UNIVERSITI TEKNOLOGI MARA

MATHEMATICS AND STATISTICS

UNDERGRADUATE RESEARCH PROCEEDINGS 2025

UiTM CAWANGAN NEGERI SEMBILAN

Predicting Factors in Financial Loss Among Malaysian Scam Victims Using Machine Learning

Nur Alisa Binti Azian[1], Che Norhalila Binti Che Mohamed[1, *]
[1]*Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawanagn Negeri Sembilan, Kampus Seremban, 70300 Seremban, Negeri Sembilan.*
[*]cheno849@uitm.edu.my

Abstract

This study compares decision tree and logistic regression models to predict financial losses among 394 Malaysian scam victims and to identify key predictors. Model accuracy refers to overall classification accuracy on the validation set, alongside AUC, sensitivity, specificity, and F1 score. Model interpretability was operationalized as the transparency of decision rules in the fitted tree, that is, the ease of tracing split conditions along each decision path. On validation, the decision tree achieved higher AUC and accuracy than logistic regression (AUC 0.838 versus 0.797; accuracy 83.05 percent versus 72.88 percent), with gains in sensitivity and F1 score as well. Emotional harm emerged as the strongest predictor of financial loss, followed by cybersecurity knowledge and age, whereas gender, urbanity, education, and internet use contributed modestly. These findings support targeted victim support, stronger fraud detection, and population-level digital literacy initiatives.

Keywords*:* Decision Tree, Financial Loss, Logistic Regression, Scam Victims, Malaysia

Introduction

The Internet is widely regarded as one of humanity's most remarkable inventions, having permanently transformed communication, information access, and commerce in daily life [1]. In Malaysia, the number of Internet users has continued to rise, from 96.8% in 2021 to 97.4% in 2022 [2]. This increasing dependence on digital platforms and social networking sites has created new opportunities but also heightened exposure to scams and fraudulent activities.

Malaysia has experienced a rapid escalation in online scams, with 95,837 incidents reported between 2021 and April 2024, amounting to financial losses of RM3.18 billion. Reports from the National Scam Response Centre (NSRC) indicate that these figures are likely underestimated, as many victims do not file official complaints [3]. The growing prevalence of scams underscores the urgent need for evidence-based strategies to protect vulnerable groups and minimise financial harm.

This study aims to classify scam victims based on their demographic and knowledge-related characteristics and to identify predictors of financial loss using machine learning techniques. By addressing this gap, the research contributes to understanding the dynamics of scam victimisation in Malaysia and supports the design of more effective preventive measures.

Background

Scams or fraudulent activities, often referred to as confidence tricks, have existed long before the advent of the Internet [4]. Susceptibility to scams is shaped by multiple factors, including the persuasive strategies of perpetrators, cognitive processes of victims, and individual differences in vulnerability [5]. In Malaysia, common scam types include love scams, Macau scams, job scams, parcel scams, and online loan scams [6].

Prior research has examined demographic and psychological risk factors associated with scam victimisation. However, limited attention has been devoted to applying advanced analytical methods to predict financial loss. Machine learning techniques, such as decision trees and logistic regression, are well-suited to this task because they can manage complex interactions among variables and generate accurate predictions [7,8]. Decision trees, in particular, provide intuitive visualisations that enhance interpretability for policymakers and stakeholders [9]. Recent studies highlight their potential in identifying how demographic, behavioural, and knowledge-based factors contribute to vulnerability [10].

Despite this progress, research employing machine learning for scam prediction in Malaysia remains scarce. Machine learning, as a branch of artificial intelligence (AI), offers robust tools capable of detecting subtle patterns those traditional statistical methods may overlook [11]. Leveraging these approaches may generate novel insights into predictors of financial loss and inform targeted interventions to reduce the burden of scams.

Methodology

Research Design

This study used both descriptive and exploratory research designs. A descriptive design is used to describe the type of scam experienced by the victims to find out the amount of losses incurred by the victims of scams in Malaysia. An exploratory design tests the hypotheses based on the theoretical framework. A cross-sectional design was implemented in this study because it captures data at a single point in time, and the conclusions drawn are valid only during the period in which the research was conducted.

Population and Sample

The study targeted Malaysians who had experienced scams, irrespective of the type of scam. Due to the unknown population of scam victims, the minimum sample size for categorical data was computed using Cochran's formula with $p = 0.50$ and a 5% margin of error, which yielded 384 respondents at a 95% confidence level. This decision also falls within the recommended sample range of 30 to 500 for behavioral research [12], supporting its suitability for the present study.

Sampling Method

Sampling methods are categorised into probability sampling, where every individual has an equal chance of selection, and non-probability sampling, which relies on subjective judgment [13]. Due to the confidentiality of scam victims' data from the Royal Malaysia Police (Polis Diraja Malaysia, PDRM), no sampling frame was available, necessitating the use of non-probability sampling. Convenience sampling was chosen, relying on accessible primary data sources for faster, cost-effective data collection. However, convenience sampling is inherently prone to bias because it does not ensure that the sample is representative of the entire population, as it selects participants based on their availability and willingness to participate [14,15]. This method can overrepresent certain groups and underrepresent others, leading to results that may not accurately reflect the broader population. Consequently, findings are generalisable only to the sampled subpopulation, limiting their applicability to broader contexts [16,17].

Data Collection Method

Primary data were collected directly from scam victims using a quantitative approach. An online survey was administered through Google Forms and distributed via multiple social media platforms, including Facebook, WhatsApp, Telegram, and Instagram. The recruitment period lasted two months after obtaining ethical approval, during which the questionnaire links were actively disseminated to family members, friends, and members of scam victims' groups on Facebook to maximise participation. This strategy enabled the collection of firsthand information on demographic characteristics, internet safety knowledge, and the financial, emotional, and behavioural impacts of scams. Online surveys were selected due to their speed, accuracy, cost-effectiveness, and flexibility, making them well-suited for this research [18].

Research Instrument

The research instrument used in this study was a 29-question questionnaire divided into three parts. Part A included demographic questions such as gender, age, education level, internet usage, urbanity, and scam type. Part B consisted of 12 questions on internet safety knowledge, based on the Knowledge and Attitude Towards Internet Safety Questionnaire (KATISQ) [19], and three additional questions about the scamming channels, reporting behaviour, and recipients of reports [20]. Part C focused on the negative impacts of scams, covering financial, emotional, and behavioural effects, including financial losses, with seven questions in total [20]. Overall, the designed questionnaire, with detailed descriptions of the variables provided in Table 1.

Table 1: Questionnaire Item.

Part	Item	Number of Questions	Question Number
A	Demographic	6	1 – 6
B	Knowledge about internet safety	12	7 - 18
	Scammed channel	1	19
	Reported scam	1	20
	Reported a scam to	1	21
C	Negative impact	6	22 - 28
	Negative attitude	1	29

Theoretical Framework

Figure 1 illustrates the research theoretical framework of this study, in which the independent variables are demographic factors (age, gender, level of education, internet usage and urbanity), knowledge about internet safety and negative impacts (emotional impact and physical impact) are the factors that influence the financial loss experienced by scam victims.

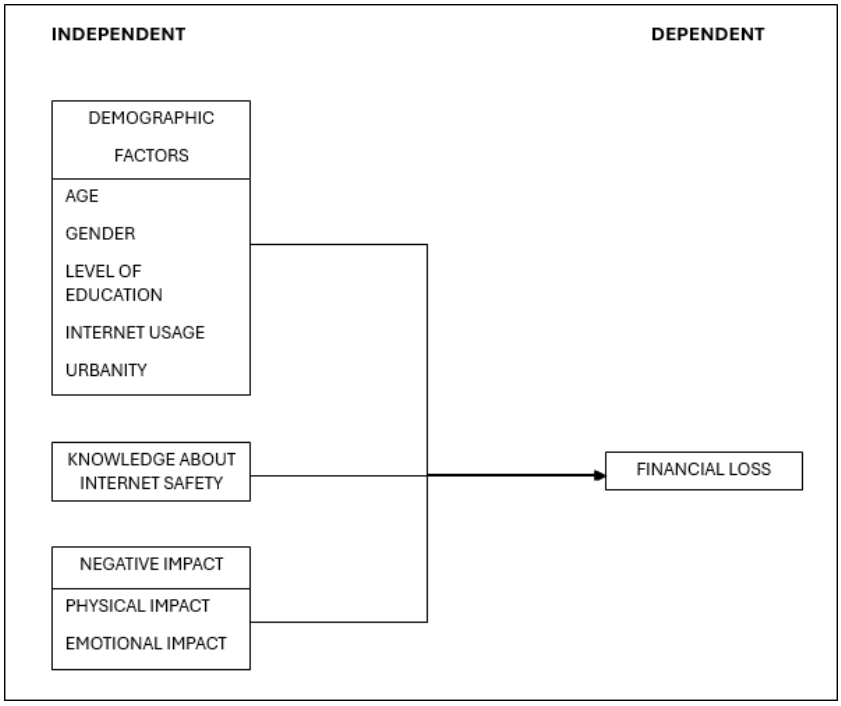

Figure1: Research Theoretical Framework.

Data Analysis

Descriptive analysis was employed to summarise the traits of individuals who have fallen victim to scams. By identifying shared characteristics and vulnerabilities among scam victims, this analysis provides insights into those most at risk and informs tailored interventions to protect them. This examination highlights the demographic trends of scam victimisation in Malaysia, supporting a more effective strategy to combat scams and mitigate their impact on individuals and society. Additionally, the analysis focused on identifying the best predictive model and the factors influencing financial losses among scam victims. Logistic regression and decision tree analysis were utilised, with model performance evaluated using metrics such as classification accuracy, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) to identify the factors influencing the financial losses experienced by scam victims in Malaysia. Logistic regression highlighted significant predictors ($p < 0.05$), while decision tree analysis ranked variables based on importance scores. This approach provided a comprehensive understanding of the key factors contributing to financial losses.

Machine Learning Models and Evaluation

Two supervised learning algorithms were trained to classify financial loss status, namely logistic regression and a Classification and Regression Tree (CART)-style decision tree using Gini impurity as the splitting criterion. Logistic regression estimates the log-odds of loss as a linear function of predictors and yields adjusted odds ratios for inference and communication [21]. In contrast, decision trees iteratively partition the predictor space into a series of decision rules that can be directly inspected, offering greater transparency and interpretability for stakeholders and policymakers [22]. Model performance was evaluated on the validation dataset using several complementary measures: the area under the receiver operating characteristic curve (AUC), overall classification accuracy, sensitivity, specificity, and the F1 score [23]. Interpretability of the decision tree was operationalized in terms of rule transparency, defined as the ease of tracing conditions from root to leaf, which provides an intuitive supplement to quantitative performance metrics. Variable importance values were also extracted from the tree to rank predictors according to their relative contribution to classification.

Result and Discussion

Descriptive Analysis

The demographic findings in Table 2 show that females are slightly more susceptible to scams, though the difference is not significant. Older individuals aged 55 and above are more likely to experience financial loss, while those with only primary or secondary education are particularly vulnerable. This highlights the critical need for financial literacy and scam awareness programs to equip these groups better. Moderate internet users are at a higher risk of financial loss compared to light or heavy users, possibly due to inconsistent online habits. Urban residents are also more prone to scams, likely because of greater exposure and higher levels of online activity. These findings underscore the importance of developing tailored, urban-focused scam

prevention strategies and targeted educational initiatives to reduce vulnerabilities across these demographic groups in Malaysia.

Table 2: Summary of Demographics of Scam Victims

Variable		Financial Loss (Yes)		Financial Loss (No)	
		N	%	N	%
Gender					
	Female	129	32.7	107	27.2
	Male	104	26.4	54	13.7
Age (years)					
	18–24	100	25.4	100	25.4
	25–34	51	12.9	22	5.6
	35–54	62	15.7	35	8.9
	≥55	20	5.1	4	1
Education level					
	Primary School	1	0.3	0	0
	High School	28	7.1	16	4.1
	Diploma	87	22.1	45	11.4
	Bachelor	103	26.1	89	22.6
	Postgraduate	14	3.6	11	2.8
Internet Usage					
	Light	35	8.9	17	4.3
	Moderate	143	36.3	102	25.9
	Heavy	55	14	42	10.7
Urbanity					
	Urban	136	34.5	89	22.6
	Suburban	61	15.5	47	11.9
	Rural	36	9.1	25	6.3
Total		233	59.1	161	40.9

Based on the Figure 2, online sales scams stand out as the most common, with a significant number of victims experiencing financial loss. Short Message Service (SMS) scams also demonstrate a concerning prevalence, followed by Macau scams and job scams that contribute notably to financial loss. Additionally, the chart also reveals that African scams and business email compromise have relatively fewer victims but are still areas of concern. The findings highlight the necessity for targeted interventions and educational programs tailored to specific scam types. The findings highlight that online sales and SMS scams are the most prevalent, posing significant risks due to their digital and mobile nature. Macau and job scams also

contribute notably to financial losses, while African scams and business email compromise remain areas of concern.

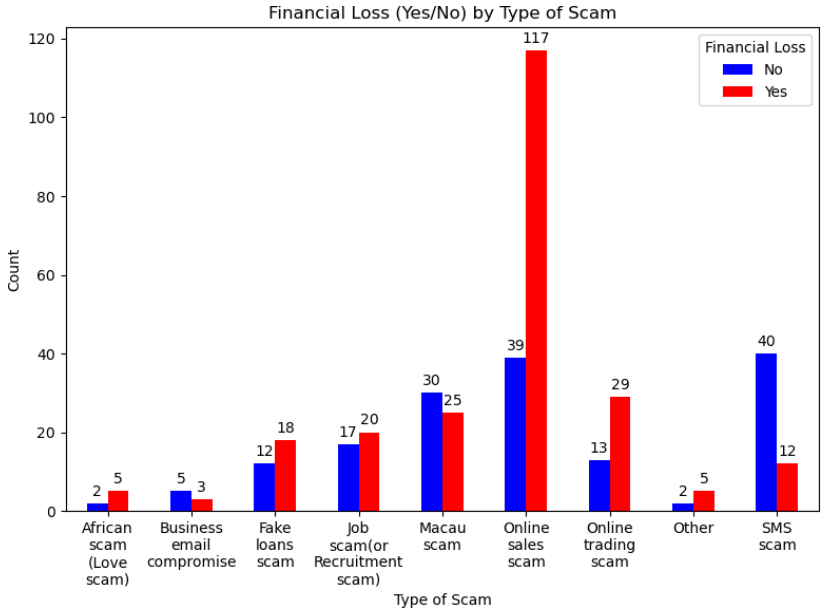

Figure 2: The Distribution of Scam Types by Financial Loss (Yes/No).

Best Model for Predicting the Financial Losses Experienced by The Scam Victims in Malaysia

According to Table 3, the decision tree model outperformed logistic regression in predicting financial loss, as evidenced by higher AUC, accuracy, sensitivity and F1 score. Firstly, the decision tree model demonstrated a higher AUC value on the validation dataset (AUC = 0.838) compared to the logistic regression model (AUC= 0.797). This indicates that the decision tree model generalises better to unseen data and is more likely to predict financial loss in real-world scenarios accurately. Furthermore, the decision tree model outperformed the logistic regression model in terms of accuracy (83.05% vs. 72.88%) and sensitivity (87.14% vs. 72.86%), highlighting its effectiveness in correctly identifying individuals who experienced financial loss. While the logistic regression model showed slightly higher specificity (72.92% vs. 77.08%), the overall performance of the decision tree model, as reflected by the F1 score (85.92% vs. 76.12%), suggests its superiority in predicting financial loss among scam victims. These insights reinforce the decision tree model's potential as a valuable method for mitigating financial loss among the scam victims in Malaysia.

Table 3: Model Performance Comparison

Comparison Metrics	Decision Tree (GINI)	Logistic Regression (Main Effect)
AUC	0.838	0.797
Accuracy	83.05%	72.88%
Sensitivity (Recall)	87.14%	72.86&
Specificity	72.92%	77.08%
F1 Score	85.92%	76.12%

Factors That Influence the Financial Losses Experienced by The Scam Victims in Malaysia

Given that the decision tree model outperformed the logistic regression model in predictive accuracy and reliability, Table 4 summarizes the key results of the decision tree analysis. Emotional harm was identified as the most influential variable with an importance score of 1.0000, indicating its critical role in distinguishing between victims who experienced financial losses and those who did not. Knowledge score was the second most important factor (0.5123), suggesting that individuals with higher internet knowledge were better classified by the model. Age ranked third (0.3947), reflecting a moderate influence likely tied to variations in exposure or susceptibility across age groups. Gender contributed less significantly (0.1730), while urbanity (0.1152), education level (0.0947), and internet usage (0.0821) showed limited predictive power. Physical harm had an importance score of 0.0000, indicating no contribution to the classification. These results suggest that emotional harm is the primary predictor, followed by knowledge and age, while other factors play minor roles.

The predominance of emotional harm as the strongest predictor aligns with growing evidence that psychological distress plays a central role in shaping victims' financial decision-making under fraudulent influence. Victims who experience heightened emotional states such as fear, urgency, or shame are more vulnerable to persuasive tactics and more likely to comply with scammers' demands, thereby increasing the probability and magnitude of financial loss. Prior research has shown that the severity of emotional harm is strongly correlated with financial vulnerability: identity theft victims reporting higher distress tend to incur larger losses [24], while more recent qualitative studies of romance scams demonstrate that emotional manipulation intensifies susceptibility to substantial monetary harm [25,26]. These findings suggest that emotional harm is not merely a consequence of fraud but functions as a key mechanism of victimization, underscoring the need for integrated psychosocial support in addition to financial protection interventions.

Table 4: Feature Importance.

Variable	Importance Score
Emotional harm	1.0000
Knowledge score	0.5123
Age	0.3947
Gender	0.1730
Urbanity	0.1152
Level of education	0.0947
Internet usage	0.0821
Physical harm	0.0000

Conclusion

The main aim of this study was to determine the best predictive model between the decision tree and logistic regression and to identify the factors influencing financial losses among scam victims in Malaysia. The findings revealed that the decision tree model outperformed logistic regression in predictive accuracy and classification performance, making it the superior choice for this analysis. The study highlighted emotional harm as the most significant predictor of financial loss, followed by internet safety knowledge and age, while gender, urbanity, education, and internet usage played minor roles. Notably, physical harm showed no influence on financial losses.

Addressing emotional harm and improving digital literacy are essential strategies for mitigating financial losses. These findings underline the need for targeted public education campaigns, robust fraud detection through collaboration between the government and financial institutions, and expanded victim support services.

Future research should further extend this work by integrating advanced machine learning techniques, such as neural networks and ensemble learning models, to capture more complex and nonlinear relationships between predictors of financial loss. Additionally, longitudinal datasets on scam victimization would enable the examination of temporal dynamics, such as changes in vulnerability and recovery patterns over time, which cannot be captured in cross-sectional studies. Incorporating these approaches could provide a deeper understanding of the causal mechanisms underlying scam victimization and improve the predictive validity of prevention models.

Acknowledgements

The authors would like to express their sincere gratitude to all respondents who generously shared their time and experiences, without whom this study would not have been possible. Appreciation is also extended to Universiti Teknologi MARA (UiTM) and the College of Computing, Informatics, and Mathematics for providing the necessary resources and support.

The authors are deeply thankful to the supervisor for her invaluable guidance and encouragement throughout the study, and to colleagues and lecturers for their constructive input. Finally, heartfelt thanks are given to family and friends for their continuous support and motivation.

References

[1] Brooks, R. (2022). *The invention and evolution of the Internet*. Wrexham University Online. https://online.wrexham.ac.uk/the-invention-and-evolution-of-the-internet

[2] Department of Statistics Malaysia. (2022). *ICT use and access by individuals and households survey report, Malaysia*. Putrajaya: DOSM.

[3] Mahaizura, A. M. (2024, August 1). *Kes tipu direkod paling tinggi dalam jenayah atas talian*. Harian Metro. Retrieved September 9, 2024, from https://www.hmetro.com.my/mutakhir/2024/08/1121860/kes-tipu-direkod-paling-tinggi-dalam-jenayah-atas-talian

[4] Whitty, M. T. (2020). Is there a scam for everyone? Psychologically profiling cyberscam victims. *European Journal on Criminal Policy and Research, 26*(3), 399–409. https://doi.org/10.1007/s10610-019-09425-9

[5] Hanoch, Y., & Wood, S. (2021). The scams among us: Who falls prey and why. *Current Directions in Psychological Science, 30*(3), 260–266. https://doi.org/10.1177/09637214211043460

[6] Majlis Keselamatan Negara. (2023, March 28). *Jenis penipuan atas talian*. Laman Web Rasmi MKN. Retrieved August 27, 2024, from https://www.mkn.gov.my/web/ms/2023/03/28/jenis-penipuan-atas-talian/

[7] Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

[8] Zhang, Y., Chen, J., & Hu, Y. (2021). Exploring machine learning methods for financial fraud detection: A review. *Financial Innovation, 7*(1), 1–20. https://doi.org/10.1186/s40854-021-00232-9

[9] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

[10] Nguyen, T., Le, H. D., & Tran, Q. (2023). Decision tree models in cybersecurity risk analysis: Applications and future directions. *Journal of Cybersecurity Research, 12*(2), 56–73.

[11] Brown, S. (2021). *Machine learning, explained*. *MIT Sloan School of Management*. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[12] Sekaran, U., & Bougie, R. (2016). *Research methods for business: A skill-building approach* (7th ed.). Wiley.

[13] Bhardwaj, P. (2019). Types of sampling in research. *Journal of the Practice of Cardiovascular Sciences, 5*(3), 157–163. https://doi.org/10.4103/jpcs.jpcs_62_19

[14] Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review, 33*(4), 357–370. https://doi.org/10.1016/j.dr.2013.08.003

[15] Sedgwick, P. (2013). Convenience sampling. *BMJ, 347*, f6304. https://doi.org/10.1136/bmj.f6304

[16] Andrade, C. (2021). The inconvenient truth about convenience and purposive samples. *Indian Journal of Psychological Medicine, 43*(1), 86–88. https://doi.org/10.1177/0253717620977000

[17] Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). More than just convenient: The scientific merits of homogeneous convenience samples. *Monographs of the Society for Research in Child Development, 82*(2), 13–30. https://doi.org/10.1111/mono.12296

[18] Salkind, N. J. (2010). *Encyclopedia of research design* (Vol. 1). SAGE.

[19] Okanlawon, A. E., Yusuf, F. A., & Abanikannda, M. O. (2015). University students' knowledge and attitude towards Internet safety: A preliminary study. *Journal of Emerging Trends in Educational Research and Policy Studies, 6*(3), 279–286.

[20] Bijwaard, D. (2020). *Survey on "scams and fraud experienced by consumers" — Final report*. European Institute for Gender Equality. Lithuania. Retrieved January 15, 2025, from https://commission.europa.eu/system/files/2020-01/survey_on_scams_and_fraud_experienced_by_consumers_-_final_report.pdf

[21] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

[22] Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: From efficient prediction to responsible AI. *Frontiers in Artificial Intelligence, 6*, 1124553. https://doi.org/10.3389/frai.2023.1124553

[23] Çorbacıoğlu, Ş. K., & Delil, Ş. (2023). A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine, 23*(4), 165–169. https://doi.org/10.4103/tjem.tjem_155_23

[24] DeLiema, M., Shadel, D., Pak, K., & Niemiec, R. (2021). The financial and psychological impact of identity theft. *Frontiers in Psychology, 12*, 777043. https://doi.org/10.3389/fpsyg.2021.777043

[25] Cole, R., Anderson, S., & Ward, C. (2024). A qualitative investigation of the emotional, physiological, and financial impacts of online romance scams. *Social Science & Humanities Open, 10*(1), 100060. https://doi.org/10.1016/j.ssaho.2024.100060

[26] Wang, J., Liu, X., & Li, H. (2024). The dynamic emotional experience of online fraud victims. *Journal of Criminal Justice, 90*, 102112. https://doi.org/10.1016/j.jcrimjus.2024.102112