

THE ANALYSIS OF ASSOCIATION BETWEEN VICTIM IDENTITIES AND EXECUTION FROM GENOCIDE DATA SET USING QUANTITATIVE ANALYSIS AND OCR TECHNIQUE

¹ TODSANAI CHUMWATANA, ² PUNNITA SAMBATH, ³ SOPHEAKNINE CHIM

^{1,2,3} Information and Communication Technology, International College, Rangsit University, Pathumthani, Thailand

E-mail: ¹ Todsanai.c@rsu.ac.th, ² Punnita.s63@rsu.ac.th, ³ Sopheaknine.c63@rsu.ac.th

ABSTRACT

Over centuries, there have been so many conflicts happening around the world both domestic and international such as war crises, separatism, or genocide. The causes of conflicts are diverse, and rooted in historical grievances, geopolitical ambitions, economic disparities, and cultural differences. These conflicts ravage countries in various ways: economic growth, infrastructure, food system, and depopulation. Genocide is one of the conflicts mainly caused by racial differences which impact a country's development as a decrease in the population. This is because human capital is a key driver of economic growth as it directly influences the productivity of the workforce. This study aims to analyze the relationship between victim identities (gender, age, occupation, nationality, education) and their execution in prison. This research utilizes data from the genocide biographic database, developed by Yale University, and the documentation center in which documents have been stored in the format of hard copies, scanned documents, images, and PDFs which need to be transformed into digital format for future analysis. The techniques used in this proposed research are web scraping and Optical Character Recognition, also called OCR, for the extraction process. For the analysis process, the research employs statistical models like the Chi-Square Test and Logistic Regression. And also, reveals data insight by using data visualization techniques to enhance the presentation of findings. The experimental studies showed that most of the victims are male and the majority of the victims are students, military, and higher education people. This might highlight the regime's effort to change the social fabric, and also downgrade occupational status, especially among students and intellectuals by forcing them to change their status from upper class to lower class, to avoid the difficulty of controlling these people. As a result, this significant finding showed that this problem leads to capacity reduction for country development because higher education and the young generation have been regarded as valuable and potential human resources for country development in quality and quantity. This is the reason to encourage people over the world to realize on stopping conflict, which causes the world has take a turn for the worse.

Keywords: *Web scraping, Optical Character Recognition (OCR), Genocide studies, Data visualization, Data analytics*

1. INTRODUCTION

War crises and world conflicts are complex events with deep historical roots and wide-ranging impacts on humanity and the planet. While the international community has made strides in conflict resolution and prevention, the ongoing challenges underscore the need for continued vigilance, cooperation, and innovation in the quest for peace. Understanding the dynamics of these conflicts is essential for policymakers, scholars, and citizens.

Genocide is one of the conflicts, which is deliberate and systematic destruction of a racial, political, or cultural group, stands as one of the gravest phenomena in human history. Its study encompasses various disciplines, including history, sociology, psychology, law, and international relations, reflecting its complex nature and the multifaceted approaches needed to understand and prevent it. The term "genocide" itself was coined by Raphael Lemkin in 1944 [1], [2], in response to the atrocities of the Holocaust. Studying genocide is crucial for acknowledging and understanding the historical

truths of affected communities which is vital for healing and reconciliation processes in post-genocide societies.

There is a trend that occurs in the political research on how politics played out during the regime, and an enormous amount of research on post-regime stories and biography. Furthermore, several pieces of research focus on gender and genocide but with a research scope on women’s rights. Some researchers provide an overview of the connection between climate change and genocide intensity due to migration and resource scarcity. As for recent studies, researchers have produced a prototype model capable of forecasting genocide one year into the future using a global genocide dataset, and also have conducted in-depth research on the identification of genocide risk factors to determine why leaders of some countries are likely to perform genocides.

Moreover, there has been a global trend in genocide research focusing on future genocides and when and where genocides are likely to occur, which is one of the primary reasons for studying genocide in order to prevent future atrocities. By understanding the warning signs and stages leading up to genocide, scholars, policymakers, and the international community can develop strategies and interventions to prevent such crimes from occurring [3], [4].

From above, majority of research has been conducted on the relationship between genocide and gender or another factors such as climate change, migration, policymakers. Almost no research has been conducted on the relationship between genocide and another form of identity, such as occupation, Gender, Nationality, and Educational Background. This is an extensive research gap that has made it very crucial for this research to be proposed as it will analyze and prove the involvement of gender and other forms of identity, such as Occupation, Gender, Nationality, and Educational Background, using data analysis methodology and NLP techniques.

Due to the importance of the study of Genocide as mentioned above, This research aims to analyze the relationship between the execution of victims and their identification such as Occupation, Gender, Nationality, and Educational Background, focusing on the public dataset of victims available in the genocide database developed by Yale University and the genocide museum document archives. The first

part of the data has been stored and structured into a digital and usable format for analytics propose, which is available on the public website. (source: <https://gsp.yale.edu/cambodian-genocide-databases-cgdb>).

However, the second part of the data has been stored in the format of hard copies, scanned documents, or images which need to be transformed into digital form for future use. This is because the majority of these documents are historical documents kept for over a century. These physical documents make data analysis a challenging task and time-consuming. OCR is then one of the solutions, which refers to a technology that enables to transformation of physical documents into fully editable files for further utilization. Applying this technology can contribute to them and presenting data in such a way which is simply to analyze [5], [6].

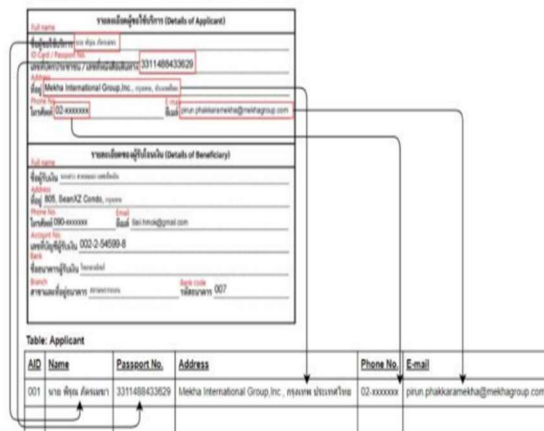


Figure 1: Example of OCR algorithms

Because of this issue, this paper applies the OCR technique for historical document digitization for data gathering and preparation processes before further analysis. Then, the research will employ quantitative data analysis methods, including descriptive correlational analysis, using statistical tests such as the Chi-Square test of independence and Logistic Regression. In addition, Business Intelligent, also called BI, was applied to reveal data insight.

The research aims to contribute to the understanding of genocide by highlighting potential patterns and associations in victim execution based on their identity factors. The study will provide a more nuanced understanding how the regime’s effort to change the social fabric.

2. LITERATURE REVIEW

In the last decade, there have been many research studies on genocide such as the paper “We Planted Rice and Killed People:” [7]. Woolford Andrew provides an overview of the connection between climate change and genocide intensity. As challenges such as migration, resource scarcity, and famine increase, there is a future risk that genocide may occur. The study states that the nature of genocide and its cultural or social values are inseparable. This study examines the relationship between the people and the consumption and cultivation of rice to highlight the consequences of the inseparable connection between social and natural death during the genocide. In 2021, Randall published the book: “Genocide and Gender in the Twentieth Century: A Comparative Survey” which examined the relationship between genocide and gender by acknowledging the difference in women’s and men’s experiences [8]. This study uses the concept of gender as a lens to understand genocide, which primarily focuses on ideologies and the intentions of regimes. As for predicting studies, a study titled, “A Two-Stage Approach to Predicting Genocide and Politicide Onset in a Global Dataset” produced a prototype model capable of forecasting genocide one year into the future using a global genocide dataset [9]. In 2023, Mitchell, S. M. conducted in-depth research on the identification of genocide risk factors to determine why leaders of some countries are likely to perform genocides [10]. This study uses the genocide that occurred in Rwanda and Burundi in the early 1990s. The research focuses on the content analysis of public statements.

In addition, there is a few researches apply data analytics to genocide data represents a significant advancement in the fields of human rights and international security. Through the collection, analysis, and interpretation of relevant data, researchers aim to identify the patterns or signs of genocidal actions for example:

Verdeja’s work examines the challenges and methodologies for forecasting genocidal violence and focuses on the risk assessment regarding a country’s long-term structural conditions such as regime type, and state-led discrimination [11]. Meanwhile, Harff, who discusses the development and application of risk assessment models for genocide prevention, reports a test of a structural model of the antecedents of genocide which consist of 6 factors; political upheaval, prior genocide, the

ideological orientation of the ruling elite, regime type, ethnic character of the ruling elite, and trade openness [12].

However, genocide data can be from many different sources, and some of documents are historical documents that has been stored in the format of hard copies, scanned documents, images or PDF which need to be transformed into machine-readable format by using OCR technique in order to avoid time-consuming and labour intensive [13]. Nowadays, OCR technique has been widely used and become important task for many areas and languages, and also broadly used in a large enterprise. Many researches apply OCR tool Tesseract which is open source for pre-processing step before data is used for further analyst [14].

For example, a Complete Optical Character Recognition Methodology for Historical Documents was proposed by G. Vamvakas [15]. This research employed OCR for recognizing historical documents with two steps: first step is using a collection of documents to generate a database for training, and the second step is documents recognition. In addition, OCR is also used in the other area such as license plate number detection. The research outcomes indicate that OCR is able to scan and retrieve the text and number from the vehicle license plate. [16].

Although, as there are not much research employ OCR technique to non-English compared to English language, due to the complexity and the difficulty to recognize texts. There are still many researches apply OCR approaches for non-English language, e.g., Chinese, Cambodian, Korean and Thai. In 2021, “Using OCR Framework and Information Extraction for Thai Documents Digitization” was proposed to extract all Thai texts from photocopies into database structure by using OCR [13]. Meanwhile, Michael Arrigo proposed “A Corpus for OCR in Multiple Languages” called CAMIO corpus, which created by Linguistic Data Consortium to support the development and evaluation of OCR and related technologies for 35 languages and comprises nearly 70,000 images of machine printed text, [17]. For Khmer language, Rina Buoy presented a convolutional Transformer-based text recognition method for low-resource non-Latin scripts. This method can handle images with non-Latin writing without explicit word boundaries by using OCR approach [18].

3. METHODOLOGY

In this section, the proposed research aimed to explore the influence of identity factors like occupation, gender, nationality, and education on victims executed in prison. Both descriptive and causal research methods have been used, starting from data finding to analysis and discussion on the result at the end.

There are five main steps operated in the proposed technique: (A) *Data finding and collection*, (B) *Data preparation (including cleaning and integration)*, (C) *Data analysis*, (D) *Data visualization*, and (E) *Result discussion*, as signified in the figure given below.

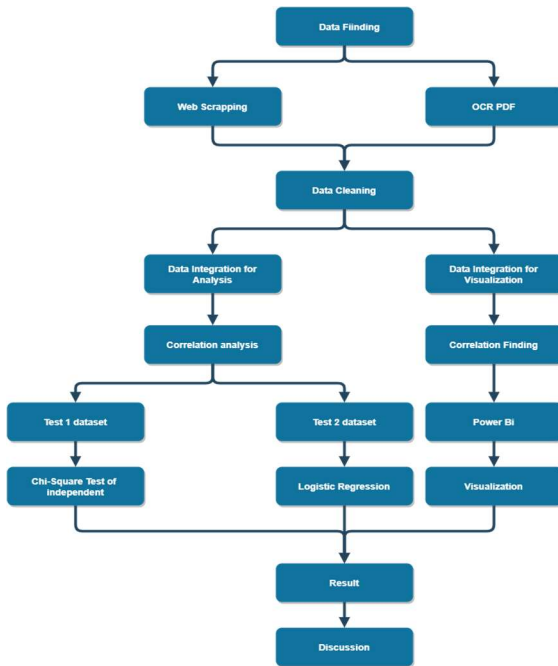


Figure 2: The workflow process of the proposed technique

(A) Data finding and collection

The first step begins with data finding from various sources like the genocide database developed by Yale University, which publish on the website (source: <https://gsp.yale.edu/cambodian-genocide-databases-cgdb>) collected by using web scrapping as shown in figure 3, and the genocide

museum document archives extracted by using OCR techniques as shown in figure 4.

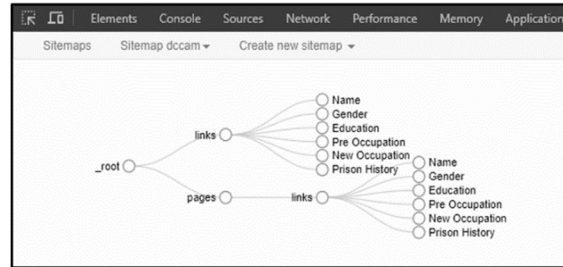


Figure 3: Sitemap of Web scraper tool to extract relevant key fields from genocide database website

The data collection for this research involves two main sources. The Genocide Database from Yale University provides a comprehensive set of fields, including demographic and biographical information of the victims. The documentation center of E-document archives offers additional key fields relevant to the study, accessed through OCR tools, contributing further data on victims. These sources collectively provide a rich dataset, enabling a detailed analysis of the impact of various identity factors on the victims.

The use of Optical Character Recognition (OCR) in the research, particularly focusing on Python Tesseract, which is an open-source OCR engine managed by Google. PyTesseract library is used to extract the text from images or PDF documents as the code shown below.

```

import pdfplumber
from pdf import Page
from pdfminer.converter import TextConverter
import pytz

# function to convert a table to a markdown-like format string
def table_to_string(table):
    table_string = ""
    for row in table:
        cleaned_row = []
        for cell in row:
            if cell is None else cell.replace('\n', '').strip()
        table_string += '| ' + '| '.join(cleaned_row) + '| \n'
    return table_string.strip()

# function to extract text and format from a pdf element using pdfminer
def extract_text_and_format(element):
    if not isinstance(element, TextContainer):
        return ""
    text_content = element.get_text()
    fonts = set()
    for text_line in element:
        if isinstance(character, TChar):
            fonts.add(character.fontname, character.size)
    return text_content, list(fonts)

# function to check if an element is inside any of the tables
def is_inside_any_table(element, tables):
    for table in tables:
        if table.bbox[0] <- element.bbox[0] and table.bbox[1] <- element.bbox[1] <- table.bbox[2]:
            return True
    
```

Figure 4: Python code implementation for OCR and PDF analysis

The process involves inputting PDF documents into Tesseract, which identifies and converts text within these documents into a machine-readable format. Python Tesseract offers extensive customization options, enabling adjustments to

parameters like page segmentation for improved accuracy. Additionally, Python libraries such as Pillow for image processing and PDF Plumber for PDF parsing are utilized to prepare documents for OCR. These steps include converting PDF pages into images and enhancing image quality for more effective text extraction. The combination of Python Tesseract and these libraries ensures not only the extraction of text but also its precision and efficiency, crucial for the quality of data in subsequent analysis phases.

(B) Data Preprocessing

Data cleaning

From two sources, datasets include six entities: name, record ID, occupation, gender, nationality, and prison history. However, there is a significant amount of missing and noisy data. For example, data is missing for occupations around 20%, or gender around 25%.

For handling dirty data, the missing gender data in the dataset is identified as Missing Completely at Random (MCAR). This means that the absence of gender information is random and unrelated to any other data in the dataset, such as age, occupation, or nationality. The researchers believe that the primary reason for this missing data is the destruction of records before the end of the regime.

Given the high level of missing data (25% in the dataset of 458 victims), traditional methods like simple or multiple imputation seemed unsuitable. The deletion of datasets with missing gender information is also not considered viable, as the research aims to examine the influence of gender on execution rates.

To address this issue, this research opts for predictive modeling using software called 'Gender API' from genderize.io.



Figure 5: 'Gender API' from genderize.io

This tool predicts gender based on first and last names, utilizing global lookup, normalization techniques, and correlations between names and genders as shown in figure 6. However, this method has limitations, such as reduced accuracy with unisex names and cultural variability in name-gender associations, which may affect prediction accuracy.

ID	ga_first_name	ga_last_name	ga_gender	ga_accuracy	ga_samples
Y06286			Female	98	76776
Y06291			Female	68	48002
Y06290			Male	75	4
Y06280			Male	90	1838

Figure 6: An example of how missing Gender data is handled through the API tool.

Another issue is data inconsistencies. The dataset of victim selectively scraped and extracted from the databases and documents, had issues like data inconsistencies from free text data entry. For example, the gender variable included unnecessary details like "Male (Source: K08250, P.1)", and the occupation was overly detailed, e.g., "HURIDOCS code: .62 Notes: Agriculture Worker (Source: K08250, p.1)". To rectify these, standardization was performed to simplify and correct the data format, removing irrelevant information or spelling consistency, for instance, sometimes lieutenant was spelled with no capital letter, but in some cells, it was spelled in all capitals, "LIEUTENANT". In addition, the major step of cleaning which is labor-intensive and time-consuming is data categorization. This research categorizes the variety format of occupations in the same category in which members in the group are similar to each other as shown in figure 7.

Inconsistencies Correction Format		
Category of Data	Original Data	Cleaned Data
PK Occupation	HURIDOCS code: .62 Notes: Agriculture Worker (Source: K08250,p.1)	Agriculture Worker
Education	Class 1. Angduong High School (Source: K05877, p.1)	High Schooler
Name	Pi set (Source: K05877,p.1)	Pi Set
Gender	Male (source: I10258, p.1)	Male

Figure 7: Data transformation of occupation, education, name and gender following the above inconsistencies correction Format.

Furthermore, data deletion is another necessary step in this research due to incomplete victim datasets, where imputation methods are ineffective. For example, missing data on occupation cannot be inferred from identifiers like name, nationality, or gender. Therefore, datasets with missing occupation information are deleted for correlation tests. However, this deleted data is not entirely discarded; it is transferred to a separate data set for potential use in other analyses, such as correlational studies based on gender or data visualization. In summary, while some data is removed from certain analyses due to irrelevance, it is still retained for alternative research purposes.

Data Integration

Data integration in this research involved merging victim datasets from two sources: the Genocide Database from Yale University collected by web scraping and the PDF files from the documentation Center of E-document archives extracted by OCR technique. From the integration process, the dataset is mainly in CSV format, contained variables like name, record ID, occupation, gender, nationality, prison history, and arrest location. After cleansing, the focus was narrowed down to occupation and gender. Two main integration tests were conducted: Test 1 combined occupation and gender data from 461 victims, while Test 2 merged datasets with gender, nationality, and survival status for 582 victims.

The final phase involved integrating these datasets for correlation analysis. This step combined datasets, aligned similar categories, and performed statistical analysis to uncover relationships as shown in the figure below.

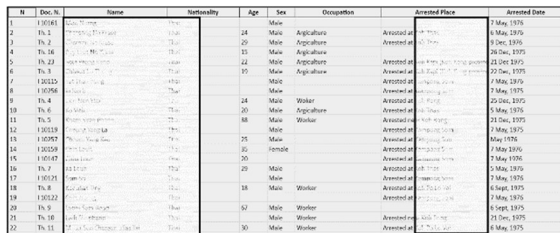


Figure 8: the integrated datasets for analysis.

4. EXPERIMENTAL STUDIES AND RESULT

In this section, the experimental studies have been divided into two main parts: data analysis and data visualization.

(C) Data analysis

This part focuses on the results from the correlational analysis conducted on the victim datasets. The first section of the data analysis part of this research presents the demographic information of the two primary victim sampled and cleaned datasets that were used to conduct a correlational test: Test 1 and Test 2. The second section will describe the analytic result of Test 1 and the results of its chosen correlation.

The experimental study summarizes the results of correlational analysis conducted on victim datasets. The analysis is divided into two main tests: Test 1 and Test 2.

Test 1 Overview:

- Uses datasets focusing on 461 victims with occupation and gender.
- Occupation data is transformed into nominal data based on a ranking system reflective of the social hierarchy.

Table 1: A table showing the occupation ranking system for data preprocessing

Ranking	Occupation Type	Occupation Example
1	Military	Deputy Chief, Lieutenant Colonel, Capitan, etc.
2	Government and High Education Level Occupation	Police, Minister, Engineer, Doctor, etc.
3	White Collar Worker and Student	High Schooler, University Student, Merchant, etc.
4	Low Education Level Occupation	Farmer, Livestock Breeder, Shop Keeper, etc.

The Table above shows the ranking system of the victim's occupation; from datasets, there are a total of 256 different occupations found. Through transformation and standardization based on the ranking system depicted in Table 1 above, the occupation has been transformed into nominal data

that only consists of four main categories based on their ranking.

- Gender is represented as a variable (Male = 0, Female = 1).

Test 1 Chi-Square Test of Independence:

- Analyzes the association between victims' occupation, gender, and execution.
- A significant association is found, with Pearson Chi-Square value of 17.116 ($p < .001$).

Table 2: Pearson Chi-Square value

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	17.116 ^a	3	<.001
Likelihood Ratio	23.946	3	<.001
Linear-by-Linear Association	5.598	1	.018
N of Valid Cases	461		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.44.

- Symmetric measures suggest a weak positive correlation between gender and occupation.

Test 2 Overview:

- Uses datasets focusing on gender, nationality, and survival status of 582 victims.
- Nationality and survival status are numerically encoded for analysis. There are a total of 7 different nationalities found in our data set represented by 0 to 6. As for survival status, it is defined by two numbers, 0 and 1. The number 0 indicates that the victims have not survived the imprisonment and the number 1 indicates that the victims survived the imprisonment.
- Gender is represented as a variable (Male = 0, Female = 1).

Test 2 Logistic Regression analysis:

- Explores further the hypothesis using Logistic Regression, predicting relationships between variables like gender, nationality, and the survival rate.

- Results indicate significant associations, but with limitations due to the imbalance in the dataset.

This research highlights the disparity in gender distribution across occupation ranks and significant but weak associations between victims' identities and their execution or survival. It underscores the challenges in analyzing such a complex historical data.

(D) Data visualization

In the second part of the experimental studies, the data visualization technique is also employed to represent the information. The Power BI Desktop application was used to implement a dashboard based on the demographic data of victims as depicted in figure 9.



Figure 9: Data visualization using PowerBI to represent data insight

From the above dashboard, it includes a gender distribution represented by a pie chart, showing the impact on both males and females. A Treemap illustrates the victim's occupations, highlighting targeted groups. The bar chart on prison history emphasizes the extent of detention and execution. The nationality column chart reveals the ethnic diversity of victims, indicating a widespread impact. These visuals collectively narrate the profound societal changes and the regime's broad-reaching influence on various demographics.

5. DISCUSSION

(E) Result discussion

This research aims to reveal a correlation between victims' identities (Gender, Occupation, Education, Nationality) and their execution, using two tests: Chi-Square test of independence and Logistic Regression.

For Chi-Square test of independence, It checks for associations between victims' occupation, gender and their execution. Significant findings were identified (Pearson Chi-Square p-value < .001), indicating a statistically significant but weak association between these variables. Both Pearson's R and Spearman's Correlation suggest a significant correlation, though not strong, implying other factors could also influence executions.

For Logistic Regression, this model was conducted due to Chi-Square results, especially considering the limited survivors in the dataset. It showed a significant relationship between gender and survival likelihood.

However, this experimental study faces some constraints, mainly due to imbalanced and incomplete datasets. The available data is not ideal for quantitative analysis due to its structure and loss of data. Further research is encouraged, utilizing survivor insights for a more comprehensive understanding of the genocide, including aspects of emotional well-being.

For the data visualization part, the dashboard illuminates the widespread impact of across genders, occupations, and nationalities. It highlights the regime's concerned effort to suppress intellectuals to establish a classless society. The data reflects that regime's effort to destruct the social fabric and also downgrade occupational status, especially among the educated elite by forcing them to change their status from upper class to lower class, in order to avoid the difficulty of controlling these people. The dashboard significantly shows that the first rank of victim's occupation is student and military, the second rank is intellectuals (like higher education people, workers and government workers), the third rank of victim occupation is lower education people (like farmer and livestock breeder) and so on, as shown in the ranking and occupation treemaps in figure 9.

6. PROBLEM AND RESEARCH ISSUES

The scope of this research is to implement quantitative research on the relationship between victim identities (gender, ages, occupation, nationality, education) and their execution. This study will focus on the dataset containing names, ages, occupations, nationality and other variables of the victims that are available for extraction from the genocide biographic database, developed by Yale

University, and the documentation center in which documents have been stored in the format of hard copies, scanned documents, images, and PDFs.

These historical documents have been stored in the format of hard copies and scanned documents over many centuries. This is clear that there are limitations to this research due to historical document character. The regime was known to destroy documents and execute anyone at any time as they wished. As a result, there is limited information on the data of each victim. Another limitation is that most of the documentation is confidential and in the form of scanned documents; therefore, the need to do a data cleansing and OCR is mandatory. However, the performance of text extraction is depended on OCR technique and language.

While framework is robust, several limitations have been founded:

- Data quality: variability in the quality of scanned documents may impact OCR accuracy and subsequent analysis.
- Language barriers: challenges related to the language, including limited NLP tools, necessitate manual verification and introduce language-specific complexities.

However, despite these challenges, research framework was designed to maximize the potential for extracting meaningful insights from the available data set.

Another research issue is ethical considerations. Throughout this research, a steadfast commitment to ethical considerations has been maintained:

- Data privacy and anonymity: to protect privacy and anonymity, all survivor testimonies and victim data are anonymized, and any identifying information is redacted before analysis.
- Cultural sensitivity: the sensitivity of the topic and the potential cultural implications has been recognized. Ethical approvals have been obtained, and interviews, data handling, and analysis are conducted with the utmost respect for cultural norms and survivor experiences.

Meanwhile, this research proposed the efficiency framework, some challenge stills remain as discussed below.

The main challenge is the quality and completeness of victim's dataset; there are indeed a lot of missing gender, nationality, and occupation data of the victims, and data such as nationality and occupation could not be handled as they are personal data that concern the real identity of each victim, therefore methods such as data exploration, imputation are out of the picture. From the experimental studies in this research, analysis result has revealed that there is a statistically significant correlation between gender, ethnicity, and the victim's execution or survival. However, due to some bias in the model and imbalance in the datasets, the key takeaway for this research is that the proposed technique has proved an association but the correlational strength is weak. Consequently, the problem of quality, completeness and imbalance in victim's dataset is needed to be solved for future work, by seeking for better quality of victim's datasets for further genocide studies.

7. CONCLUSION

This research aims to investigate the association between the execution of victims and their identification factors like gender, occupation, nationality, and educational background. Previously, understanding of the genocide was primarily derived from qualitative sources like oral investigations and survivor testimonies. This study uses descriptive correlation analysis on victim data to support the hypothesis that executions were based on these identification factors, challenging the regime leaders' claims that only ethnic minorities were targeted.

The research utilized datasets from the Biographic Database, developed by Yale University and the Documentation Center, which underwent various sampling and data-cleaning methods. Two correlational tests, the Chi-Square test of independence and Logistic Regression test, were conducted. These tests showed a statistically significant but weak association between the victim's gender, Occupation and their execution, suggesting other influencing factors.

Further analysis indicated a significant correlation between gender, nationality and the execution or survival of victims. However, this finding is limited by imbalance in the datasets. The

primary conclusion is that while an association was identified, the connection's strength is weak due to data limitations.

Furthermore, the data visualization provides a conclusion to the understanding of the human dimension in this research. The dashboard shows that the majority of victims are male gender, and also it reveals that the first rank of victim's occupation is student and military, the second rank is intellectuals (like higher education people, workers and government workers), the third rank of victim occupation is lower education people. This might highlight that regime effort to destroy the social fabric, and also downgrade occupational status, especially among students and intellectuals by forcing them to change the status from upper class to become lower class people, in order to be easy to control.

As a result, the subsequent "brain drain" and lack of the higher education people and young generation had long-lasting impacts on country's development. This is the reason to encourage people all over the world to be concerned about stopping conflict before it starts again and again, which causes the world has take a turn for the worse.

REFERENCES:

- [1] Schaller, D .J., & Zimmerer, J.,) Eds.(.)2013.(*The origins of genocide :Raphael Lemkin as a historian of mass violence* .Routledge, 2013.
- [2] Jones, A., *Genocide: A comprehensive introduction*. Routledge, 2016.
- [3] Evans, G., *The responsibility to protect : ending mass atrocity crimes once and for all*. Rowman & Littlefield, 2009.
- [4] Hamburg, D .A., *Preventing genocide : Practical steps toward early detection and effective action* .Routledge, 2015.
- [5] Vamvakas, G., Gatos, B., Stamatopoulos, N., and Perantonis, S., "A Complete Optical Character Recognition Methodology for Historical Documents," *The Eighth IAPR International Workshop on Document Analysis Systems*, 2008.
- [6] Pai, N., and amp; Kolkure, V., S., "Optical Character Recognition: An Encompassing Review," *International Journal of Research in Engineering and Technology (IJRET)*, Volume-4(Issue-1), 407-409.
- [7] Woolford, A., June, W., and Um, S, "We Planted Rice and Killed People:"

- Symbiogenetic Destruction in the Cambodian Genocide. *Genocide Studies and Prevention: An International Journal*, 15(1), 7, 2021.
- [8] Randall, A. E. (Ed.), “Genocide and gender in the twentieth century: A comparative survey”, *Bloomsbury Publishing*, 2021.
- [9] Goldsmith, B. E., Butcher, C. R., Semenovitch, D., and Sowmya, A., “A Two-Stage Approach to Predicting Genocide and Politicide Onset in a Global Dataset”, *Available at SSRN 2027396*, 2012.
- [10] Mitchell, S. M., “Institutional Legacies and the Decision to Commit Genocide. *Genocide Studies and Prevention*”, *An International Journal*, 17(1), 1-24.
- [11] Verdeja, E., “Predicting genocide and mass atrocities,” *Genocide Studies and Prevention: An International Journal*, 9(3), 5, 2016
- [12] Harff, B., “No lessons learned from the Holocaust? Assessing risks of genocide and political mass murder since 1955,” *In Genocide and Human Rights*, (pp. 329-345), Routledge, 2017.
- [13] Todsanai, C., and Waramporn, R., “Using OCR Framework and Information Extraction for Thai Documents Digitization,” *2021 International Electrical Engineering Congress (iEECON2021)*, March 10-12, 2021, Pattaya, THAILAND.
- [14] Patel, C., Patel, A., and Patel, D., “Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study,” *International Journal of Computer Applications*, 55(10), 50–56, 2022.
- [15] Vamvakas, G., Gatos, B., Stamatopoulos, N., and Perantonis, S., “A Complete Optical Character Recognition Methodology for Historical Document,” *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 2008.
- [16] Aljelawy, Q. M., and Salman, T. M., “Detecting license plate number using OCR technique and Raspberry Pi 4 with camera,” *In 2022 2nd International Conference on Computing and Machine Intelligence (ICMI)* (pp. 1-5). IEEE, April, 2022.
- [17] Arrigo, M., Strassel, S., King, N., Tran, T., and Mason, L., “CAMIO: A Corpus for OCR in Multiple Languages,” *In Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp. 1209-1216), June, 2022.
- [18] Buoy, R., Iwamura, M., Srun, S., and Kise, K. , “Toward a Low-Resource Non-Latin-Complete Baseline: An Exploration of Khmer Optical Character Recognition,” *IEEE Access*, 11, 128044-128060.