

Department of Informatics

INDIVIDUAL ASSIGNMENT

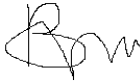
Surname	Bhunu							
Initials	M							
Student Number	2	3	9	5	3	0	0	5
Module Code	INF				7	9	1	
Assignment number	1							
Name of Lecturer	Dr. Mike Nkongolo							
Date of Submission	2024/09/20							
<p>Declaration:</p> <p>I declare that this assignment, submitted by me, is my own work and that I have referenced all the sources that I have used.</p>								
Signature of Student								

TABLE OF CONTENTS

1. INTRODUCTION.....	1
2. LITERATURE REVIEW.....	1
3. DATA COLLECTION AND PREPARATION	2
3.1 DATA SOURCES.....	2
3.2 DATA DESCRIPTION.....	2
3.3 QUALITY ASSESSMENT CRITERIA.....	2
3.4 PREPROCESSING.....	3
4. METHODOLOGY	5
4.1 ANALYTICAL APPROACH	5
4.1.1 Relationship between the numerical features.....	5
4.2 MODELING	5
4.3 TOOLS AND LIBRARIES	6
5. RESULTS.....	6
5.1 MODEL PERFORMANCE	6
5.2 INSIGHTS.....	7
5.3 MODEL PERFORMANCE VISUALIZATIONS	7
5.3.1 Model accuracies	7
5.3.1 Precision, recall and f1-score fluctuations.....	8
5.3.2 Mean cross validation accuracy.....	9
5.4 RESEARCH SUPPORTING VISUALIZATIONS	10
5.4.1 Time Distribution between roles and difficulty levels	10
5.4.2 Win and loss ratios across different levels of the game.....	11
6. DISCUSSION.....	12

6.1	INTEPRETATION	12
6.2	LIMITATIONS	12
7.	CONCLUSION.....	13
7.1	SUMMARY.....	13
7.2	RECOMENDATIONS	13
7.3	FUTURE WORK.....	13
8.	REFERENCES.....	14

ANALYSIS FOCUS

Predicting game outcomes in cybergaming: The role of game duration and player performance.

LIST OF FIGURES

Figure 1 Time Distribution Before SRT and Outlier Removal.	3
Figure 2 Time Distribution after SRT and Outlier Removal.	4
Figure 3 Time, Defender and Attacker Score After Outlier Removal.	4
Figure 4 Corelations between numerical features of the dataset.	5
Figure 5 Model Accuracy	7
Figure 6 Lazy Predict Results.....	8
Figure 7 Precession, Recall and F1-Scores of the Models.	8
Figure 8 Confusion matrices for the models.	9
Figure 9 Mean Cross Validation Accuracy of Models.	9
Figure 10 Time Distribution of Game Duration by Role (Defender vs Attacker)	10
Figure 11 Time Distribution across each level.....	10
Figure 12 Win and loss ratios at different levels of the game.....	11

LIST OF TABLES

Table 1 Table Showing Performance of Machine Learning Models.	7
--	---

ABSTRACT

The study aims to explore the relationship of game durations, player performance, and game outcome in a controlled cybersecurity environment through gaming. It makes use of a dataset that monitored the outcome, the score of the players, and game duration to predict the winner by using different machine learning models such as Random Forest, Support Vector Machine Learning, and Naive Bayes to predict the winner in each match. This research aims to understand how these features are tipping the balances of a match and how well we could influence the outcomes to raise awareness among people of the importance of cybersecurity in order to avoid online frauds and create a better environment for organizations.

The data were prepared for analysis by extensive processing, such as removing the impact of outliers with Interquartile ranges and transforming game duration with a square root transformation (SRT) to reduce the skewness. Each model is validated using a 5-fold cross-validation test to make sure that the model accuracy is not due to over fitting or under fitting. In accuracy, the three machine learning models performed the same, with Ensemble learning performing slightly better. From the results of this study, aside from player scores, game duration determines the rate of becoming a match winner; longer games favour the Defenders, while the shorter games favour the Attackers. These results are indicative of the fact that machine learning models can be applied in such a way that they can predict the outcomes of games using just a few features to provide more insights about player behavior and game balance.

1. INTRODUCTION

The concept of serious gaming has proven to be quite effective in educating individuals in different industries (Pramod, 2024). Serious gaming is not just designed to entertain but is meant to educate individuals especially in concepts that are related to cybersecurity awareness and training. The concept of cybergaming refers to the use of game-based methods to teach and enhance cybersecurity skills to reduce the susceptibility to online threats to individuals. The combination of gaming concepts and education keeps players engaged enhancing their knowledge through the gaming experience (Awojana and Chou, 2019).

The study is guided by the research question, predicting player behaviour, and enhancing game balance. The study aims to determine how does game duration and other gameplay factors influence the outcome of a cybergaming matches between the (Defender and Attacker) leveraging gameplay data from a number of players in a cybersecurity game, CybeVigillance and also understanding player behaviour. The defender and the attacker have a fair chance of actually winning the game. The use of machine learning models can be used to determine patterns in how players actually make strategic decisions, providing valuable insights the games design and also training on cybersecurity. The main objective of the study is to evaluate game duration, defender and attacker scores and other gameplay factors influence on the outcome of a game.

Participants were requested to prepare a dataset from 10 consecutive gameplays and transfer the gaming data to the central repository as a CSV file. Exploratory Data Analysis techniques using various charts were implemented to show the distribution of the data. Several machine learning models were used such as Random Forest, Support Vector Machine learning (SVM), and Naïve Bayes were implemented to predict the game outcomes and factors which influence the result. The research enables us to better understand game balance, which in turn helps us come up with better design of the cybersecurity games.

2. LITERATURE REVIEW

Various research has been carried out to try and mitigate the risks that individuals may face from hackers. According to the research carried out by Nkongolo, 2024, while maintaining a cultural heritage through the implementation of a Cybersecurity game CybeVigillance based on the traditional South

African game Morabaraba, confirms that Cybersecurity awareness games such as these can enhance the security of organizations by educating users about cybersecurity.

Research carried out by Hafeez *et al.*, 2021 which leveraged the use of Brute Force which is a tower defence game that is designed to educate players about cybersecurity, by challenging them at each and every level resulted in players having a more responsible attitude towards the use of strong and unique passwords after playing the game.

One of the issues that is coupled by research on Cybersecurity awareness is the lack of longitudinal data. Both of the cyber security awareness games mentioned earlier aim at accessing the immediate effect that they have on awareness. The research that is being carried out in this paper tends to uncover other gaps such as player profiling based on timing and game levels to determine how well players understand the game being used to generate the data entailing how well they apply cyber security concepts in the real world at after the research.

3. DATA COLLECTION AND PREPARATION

3.1 DATA SOURCES

The research participants were asked to upload a Comma Separated Value (CSV) file to a google drive central repository that contained 10 consecutive gameplays. Participants were asked to complete the gameplays over a period of three days and upload the file.

3.2 DATA DESCRIPTION

The data collected was in the form of Comma Separated Values that were generated from a player's 10 consecutive gameplays. The columns from the data included a player's Nickname, the Defenders Score, Attackers Score, Time in Seconds, the Winner, and the Level of the player based on their gameplay statistics. The Defender, Attacker Scores and Time were the only numerical features recorded. A gameplays time was recorded over a period of time where the user was engaged in the gameplay. The score tallies were kept by the fields Defender and Attacker Scores. The data seems to indicate that there is always a winner but in certain circumstances of rare occurrences a draw would emerge depicting an equal understanding of a player's defender and attacker ticks.

3.3 QUALITY ASSESSMENT CRITERIA

To assess the quality of data, the data had to include 10 consecutive gameplays from each single player. This was difficult to tell if a player had 10 gameplays due to some gameplays having 10 different

Nicknames from a single file. Files that were uploaded in formats other than CSV files were excluded from the dataset. To have high quality data, all of the fields that had duplicates or even missing features were dropped from the original dataset. After data processing, draws were quite scarce in the combined dataset with 11, in a data shape of over 1200 gameplays. To avoid class imbalance, the draws were discarded.

3.4 PREPROCESSING

To prepare the data for machine learning, various methods were implemented to counter the underlying issues that were in the data. The Time in seconds variable for instance, possessed a positive skew as opposed to other distributions of the data. Before performing the square root transformation, which was slightly less harsh on the time values as opposed to other methods of transformation. All the numeric values that had values that were over or below the 1.5 interquartile range (IQR) threshold were discarded. Finally, there were only 11 draws that remained in the dataset, they were discarded to avoid class imbalance because of underrepresentation. The following charts show the time value before the square root and IQR outlier removal were applied to the time distribution so as to reduce the level of skewness of the time distribution.

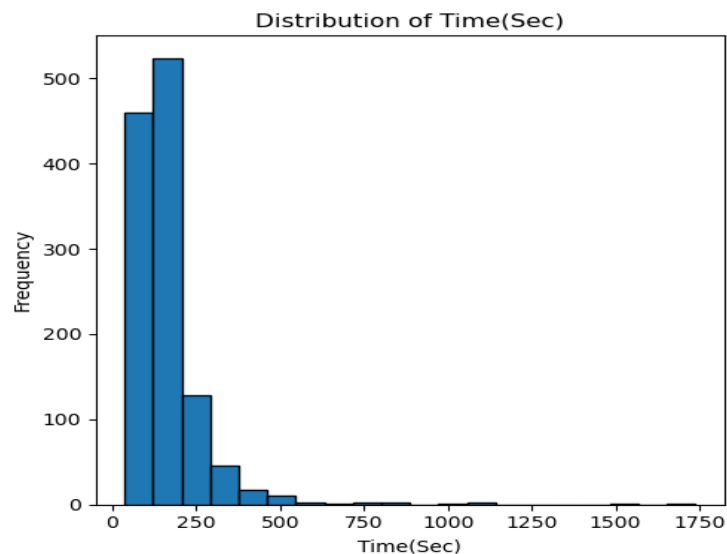


Figure 1 Time Distribution Before SRT and Outlier Removal.

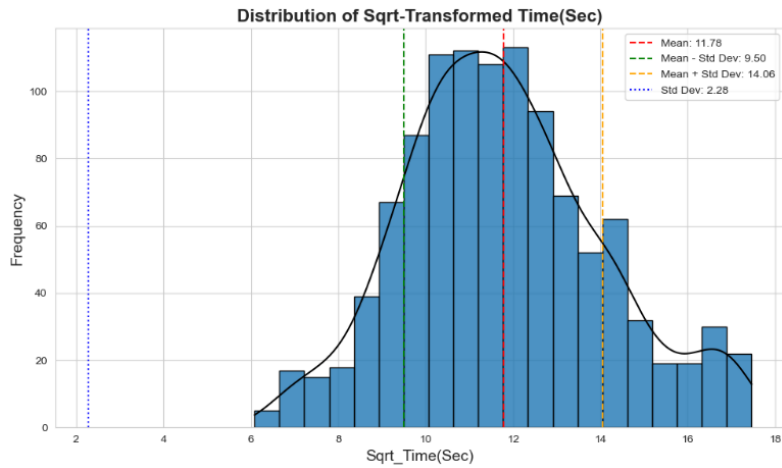


Figure 2 Time Distribution after SRT and Outlier Removal.

The SRT played a huge part in removing the skewness of the Time (Sec) distribution. Outliers were removed by excluding any data which were outside the 1.5 threshold set for the IQR. The square root time was used throughout the study to build machine learning models and understand different features of the dataset. Fortunately, the other features from the data that was collected had data that was normally distributed there was no need to perform any adjustments to the data except for outlier removal. Figure 3 shows the distribution of the three numeric features the Attacker, Defender Scores and Time Distribution after outlier removal.

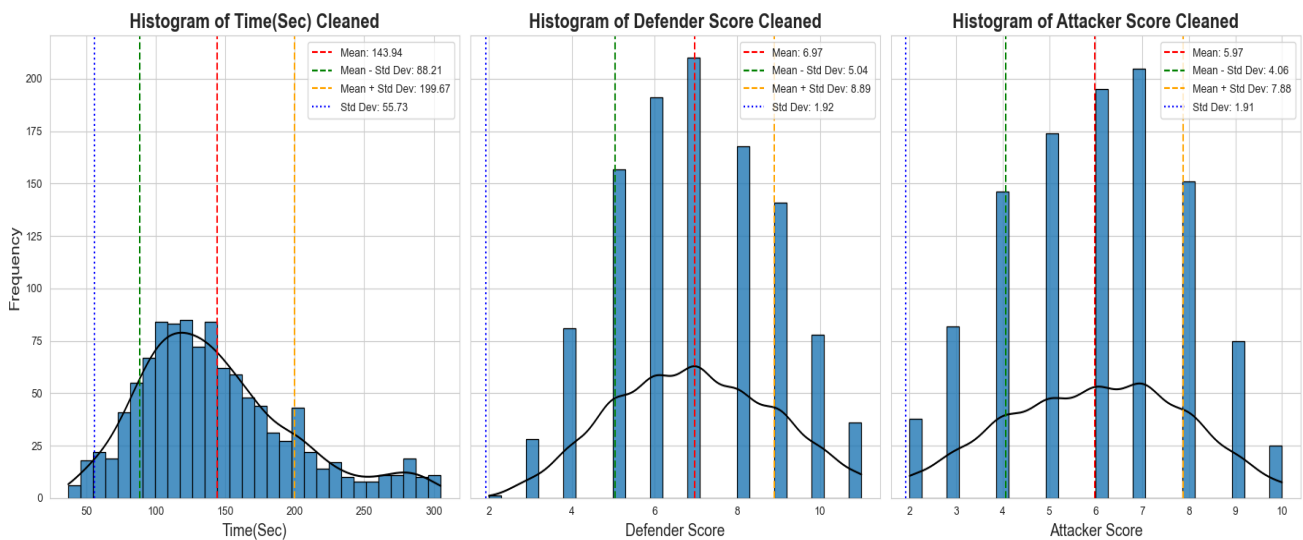


Figure 3 Time, Defender and Attacker Score After Outlier Removal.

4. METHODOLOGY

4.1 ANALYTICAL APPROACH

Exploratory Data Analysis (EDA) was carried out to understand the relationship between the different features of the dataset and how they affect the outcomes of the game.

4.1.1 RELATIONSHIP BETWEEN THE NUMERICAL FEATURES

The relationship between the different features such as Sqrt_Time (Sec), Attacker and Defender scores were analysed with a correlation matrix showing a high negative correlation between the Attacker and Defender. As Attacker Scores increases the Defender Scores decrease. Surprisingly, the time does not seem to have any effect on the player's score with a positive correlation of **0.13** on Attacker Scores and a negative correlation of **-0.13** on Defender Scores. Both of these correlations are very close to 0, thus the length of the game does not seem to have any effect on a player's score. In other words, player performance is not related to the duration of the game, however the time can still be used in predictive analysis in training models to determine the winner between the defender and attacker. The following correlation map shows the relationship between the numerical features used in training the machine learning models.

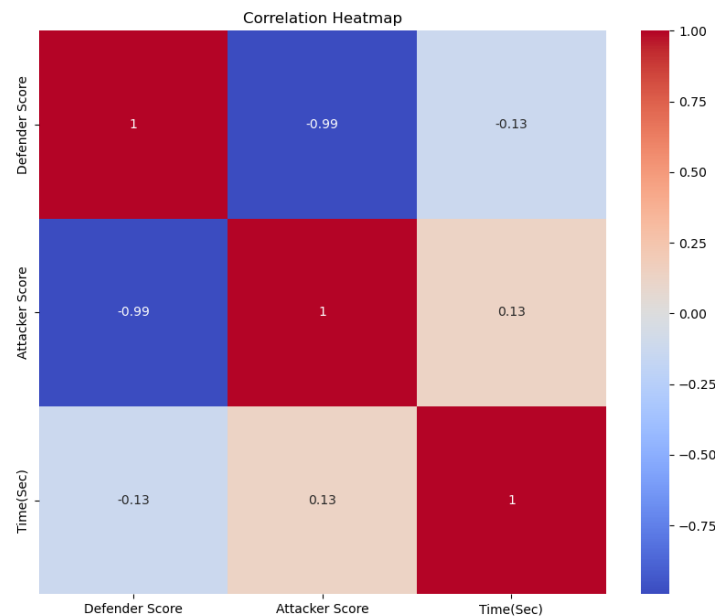


Figure 4 Correlations between numerical features of the dataset.

4.2 MODELING

Three supervised machine learning models were implemented for predictive analytics. The Random Forest which is a non-linear model that excels mostly with unbalanced data and is useful in determining feature importance, Support Vector Machine (SVM) which is strong in classification of tasks with non-

linear boundaries and Finally Naïve Bayes, a probabilistic model that is quick over large datasets resulting in faster computations. The evaluation of the models was based on the Accuracy which is the total number of predictions over a certain number of predictions. The Precision, Recall and F1-Score were used to measure the quality of predictions. Confusion matrixes were also visualized for each and every model showing the true positives, true negatives, false positives, and false negatives. Finally, to determine that the models are not overfitting or underfitting, the 5-fold cross-validation was implemented to ensuring the models robustness.

The three models were selected for various reasons. The Random Forest was selected for its ability to handle large datasets with non-linear relationships and identifying important features. The SVM was selected for its ability to find non-linear decision boundaries. For simplicity when working with a smaller dataset and speed, the Naïve Bayes was selected. The implementation of the models was based on 80% training and 20% testing data split the train_test_split method.

For further analysis, a number of machine learning models were testing using a Python package called Lazy predict, that automatically trains and compares multiple machine learning models.

4.3 TOOLS AND LIBRARIES

The programming language that was used for this analysis was Python. Pandas, a library that is used in data analysis was used manipulation of data and preprocessing. This includes feature engineering and handling missing values. To get a sense of the data and how it looks for a quick EDA overview a Python Library, Sweetviz was implemented to provide a quick insight on the data that was generated. NumPy a Python library was used for Arithmetic calculations as well as array manipulations. For visualization of data, Matplotlib and Seaborn were used for EDA and data visualizations. The Scikit-learn was used for training and evaluation of the three models that were implemented.

5. RESULTS

The section explains the different models that were tested for predicting game outcomes which are the Defender and Attacker wins. This research evaluates a number of machine learning models, Ensemble Learning, Random Forest, SVM and Naïve Bayes.

5.1 MODEL PERFORMANCE

Cross validation accuracy was carried out to determine if the machine learning models were overfitting or underfitting. The best performing model was the ensemble learning with an accuracy of 95% also achieving very high precision recall and then F1 score. The cross validation of the models was carried

out over a 5-fold validation, and still yielded high accuracy. The following table shows the performance of the different machine learning models that were tested in this research.

Table 1 Table Showing Performance of Machine Learning Models.

Model	Accuracy	Precession	Recall	F1 Score	Cross Validation Accuracy
Random Forest	0.94	0.95	0.93	0.95	0.93
SVM	0.94	1.00	0.90	0.95	0.91
Naïve Byes	0.94	1.00	0.90	0.95	0.91
Ensemble Learning	0.95	0.98	0.93	0.95	0.92

5.2 INSIGHTS

When it comes to comparing the different models all the models performed slightly the same the Random Forest, SVM and Naive Bayes yielding an accuracy of 94%. The only different difference lies in the precision and recall values with Naïve Bayes and SVM having a precession of 100% but slightly lower recall.

5.3 MODEL PERFORMANCE VISUALIZATIONS

All the machine learning models performed slightly the same. The accuracy of the machine learning models was very high which indicated that the data may be facing some issues of overfitting and underfitting. Cross validation was carried out to verify if the machine learning model was underfitting or over fitting.

5.3.1 MODEL ACCURACIES

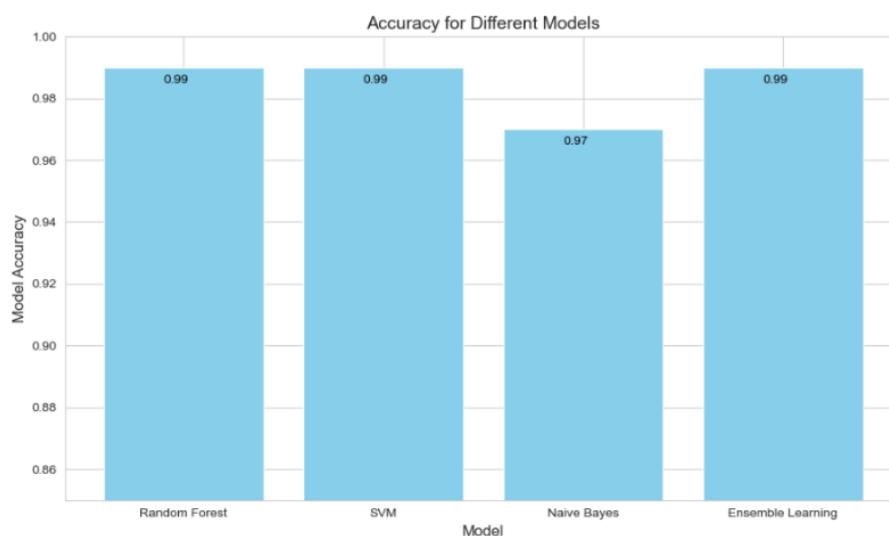


Figure 5 Model Accuracy

As highlighted earlier the models performed quite well. Three of the models Random Forest, SVM and Ensemble learning performed quite similarly. Naïve Bayes had the lowest accuracy but still it performed well.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
PassiveAggressiveClassifier	0.94	0.95	0.95	0.94	0.02
BernoulliNB	0.94	0.95	0.95	0.94	0.02
NuSVC	0.94	0.95	0.95	0.94	0.04
NearestCentroid	0.94	0.95	0.95	0.94	0.03
SVC	0.94	0.95	0.95	0.94	0.02
GaussianNB	0.94	0.95	0.95	0.94	0.02
QuadraticDiscriminantAnalysis	0.94	0.95	0.95	0.94	0.05
RidgeClassifier	0.94	0.95	0.95	0.94	0.03
RidgeClassifierCV	0.94	0.95	0.95	0.94	0.02
LinearDiscriminantAnalysis	0.94	0.95	0.95	0.94	0.04
LabelSpreading	0.95	0.94	0.94	0.95	0.09
KNeighborsClassifier	0.95	0.94	0.94	0.95	0.03
LabelPropagation	0.94	0.94	0.94	0.94	0.06
ExtraTreesClassifier	0.94	0.94	0.94	0.94	0.14
XGBClassifier	0.94	0.94	0.94	0.94	0.15
AdaBoostClassifier	0.94	0.93	0.93	0.94	0.15
ExtraTreeClassifier	0.94	0.93	0.93	0.94	0.02
DecisionTreeClassifier	0.94	0.93	0.93	0.94	0.03
LGBMClassifier	0.93	0.93	0.93	0.93	0.15
RandomForestClassifier	0.93	0.92	0.92	0.93	0.19
BaggingClassifier	0.93	0.92	0.92	0.93	0.06
CalibratedClassifierCV	0.91	0.91	0.91	0.91	0.06
LinearSVC	0.91	0.91	0.91	0.91	0.02
LogisticRegression	0.91	0.91	0.91	0.91	0.03
SGDClassifier	0.89	0.86	0.86	0.89	0.02
Perceptron	0.89	0.85	0.85	0.88	0.02
DummyClassifier	0.61	0.50	0.50	0.46	0.01

Figure 6 Lazy Predict Results

5.3.1 PRECISION, RECALL AND F1-SCORE FLUCTUATIONS

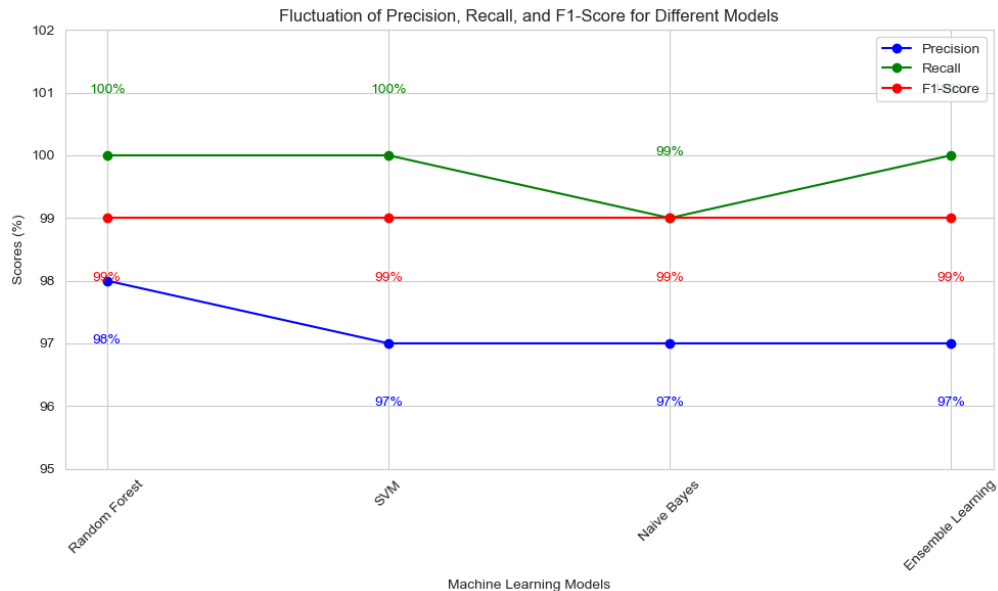


Figure 7 Precession, Recall and F1-Scores of the Models.

Figure 7 highlights that all the models perform the same. There are slight fluctuations in the Precession and Recall, however the F1 scores are constant throughout. Figure 8 shows all the confusion matrices from the performance of the different machine learning models. All the models seem to have performed well.

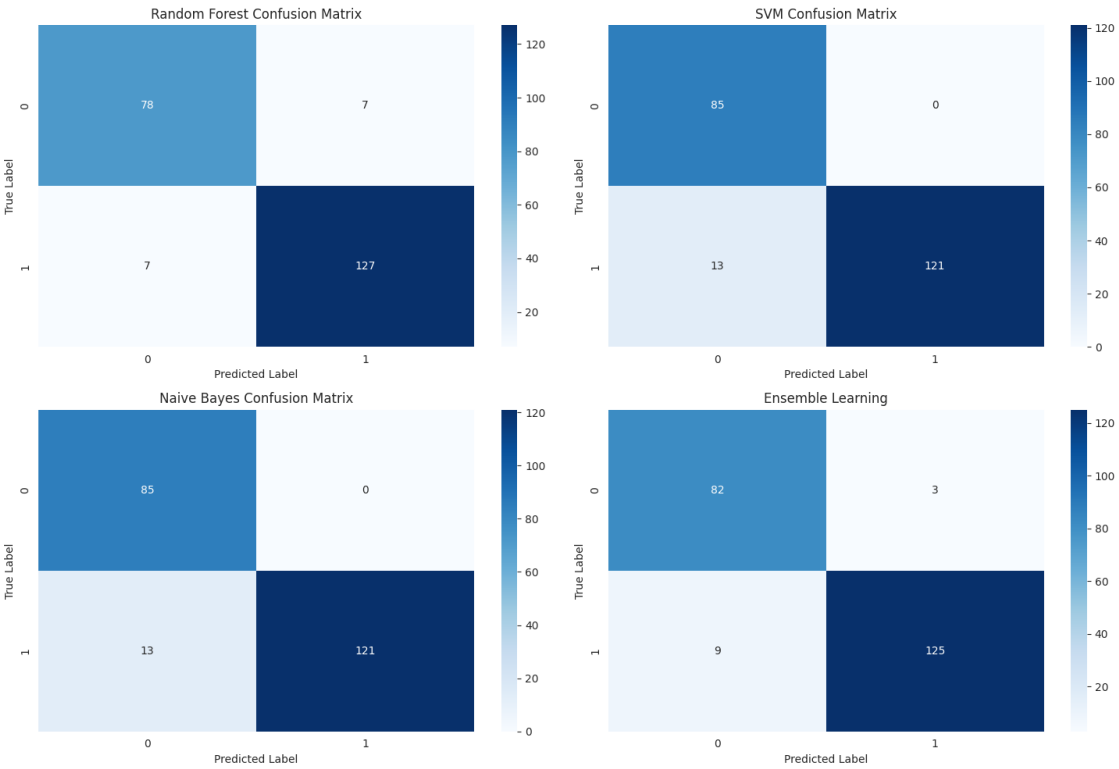


Figure 8 Confusion matrices for the models.

5.3.2 MEAN CROSS VALIDATION ACCURACY

The performance of the models was quite high and thus a 5-fold cross validation was carried out and the mean average was still high indicating that the model still performed well.

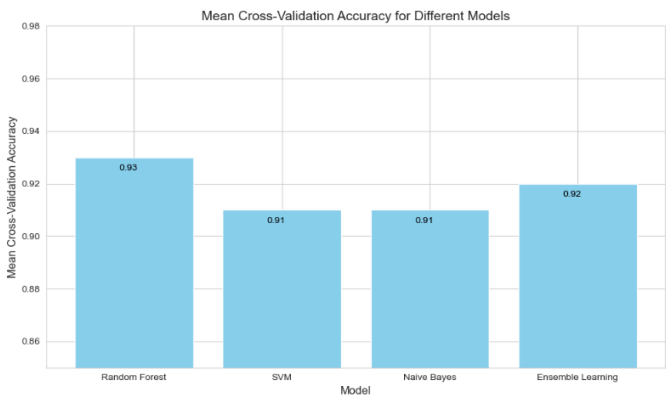


Figure 9 Mean Cross Validation Accuracy of Models.

5.4 RESEARCH SUPPORTING VISUALIZATIONS

5.4.1 TIME DISTRIBUTION BETWEEN ROLES AND DIFFICULTY LEVELS

Since time is important in training our models to determine the outcome of the game a further analysis was carried out to determine the mean amount of time required for each game level and role.

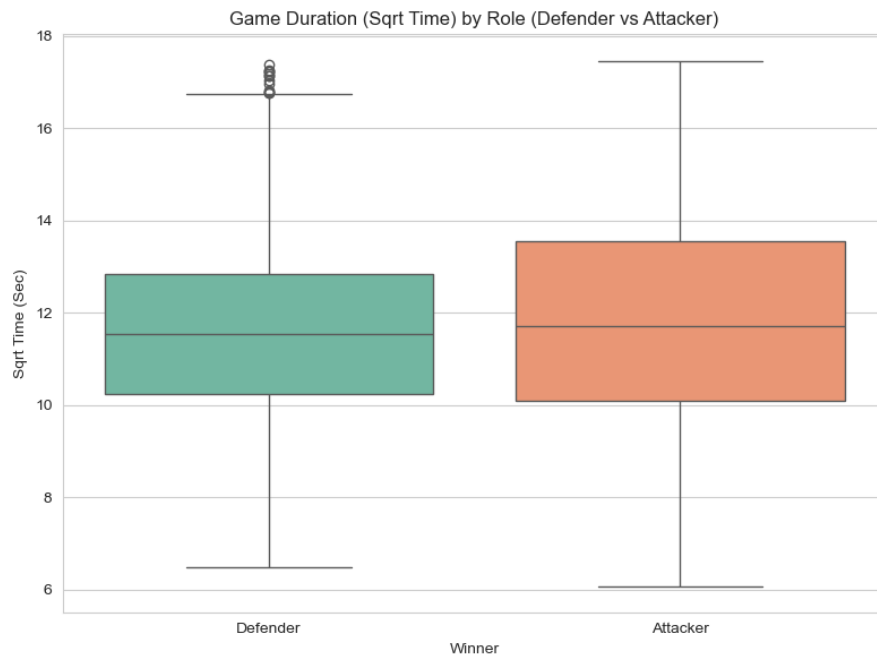


Figure 10 Time Distribution of Game Duration by Role (Defender vs Attacker)

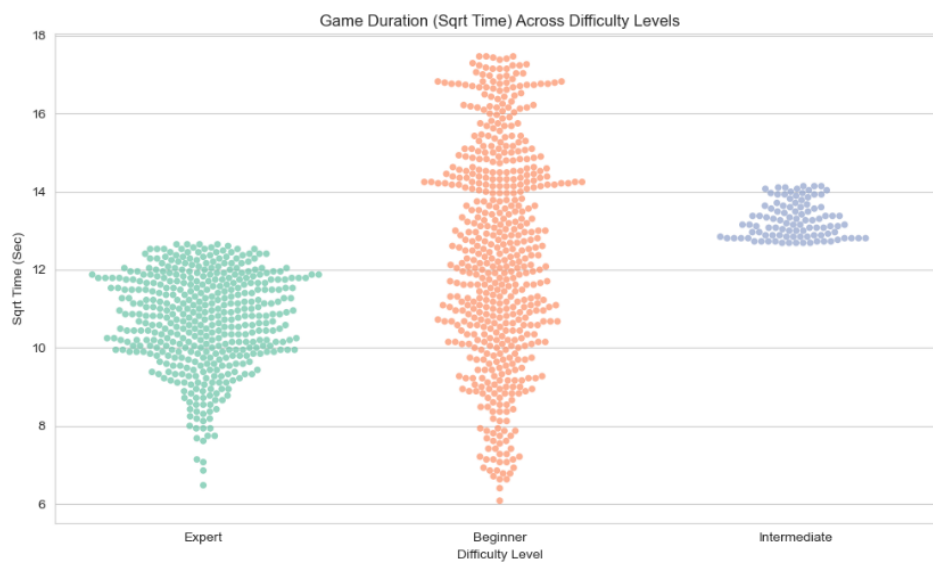


Figure 11 Time Distribution across each level.

Figure 10 shows that the defender wins tend to be slightly faster with a slightly lower game duration. The outliers may indicate very fast and slow wins for the defender and inversely longer game times for the attacker. Attacker wins tend to happen in a slightly spread range with some taking a longer period of time.

Figure 11 shows that that the Expert level players take a shorter period of time to complete the game as compared to the other players. Beginner level times are spread over a wide range of time with the defenders taking most of the time for the distribution. The intermediate players take slightly less time than the beginner players, and the distribution is spread over a short period of time.

5.4.2 WIN AND LOSS RATIOS ACROSS DIFFERENT LEVELS OF THE GAME

Win loss ratios were also analysed for each and every player at different levels of the game between the Defender and Attacker in the game.

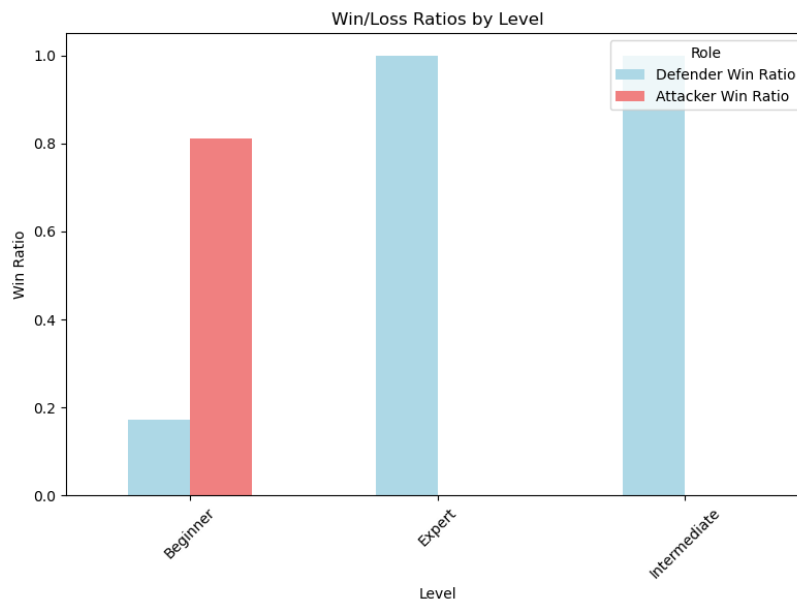


Figure 12 Win and loss ratios at different levels of the game.

Figure 12 depicts one of the weaknesses of the game that may slightly need some improvement. At the beginner a level, the human plays as a defender, their win ratio is very low, which is **0.2** as opposed to the computer that winds at an average of **0.8**. This would be expected as the user begins to play the game, however as the game progresses the Defender knowing all the different strategies that the Attacker possess always wins the game. The game may need to adjust its prompts as the attacker or defender as a player progresses.

6. DISCUSSION

6.1 INTEPRETATION

The game points out the fact that as the game progresses the players tend to win. This is quite evident that at this point the defenders would have mastered all that attacking strategies that would have been set. It is quite likely that the game attack strategies do not change when users progress throughout the game. The research uncovered that defender wins occurred mostly in shorter games. On the other hand, Attacker wins always seem to happen in longer games. This suggests that defenders tend to win as opposed to Attackers that take a longer time. This is evidence that the game play is influencing the players to know more about protecting themselves from online threats.

The research showed that machine learning models can be used to predict the game outcomes which are the Defender and Attacker wins. The best performance came from Ensemble learning machine learning model. Combining various machine learning models seems to produce the best outcome when predicting results.

6.2 LIMITATIONS

They were some class imbalances that occurred between the defender, attacker and draws. The number of draws were quite few only 11, after cleaning the entire dataset. The draws ended up being dropped to avoid any class imbalances. The wins between the defender and attacker were slightly significant and this might have skewed the model predictions. This significant difference and also not having the draws may have influenced the model's predictions intensively.

The dataset was also limited to quite a few features. Having additional features such as the players strategies would have assisted in providing better outcomes. Having strategies that the player used for instance would go a long way in actually predicting the winner of the game between the defenders and the attackers.

Finally, the model was trained on a particular dataset. This may be impossible to actually take the same dataset and test it on a different game. In other words, the model may not generalize quite well with other games.

7. CONCLUSION

7.1 SUMMARY

As the difficulty of the game increased, the skills of the players increased, there were a high number of defender wins suggesting game imbalance. Further development of the game may be required so as to change maybe attacking strategies as the players become knowledgeable of the different strategies. This may result in creating an even playing field between the defenders and the attackers.

The attacker wins are more frequent in more time. Defender wins are common in shorter period of time. This highlights that the players of the game are learning to defend themselves while enjoying the game at the same time. If we had a feature such as player strategies, it would tell more on everything that the players are learning to actually by defending themselves.

Finally, Ensemble Learning was the model that performed better than the rest in predicting game outcome with an accuracy of 99%. It may be quite beneficial to combine multiple classifiers to predict outcomes. Other models, such Random Forest, SVM and Naïve Bayes were not far off performing in a similar fashion.

7.2 RECOMENDATIONS

On the game itself, the developers may need to work more on the game balance. As the level of player intelligence increases it may be necessary to switch strategies by the attackers or defenders depending on the human players role. To also get more insight from the game, it may be necessary to actually add more features. Features such as the strategy would go a long way in predicting the winner of the game.

7.3 FUTURE WORK

It may be important to actually look into advanced models such as deep learning models or Gradient Boosting Machines to further enhance the prediction accuracy. The research ended up focusing on a binary outcome, the defender and attacker. Having a balanced dataset may assist in getting enough of the data so that we can predict features such as draws.

Also, it may be exciting to look into real-time game predictions. While the game is being played between the defender and attackers, analysing their strategies, we may be able to get the winner from the third move, giving us an early idea of how well the human player is understanding the defensive strategies.

8. REFERENCES

- Awojana, T. and Chou, T.-S. (2019) 'Overview of Learning Cybersecurity Through Game Based Systems', *2019 CIEC Proceedings*, p. 31521. Available at: <https://doi.org/10.18260/3-2-370-31521>.
- Hafeez, T. *et al.* (2021) 'EEG in game user analysis: A framework for expertise classification during gameplay', *PLoS ONE*, 16(6), p. e0246913. Available at: <https://doi.org/10.1371/journal.pone.0246913>.
- Nkongolo, M. (2024) 'CyberMoraba: A game-based approach enhancing cybersecurity awareness', *International Conference on Cyber Warfare and Security*, 19(1), pp. 240–250. Available at: <https://doi.org/10.34190/iccws.19.1.1957>.
- Pramod, D. (2024) 'Gamification in cybersecurity education; a state-of-the-art review and research agenda', *Journal of Applied Research in Higher Education*, ahead-of-print(ahead-of-print). Available at: <https://doi.org/10.1108/JARHE-02-2024-0072>.