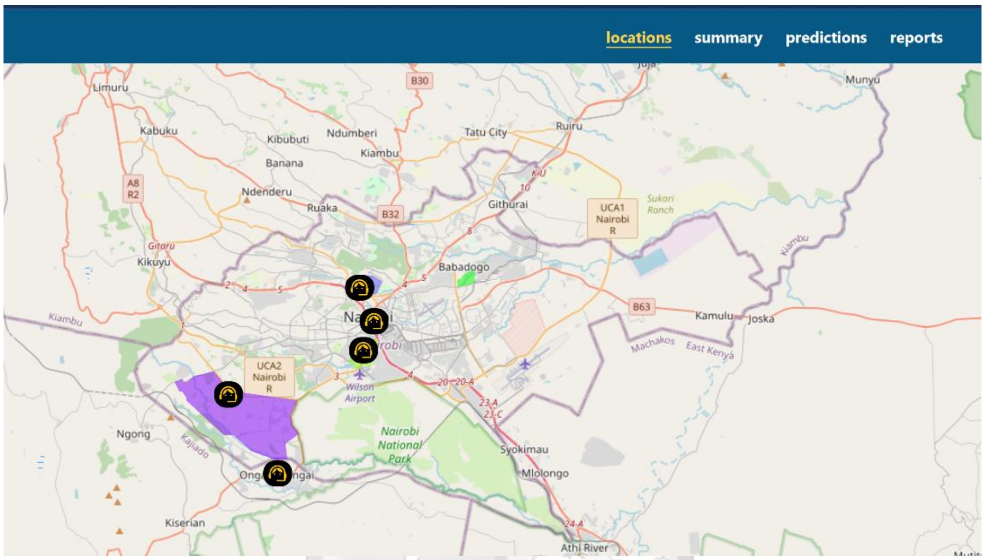

Figure 5. 9 Locations Summary

5.3.3.2 Summary Module

The summary module depicts a summary of cases reported by the victims in numbers: the total number of cases reported, the number of current cases per stage, including those handled at the hospital level, those that are being handled by the police, and those being handled by the pro-bono lawyers. The summary module is accessible to different user types and allows them to view only a summary of the reports, not the details.

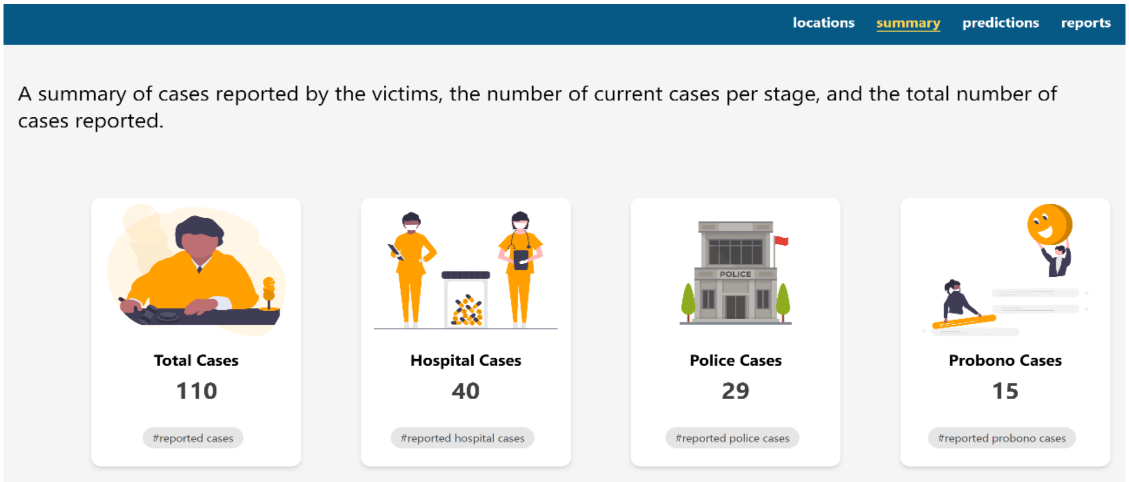

Figure 5. 10 Summaries Module

5.3.3.3 Predictions Module

The predictions module displays an overview of the collected data. It updates after every successfully reported case and keeps track of the data. Once it reaches 1000, then the dataset can be used to make good predictions, and training of the Poisson Regression Model is possible. The target outcome will eventually be to map hotspot areas, predict the number of SGBV cases and timing of violations.

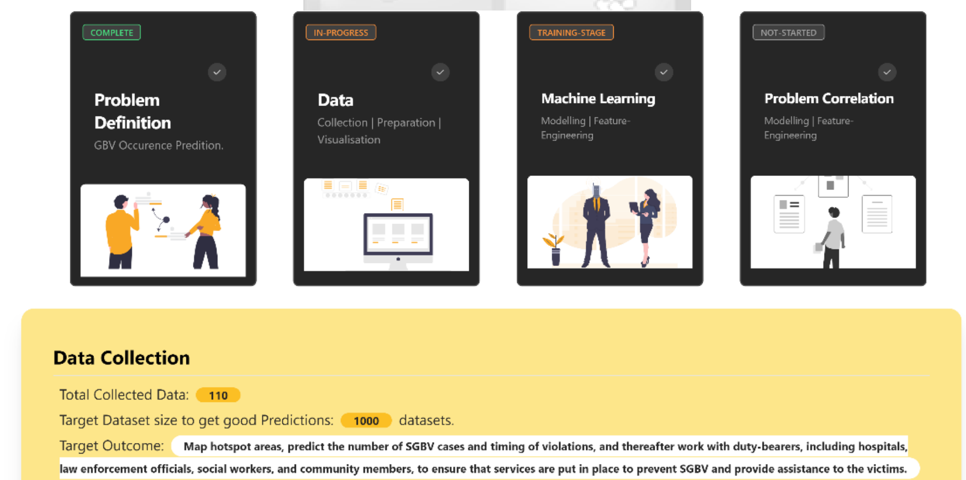

Figure 5. 11 Predictions Module

5.3.3.4 Reports Module

The reports module displays a summary of the collected data in chart format to better understand the trends and patterns of the reported violations. The key analysis of patterns looked out for includes age, gender, type of violation, and location. These patterns will help formulate policies that mitigate these violations with respect to the type of violations, the age groups of the victims, the gender mostly affected, and which locations have a higher frequency of violations.

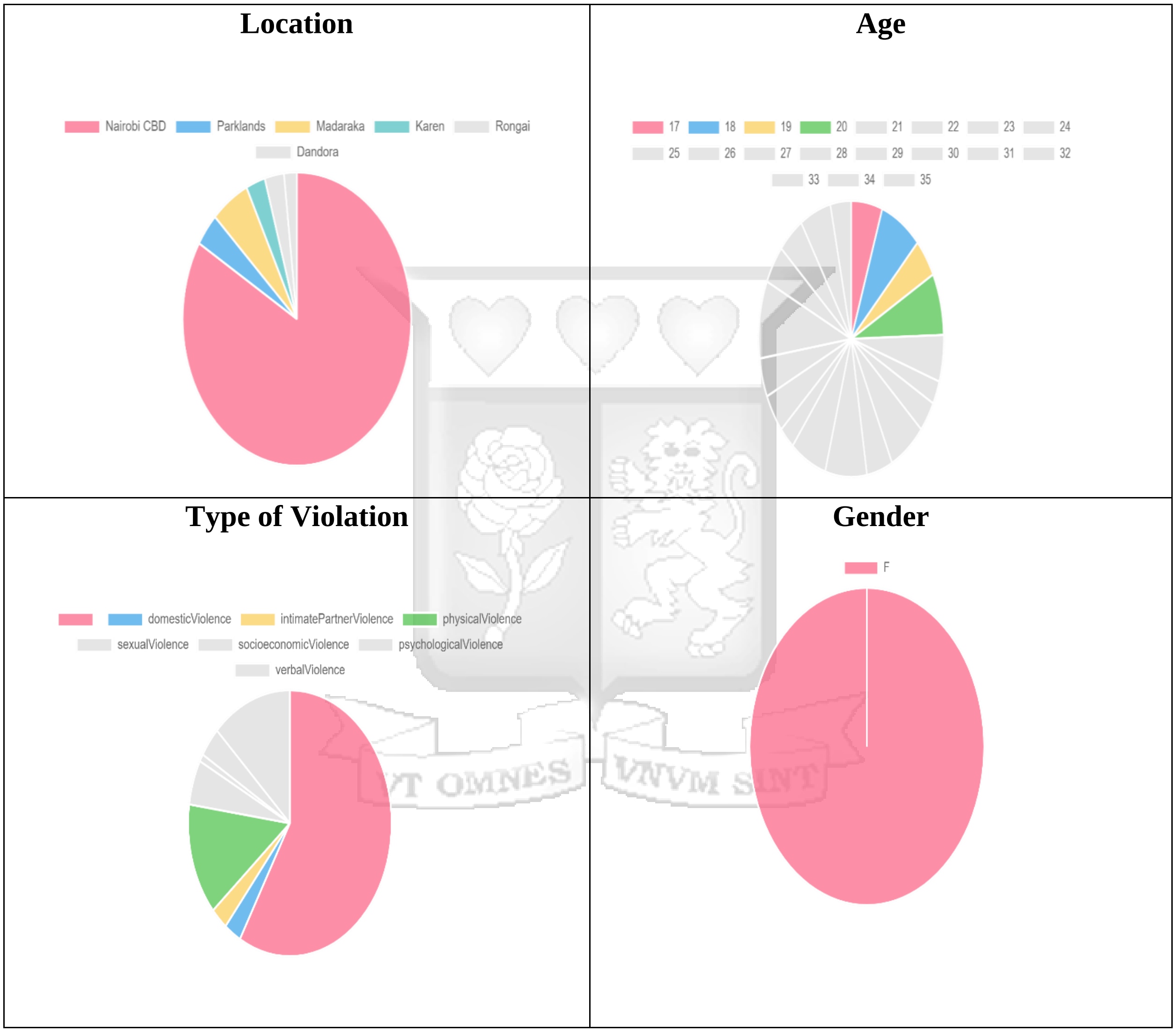

Figure 5. 12 Reports Charts

5.3.4 Implementing the Poisson Regression Algorithm

The steps for putting the algorithm into place were collecting data, pre-processing the data, exploratory data analysis, LDA modeling, and fitting the Poisson regression model. For the data collection process, the data set from the Kenya National Commission on Human Rights (KNCHR) was used for training the model. The data set was downloaded in CSV format.

5.3.4.1 Loading of the CSV file

```
#Loading CSV file
df1=pd.ExcelFile("crystalrep2.xls")
df=pd.read_excel(df1)
analysis_data=df[['Age', 'AllegId', 'County', 'FULLNAMES', 'Gender', 'HandlingOfficer',
'ReferalStatus', 'Remarks', 'ReportDate']].dropna(axis=0,how="all")
```

Figure 5. 13 Load CSV file

5.3.4.2 Cleaning of the data

```
# Remove stop words and lemmatize
nltk.download('punkt')
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()
def process_text(text):
    tokens = word_tokenize(text)
    tokens = [token.lower() for token in tokens if token.lower() not in stop_words]
    tokens = [lemmatizer.lemmatize(token) for token in tokens]
    return ' '.join(tokens)
text_data = pd.DataFrame(duplicate_free['Remarks'].dropna().apply(process_text))
```

Figure 5. 14 Cleaning of the data

5.3.4.3 Word Cloud

It was important to create a word cloud from the dataset to find out what words are commonly used. Figure 5.15 depicts the words frequently used to highlight either perpetrator, victims, the type of violation, location of the violation, age of the victims, and whether they reported at the police station.

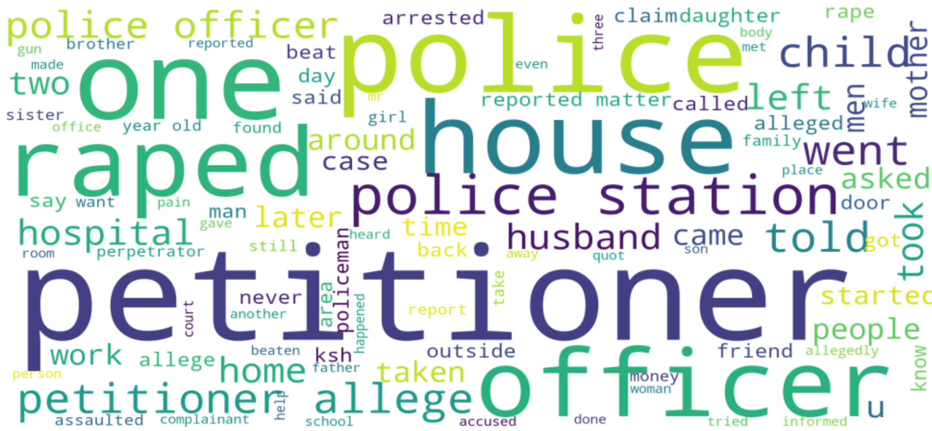

Figure 5. 15 Word Cloud

5.3.4.4 Exploratory Data Analysis

Exploratory data analysis was performed on the dataset to understand the trends and patterns of the occurrences of the cases with regard to the months and years from the year 2012 to 2022. A time-frequency analysis of SGBV cases was performed to understand the patterns of the yearly evolution of these cases. Further, an analysis of the months that violation cases are mostly reported and the results depicted that during holiday periods violations are prone to be high. Figure 5.16 is an example of the above statement, a comparison of the years 2021 and 2022, with the months of November, December, and August having the highest numbers; these months are also calendar months for school closure.

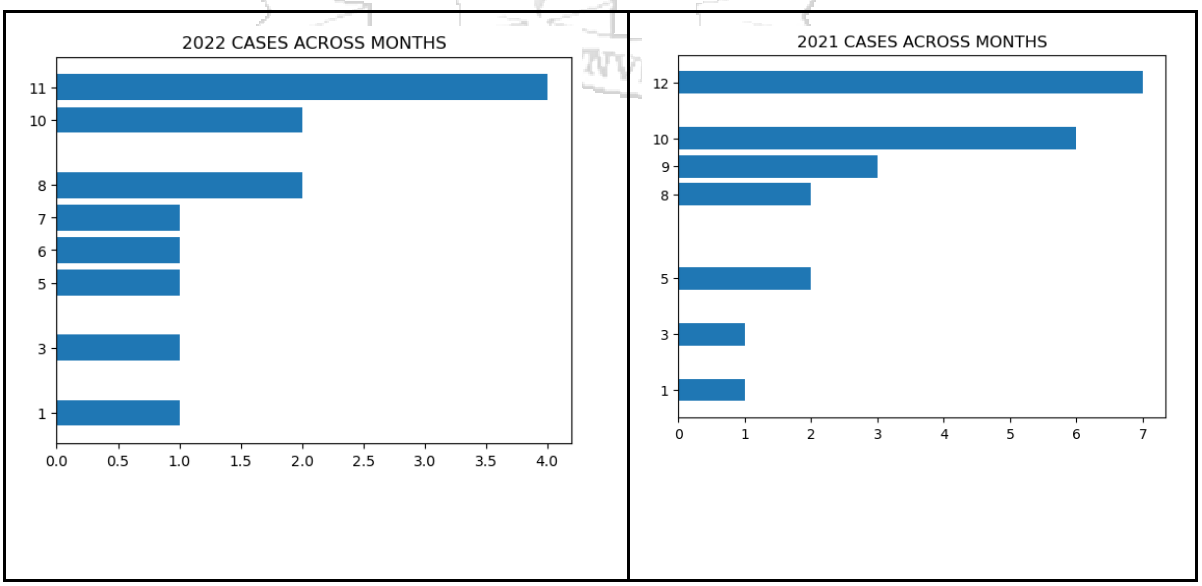

Figure 5. 16 SGBV cases comparison

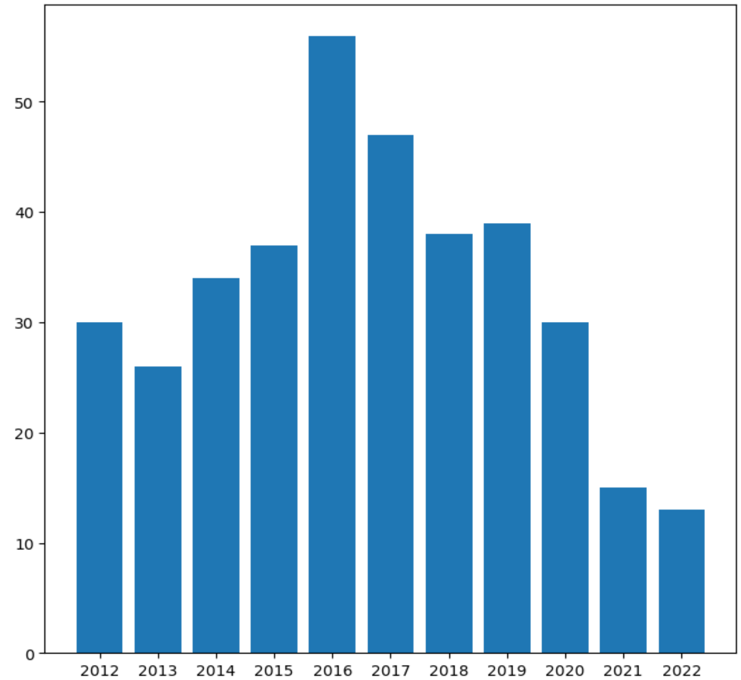

Figure 5. 17 Yearly Evolution of SGBV cases

Based on the dataset, Figure 5.17 shows how the number of SGBV cases has changed over time and in what ways. The year 2013 was an election year, and the number wasn't as high as the subsequent election years, including 2017 and 2018, which recorded the highest number of cases over the years. These two years were election years, and cases of post-election violence were experienced. 2022 was also an election year but had the lowest number of SGBV cases reported, this is a positive indication that policies have been put in place by the government to improve security during these periods.

It was worth analyzing the monthly comparison of SGBV cases across the years. This informed me about the patterns that were emerging in similar months but different years when it came to occurrences of violation cases. Figure 5.18 depicts the months of June, September, and October as having a higher number of cases compared to the rest of the months between the years 2012 and 2022. Further analysis might show reasons why this is the case.

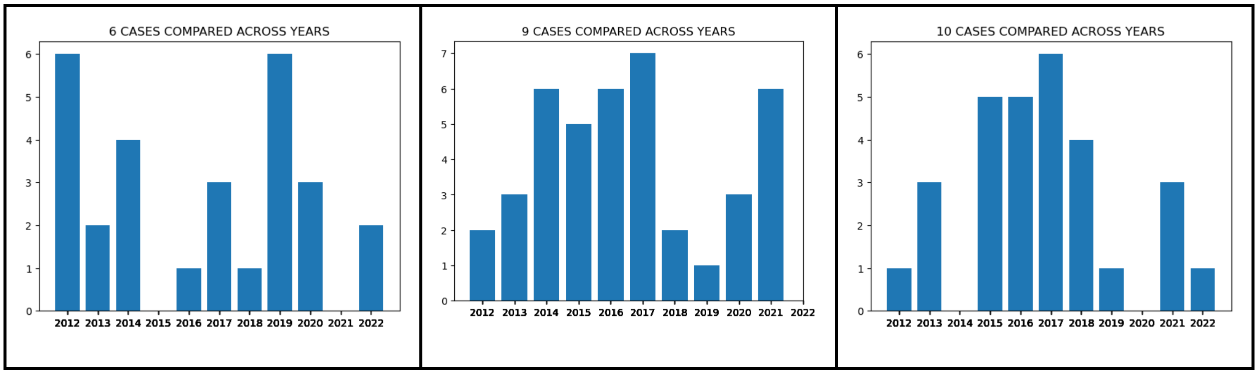

Figure 5. 18 Similar Monthly Evolution in Different Years

The dataset was looked at to see which counties in the country reported the most cases. Figure 5.19 illustrates this further.

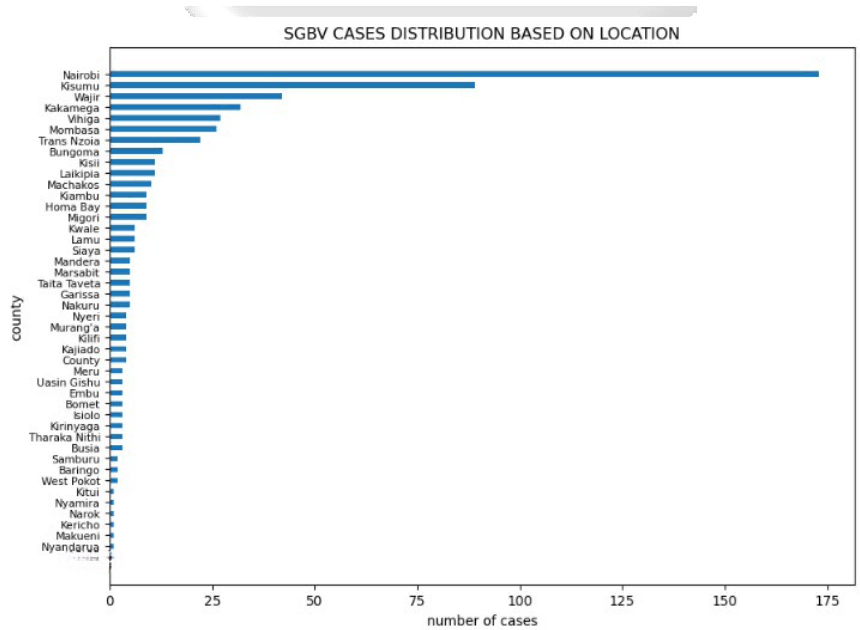

Figure 5. 19 SGBV Cases distribution based on location

SGBV affects all genders, and many times the assumption is made that women are the only victims of SGBV. However, that isn't the true representation; male victims suffer from stigma and therefore don’t report when the violations occur. The same is reiterated in Section 4.1.2. Figure 5.20 depicts the distribution of the cases based on gender, with 58.6% females, 41.2% males, and 0.17% intersex persons.

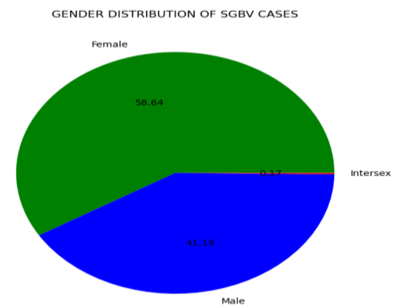

Figure 5. 20 Gender distribution of SGBV cases

It is noted with great concern that victims of SGBV are mostly children, and these cases occur mostly during holiday periods when schools are closed, with the alleged perpetrators being relatives and neighbours. Therefore, the researcher found it useful to do a holiday analysis, and from Figure 5.21, the researcher concluded that for holidays, especially those in 2017, most of the SGBV cases occurred during this period.

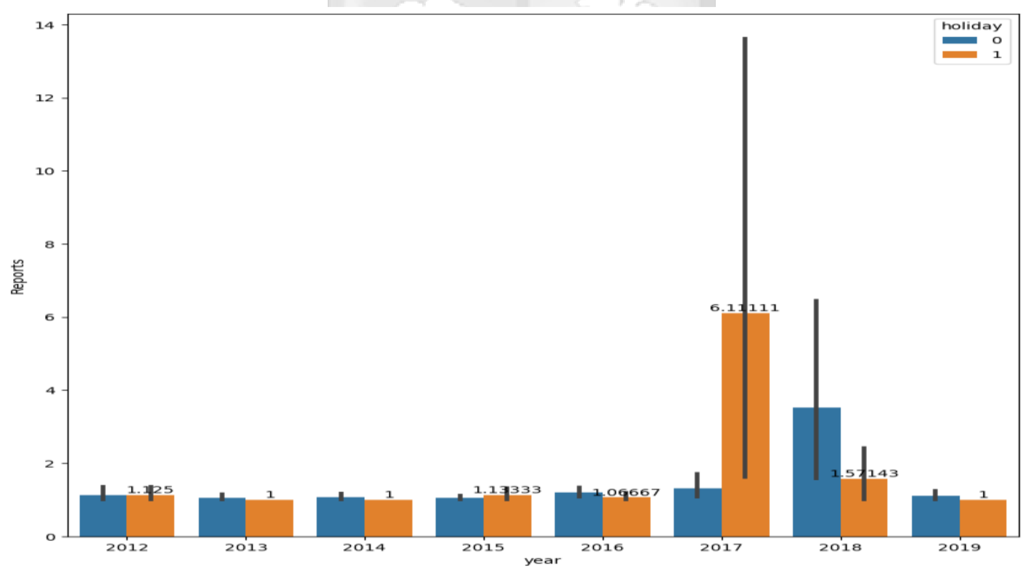

Figure 5. 21 Holiday Analysis

5.3.4.5 Estimation of Poisson Regression

The Poisson regression was used to model SGBV cases with respect to location, as a function of gender, and using month and year as predictor variables. Figure 5.22 illustrates the code for estimating SGBV cases using the month as the predictor variable; Figure 5.23 depicts the code for estimating SGBV cases using the year as the predictor variable; Figure 5.24 depicts the modeling of SGBV cases as a function of gender; and Figure 5.25 depicts the modeling of SGBV cases with respect to location.

```
day['month'] = pd.Categorical(day['month'])
day["year"]=pd.Categorical(day['year'])
# Fit the Poisson regression model
model1 = sm.GLM(np.asarray(day['Reports']), sm.tools.categorical(np.asarray(day['month'])),
family=sm.families.Poisson()).fit()
# Print the summary of the model
print(model1.summary())
```

Figure 5. 22 Estimation of SGBV cases using month as predictor variable

```
# Fit the Poisson regression model
model2 = sm.GLM(np.asarray(day['Reports']), sm.tools.categorical(np.asarray(day['year'])),
family=sm.families.Poisson()).fit()
# Print the summary of the model
print(model2.summary())
```

Figure 5. 23 Estimation of SGBV cases using year as predictor variable

```
gender_df.rename(columns={"gender":'gen_category'},inplace=True)
gender_df["gen_category"]= pd.Categorical(gender_df["gen_category"])
gender_df['gen_category'] = pd.Categorical(gender_df['gen_category'])
gender_df.dropna(inplace=True)
encoded_gender_df = pd.get_dummies(gender_df,columns=["gen_category"])
# calculate VIF values for each variable
vif2 = [variance_inflation_factor(encoded_gender_df.loc[:,'gen_category_Female':].values, i)
for i in range((encoded_gender_df.loc[:,'gen_category_Female':].shape[1]))]
# identify variables with high VIF values
high_vif1 = [encoded_gender_df.loc[:,'gen_category_Female':].columns[i] for i, v in
enumerate(vif2) if v > 5]
# fit Poisson regression with Firth's correction
model2=sm.GLM(encoded_gender_df[["cases"]],X1,
family=sm.families.Poisson(link=sm.families.links.log()))
model_fit2 = model2.fit_regularized(method='elastic_net', alpha=0.5, L1_wt=1.0)
```

Figure 5. 24 Modeling of SGBV cases as a function of gender

Table 5. 1 Gender Coefficients

Gender	Coefficient
Female	5.812637
Male	5.457456

Table 5.1 confirms that the female gender is more susceptible to SGBV occurrences such that by being of the female gender 5.8 SGBV cases occur, whereas for males, 5.5 SGBV cases occur.

```
location_df['location'] = pd.Categorical(location_df['location'])
location_df.dropna(inplace=True)
encoded_location_df = pd.get_dummies(location_df,columns=["location"])
from statsmodels.stats.outliers_influence import variance_inflation_factor
# calculate VIF values for each variable
vif = [variance_inflation_factor(encoded_location_df.loc[:,'location_Baringo':].values, i) for i in range((encoded_location_df.loc[:,'location_Baringo':].shape[1]))]
# identify variables with high VIF values
high_vif = [encoded_location_df.loc[:,'location_Baringo':].columns[i] for i, v in enumerate(vif) if v > 5]
# fit Poisson regression with Firth's correction
model=sm.GLM(encoded_location_df[["cases"]],X,
family=sm.families.Poisson(link=sm.families.links.log()))
model_fit = model.fit_regularized(method='elastic_net', alpha=0.5, L1_wt=1.0)
```

Figure 5. 25 Modeling of SGBV cases with respect to location

Table 5. 2 Location Coefficients

Location	Coefficient
Kakamega	2.302585
Kisumu	4.204693
Mombasa	1.386294
Nairobi	5.017280
Vihiga	1.609438
Wajir	2.995732

From Table 5.2 the researcher concluded that when it comes to these 6 counties SGBV cases are more prevalent, the estimated regression would be best summarised as:

cases = 1+5.02 Nairobi+4.20 Kisumu+2.996 Wajir+2.30 Kakagema+1.61 Vihiga+1.39 Mombasa

The above regression made the researcher conclude that by virtue of being in Nairobi, Kisumu, Wajir, Kakamega and Mombasa 5, 4, 3, 2 and 1.4 SGBV cases occur.

5.4 System Testing

System testing's purpose is to examine the full workings of the system against the requirements and objectives before delivery of the final product.

5.4.1 Functionality Testing

Functional testing was done to make sure that the prototype of the application meets the functional requirements that were set up for. The following tests were carried out:

Table 5. 3 Functional Testing

Test Case Description	Action	Expected Output	Result
User Registration	1. User launches the application. 2. User clicks on the account type they are creating an account as, then clicks on the No Account button to create an account in the system. 3. The user inputs the required details and click on Register button.	1. The system will verify user details provided and thereafter provide access. 2. The user registers and logs in successfully.	Pass
User Login	1. User initiates the application. 2. User clicks on the Login button.	1. The system will verify the username and password validity and thereafter provide access.	Pass

		2. The user logs in successfully. .	
Victim Reporting a violation	1. User (Victim) launches the application. 2. User (Victim) clicks on the Login button. 3. User (Victim) can view the reporting module and be able to fill the section with required information.	Users are able to report a violation successfully.	Pass
Medical Practitioner viewing and updating information	1. User (Medical Practitioner) launches the application. 2. User clicks on the Login button. 3. Users can view the information documented in the system and can update the reports with the necessary documentation i.e PRC forms.	User is able to view and update violation information	Pass
Law Enforcement Officer viewing and updating information	1. User (Law Enforcement Officer) launches the application. 2. User clicks on the Login button. 3. Users can view the information documented in the system and can update	User is able to view and update violation information	Pass

	the reports with the necessary documentation i.e P3 forms and OB Numbers.		
Pro-bono Lawyers viewing information	1. User (Pro-bono Lawyers) launches the application. 2. User clicks on the Login button. 3. Users can view the information documented in the system and can assist the victims access justice.	User is able to view the violation information	Pass
Send Recommendations/ Feedback	1. The application is launched by the user. 2. User clicks on the Login button. 3. The system should be able to send feedback to the victim based on the reported violation.	User is able to receive recommendations/ feedback	Pass
Reports Generation	1. The application is launched by the user. 2. User clicks on the Login button. 3. Users should be able to use the system to view the information and reports for each center.	User is able to generate reports successfully	Pass

5.4.2 Compatibility Testing

Different browsers were used to test the web application and they included Microsoft Edge, Chrome, and Mozilla Firefox. The mobile application was tested on different mobile devices to confirm its compatibility such as Nokia, Samsung, and Oppo.

Chapter 6: Discussion

6.1 Introduction

This chapter discusses how the research objectives contributed to the research topic. The overarching goal of the study was to design an application that outlines the developed solution with regard to how the research questions were addressed, how timely reporting of sexual and gender-based violence (SGBV) violations is achieved, and to employ machine learning to do pattern analysis of SGBV occurrences in order to support duty bearers in improving their policies.

6.2 Review of Research Objectives

Referring to Section 1.3, the first goal was to find the problems with the current ways of collecting data on SGBV. The results show that it is hard to collect, manage, and store data on SGBV cases because most of the processes are done manually. Based on the answers to questionnaires sent to a sample of police stations and hospitals in Nairobi, the respondents said that data collection is done manually because victims have to go to the police station in person to report a violation. The records are kept as case files in registries and exhibit rooms, and they are also written down in occurrence books. This makes data management a time-consuming process. As for the hospitals, the situation is somewhat similar; the information collected is documented in Post Rape Care (PRC) forms, and copies are made and stored in the registry and on the system. The respondents mentioned that the copies documented in the system tend to be less clearly visible, and for the physical copies, as the years go by, they also lose clarity. Therefore, if a victim walks into the centers in ten years and asks for their records, it may be a challenge to access them. The mobile application is as beneficial to the victim as it is to the police stations and hospitals, as it integrates the three most crucial entities needed to complete the cycle of SGBV evidence collection by ensuring data collection and management are improved to ensure efficiency and effectiveness in handling these sensitive cases and makes it easier to track down the perpetrators.

The second goal was to look at the methods that are already in place for reporting cases of sexual and gender-based violence. The findings illustrated that victims usually report the violations physically at the police stations. Some victims also opt not to report the violations as they are discouraged due to certain factors such as bribery by the perpetrators, harsh judgments, stigma, fear, and alternative dispute resolution mechanisms. The application will encourage victims of

SGBV to report the violations timely and bypass some of the challenges of reporting mentioned above.

The third goal was to develop a mobile app so that cases of sexual and gender-based violence could be reported quickly. In order to design and develop an appropriate solution, the findings of the study were used to identify the appropriate technologies to use based on the literature review that was conducted in Section 2.2 onwards. Use case diagrams, sequence diagrams, data flow diagrams, entity relationship diagrams, and class diagrams were utilized as design tools. The application developed can help victims of SGBV report these violations. The application can be used by duty-bearers, including hospitals and police stations, to improve their data management process, which is currently a manual process.

The fourth goal was to make a machine learning model for looking for patterns in cases of sexual and gender-based violence. Chapter 2 included the review of various algorithms, and the research made use of topic modeling of the data to see the major topics arising, then modeled a Poisson Regression algorithm to form the pattern analysis. The model illustrated various trends and patterns, which are explained further in Section 5.3.3. The model used only location, year, month, and gender as its predictor variables.

The fifth objective was to test and validate the proposed tool. The Android platform was used to develop the mobile application, JavaScript programming language for the web application and the model was designed in Python using the Jupyter Notebook IDE. The tests that were conducted on the system included functional testing and compatibility testing including various web browsers accessing the web application and different devices accessing the mobile application.

6.3 Contribution of the Study

There are currently challenges to the under-reporting of SGBV violations due to a variety of reasons, as outlined in Section 4.1.5. The mobile application will increase the number of documented SGBV reports because it addresses the majority of the issues raised. When dealing with the SGBV issue, duty-bearers can make more informed decisions and plan better with real-time data. The ML model assists duty-bearers in maximizing the strategies they implement by demonstrating the trends and patterns of SGBV cases, allowing them to build improved policies to respond to and prevent SGBV.

6.4 Research Limitations

Due to the sensitive nature of the research study, the researcher had trouble getting information when collecting data. Then, the researcher had to write formal letters to the hospitals' management officers and the deputy inspector general to get permission to give the questionnaires at their centres. However, the approval from some of the hospitals took weeks, whereas other centres had many requirements before they approved the request, which caused a delay for the researcher. Many of the institutions have many requirements that don’t align with the academic processes, as they require approval from their internal ethics committee, which takes months, and approval at the county level which require payment to be made. Since this kind of information is sensitive, the information isn’t easily and readily available therefore the researcher had to work with a small dataset, which was a challenge. This meant that the researcher had to use the data that was available to model the patterns of the cases that had already been reported.

VT OMNES VNVM SINT

Chapter 7: Conclusion and Recommendations

7.1 Introduction

The researcher developed an Android-based mobile application that helps victims of sexual and gender-based violence report crimes as soon as they happen. The application also has an awareness platform that tells them what they need to do next. This chapter includes the recommendations, conclusions, and future work for this research.

7.2 Conclusions

The policy brief from the National Council for Population and Development and other groups talks about the main problems with SGBV and gives a multi-sectoral plan for dealing with this health and human rights issue more effectively. This policy brief shows that SGBV is common and has big effects on health and development, but not enough attention has been paid to it at all levels. So, it strongly encourages actors to take part in prevention and care, and it suggests ways to make the Sexual Offenses Act better enforced. The magnitude of SGBV is difficult to estimate. It is commonly known, however, that documented cases only reflect a small portion of the wider reality. Sexual violence goes unreported even in normal circumstances because of fear, humiliation, impotence, a lack of support, or the unreliability of public services.

The main goal of this study was to develop a mobile application that would help victims of sexual and gender-based violence report crimes to the nearest police stations and connect them to the nearest hospitals so they could get help right away. The researcher has reached this objective as well as all of the other objectives listed in Section 1.3. The SGBV victims will benefit from the mobile app because they won't have to go to the police station to report the crime. This will help solve the problem of people not reporting crimes. The ML model that shows the trends and patterns of the cases that have been reported will help governmental and non-governmental organizations plan ahead for better resource optimization and comprehend the magnitude of SGBV in order to mitigate future incidents.

7.3 Recommendations

Policymakers and people in charge of making decisions should use the patterns and trends shown by the reported cases. Addressing the serious issue of under-reporting of SGBV cases because of shame, stigmatization, fear, and re-victimization is minimized by the use of the mobile application to report the violations without physically visiting the police station. The mobile application should be publicized to the general public to ensure more people are aware of its existence and encourage them to make use of it when a need arises. Duty bearers should make use of the data collected to come up with policies that protect children at all times, especially during holidays when they are at home. These policies should also know the persons at high risk of SGBV and take them into account in everything to ensure their safety. The government and non-governmental organizations should support community-based organizations in the provision of services or initiatives that are beneficial to SGBV survivors at the community level.

7.4 Future Works

The future improvements that can be implemented include the following:

i. Since SGBV is a problem all over the country that needs to be fixed, more research should be done in Nairobi County, and the proposed solution should be put into place in other parts of the country.
ii. The proposed solution should be extended in the future to include the development of the application on other platforms, such as IOS and Windows, to target other users who don't use the Android platform.
iii. For the Poisson regression model used in this research and how to improve it to get a better answer, technological progress should be thought about.
iv. For this study, the Poisson regression was used to infer variables that influence SGBV based on the dataset; however, the model can be improved to also predict future occurrences.
v. Integrate the solution into police stations and hospitals so that the ecosystem is complete and the goal of making the world a better place is met.
vi. Disseminating the pattern analysis data to the duty bearers to aid them in key decision-making and policy formulation. The solution can be improved by using data analytics techniques.
vii. Creation of awareness of the existence of the solution among the general public. This will improve its accessibility, and its impact will be felt.

References

Abdulkareem, L. & Karan, O. (2022). Using ANN to Predict Gender-Based Violence in Iraq. *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. DOI: 10.1109/ISMSIT56059.2022.9932831

Aftab, Cheung, Kim, Thakkar, & Yeddanapudi. *Information Theory & The Digital Revolution.* Retrieved from https://web.mit.edu/6.933/www/Fall2001/Shannon2.pdf

Ball, K. (2022). *New Technologies for Combatting Sexual Violence In Conflict and Non-conflict Settings. Knowledge, evidence and learning for development (K4D).* Retrieved from: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/17664/1174%20New%20Tech%20and%20CRSV.pdf?sequence=1

Bank, M. (2012). *A Survey of Text Mining Architectures and the UIMA Standard.* Retrieved from:https://www.researchgate.net/publication/237021021_A_Survey_of_Text_Mining_Architectures_and_the_UIMA_Standard

Bello, H., Palomar, N., Gallego, E., Navascuez, L., & Lozano, C. (2020). *Machine Learning to study the impact of gender-based violence in the news media.* Retrieved from: https://www.researchgate.net/publication/347125860_Machine_Learning_to_study_the_impact_of_gender-based_violence_in_the_news_media

Book, R. Formal Language Theory and Theoretical Computer Science. *New York: Springer International Publishing.* Retrieved from https://link.springer.com/content/pdf/10.1007%2F3-540-07407-4_1.pdf

Bplotkin, J. & Anowak, M. (2000). Language Evolution and Information Theory. *Journal of Theoretical Biology Volume 205, Issue 1, 7 July 2000, Pages 147-159.* https://doi.org/10.1006/jtbi.2000.2053

Cengage (2018). *Information Processing*. Retrieved from: https://www.encyclopedia.com/science-and-technology/computers-and-electrical-engineering/computers-and-computing/information-processing

Chakravarthy, S. (2020). *Tokenization for Natural Language Processing.* Retrieved from: https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4

Choi, I., Song, H., & Yoo, J. (2019). Deep Learning Based Pedestrian Trajectory Prediction Considering Location Relationship between Pedestrians. *1st International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2019, 449–451. Department of Statistics Malaysia Official Portal. (n.d.).* https://www.statistics.gov.my/index.php?r=column/cone&menu_id=dDM2enNvM09oTGtQemZPVzRTWENmZz09

Computer Science Field Guide (2013). *Formal Languages: Regular Expressions.* Retrieved from: https://www.csfieldguide.org.nz/en/chaters/formal-languages/regular-expressions/

Croce, D., Castellucci, G. (2020). GAN-BERT: *Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples.* Retrieved from https://aclanthology.org/2020.acl-main.191.pdf

Dang, S. (2014). Text Mining: Techniques and its Applications. *IJETI International Journal of Engineering & Technology Innovations, Vol. 1 (4). ISSN (Online): 2348-0866*

Dakalbab, F., Talib, M., Waraga, O., Nassif, A., Nassir, Q, & Abass, S. (2022). Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open 6(1), 2022 https://doi.org/10.1016/j.ssaho.2022.100342*

Dinama, D. M., A'Yun, Q., Syahroni, A. D., Adji Sulistijono, I., & Risnumawan, A. (2019). Human Detection and Tracking on Surveillance Video Footage Using Convolutional Neural Networks. *IES 2019 - International Electronics Symposium: The Role of Techno-Intelligence in Creating an Open Energy System Towards Energy Democracy, Proceedings*, 534–538.

Drew, C. (2022). S*hannon Weaver Model of Communication – 7 Key Concepts*. Retrieved from https://helpfulprofessor.com/shannon-weaver-model/#:~:text=The%20Shannon%20and%20Weaver%20Model%20of%20Communication%20is,concepts%3A%20sender%2C%20encoder%2C%20channel%2C%20noise%2C%20decoder%2C%20and%20receiver

Edeki, C. (2015). *Agile Software Development Methodology. European Journal of Mathematics and Computer Science.* Retrieved from: https://www.idpublications.org/wp-content/uploads/2015/05/Agile-Software-Development-Methodology.pdf

Eftimov, T., Seljak, B., Korosec, P. (2017). *A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations.* Retrieved from: https://doi.org/10.1371/journal.pone.0179488

Farnoud, F. (2019). *An Introduction to Information Theory.* Retrieved from http://ece.virginia.edu/~ffh8x/docs/teaching/Information%20Theory.pdf

Flowers, R., Sparks, D., Sitton, S., & Cook, J. (1980). *Prediction of Accident Likelihood using Poisson Regression.* Retrieved from https://static.tti.tamu.edu/tti.tamu.edu/documents/TTI-1980-3.pdf

Guedim, Z. (2020). *5 Theories of Language Development: How Languages Came to be.* Retrieved from https://edgy.app/theories-of-language-development

Han, J., Kim, G., Lee, C., Han, Y., Hwang, U., & Kim, S. (2019). Predictive Models of Fire via Deep learning Exploiting Colorific Variation. *1st International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2019, 1997,* 579– 581.

Ilamvaluthi, S., Bala, C., Harish, T. & Mohan, R. (2017). A Survey of Data Mining Techniques for Analyzing Crime Patterns. *International Journal of Engineering Science and Computing (IJESC) 7(3).* Retrieved from: https://ijesc.org/upload/78a7afea0768dcab05be2b37731764c5.A%20Survey%20of%20Data%20Mining%20Techniques%20for%20Analyzing%20Crime%20Patterns.pdf

Ilhamsyahids, S. (2020). *Basic Theories of Natural Language Processing (NLP).* Retrieved from https://ilhamsyahids.com/blog/2020/07/20/basics-theories-of-nlp/

JavaTpoint. *Agile Methodology*. Retrieved from: https://www.javatpoint.com/agile-vs-waterfall-model

Jilcha, K. (2019). *Research Design and Methodology.* Retrieved from: DOI:10.5772/intechopen.85731

Kenya: *Survivors of Gender-Based Violence Lack Help. (2021, September 21).* Human Rights Watch. https://www.hrw.org/news/2021/09/21/kenya-survivors-gender-based-violence-lack-help

VT OMNES VNVM SINT

Karan, O., & Abdulkareem, L. (2022). Using ANN to Predict Gender-Based Violence in Iraq. *IEEE Xplore*. https://doi.org/10.1109/ISMSIT56059.2022.9932831

Kenya National Commission on Human Rights (2017). *Silhouettes of Brutality: An account of sexual violence during the 2017 General Election.* Retrieved from: https://www.knchr.org/Publications/Thematic-Reports/Civil-and-Political-Rights/Election-Related-Reports

Kim, M. C., Koo, J. H., Cho, S. W., Baek, N. R., & Park, K. R. (2018). Convolutional neural network-based periocular recognition in surveillance environments. *IEEE Access*, *6*, 57291–57310.

Kobiruzzaman, M. (2019). *Communication Noise- Types of Noise Barriers in Communication.* Retrieved from: https://newsmoor.com/communication-noise-5-types-of-noise-in-communication-barriers/

Kumar, L. & Bhatia, P. (2013). Text Mining: Concepts, Process and Applications. *Journal of Global Research in Computer Science, 4(3) ISSN 2229-371X.Lanigan, R. (2013).* Information Theories: Theories and Models of Communication. Retrieved from: ISBN: 978-3-11-024044-3.

Lynch, A. (2022). *Algorithms.* Retrieved from: https://www.edrawsoft.com/algorithm-definition.html

Manzoor, A., Hassan, S., Muazzam, A., Tuarob, S., & Nawaz, R. (2021). Social mining for sustainable cities: thematic study of gender-based violence coverage in news articles and domestic violence in relation to COVID-19. *Journal of Ambient Intelligence and Humanized Computing.* https://doi.org/10.1007/s12652-021-03401-8

Mohajan, H. (2017). *Research Methodology.* Retrieved from: https://www.researchgate.net/publication/321964409_Research_Methodology

Momoh, O. (2022). *Statistics and How to Measure It*. Retrieved from: https://www.investopedia.com/terms/p/population.asp#:~:text=A%20population%20is%20a%20distinct%20group%20of%20individuals%2C,feature%20can%20be%20said%20to%20be%20a%20population. Investopedia.

National Gender and Equality Commission (2017). *County Government Policy on Sexual and Gender Based Violence.* Retrieved from https://www.ngeckenya.org/Downloads/NGEC%20Model%20Policy%20on%20GBV%20for%20County%20Govts.pdf

Ningrum, P. & Pansombut, T. (2020). *Text mining of online job advertisements to identify direct discrimination during the job-hunting process: A case study in Indonesia.* Retrieved from: https://doi.org/10.1371/journal.pone.0233746

Nnoko-Mewanu, J. (2021). *"I Had Nowhere to Go."* HumanRights Watch. https://www.hrw.org/report/2021/09/21/i-had-nowhere-go/violence-against-women-and-girls-during-covid-19-pandemic-kenya

Ozdenizci, B., Kerem, K. & Coskun, V. (2016) *A Tokenization-Based Communication Architecture for HCE-Enabled NFC Services.* Mobile Information Systems. Retrieved from: https://doi.org/10.1155/2016/5046284

Petering, R., Um, M., Kumari, R., & Alipourfard, N. (2018). *Artificial intelligence to predict intimate partner violence perpetration.* Retrieved from https://www.researchgate.net/publication/329281099

Pramanik, M., Lau, R., Yue, W., Ye, Y., & Li, C. (2017). *Big data analytics for security and criminal investigations.* Retrieved from: https://doi.org/10.1002/widm.1208

Purwania, I., Kumara, N., & Sudarma, M. (2020). Systematic Review of Text Mining Application Using Apache UIMA. *International Journal of Engineering and Emerging Technology, Vol.5(2). P-ISSN: 2579-597X, E-ISSN: 2579-5988.*

Ramkumar, R. & Ramesh, R. (2019). Architecture and implementation of a Text Mining application for financial documents for Early Warning Systems. *International Journal of Applied Engineering Research Volume 14(5). ISSN 0973-4562.*

Rape, Abuse and Incest National Network. Victims of Sexual Violence Statistics. Retrieved from https://www.rainn.org/statistics/victims-sexual-violence

Rasanayagam, K., Kumarasiri, S. D. D. C., Tharuka, W. A. D. D., Samaranayake, N. T., Samarasinghe, P., & Siriwardana, S. E. R. (2018). CIS: An Automated Criminal Identification System. *2018 IEEE 9th International Conference on Information and Automation for Sustainability, ICIAfS 2018.*

VT OMNES VNVM SINT

Reshma, G., Dharani, D., & Rani, K., (2021). Machine Learning Based Regression Analysis on Women Safety in India. *International Journal of Advance Research in Science and Engineering (IJARSE) 10(01). ISSN 2319-8354.*

Rodriguez, I., Rodriguez, J., Quiles, D., Gonzalez, P., & Chatzigiannakis, I. (2020). Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques. *Applied Science 2020, 10(22), 8244; https://doi.org/10.3390/app10228244*

Roopa, & Prof. Khan, T. (2022). Crime Analysis and Prediction Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)* 9(9). Retrieved from: https://www.irjet.net/archives/V9/i9/IRJET-V9I9162.pdf

Rowe, W., & Johnson, J. (2020). *Top Machine Learning Frameworks to Use.* Retrieved from: https://www.bmc.com/blogs/machine-learning-ai-frameworks/

Sathya, S. & Rajendran, N. (2015). A Review on Text Mining Techniques. *International Journal of Computer Science Trends and Technology (IJCST) – Vol.3(5).* Retrieved from: http://www.ijcstjournal.org/volume-3/issue-5/IJCST-V3I5P40.pdf

ScienceDirect. *Language Theory.* Retrieved from: https://www.sciencedirect.com/topics/social-sciences/language-theory

Scott, L. (2009). *Formal Theory.* ScienceDirect. Retrieved from: https://www.sciencedirect.com/topics/computer-science/formal-language

Sexual and gender-based violence (SGBV) prevention and response UNHCR|Emergency Handbook. (n.d.). Retrieved from: https://emergency.unhcr.org/entry/60283/sexual-and-gender-based-violence-sgbv-prevention-and-response

Shaik, N., Krishna, S. & Kumar, P. (2022). Crime Analysis and Prediction using Machine Learning. *International Journal of Innovative Science and Research Technology 7(6). ISSN No:2456-2165*

Shako, K. & Kalsi, M. (2019). *Forensic observations and recommendations on sexual and gender-based violence in Kenya.* Retrieved from: https://doi.org/10.1016/j.fsisyn.2019.06.001

Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656.*

Shilpa, D. & Ahmad, H. (2015). A Review of Text Mining Techniques Associated with Various Application Areas. *International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064*

Shilpa, D. & Ahmad, H. (2014). Text Mining: Techniques and its Application. *IJETI International Journal of Engineering & Technology Innovations. ISSN (Online): 2348-0866*

Snoderly, J. & Faisandier, A. (2022). *System Implementation. Systems Engineering Body of Knowledge (SEBOK).* Retrieved from https://sebokwiki.org/wiki/System_Implementation

Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal of Healthcare Engineering.* Retrieved from: https://doi.org/10.1155/2018/4302425

Thongsatapornwatana, U. (2018). *A survey of data mining techniques for analyzing crime patterns.* DOI: 10.1109/ACDT.2016.7437655

Toal, R. (2017). *Theories in Computer Science.* Retrieved from: https://cs.lmu.edu/~ray/notes/cstheories/

Violence against women. (2021). https://www.who.int/news-room/fact-sheets/detail/violence-against-women

Vladislav, P & Stanislav, P. (2017). *Comparison of methods of data mining techniques for predictive accuracy.* Retrieved from https://mpra.ub.uni-muenchen.de/79326/

Wado, Y. D. (2021). *Violence against women in Kenya: Data provides a glimpse into a grim situation. The Conversation.* http://theconversation.com/violence-against-women-in-kenya-data-provides-a- glimpse-into-a-grim-situation-170109

Wanjiru, Q. (2021). Causes and Effects of Gender-Based Violence. A Critical Literature Review. *Journal of Gender Related Studies, 2(1), Article 1. https://doi.org/10.47941/jgrs.742*

World Health Organization. Sexual Violence. Violence Info. Retrieved from: https://apps.who.int/violence-info/sexual-violence/

VT OMNES VNVM SINT

Xia, Y., Zhang, B., & Coenen, F. (2016). Face occlusion detection based on multi-task convolution neural network. *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2015*, 375– 379.

Xie, L., Buntine, W., Sanner, S., & Mehrotra, R. (2013). *Improving LDA topic models for microblogs via tweet pooling and automatic labeling.* Retrieved from: https://doi.org/10.1145/2484028.2484166

Xu, H. &Yanqing, L. (2022). *Mining and Application of Tourism Online Review Text Based on Natural Language Processing and Text Classification Technology.* Retrieved from: https://doi.org/10.1155/2022/9905114

Zamria, N., Tahira, N., Ali, M., Ashad, N., & Misrebe, A. (2021). Mini-review of Street Crime Prediction and Classification Methods. *Journal Kejuruteraan 33(3).* https://doi.org/10.17576/jkukm-2021-33(3)-02

Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., … Shen, B. (2012). Biomedical text mining and its applications in cancer research. *Journal of Biomedical Informatics.* Retrieved from: https://doi.org/10.1016/j.jbi.2012.10.007

VT OMNES VNVM SINT

Appendices

Appendix A: Use Case

Table 4.1 Use case description for adding a new user

Use Case 1	Registering as a new user.
Description	New user registers to the system.
Actor	User.
Preconditions	The user must be a SGBV survivor or assisting the survivor in reporting.
Postconditions	The user is logged into the application.
Main success scenario	User registers in the application. System records user details. User logins to the app.

Table 4.2 Use case description for reporting a violation

Use Case 2	Report a new violation.
Description	User reports a violation.
Actor	User.
Preconditions	The user must be a SGBV survivor or assisting the survivor in reporting. The user must be a registered user.
Postconditions	Reporting a violation.

VT OMNES VNVM SINT

Main success scenario	User registers in the application. System records user details. User logins to the application. User reports a violation. User receive feedback on what actions they should take.

Table 4.3 Use case description for viewing and updating information

Use Case 3	**Viewing Violation Information**
Description	The users can view information
Actor	Victim, Law Enforcement Officers, Medical Officers, and Pro-bono lawyers.
Preconditions	The user is logged in.
Postconditions	Viewing Violation Information
Main success scenario	User is registered and logged in. User (Victim) is able to view the violation information. Users (Medical officers and Law enforcement officers) are able to view the violation information. Users (Medical officers and Law enforcement officers) are able to update the information with necessary documentation.

Table 4.4 Use case description for reports generation

Use Case 4	**Generate Reports**
Description	Generation of the reports by the administrator.
Actor	Law Enforcement, Medical practitioners, and Pro-bono lawyers.
Preconditions	The actors are logged into the system, they can view reports beneficial to them.
Postconditions	Report is generated.
Main success scenario	Actors log into the application. Actors can view the customized reports specific to their centers. Actors can download the reports for their further action.

Table 4.5 Use case description for managing users

Use Case 5	**Managing Users**
Description	The administrator manages the users.
Actor	System Administrator.
Preconditions	The administrator is logged in.
Postconditions	User management

Main success scenario	Admin approves requests to add users. Admin updates/edits the system. Admin can delete users from the system. Admin sends feedback to the user on actions to take. Admin sends alerts or notifications to the users.

VT OMNES VNVM SINT

Appendix B: Questionnaire for Medical Officers

Demographic Data.

Name of Hospital or Center	
Location of the Hospital or Center	
Designation of Respondent	
Sex of Respondent	

How frequent do you receive victims of SGBV at your center? Please mark appropriate box with a “ ✓ ”

On a daily basis	
On a weekly basis	
On a monthly basis	

Which form or type of SGBV is usually experienced?

…………………………………………………………………………………

Based on the data collected, after how long can you say victims of SGBV seek medical assistance? Please mark appropriate box with a “ ✓ ”

Less than 72 hours	
Between 73hours – 1 week	
After 1 month	
Between 2 months – 6 months	
After one year or more	

After examination of the victim, how do you store the data collected?

……………………………………………………………………………………………………

How do you preserve the information and evidence collected for future access?

……

Do existing services to respond to SGBV survivors address needs of survivors who are living with disability? Please explain

……

Have you ever been asked to present the medical examinations in court and what was the outcome? Please explain. …………………………………………….

Appendix C: Questionnaire for Law Enforcement Officers

Demographic Data

Name of Police Station	
Location of Police Station	
Designation of Respondent	
Sex of Respondent	

How frequent do you receive victims of SGBV at your center? Please mark appropriate box with a " ✓ "

On a daily basis	
On a weekly basis	
On a monthly basis	

Which form or type of SGBV is usually reported?

………………………………………………………………………………

Based on the data collected, after how long can you say victims of SGBV report the violation? Please mark appropriate box with a " ✓ "

Less than 72 hours	
Between 73hours – 1 week	
After 1 month	
Between 2 months – 6 months	
After one year or more	

When a victim reports the violation, how do you store the data collected?……………………………………………………………………………………………………

How do you preserve the information and evidence collected for future access?

……

Do existing services to respond to SGBV survivors address needs of survivors who are living with disability? Please explain

Do the alleged perpetrators get arrested and undergo prosecution?

..…………………………………

Appendix D: Questionnaire for Community-Based Organizations

1. Demographic Data

Name of Community	
Designation of Respondent	
Sex of Respondent	

2. What are the situations that pre-dispose people to Sexual and Gender-Based Violence (SGBV) in this community?

   ……………………………………………………………………………………………

3. What forms of SGBV do you think occur the most in this community?………………………………………………………………………………………

4. Which people are mostly at risk of gender base violence in this community?………………………………………………………………………………………

5. What are the specific forms of SGBV against women and girls in this community?

   ………………………………………………………………………………………………

6. What are the specific forms of SGBV against men and boys in this community?

   ………………………………………………………………………………………………

7. What do people in this community do to protect themselves from gender-based violence and how do they respond to these incidents?

   ………………………………………………………………………………………………

8. Has the problem of SGBV in this community gotten worse, better, or stayed the same in the last year? ……………………………………………………………………………….

9. Do survivors seek for help when they experience SGBV? Do they tell anyone?

   ………………………………………………………………………………..

10. What barriers do the survivors face in seeking care?

    …………………………………………………………………………………..

11. What barriers survivors face in reporting SGBV in this community?

    ……………………………………………………………………………………………

12. What services are typically available to SGBV survivors in your community?

    ………………………………………………………………………………..

13. Do existing services to prevent or respond to SGBV survivors address needs of survivors who are living with disability? Please explain

    ………………………………………………………………………………

14. What is the quality and accessibility of the existing SGBV services in this community?

Appendix E: Questionnaire for Victims

Section A: Demographic Data. Please mark appropriate box with a " ✓ "

Age in years

Less than 18 years	
18- 35 years	
36- 60 years	
60 and above	

Gender

Male	
Female	

Marital Status

Single	
Married	
Separated	
Divorced	
Widowed	

Education Level

Primary Level	
Secondary Level	
Tertiary College Level	
University Level	

Do you have any form of disability?

Yes	
No	

Nationality

Kenyan	
Non- Kenyan	

Have you ever experienced any form of sexual violence?

Yes	

VT OMNES VNVM SINT

No	

If yes, what type or form of SGBV was it?

………………………………………………………………………………………

How long ago did this happen?

Less than 1 year	
2-5 years	
6-10 years	
More than 10 years	

Did you seek medical assistance?

Yes	
No	

Did you report the occurrence to the police?

Yes	
No	

After how long did you seek medical assistance?

Less than 72 hours	
Between 73hours – 1 week	
After 1 month	
Between 2 months – 6 months	
After one year or more	

After how long did you report the violation to the police?

Less than 72 hours	
Between 73hours – 1 week	
After 1 month	
Between 2 months – 6 months	
After one year or more	

Section B: Knowledge about Prevention and Control

If you didn't report the violation to the police, what are your reasons? ………………………

If you didn't seek medical assistance, what are your reasons? …………………………………..

If you reported, which methods or channels did you use? ………………………………………………..

Do you have basic awareness on what to expect from the medical officers or police officers handling you at first contact?...

Mobile Phone Access

Do you have a mobile phone? Tick as appropriate.

Yes	
No	

VT OMNES VNVM SINT

Appendix F: Ethical Clearance Confirmation

22nd February 2023

Ms Makau Agnes Nzembi,

Dear Ms Makau,

RE: Sexual and Gender-Based Violence Pattern Analysis using Machine Learning: A Case of Nairobi County

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU- master's** research proposal. Your application reference number is **SU-ISERC1582/23.** The approval period is from **22nd February 2023 to 21st February 2024.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, and MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 48 hours of notification
iv. Any changes, anticipated or otherwise, that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 48 hours
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU- ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research- portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

for: **Dr Ben Ngoye, Secretary;SU-ISERC**

STRATHMORE UNIVERSITY INSTITUTIONAL SCIENTIFIC AND ETHICAL REVIEW COMMITTEE
(SU-ISERC)
Email:ethicsreview@strathmore.edu
P.O BOX 59857-00200
NAIROBI-KENYA
22-Feb-2023

Appendix G: NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Ref No: **186926**

Date of Issue: **02/March/2023**

RESEARCH LICENSE

This is to Certify that Ms.. Agnes Nzembi Makau of Strathmore University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Nairobi on the topic: Sexual and Gender-Based Violence Pattern Analysis using Machine Learning: A Case of Nairobi County for the period ending : 02/March/2024.

License No: **NACOSTI/P/23/24077**

186926

Applicant Identification Number

Director General

NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

National Commission for Science, Technology and Innovation -

Appendix H: Approval Letter by KNCHR

Kenya National Commission on Human Rights

1st Floor, CVS Plaza, Lenana Rd, P.O. Box 74359 - 00200 Nairobi - Kenya. Tel: +254-020-3969000
Mobile: 0733 780 000; 0736 780 000; 0724 256 448; 0726 610 159 SMS: 22359

Website: www.knchr.org Email: haki@knchr.org Complaints: complaint@knchr.org.

Coast Region Office-Mombasa
Imaara building 10th Floor unit 10D
Dedan Kimathi Avenue Kizingo
Opposite - Pandya Memorial Hospital
P.O. Box 90171-80100, Mombasa
Email:coast@knchr.org
Twitter:@KNCHRMombasa

Western Region Office-Kisumu
3rd Floor, Re Insurance plaza
Oginga Odinga Street
P.O Box 7768-40100, Kisumu
Tel: 057-2020078
Email:kisumu@knchr.org
Twitter:@KNCHRkisumu

North Rift Regional Office – Kitale
AFC Building
Opposite Mega center – Nakumatt
P.O Box 299 - 30200, Kitale
Tel: 054- 31773/0203969057
Emial: northrift@knchr.org
Twitter: @KNCHRKitale

North Eastern Region – Wajir
Wajir-Bor Rd, near Public Works
P.O Box 363 - 70200, Wajir
Tel: 046-4421512
Email: nothernkenya@knchr.org
Twitter: @KNCHRWajir

Central Region Office
Agriculture Finance
Near Nyahururu Law Courts
P.O Box 650 - 20300
Tel: 0705 982 617
Email: laikipia@knchr.org
Twitter: @KNCHRLaikipia

REF: KNCHR/RECORDS/VOL01

FROM: COMMISSION SECRETARY

TO: MAKAU AGNES

DATE: 3rd March 2023

SUBJECT: REQUEST FOR APPROVAL TO USE KNCHR DATASET ON SEXUAL AND GENDER BASED VIOLENCE

The above subject matter refers.

Kindly be informed that your request to utilize the KNCHR dataset has been approved.

You will be expected to follow the guidelines, ensure anonymity, privacy and confidentiality of the information collected and use it only for purposes of the research.

Dr. Bernard Mogesa, PhD, CPM

Commission Secretary/ CEO

An accredited "A" Status National Human Rights Institution **Chairperson:** Roseline Odede HSC, **Vice Chair:** Dr. Nyeris Raymond, PhD **Commissioners:** Hon. Sara Bonaya, Dr. Dennis N. Wamalwa, PhD, Prof. Marion Mutugi, EBS

Commission Secretary / C.E.O: Dr. Bernard Mogesa, PhD, CPM

Appendix I: Similarity Index

SEXUAL AND GENDER-BASED VIOLENCE PATTERN ANALYSIS USING MACHINE LEARNING A CASE OF NAIROBI COUNTY.pdf

ORIGINALITY REPORT

23%	20%	8%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	su-plus.strathmore.edu Internet Source	5%
2	www.ukm.my Internet Source	2%
3	www.genderinkenya.org Internet Source	1%