Towards a data-driven behavioral approach to prediction of insider-threat

Subhasree Basu*, Yi Han Victoria Chua*, Mei Wah Lee† , Wanyu Geraldine Lim†,
Tomasz Maszczyk†, Zheng Guo*, Justin Dauwels*

*School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore,
{ subhasree.basu, victoriachua, guozheng, jdauwels}@ntu.edu.sg

†{meiwah1992, geraldinelwy, tomasz.maszczyk}@gmail.com

***Abstract*—Insider threats pose a challenge to all companies and organizations. Identification of culprit after an attack is often too late and result in detrimental consequences for the organization. Majority of past research on insider threat has focused on post-hoc personality analysis of known insider threats to identify personality vulnerabilities. It has been proposed that certain personality vulnerabilities place individuals to be at risk to perpetuating insider threats should the environment and opportunity arise. To that end, this study utilizes a game-based approach to simulate a scenario of intellectual property theft and investigate behavioral and personality differences of individuals who exhibit insider-threat related behavior. Features were extracted from games, text collected through implicit and explicit measures, simultaneous facial expression recordings, and personality variables (HEXACO, Dark Triad and Entitlement Attitudes) calculated from questionnaire. We applied ensemble machine learning algorithms and show that they produce an acceptable balance of precision and recall. Our results showcase the possibility of harnessing personality variables, facial expressions and linguistic features in the modeling and prediction of insider-threat.**

***Keywords*-Insider Threat Detection, Behavioral Analysis, Game-based Approach**

I. Introduction

Detecting malicious insiders is a challenging endeavor. Insider attacks are hard to anticipate and counter as they are perpetuated by trusted individuals who are granted access to the organization's assets and are strongly motivated to conceal their actions [1]. Insider crime related to intellectual property (IP) theft cause heavy financial damages related to legal embroilments, loss of patent opportunity, trade secrets and organizational reputation. In Shaw & Stock's critical pathway model, the trajectory towards an insider threat from these individuals can be exacerbated by personal stressors (e.g., financial instability), organizational factors (e.g., lack of safety culture) and professional stressors (e.g., poor performance feedback) or negated by protective factors (e.g., social support) [2], [3].

In this paper, we outline a deception study that provides a low-risk, virtual environment platform to simulate IP theft and collect real-time behavioral data. Using this platform, we collect various real-time behavioral data and investigate if affective, linguistic, and personality features can help in the automated prediction of insider-threat related behaviors. This study is part of a long-term research aiming to develop an objective and automated screening tool for insider threat prevention and intervention. Being able to identify and cluster personality vulnerabilities and relevant features can improve effective intervention for at-risk individuals and prevent these insider attacks. In this respect, our ensemble learning algorithms provide acceptable F-scores while predicting the insider threat behaviors among participants.

The rest of the paper is organized as follows. In section II we provide a brief overview of the related works. Section III describes our game framework used to collect data for our experiments and we elaborate on the experimental setup in section IV. We then report the results and analysis of the data collected in section V and conclude with the summary of our work in section VI.

II. Related Work

Researchers have outlined suspicious behaviors such as counterproductive work behaviors (CWBs) [4] and profiled psychological vulnerabilities such as resentment and entitlement [5]. These profiles suggest the potential of preventing insider attacks, with the premise that the organization is in the position to detect and report these behavior. However, such a method may become unrealistic for a large organization. As such, some researchers have observed the need to develop computer models to predict insider attacks and aid human decision-making when flagging potential insider employees. Schultz [6], outlined a broad framework that addresses cyber, behavioral and psychological elements, while Greitzer et al. [7], evaluated a comprehensive Bayesian net-based model, which identified and ranked 10 simulated scenarios, achieving a high level of agreement with human expert judgements (R^2 = 0.94). Despite providing important indicators to identify potential insiders, the fact that the data is synthetic limits the ecological validity of the study.

Very few studies, to date, managed to apply and model their frameworks on ecological data. Brdiczka and colleagues [8] developed an architecture for insider-threat detection by combining Big Five traits and structural anomaly detection of social and information networks. They predicted

malicious sabotage behaviors (defined as leaving a player guild and harming the progress of the guild) within a dataset of actual players in a multi-player online game, War of Warcraft.

Due to the difficulty of acquiring real-world behavioral data of insiders, actual empirical evidence of insider behaviors is limited [9]. Studies have turned to simulating insider threat scenarios by asking participants to role-play as an insider. These studies found significant differences between insiders and control groups (for a review see [10]). In studies by Azaria et al. [10] and Taylor et al. [11], participants were instructed to role-play and perform data exfiltration tasks (i.e., act in specific ways to obtain some information surreptitiously). These approaches, while closer to real world data, still presents a degree of artificiality in the data as participants were explicitly instructed to act as insiders and may not truly be motivated to do so.

As an effort to create more ecologically valid datasets, recent studies utilized a more unstructured environment where participants are enticed or incentivized to engage in malicious activities in a multi-player game environment. Ho et al. [12] found that participants who chose to betray and deceive their group members exhibited significant linguistic differences during their chat-room interactions. Additionally, a study by Rizzo et al. [13] predefined a list of behavioral indicators and found that engagement and feedback-related behaviors during online social interactions can contribute to the distinguishing between betrayers and non-betrayers.

Beyond deceptive communication and betrayal behaviors, Harilal and colleagues [14] designed a gamified competition, enticing participants to explore malicious strategies to get ahead of the competition. With the aim to observe a wider range of insider-threat related behaviors, they collected a dataset that includes observations of malicious insider instances from multiple data sources such as emails, keystrokes, network traffic, etc. On top of that, the dataset consists of personality traits of participants. However, prediction and detection of insider threat was beyond the scope of their study.

While research on understanding the social interactions of insiders is relevant, our work focuses on IP-theft related behavior and investigates behavioral and personality dispositions of perpetrators who act in isolation when left alone or avoid social communication with others while perpetuating a malicious act. We utilize a covert game environment to study personality and behavioral dispositions associated with individuals who exhibit insider-related behavior in face of organizational stressors and extrinsic rewards. We also showcase the potential of harnessing these behavioral manifestations and personality dispositions as features to distinguish individuals who exhibit IP-theft behaviors and those who do not.

III. System Overview

The game (see Fig. 1) is modeled after a real-world insider threat situation, where perpetrators are trusted with access to a shared information system, faced with stressors and a promising reward for extrinsic motivation to react maliciously. The game framework consists of 5 mini games, the screenshots of which are provided in Fig. 2. Participants have to complete the mini games in order to collect clues and solve a final riddle.

A. Mimicking organizational stressors

Case studies of known insiders reveal that these individuals, when faced with negative feedback and under-appreciation, are more likely to feel disgruntled and susceptible to act in revenge against the organization, should the opportunity arise [15]. In our game platform, we include in-game events and elements designed to mimic organizational stressors that precipitate vulnerable individuals to perpetuate insider attacks. Common stressors include poor performance feedback, under-appreciation, and work pressure [16].

We simulate work pressure through a timer at the top of the screen and the emphasis of time-based performance (see Fig. 1). Moreover, three mini games are cognitive puzzles which are set at a medium difficulty level so that the participants are required to spend considerable amount of time on them. The motivation for these is to test whether the participants are enticed to resort to unfair means while faced with stressful and frustrating situations.

In addition, when playing the game, participants receive periodic notifications that:

1) other dummy participants have completed their tasks,
2) falsely inform that they are not performing up to standards,
3) undermine their efforts in completing the game.

These notifications are designed to mimic poor performance feedback and under appreciation of efforts (see Fig. 1 for an example).

B. Honeypots

There are 2 honeypots which participants can steal information from to boost the speed and accuracy of their performance. The motivation for the placement and ease of access of these honeypots is to model scenarios where the individuals have easy access to insider information and can steal or pry on such information if the need arises.

Upon completion of each mini game, a clue (puzzle piece) to solve the final riddle is uploaded into a folder designated to the participant. Participants are allowed access to the central folder containing all participants' folder, including their own. The participants are explicitly instructed to only access their own folder. However, participants may choose to open other folders and view or steal the clues to solve the mystery. This is the first honeypot accessible to the participants, where the participant can steal clues from other

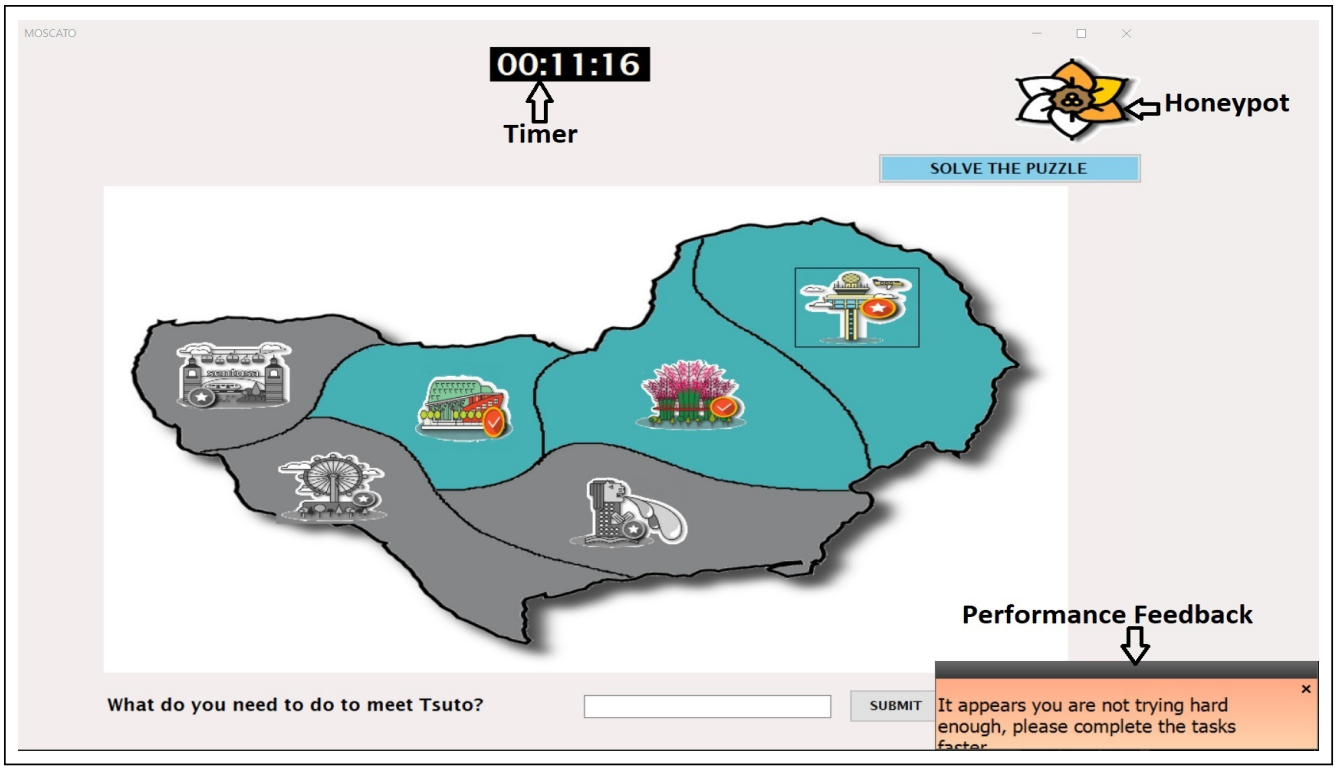


Figure 1: Main screen of the game framework.

participants' folders and complete the final puzzle without playing the mini games.

In addition, the solutions to the three cognitive puzzles are uploaded in the central folder. This is defined as the second honeypot, where participants may access the answers, submit it as their own to achieve the highest number of accurate answers.

C. Implicit and explicit personality measures

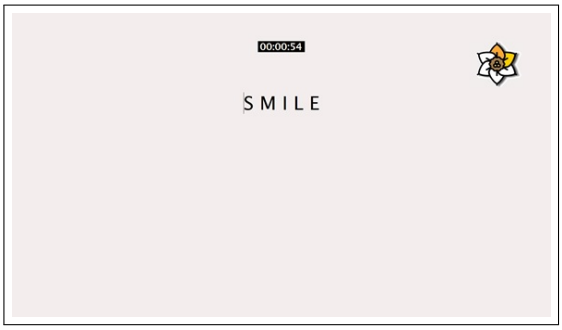
Two implicit tests are disguised as mini games aimed at measuring the participants' emotional predisposition and implicit cognition. The first is a writing exercise (see Fig. 2), where participants are tasked to generate stories for 4 pictures, and the second is a word completion task (WCT in Fig. 2) designed to measure trait affectivity, i.e., an individual's emotional predisposition. High negative trait affectivity has been found to correlate positively with counterproductive work behaviors (CWBs) [17], a strong precursor of insider actions. Moreover, as previous studies have preliminarily shown the possibility of leveraging linguistic cues in chat messages of players acting as insiders [11], we administered these implicit tasks to explore if individual differences in linguistics and trait affectivity could be leveraged as features in the classification of insider threat behavior.

Upon completion of the game i.e., successfully solving the final riddle, participants were electronically administered 3 personality questionnaires, where participants are presented with statements and tasked to rate the extent to which they agree or disagree with the statements.

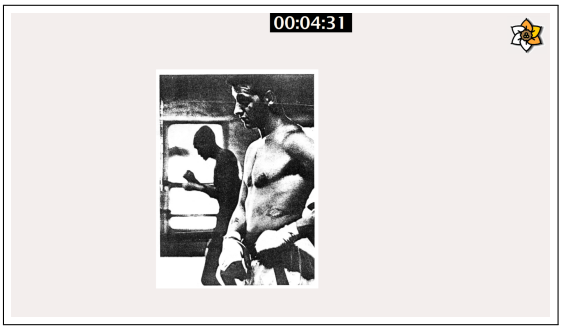
The first questionnaire is the HEXACO Personality Inventory [18], which measures the Big Five traits of Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism, as well as Honesty-Humility. This questionnaire measures six personality traits, namely Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. It consists of 60 statements that reflect the different traits and participants will be required to rate the extent to which they agree or disagree with the statements.

Next, the Dark Triad (DT) of Personality [19] consists of 3 subscales – Narcissism, Psychopathy and Machiavellianism. An example of an item measuring Machiavellianism is "It's wise to keep track of information that you can use against people later". An example of an item measuring Narcissism is "I know that I am special because everyone keeps telling me so", while "Payback needs to be quick and nasty" measures Psychopathy.

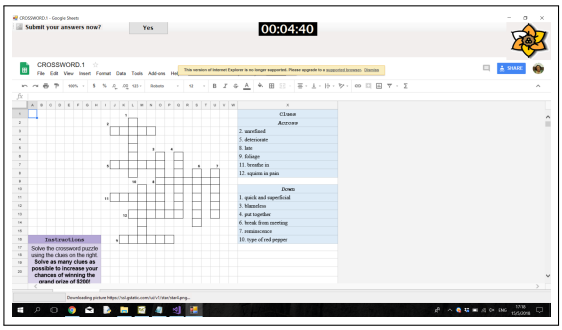
Lastly, the Entitlement Attitudes (EA) Questionnaire measures an individual's level of psychological entitlement [20]. Participants are presented with statements and tasked to rate the extent to which they agree or disagree with the statements. The questionnaire consists of three subscales – Active, Passive and Revenge [21], [22]. Active Entitlement



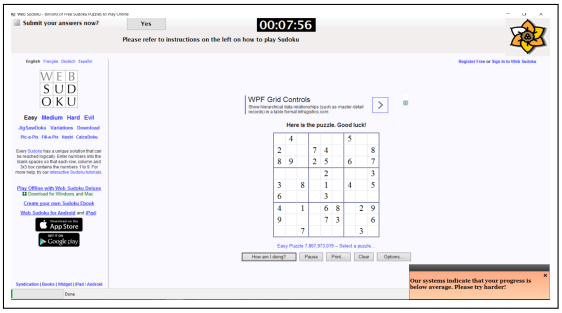
(a) Screenshot for WCT.



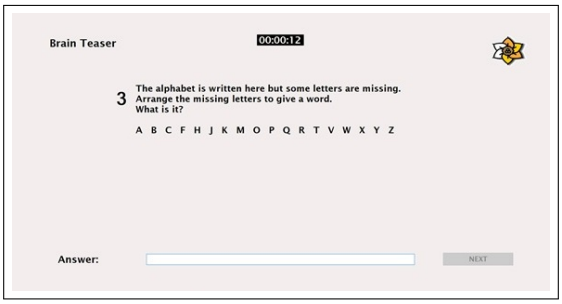
(b) Screenshot for Writing Task.



(c) Screenshot for Crossword.



(d) Screenshot for Sudoku.



(e) Screenshot for Brain Teaser.

Figure 2: Screenshots of minigames.

is based on the promotion of self-interest and self-reliance in achieving life goals and is measured by items such as "I deserve the best". Passive Entitlement is conceptualized as the belief that other people, institutions and social groups that the individual belongs in must serve the interest of the individual. An example of an item measuring Passive Entitlement is "It is the duty of the state to care for all citizens". Lastly, Revenge Entitlement is defined as the tendency to insist on revenge and the inability to forgive prior harms or insults [10]. An example of an item is "I do not forgive sustained insults".

Post-hoc personality analysis of known insiders found that insiders tend to have a strong sense of entitlement, exhibit the Dark Triad personality traits [23] and decreased levels of HEXACO traits [24]. Taken together, we collect a variety of personality variables through implicit and explicit means as a comprehensive effort to delineate potential disposition indicators for insider-threat detection.

IV. Experimental Design

In order to collect data for the analysis of the effectiveness of the game framework, we performed some experiments with the game platform described in Section III. Participants were recruited and briefed that the study was a language and memory game and, they had no knowledge of the actual purpose of the study. Each participant was administered the consent procedure and agreed for their video and audio recordings to be taken. To complete the game, participants had to solve a mystery by collecting puzzle pieces awarded upon completion of each mini game. Participants are in-

formed that that the speed and accuracy of their performance in each mini game will be pitted against other participants for a grand prize of SGD$200. Participants are left alone in the room to complete the game, mimicking the independence and implicit organizational trust awarded to employees [25]. Upon completion of the game, participants were electronically administered the HEXACO questionnaire [18], the Dark Triad Questionnaire [19], the Entitlement Attitudes Questionnaire [20] and a demographics questionnaire. We present a short diagram depicting the flow for experiment in Fig. 3.

V. Analysis and Results

In this section, we describe the dataset we have collected as well as our data analysis pipeline and results.

A. Dataset

The dataset presented in this study is collected from 40 participants (15 males, 25 females). For each subject, we have collected the following:

1) HEXACO, DT and EA scores,
2) WCT scores,
3) linguistic cues extracted from the text generated in the writing task,
4) Facial emotions extracted using Affdex SDK during game and debriefing interview.

We computed the scores for the questionnaires (HEXACO, DT, EA) and WCT as described in [17]–[20] and used the scores as features. For the text generated in the writing task, we applied Linguistic Inquiry and Word Count (LIWC2015) to extract linguistic features. For each set of text, this dictionary-based tool categorizes and counts the number of words that corresponds to several subsets of words representing different linguistic dimensions, psychological states, and affective, social and cognitive processes, providing an 80-dimensional feature vector. A detailed description of LIWC 2015 and the word-subsets are available at [26]. All word counts were normalized by the number of words written by the participant and served as linguistic features for classification.

For the video recordings, we applied Affectiva's expression recognition toolkit, Affdex SDK [27], to extract participants' facial expressions. The feature set comprises 43 cues, including emotions (e.g., joy, fear, sadness, surprise, anger, disgust), expressions from cheek, eye, lip, brow, dimple, etc., and emojis (e.g., flushed, kissing, rage, relaxed, scream, smirk, etc.). Videos are scanned at 3 frames per second. Subsequently, the means and standard deviations of these cues over the whole duration of the video were computed and 86 video features were used for classification.

For the observation of IP-theft related behavior, we categorized participants into the following 2 classes based on their behavior during the game:

- *Class 0*: did not steal clues or solutions to complete the game (n = 30)
- *Class 1*: stole clues and/or solutions to complete the game (n= 10).

The candidates categorized as *Class 1* are are considered to have exhibited IP-theft related behavior.

B. Analysis

In our dataset, we had several missing data points as WCT scores were missing for 7 participants, video recordings were missing for 3 participants due to system failure and lastly, 1 participant did not have WCT scores and video recording.

Thus, for analysis, we have grouped the participants into four categories as follows:

- **Group 1:** Participants with questionnaire and LIWC scores - this includes all 40 participants.
- **Group 2:** Participants with questionnaire and LIWC scores and Affectiva features - this includes 37 participants.
- **Group 3:** Participants with questionnaire, LIWC and WCT scores - this includes 33 participants.
- **Group 4:** Participants with questionnaire and LIWC and WCT scores and Affectiva features - this includes 31 participants.

We perform an early fusion of the features for each of the groups described above and then apply our machine learning model on them (see Fig. 4).

C. Results

For classification, we designed an ensemble learning based classifier for binary classification of the data into the two classes mentioned in Section V-A. We used the Scikit-learn toolkit [28] to perform leave-one-out crossvalidation for classification and prediction tasks. In each crossvalidation loop, one sample was held as the testing set, and the rest of the samples made up the training set. We applied 3-fold cross-validation gridsearch on the training set to select significant features (sorted by ANOVA F-value) and to optimize the parameters of 5 classifiers (Support Vector Machine (SVM), Logistic Regression (LR), GradientBoost, AdaBoost, and RandomForest). Additionally, in the 3-fold cross-validation grid-search, we applied standardization to balance and scale the training set. Finally, we used these 5 optimized classifiers to predict the testing data and soft vote the prediction scores to provide the final prediction of the testing set. We report the confusion matrix, precision, recall, F-score as well as the overall accuracy for the classification method in Tables I– IV. We have reported the results based on the groups enumerated in Section V. Since we have facial emotions extracted for two cases - while participants were playing the game and while the participants were being debriefed, we have presented results for *Group 2* and *Group 4* for the following two cases:-

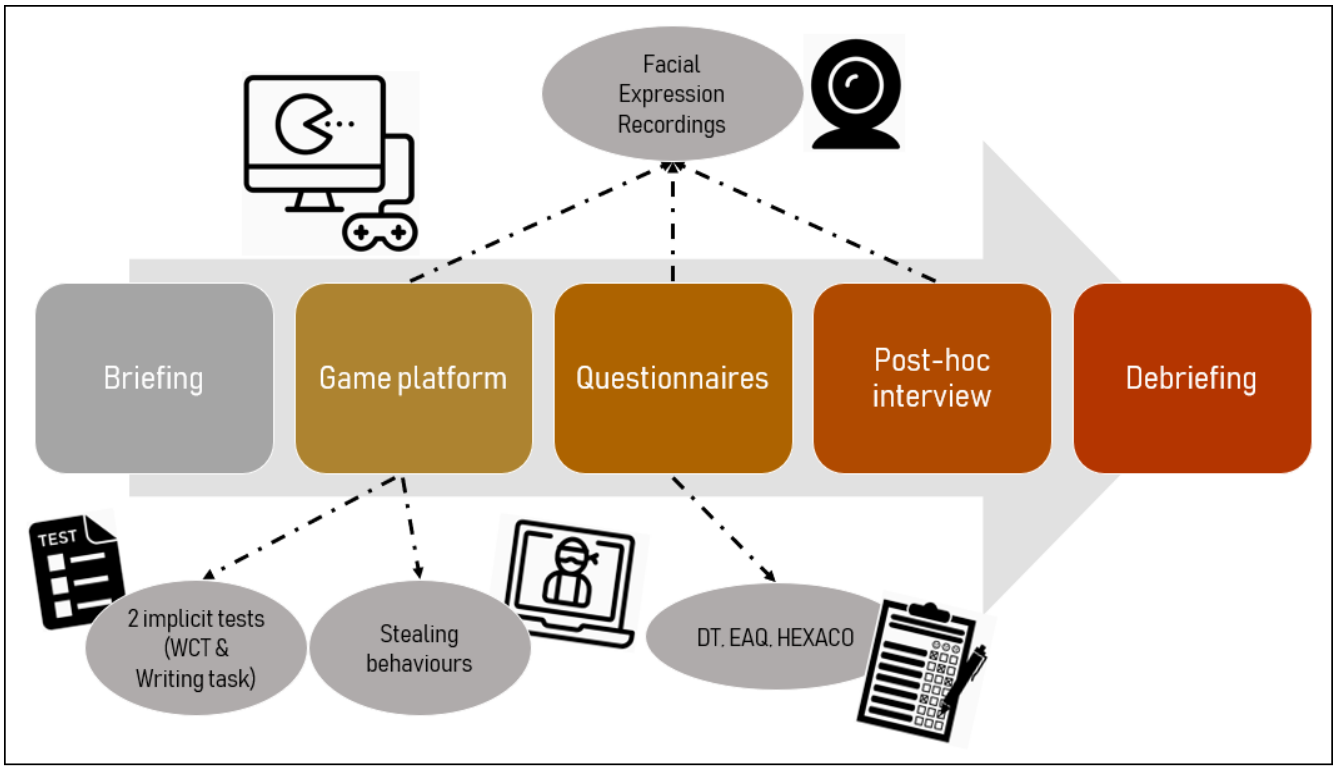


Figure 3: Experimental Design.

1) **Game Videos** - facial emotions extracted while participants were playing the game.
2) **Interview Videos** - facial emotions extracted while participants were being debriefed.

Feature	Confusion Matrix	Class 1	Class 0	Precision	Recall	F-Score	Accuracy
Group 1							
Questionnaire Only	**Class 1**	7	3	0.70	0.7778	0.7370	0.875
	Class 0	2	28	0.9333	0.9032	0.8946	
LIWC Only	**Class 1**	4	6	0.4	0.3333	0.3636	0.650
	Class 0	8	22	0.7333	0.7857	0.7586	
Questionnaire + LIWC	**Class 1**	6	4	0.60	0.75	0.6667	0.85
	Class 0	2	28	0.9333	0.9032	0.8946	

Table I: Performance metrics of classifying participants' behavior for different combinations of features in Group 1.

Feature	Confusion Matrix	Class 1	Class 0	Precision	Recall	F-Score	Accuracy
Group 2							
Game Videos							
LIWC + Affectiva	**Class 1**	5	2	0.3333	0.7142	0.4545	0.67567
	Class 0	10	20	0.9090	0.6667	0.7692	
Questionnaire + Affectiva	**Class 1**	6	1	0.4	0.8571	0.54545	0.0.7297
	Class 0	9	21	0.9545	0.7	0.8077	
Questionnaire + LIWC + Affectiva	**Class 1**	7	0	0.3684	1	0.5385	0.6757
	Class 0	12	18	1	0.6	0.75	
Interview Videos							
LIWC + Affectiva	**Class 1**	4	3	0.3333	0.5714	0.4210	0.7027
	Class 0	8	22	0.88	0.7333	0.80	
Questionnaire + Affectiva	**Class 1**	5	2	0.5556	0.7143	0.625	0.8378
	Class 0	4	26	0.9286	0.8667	0.8966	
Questionnaire + LIWC + Affectiva	**Class 1**	7	0	0.35	1	0.5185	0.6486
	Class 0	13	17	1	0.5667	0.7234	

Table II: Performance metrics of classifying participants' behavior for different combinations of features in Group 2.

Feature	Confusion Matrix	Class 1	Class 0	Precision	Recall	F-Score	Accuracy
Group 3							
LIWC + WCT	**Class 1**	3	4	0.2308	0.4286	0.3	0.5758
	Class 0	10	16	0.8	0.6154	0.6957	
Questionnaire + LIWC	**Class 1**	3	4	0.3333	0.4286	0.375	0.6970
	Class 0	6	20	0.8333	0.7692	0.8	
Questionnaire + WCT	**Class 1**	3	4	0.6	0.4286	0.5	0.8182
	Class 0	2	24	0.8571	0.9231	0.8889	
Questionnaire + WCT + LIWC	**Class 1**	3	4	0.3	0.4286	0.3529	0.6667
	Class 0	7	19	0.8261	0.7308	0.7755	

Table III: Performance metrics of classifying participants' behavior for different combinations of features in Group 3.

Feature	Confusion Matrix	Class 1	Class 0	Precision	Recall	F-Score	Accuracy
Group 4							
Game Videos							
WCT + Affectiva	**Class 1**	4	2	0.3636	0.6667	0.4706	0.7097
	Class 0	7	18	0.9	0.72	0.8	
Questionnaire + WCT + Affectiva	**Class 1**	4	2	0.2667	0.6667	0.3809	0.5806
	Class 0	11	14	0.875	0.56	0.6829	
LIWC + WCT + Affectiva	**Class 1**	3	3	0.1765	0.5	0.2609	0.4516
	Class 0	14	11	0.7857	0.44	0.5641	
Questionnaire+LIWC + WCT + Affectiva	**Class 1**	6	0	0.3158	1.0	0.48	0.5806
	Class 0	13	12	1.0	0.48	0.6486	
Interview Videos							
WCT + Affectiva	**Class 1**	3	3	0.2727	0.5	0.3529	0.6451
	Class 0	8	17	0.85	0.68	0.7556	
Questionnaire + WCT + Affectiva	**Class 1**	3	3	0.2727	0.5	0.3529	0.6452
	Class 0	8	17	0.85	0.68	0.7556	
LIWC + WCT + Affectiva	**Class 1**	3	3	0.2	0.5	0.2857	0.5161
	Class 0	12	13	0.8125	0.52	0.6341	
Questionnaire + LIWC + WCT + Affectiva	**Class 1**	3	3	0.1875	0.5	0.2727	0.4839
	Class 0	13	12	0.8	0.48	0.6	

Table IV: Performance metrics of classifying participants' behavior for different combinations of features in Group 4.

As our data is highly imbalanced, accuracy is not an appropriate and sufficient performance metric to evaluate our classification model. In the context of insider threats,

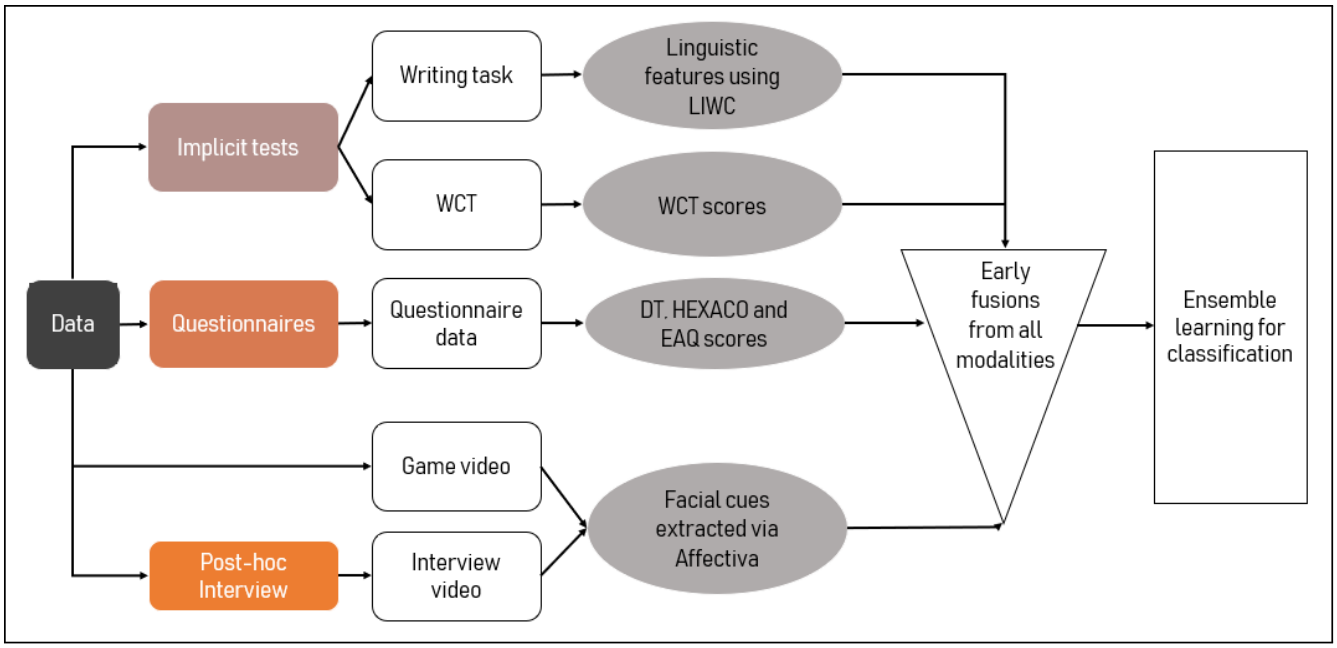


Figure 4: Data Analysis Pipeline.

favoring precision over recall (or vice versa) is much more nuanced as the issues of false positives could create mistrust in the organization, loss of resources channeled at innocent employees and delay in dealing with actual insiders. On the other hand, the heavy damages related to insider threat highlight the need to minimize false negatives as well. Thus, in this case, an F-score assessment, indicating a harmonic balance between precision and recall, is more appropriate.

To evaluate the fidelity of using each group of features for binary classification, we computed baseline values using dummy classifier that classifies all participants into the majority class (i.e., *Class 0*: No observation of IP-theft related behavior). The baseline accuracy for *Group 1* is 0.75 and the baseline average F-score is 0.64. The baseline accuracy for *Group 2* is 0.81 and the baseline average F-score is 0.726. The baseline accuracy for *Group 3* is 0.7878 and the baseline average F-score is 0.654. The baseline accuracy for *Group 4* is 0.806 and the baseline average F-score is 0.720. As seen from the tables, our ensemble classifier achieved higher F-scores than baseline F-scores in quite a few cases using each group of features individually and in combinations.

VI. Conclusion

In this paper, we presented a novel approach towards collecting real-world data of insider-like behavior and explored the potential of visual cues, personality traits and linguistic features to classify individuals who exhibit risky behaviors of interest (i.e., engage in stealing of information) and individuals who do not. We collected personality variables, facial expressions, and linguistic features of individuals who exhibited IP theft-related behavior, specifically stealing information for personal gain. As a preliminary study, we have shown that our game elicits different behavioral responses from individuals. When faced with obstacles, stress, and a tantalizing reward, participants in the study reacted in a myriad of diverse ways, specifically, a few participants engaged in the act of stealing to achieve their goals. More importantly, our results are promising and show that there are significant personality and behavioral differences between individuals that engage in stealing behaviors and those who do not. Understanding and harnessing these differences can be valuable in the prevention and intervention of insider threats. We intend to collect a larger dataset, explore more demographics, personality dispositions and behavioral cues as features. We will also improve our classification and feature selection methods to delineate discriminating features and detect common patterns among individuals who exhibit insider-threat behaviors.

VII. Acknowledgment

Subhasree Basu and Yi Han Victoria Chua contributed equally to this paper.

References

[1] J. R. C. Nurse *et al.*, "Understanding Insider Threat: A Framework for Characterising Attacks." in *IEEE Security and Privacy Workshops*, 2014, pp. 214–228.

[2] E. D. Shaw, J. M. Post, and K. G. Ruby, "Inside the Mind of the Insider." *Security Management*, vol. 43, no. 12, pp. 34–42, 1999.

[3] D. Charney, "True psychology of the insider spy." *Intelligencer: Journal of the US Intelligence Studies*, vol. 18, no. 1, pp. 47–54, 2010.

[4] R. H. Searle and C. Rice, "Assessing and Mitigating the Impact of Organisational Change on Counterproductive Work Behaviour: an Operational (Dis) trust Based Framework." 2018.

[5] A. P. Moore *et al.*, "A preliminary model of insider theft of intellectual property." 2011.

[6] E. E. Schultz, "A framework for understanding and predicting insider attacks." *Computers & Security*, vol. 21, no. 6, pp. 526–531, 2002.

[7] F. L. Greitzer and D. A. Frincke, "Combining traditional cyber security audit data with psychosocial data: towards predictive modeling for insider threat mitigation." in *Insider Threats in Cyber Security*, 2010, pp. 85–113.

[8] O. Brdiczka, J. Liu, B. Price, J. Shen, A. Patil, R. Chow, E. Bart, and N. Ducheneaut, "Proactive insider threat detection through graph learning and psychological context," pp. 142–149, 2012.

[9] R. Willison and M. Warkentin, "Beyond deterrence: An expanded view of employee computer abuse." *MIS quarterly*, vol. 37, no. 1, 2013.

[10] A. Azaria, A. Richardson, S. Kraus, and V. Subrahmanian, "Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data." *IEEE Transactions on Computational Social Systems*, vol. 1, no. 2, pp. 135–155, 2014.

[11] P. J. Taylor *et al.*, "Detecting insider threats through language change." *Law and human behavior*, vol. 37, no. 4, p. 267, 2013.

[12] S. M. Ho, J. T. Hancock, C. Booth, M. Burmester, X. Liu, and S. S. Timmarajus, "Demystifying insider threat: Language-action cues in group dynamics," in *49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp. 2729–2738.

[13] P. Rizzo, C. Jemmali, A. Leung, K. Haigh, and M. S. El-Nasr, "Detecting betrayers in online environments using active indicators," in *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2018, pp. 16–27.

[14] A. Harilal, F. Toffalini, J. Castellanos, J. Guarnizo, I. Homoliak, and M. Ochoa, "Twos: A dataset of malicious insider threat behavior based on a gamified competition," in *Proceedings of the International Workshop on Managing Insider Security Threats*, 2017, pp. 45–56.

[15] E. Shaw, K. Ruby, and J. Post, "Insider Threats to Critical Information Systems: Typology of Perpetrators, Security Vulnerabilities, Recommendations." *Task Letter Number 001: Insider Threat Profile*, 1999.

[16] D. M. Cappelli, A. P. Moore, and R. F. Trzeciak, *The CERT guide to insider threats: how to prevent, detect, and respond to information technology crimes (Theft, Sabotage, Fraud).*, 2012.

[17] R. E. Johnson, A. L. Tolentino, O. B. Rodopman, and E. Cho, "We (sometimes) know not how we feel: Predicting job performance with an implicit measure of trait affectivity." *Personnel Psychology*, vol. 63, no. 1, pp. 197–219, 2010.

[18] K. Lee and M. C. Ashton, "The HEXACO personality factors in the indigenous personality lexicons of English and 11 other languages." *Journal of personality*, vol. 76, no. 5, pp. 1001–1054, 2008.

[19] D. L. Paulhus and K. M. Williams, "The dark triad of personality: Narcissism, Machiavellianism, and psychopathy." *Journal of research in personality*, vol. 36, no. 6, pp. 556–563, 2002.

[20] M. A. Żemojtel-Piotrowska *et al.*, "Measurement of psychological entitlement in 28 countries." *European Journal of Psychological Assessment*, 2015.

[21] M. Żemojtel-Piotrowska, T. Baran, A. Clinton, J. Piotrowski, S. Bălţătescu, and A. Van Hiel, "Materialism, subjective well-being, and entitlement." *Journal of Social Research & Policy*, vol. 4, no. 2, 2013.

[22] M. A. Żemojtel-Piotrowska, J. P. Piotrowski, and A. Clinton, "Agency, communion and entitlement," *International Journal of Psychology*, vol. 51, no. 3, pp. 196–204, 2016.

[23] M. Maasberg, J. Warren, and N. L. Beebe, "The dark side of the insider: detecting the insider threat through examination of dark triad personality traits." in *48th Hawaii International Conference on System Sciences (HICSS)*, 2015, pp. 3518–3526.

[24] R. E. De Vries and D. van Kampen, "The HEXACO and 5DPT models of personality: A comparison and their relationships with psychopathy, egoism, pretentiousness, immorality, and Machiavellianism." *Journal of Personality Disorders*, vol. 24, no. 2, pp. 244–257, 2010.

[25] Intelligence and N. S. Alliance, "Assessing the Mind of the Malicious Insider: Using Behavioral Analytics to Improve Continuous Evaluation." 2017.

[26] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015." Tech. Rep., 2015.

[27] D. McDuff *et al.*, "Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected" In-the-Wild"." in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2013, pp. 881–888.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.