National College of Ireland

Enhancing Customer Retention in Online Games Using Customer Lifetime Value

MSc Research Project
MSCDAD_JAN24A_O

AASIM INAMDAR
Student ID: 23236108

School of Computing
National College of Ireland

Supervisor: JASWINDER SINGH

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	AASIM INAMDAR
Student ID:	23236108
Programme:	MSCDAD_JAN24A_O **Year:** 2024-25
Module:	Research Project
Supervisor:	JASWINDER SINGH
Submission Due Date:	29/01/2025
Project Title:	Enhancing Customer Retention in Online Games Using Customer Lifetime Value
Word Count:	7830 **Page Count**: 22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Aasim Inamdar

Date: 29/01/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Enhancing Customer Retention in Online Games Using Customer Lifetime Value

AASIM INAMDAR
23236108

Abstract

Customer Lifetime Value (CLV) is a crucial metric in the online gaming industry, and proper pathway strategies to enhance user engagement, monetisation, and retention. The research dug deeper to find the key factors influencing the CLV in freemium gaming environments and the application of advanced machine learning techniques to predict the CLV accurately. The analysis of the player data, including gameplay behaviour, in-app purchases and churn likelihood, aims to identify the patterns that categorise high-value players from casual users. Predictive modelling is a hybrid approach of convolutional neural network (CNN) and recurrent neural network (RNN) where the output recommends optimising the game design, user retention strategies and personalised recommendations to retain users. The model achieved a testing accuracy of 90 per cent and a minimal loss of 0.36. Personalised interventions based on the CLV predictions lead to enhanced retention rates and faster revenue growth, with the practical applicability of the frameworks. This research bridges the gap between the theoretical CLV research and its application in dynamic ecosystems, offering insights for the developers and stakeholders in the online gaming industry, with contributions to the ever-evolving discourse on the customer centric design to provide a data-driven road map to improve player experience and long term value.

1 Introduction

The online gaming industry has seen significant growth in the past years, and the rise of the freemium business model has gained popularity in the monetisation approach. The business model approach is that the games are free to play initially, allowing users to access the basic content without any upfront payment and later in-game purchases to unlock the premium content of the games leading to revenue generation. This has been successful in gaining a lot of initial users but faces a backslash after the short beginning. This leads to a challenge to retain players over time leading to a high churn rate, affecting the sustainability and profitability of the business.

Customer lifetime value (CLV) plays an important aspect in the online gaming industry to determine the player retention. The prediction of accurate customer lifetime value enables the companies to identify the high-value players, who are much more likely to make in-game purchases targeting tailored marketing strategies to engage and retain the users, thus increasing the overall profit.

Earlier research on the prediction of customer lifetime value has shown positive values which have helped in customer retention and the applications of machine learning techniques like the random forest, gradient boosting, etc. (Tapper, T. (2022)) for predictions. While these have been effective on the historical data provided, but struggle with the behavioural analysis subject to user's preferences such as spending habits, change in engagement level, new updates response, events, and other external factors affecting the user activity which could potentially result in less accurate long-term predictions.

Insights into the player behaviour, preferences, and future actions are important to gain long-term engagement and profit. The use of machine learning and data analytics can help develop better predictive modelling techniques, optimized targeted marketing strategies, improve user experience and attain higher retention. The research aims at the exploration of using ensemble learning methods, neural network models, convolutional neural network (CNN) and recurrent neural network (RNN), to predict the customer lifetime value (CLV) and enhance the player retention strategies in the online gaming industry.

The objectives of this research are to analyse the **key features that affect the customer lifetime value in online gaming**. To systematically investigate factors such as player demographics, in-game behaviour and purchase histories to get insights that guide in better development of effective retention strategies, marketing approaches and increased revenue. **To build an advanced machine learning model to predict the customer lifetime value** with high accuracy, using the techniques of neural networks and ensemble learning methods, which can capture the complex patterns in the player data. These methods will allow for a better understanding of the player's behaviour and spending patterns, which in turn lead to a precise CLV forecast. To evaluate the model's reliability and performance, we use the metrics Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared values, ensuring robustness and predictive accuracy. The final objective is to **Interpret the practical tips for improving player retention based on the predicted customer lifetime value**. Using the prediction generated from the model translates into valuable insights to use in the real world to enhance player engagement and retain players longer on the platform. Based on the predicted CLV, segmenting the customers to create strategies for groups based on their value to the company. Giving out special rewards and tailored messages to the high-value players based on their interests who would be willing to quit soon. These strategies would help in retaining the customers and keeping them interested for a longer time.

How can sophisticated machine learning models accurately forecast Customer Lifetime Value (CLV), uncover key determinants and enhance retention strategies?

The research aims to develop an advanced machine learning model to predict and forecast the Customer Lifetime Value (CLV) for the online gaming industry, utilising the publicly available data from **Kaggle**. This study identifies the key factors that will help in the improvement of the accuracy of the CLV prediction and gives an insight into the key ranges of stakeholders. Predicting the CLV is an essential criterion for gaming companies to enhance player retention strategies, which leads to longer engagement and higher revenue generation from the players.

The game developers can use the findings to enhance the user interface and interaction which would ultimately help in the extended customer lifetime.
These insights into the predicted behaviours and the actions of players will help the marketing team to produce enhanced, data-driven strategies to engage new users. With the predicted CLV, players can be segmented to provide better marketing campaigns to the segments, resulting in better market-effective adverts and increased return on investment (ROI). These findings will benefit the players as well with better-aligned gameplay in line with their preferences, thereby enhancing overall satisfaction and enjoyment. These insights will also contribute to academics with newer insights into the application of the applications of advanced machine learning techniques for the prediction of CLV. Also, giving out the methodologies for the evaluation of the long-term value of individual players, which can be used for similar industries which rely on customer engagement and retention.

In the section 2, reviews of the existing literature for the prediction of the customer lifetime management and the customer retention techniques which are discussed with the traditional and the cutting-edge ways to model the CLV, highlighting the gaps and steps of approach, resources required and the project plans are discussed.

2 Related Work

The usage of customer lifetime value and prediction of user churn has grown significantly over the past years due to the industry's expansion and increased competition, particularly with the surge in gaming after the COVID-19 pandemic. The machine learning and game data analysis play an important role in the optimisation of player retention, enhancing the user experience and enhancing the revenues generated from them. There have been research focusing on the prediction models using random forest, gradient boosting, and other regression models based on CLV for analysing player churn rates and means to reduce them and enhancing the data analytics techniques to gain insights into the player behaviour. However, these models lack in capturing the complexity of game behaviour as the user preference and the actions can change constantly. There exists churn prediction models which provide with insights but with the constant evolution of the games drives up the player churn rates, leading to a newer predictive model which includes behavioural and dynamic data for long term engagement.

2.1 Customer Lifetime Value (CLV) Prediction in Online Gaming

The customer lifetime value is a crucial metric in the evaluation and segmentation of the users, earlier research in the field to basic segmentation models to categorise the high-value players to get personalised marketing strategies, these segmentations helped in setting up the foundation, but these earlier models relied on demographic and transactional data which lack the depth of behavioural insights (Khajvand and Tarokh (2010)), with another contribution in predictive modelling with the inclusion of the player engagement metrics of their time spent on the game which gives a wider perspective of CLV (Valdivia(2021)). However, the traditional segmentation models limit the ability to give an output for the dynamic behaviour of the player in real time. The CLV models in earlier stages were simple and had ease of implementation leading to access to initial user segmentations.

The Formula used for the calculation of CLV is described as the discounted sum of the profits (revenue minus costs) generated by the customer in each period, considering the time value of money (Farris et al. (2010))

$$CLV = \sum_{t=1}^{T} \frac{\text{Revenue}_t - \text{Cost}_t}{(1+r)^t}$$

However, these models still lack in their approach where they often fail to capture the ever-changing player behaviour with their static approach. The lack of real-time data integration in these studies restricts their predictive accuracy with advancements in the gaming industry.

2.2 Machine Learning Techniques in CLV Prediction

The increasing use of machine learning in the prediction of the CLV for accurate and data-driven insights. The use of multiple sources and a multi-task learning approach has enabled cross-platform compatibility and better predictive accuracy, with the models efficiently handling the data heterogeneity and usage of much higher extensive computational resources restricting the smaller developers (Zhao et al.(2023)). Similarly, the application of ensemble learning methods, like random forest and gradient boosting for the prediction of CLV achieves a sturdy result, these ensemble techniques show stability in structured data environments but lack in performance with highly unstructured data or sparse data which is very common in the gaming industry (Kumar et al. (2023)).

These models have been pretty significant in improving prediction accuracy and adaptability. However, the computational intensity and the need for large and discrete datasets are the major setbacks for smaller gaming companies, which cannot afford or gain these insights easily. The use of traditional machine learning techniques often lacks with its non-linear behavioural patterns, restricting their effectiveness in predicting long-term engagement.

2.3 Deep Learning Approaches for Player Retention

Application of Deep Learning models such as CNN and RNN have shown the potential to capture complex patterns within player behaviour data. Adaption of these model architectures in the prediction of retaining the users in the credit card industry, which lies in parallel with the gaming industry (Ming et al.(2021)). A similar application of deep learning techniques to analyse the in-game engagement and findings of the RNNs has been effective in modelling sequential behaviours. Despite the learning model's strengths, it can be challenging to interpret the high computational demand, which obstructs the implementation in resource-constrained settings (Kristensen and Burelli(2019)).

Deep learning gets a better insight into sequential user behaviour, but the reliance on large datasets and demanding performance hardware makes the application of real-time data challenging. The "black-box" nature of these models often sets back the transparency, which can be an issue for the stakeholders who require interpretable insights.

2.4 Behavioural Insights and CLV Prediction

The integration of behavioural metrics for the prediction of the Customer Lifetime value and the importance of inclusion of the time spent, in-game actions, and the spending habits of the users to shape the model predictions (Ascarza et al.(2024)) has emphasised the outcome of the models. This approach of an all-round view of player engagement allows the companies to prepare targeted, customised retention strategies, these insights have been helpful for better-targeted ads for the predictions (Drachen et al.(2018)).

The reliance on these metrics of game-specific data has restricted the model performance for a generalised approach for the prediction, and an increase in the model complexity can also lead to overfitting by the model because of this, the model performs well on the training data but struggles for the general data and reduces the effectiveness in prediction and practical applications.

2.5 Retention Strategies and Their Effectiveness

Finding better strategies for retention, sustaining player engagement and profitability of the freemium gaming model has been a crucial aspect. Traditional retention methods which were based upon reward systems, have been shown the better sustenance initially and have improved player engagement but it lacks the long-term retention plan of the players (Voigt and Hinz(2016)) where the players often move away from the platform to another once there is no other reward. Whereas the data-driven retention strategies use the CLV prediction for personalised engagement efforts and in-turn, have shown a better result (Shen et al. (2024)).

The use of real-time data can be helpful where the retention strategies would be dynamic and would instantly respond to the player's behaviour, adjusting the engagement efforts to better align the player's preference and gameplay pattern. (Tuguinay et al.(2024)). The use of these real-time feedback strategies needs a lot of computational power and resources which would restrict the applicability of these in fast-paced gaming environments, while the traditional approaches are easy to implement, but they lack sophisticated strategies to retain the users for a long time, explains the need for a better dynamic and adaptive strategy.

2.6 Limitations in Existing CLV Models and Retention Techniques

Previous research has shown significant progress in CLV prediction and retention strategies for online gaming, regardless there have been critical gaps that justify the need for further research. In the earlier model for personalised CLV prediction in online gaming, where the model was trained on the historical data which lacked in predicting the real-time changes in player behaviours and adapting to it (Zhao et al.(2023)). With CLV predictions using machine learning techniques could get cumbersome and with lots of features to choose from the model could become overly complex and less interpretable which would restrict the implementation with the non-technical users in the gaming industry (Kumar et al.(2023)).

The models all rely on the monetisation features and consider the general approaches, whereas some use the dynamic behavioural data with influences the new game updates, game events or seasonal trends and integration of these into the data to capture the real-time behaviour shifts but the model building lacks with the need of challenging resources and high computational costs (Valdivia(2021)). These traditional retention methods have been working by providing the users with rewards but this strategy often fails to maintain long-term player engagement as the players keep expecting these incentives (Voigt and Hinz (2016)).

The implementation of data-driven strategies, in addition to the prediction where the feedback from the users and similar adaptive tactics require significant high investment and major resources, which have proved unfeasible for small-scale game producers (Shen et al.(2023)).

2.7 Research Niche and Novelty

The limitations from the existing literature and previous baseline models highlight the need for a new advanced CLV model that includes the historical and behavioural data, adapts to the retention strategies, and is resource effective. The traditional segmentation has slightly evolved

to advanced models that leverage machine learning techniques to improve accuracy. However, the model faces challenges to interpretation and computational efficiency. Meanwhile, the inclusion of behavioural insights has better personalisation, but the gaps still remain in addressing the diversity of gaming genres. This research aims to bridge the gap by developing a hybrid CLV prediction model to attain actionable retention strategies and ensure player engagement and profitability in the industry.

This research aims to address the gaps by developing a hybrid model that combines ensemble learning and neural networks to predict the CLV accurately, integrating both demographic and behavioural data to get deeper insights into player engagement. With a robust and adaptable approach, this research would enhance CLV prediction accuracy and help in formulating dynamic retention strategies that can optimise user engagement, and the revenue generated from the online gaming sectors.

Further, the implementation of the implied techniques is discussed in section 3 with deeper insights of the methodology and its specifications.

3 Research Methodology

This section describes the proposed solutions, steps and activities, tools, test data, ethical values, evaluation plans and the methodology employed to achieve the research objectives and address the research question and sub-research question, as established in section 1. The initial segment outlines the comprehensive methodological approach that was adopted in the research. The approach follows a systematic framework inspired by related works in predictive modelling and the CRISP-DM (Fig 1) methodology to ensure scientific rigour and reproducibility.

3.1 Research Design

This research employs a quantitative approach to build and evaluate predictive models. This has been focused on the use of supervised learning techniques, including deep learning, to classify customer engagement into predefined levels. The study process is divided into six stages, Business understanding, data understanding, data preparation, modelling, evaluation, and deployment. It includes the steps for data acquisition, feature engineering, and application of an ensemble CNN-RNN model.

3.2 Data Collection and Descriptions

The player data is gathered from Kaggle, a publicly available online portal that includes player profiles, gameplay data, transactional records, behavioural metrics, and engagement levels. The dataset comprises of 13 features and around 40034 observations. This data is further modified in this research to make the data fit the models implemented.

The dataset contains demographic details such as age, gender and location, behavior metrics such as playtime hours, sessions per week and in-game purchases, and the engagement levels of categorical variables indicating low, medium, or high interactions. The dataset complies with all the privacy regulations and GDPR. Where the sensitive data is anonymized, so that it cannot be traced back to the user and no personally identifiable information is used.

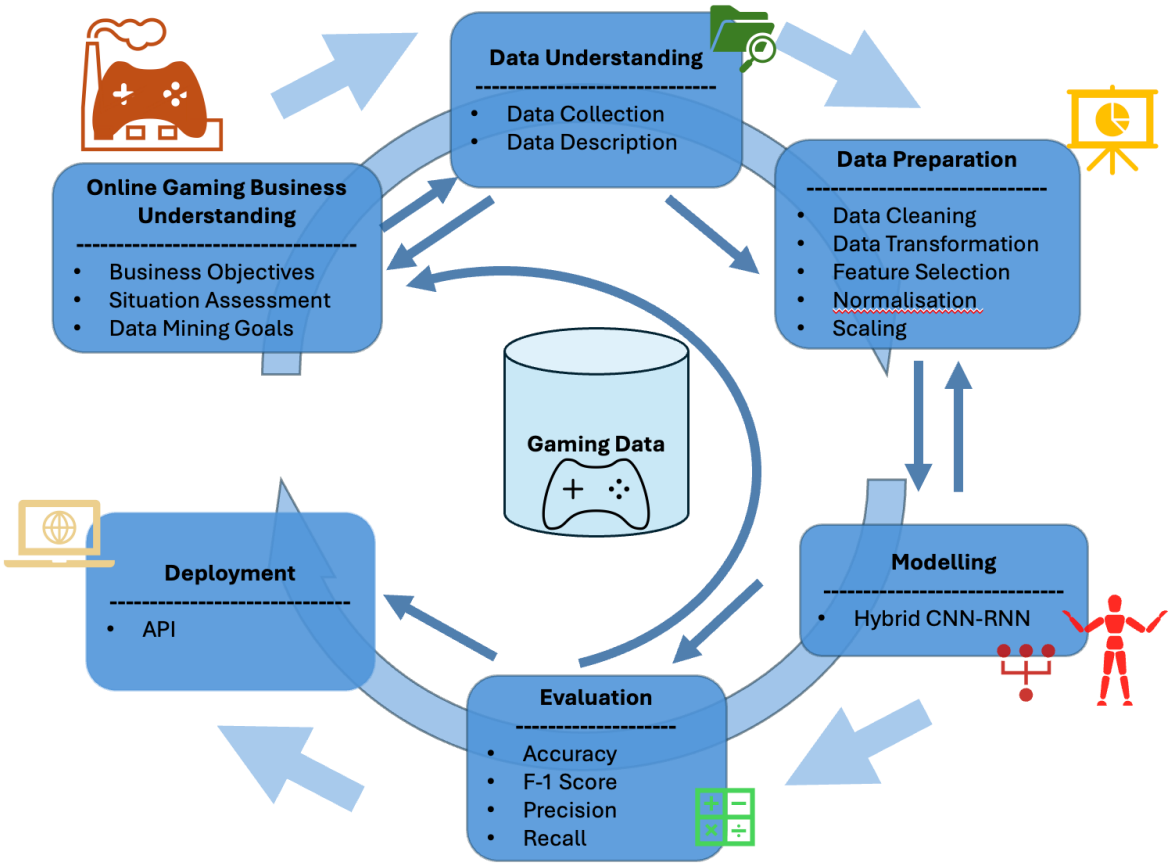


Fig 1: Crisp-DM methodology

3.3 Data Pre-processing

The data pre-processing step involves data cleaning, exploring, feature engineering and converting the datatypes to the requirement as the input needed for machine learning models.

3.3.1 Data Cleaning

The data cleaning ensures the dataset is of high quality and ensures its consistency and suitability for model training. Raw data often consists of errors, inconsistency and missing information which affects the model's performance and can result in biased results.

Missing data was checked to maintain data integrity using pandas [is.null] function. Any numerical features that were missing can be replaced with the mean or medial of the respective columns, where there were no missing values in the data that was obtained. The categorical features are replaced with mode and ensure the most frequent category to be used. This approach minimizes the loss of data and ensures that the imputed values do not introduce any additional bias. There were no missing values for the categorical features in this dataset obtained.

The duplicate data from the dataset is checked using pandas function [.duplicates()] and no duplicate values were found in the dataset. Removing any of the duplicates ensures that the data's reliability is enhanced and the risk of biasing is reduced.

The records containing inconsistent data or invalid data are filtered out to avoid distortion in the analysis and modelling. Factors like age and playtimehours with negative and unrealistic value are logically invalid, causing errors in the model building, thus these values are excluded. For categorical features, the unique values are observed and checked for any inconsistent data and filtered out.

The categorical features like gender, gamegenre, location, and game difficulty are encoded into numerical formats for better training with the model. Two approaches are used here, One-hot Encoding and Label encoding, for the variables with multiple categories such as gamegenre and location, to create a binary indicator for each category. For features such as gender and game difficulty, label encoding was used to convert the binary or ordinal features into numerical labels. This encoding is necessary to convert the non-numeric data into suitable data for machine learning algorithms, which also reduces its complexity.

The numerical features are scaled using the z-score normalization to standardize their distribution. This helps in centering the features around the zero and scales based on standard deviation. Normalization ensures that the features with larger ranges do not dominate the model's learning process which enhances the convergence during model training. There were no outliers detected in the model, these were using statistical methods using the z-score on the numerical features. This reduces the model's risk of skewness of model in predictions.

The data cleaning steps are implemented using Python and its libraries, pandas is used for handling the missing values, duplicates and basic feature transformation. NumPy was used for numerical computation, imputation, and scaling. Scikit-learn was also used for pre-processing tasks.

3.3.2 Exploratory Data Analysis

The EDA is essential to understand the dataset to uncover the patterns and identify any issues that might affect the modelling process. With the data visualizations, correlations, trends, patterns, and any potential issues such as outliers or skewed distribution can be analyzed. The first data distribution was analyzed for all the numerical and categorical variables in the data.

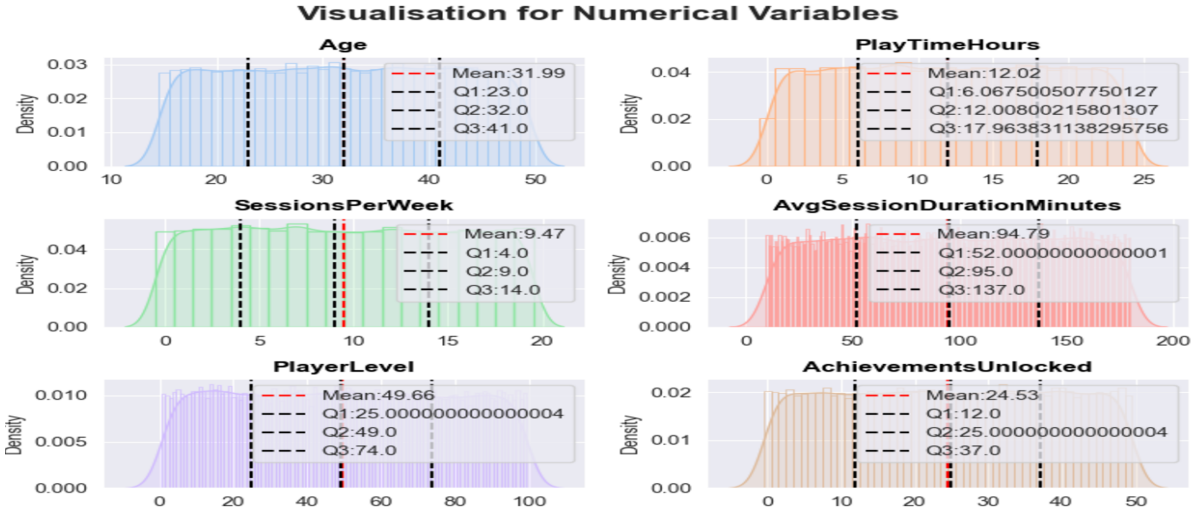


Fig 2: Visualizations of Numerical Variables

The (Fig 2) represents the distribution of data and its density, the numerical features here obsereved age, playtimehours, sessionsperweek, avgsessiondurationminutes, player level and achievementsunlocked where the density curves demonstrates the values distributed across each of the variable and indicating the data point concentration. The statistical tests here show the mean and quartile regions for 25, 50 and 75 percent.

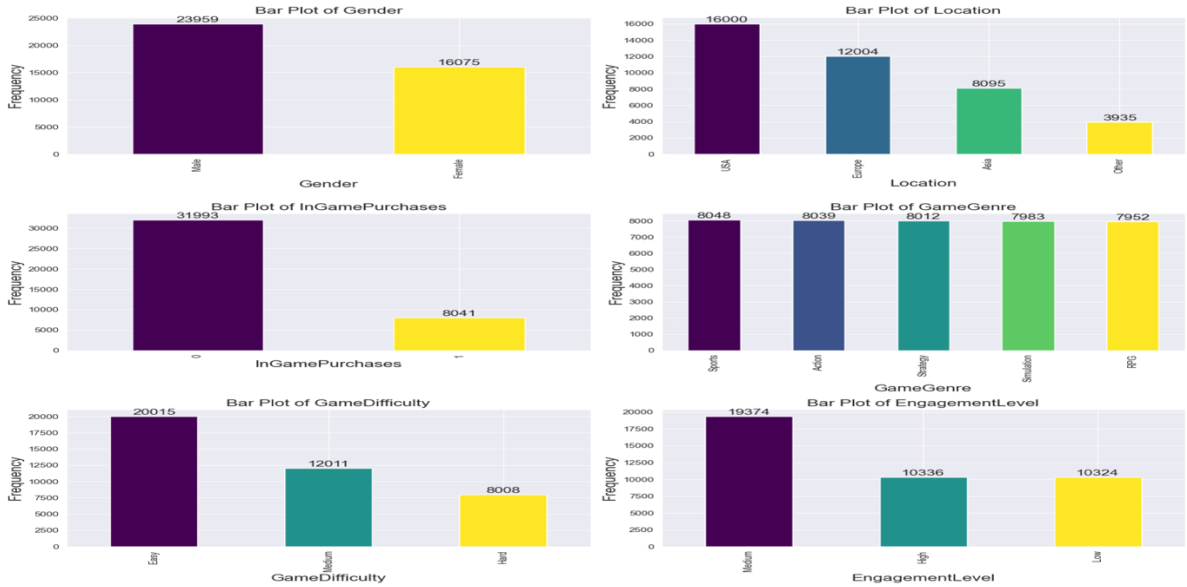


Fig 3: Bar plots of Categorical Values

From the Fig 3 visualizations, we observe that the age is distributed, with most values between 23 and 41 and an average age of around 32 years. Most of the players have about nine sessions per week, and their average duration lasts about 95 minutes. The average achievement unlocked by each player is about 24, and the player levels mean 50. These visualizations help in understanding the player's behavior and activity patterns.

The categorical visualizations shown with the bar plots give better insights into the individual categorical data, which shows the distribution among each category.

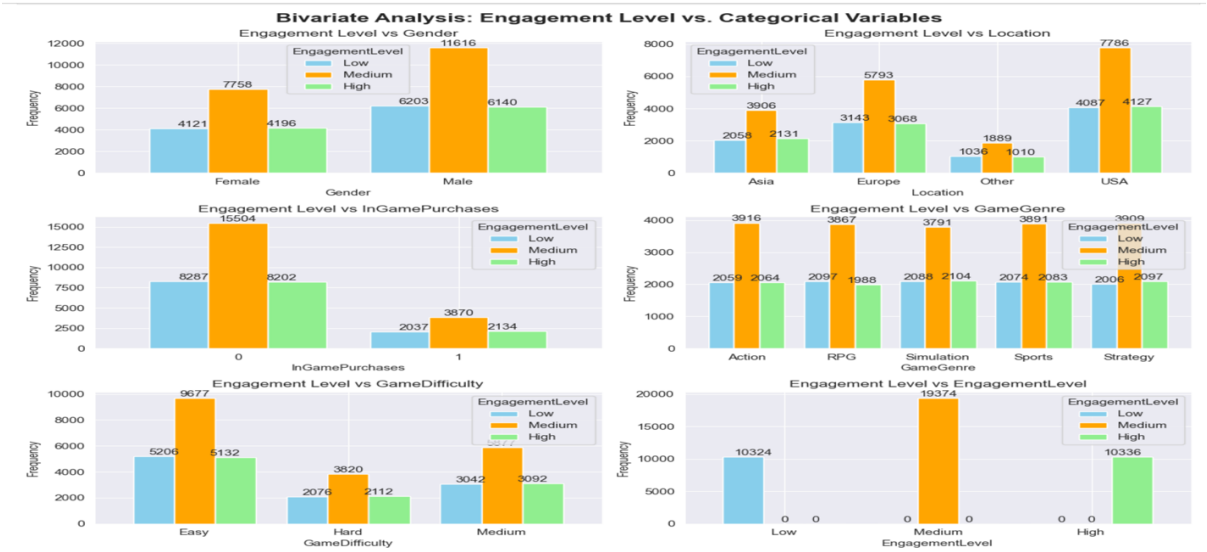


Fig 4: Engagement vs Categorical values

The (Fig 4) analysis of the categorical variables against the engagement levels (Low, Medium, High) describes the density among each categorical value, the graph vs gender shows that there are more males than females in the medium category, and there are most who don't have any in-game purchases in the medium engagement category and with the visuals we can conclude that there are the highest users in the medium engagement pattern across the groups.

The (Fig 5) Correlation matrix represents the relationships between all the variables in the dataset, each cell in the matrix shows the correlations between two variables with values between -1 and 1. A heatmap is used to show the strong positive and negative correlations among them, and the colors red and blue represent them, respectively. The diagonal is the correlations among the variable itself with the strongest correlations, and the variable's engagement level and the achievements unlocked have a strong correlation among them. It is necessary to identify the linear relationship and redundant features and select the relevant predictors for the analysis.

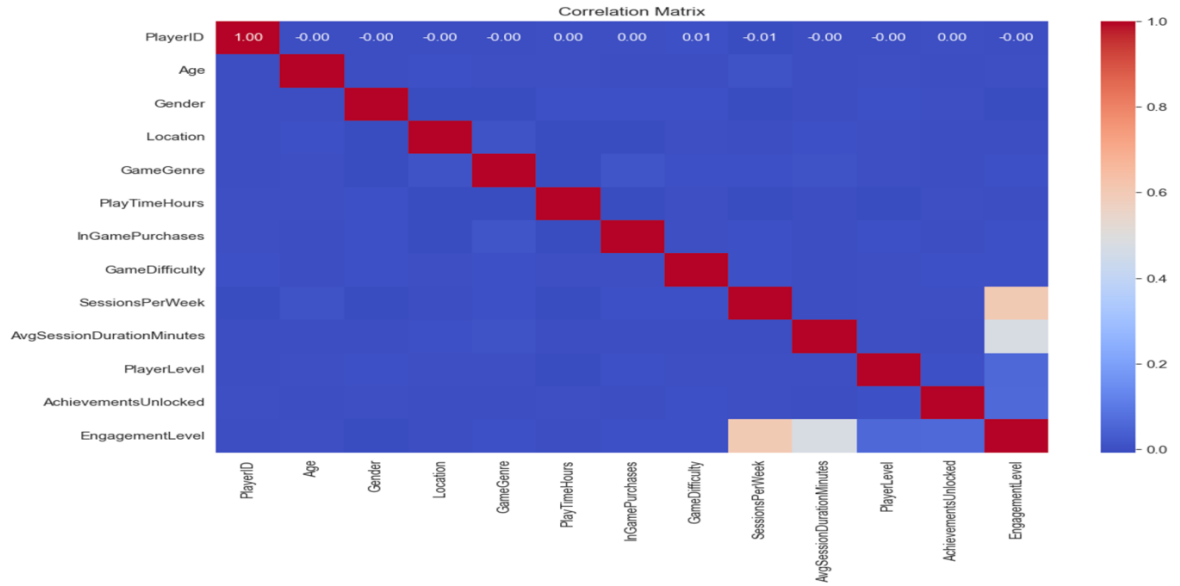


Fig 5: Correlation Matrix

The Fig 6 box plot shows the outliers in the numerical features, it is created using the z-score values and the IQR method to detect the outliers in the variables. The whiskers extended to the range of IQR and the z-score was used to normalize the data for consistent comparison. There are no significant outliers in the data here and it indicates a well distributed data.

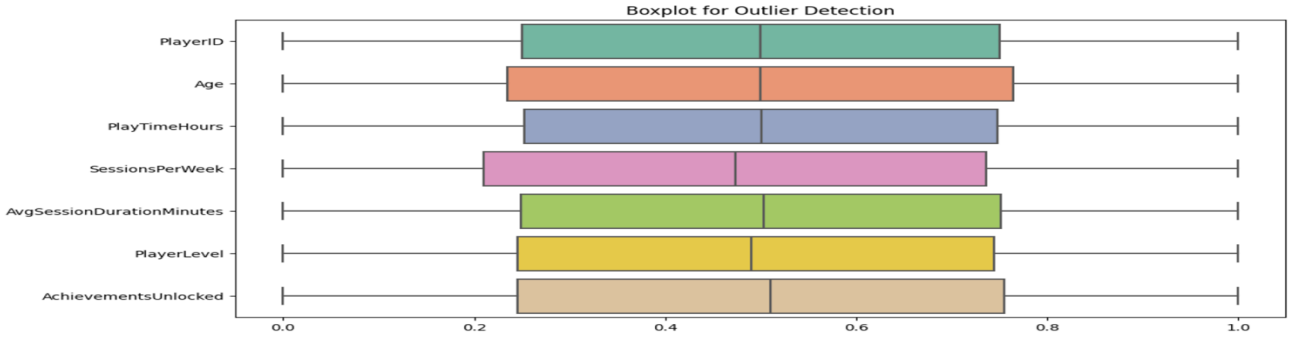


Fig 6: Outlier Detection

3.3.3 Feature Engineering¶

To generate meaningful attributes from the raw data, which enhances the predictive capability of the of the models and gives a better insight to the player behaviour. New features based on playtime and segmentation of the users based on age, session frequency, player level, and session duration are calculated.

- Total Playtime per week: calculated as the product of the PlayTimeHours and SessionsPerWeek. This would be critical for analysing the weekly activity trends of the users and help in deeper insights into the behavioural analysis.
- Age Category: categorizing the existing age into groups of teenager, young, mature and old mature. This would help in better understanding the groups which engage the most.
- SessionFrequencyCategory: categorises the existing session values into passive, occasional, regular, frequent and dedicated activity levels. This insight helps in distinguishing between casual or highly engaged players and helps in retention strategies.
- Player Level Categories: this segments the categories into beginner, intermediate, advanced, expert and master. This feature would help in better understanding of players relationship with engagement and their spending habits.
- Session Duration Categories: The grouping of session duration into short, medium and long which indicates the players gaming style and their time invested in games.

Further, the data is split into 3 segments here, training (80%), validation(10%), test(10%) subsets.

3.4 Data Mining

Data mining is an important aspect of analysing customer behaviour and predicting the customer lifetime value in the online gaming industry. With the techniques of clustering classification and predictive modelling, the business can extract actionable insights from the enormous data.

Leveraging an ensemble convolutional neural network (CNN) and recurrent neural network (RNN) hybrid model for the extraction of the spatial and temporal features from complex player data. CNN is used to identify the patterns locally and the associations between demographic attributes and gaming statistics. Inline, the RNN is applied to capture the time series patterns and sequential dependencies, which are essential for the modelling of the player's behaviour (Zhao et al. (2019)). The previous application of these models is their ability to manage non-linear and dynamic relationships, which are essential in data of the gaming industry, as explained already by (Kumar et al.(2023)) and (Valdivia(2021)).

4 Design Specification

The design specification gives the flow of architecture and methodology of the implemented model, Fig 7 shows the architecture of the hybrid deep learning model for the prediction of the customer lifetime value, leveraging both convolutional neural network (CNN) and recurrent neural network (RNN) to classify the multi-dimensional sequential data effectively. The

process begins with pre-processing and scaling the input data to ensure its compatibility with the models layers.

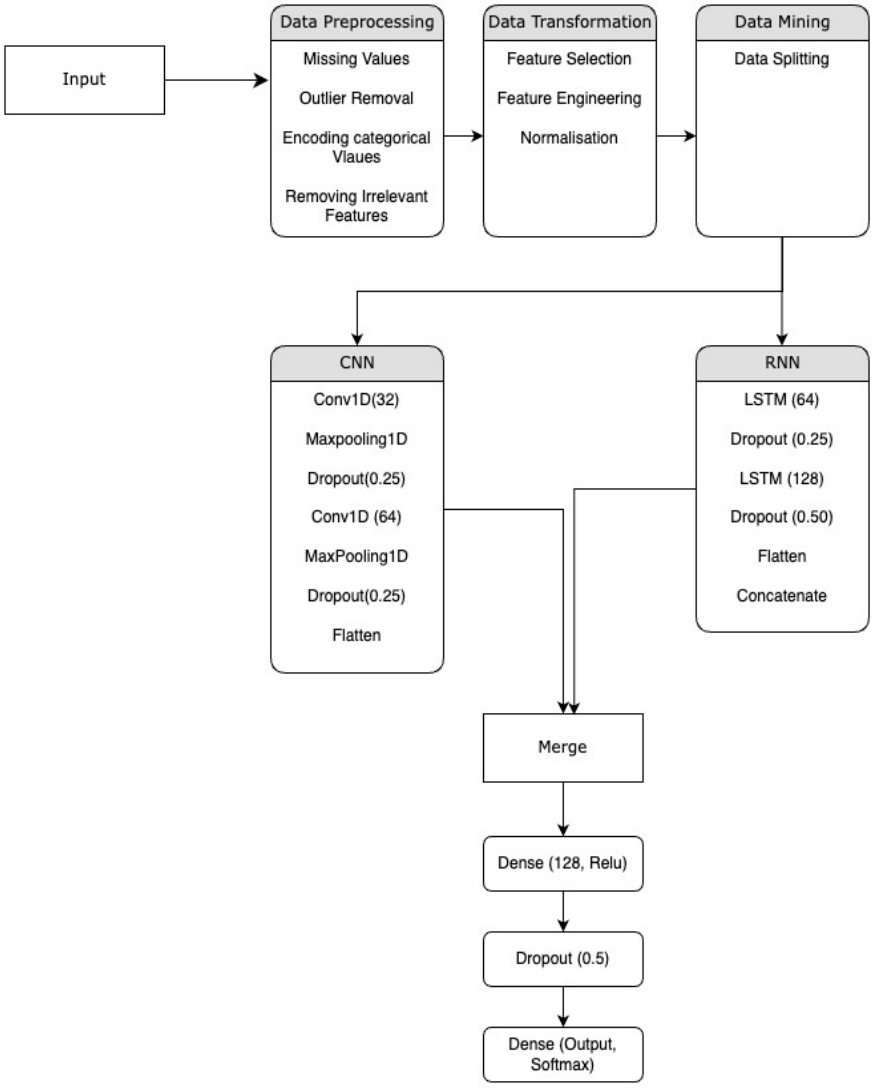


Fig 7: Design Specification Diagram

The CNN branch extracts the spatial features from the input sequence by implementing multiple convolutional layers, each with pooling and dropout layers, to reduce any overfitting and computational overheads. The RNN branch uses LSTM, long-short term memory layers, to capture the temporal dependencies and patterns in the data; both branches perform independently of each other to extract the unique feature representations.

The outputs from both branches are merged to form a unified feature representation. These combined features are passed through the dense layers, which integrate dropout layers to establish robust learning and prevent overfitting. The final output layer with softmax activation function for multi-class classification and using adam optimiser model with categorical cross-entropy loss to ensure efficient convergence.

This hybrid approach demonstrates CNN's strength to detect spatial features and RNN's proficiency in sequential data learning, which gives an accurate and scalable prediction with its adaptability to real-time scenarios where incoming data streams can be analysed dynamically.

5 Implementation

The hybrid model of CNN-RNN is implemented to predict the customer lifetime value in online gaming, and it has been ensuring a comprehensive pipeline for data pre-processing, model training, and evaluation.

5.1 Data Preparation

The transformed data from the pre-processing steps, which included handling all the categorical features being encoded using one-hot and label encoding techniques, and additional features being added to get better insights into the player behaviour such as Totalplaytimeperweek and segmenting the existing features like age and sessions for deeper insights. This data is divided into training (80%), validation (10%), and testing(10%) subsets using the scikit learn library.

5.2 Model Development

The hybrid approach is used to build the model where the CNN branch is used to extract the spatial patterns using the convolutional and pooling layers and a dropout layer after each one of them to reduce the overfitting. The RNN branch is used to capture the temporal dependencies in the sequential data; it is employed by LSTM layers to capture the essential features and dropout layers after each of them to increase robustness. Outputs from both branches are combined to process through dense layers and terminate at the softmax output layer for further classification into CLV categories.

The model has compiled with adam optimiser as it is widely used for its efficiency in training deep learning models, it also combines the RMSProp and stochastic gradient descent algorithm, which ensures dynamic adaptation of learning rates to every parameter and faster convergence, categorical entropy loss is well suited for the multiclass classification and is best suited for the production of CLV, and accuracy is set as primary evaluation metric as it gives an intuitive measure of the model's performance during and after training.

The training process is spanned over 30 epochs, with entire training data passed through the model in iterative cycles with batch sizes of 64 and updating the weights. The batch size is balanced, has computational stability, and is efficient in grading the updates. The data was divided earlier for validation, which monitored the model's performance during the training and was checked for validation loss and accuracy to check for any overfitting of the training data. Implementation of an early stopping mechanism into the training pipeline to keep a check on the validation performance would prevent any data from being overfitted by the model and introduce any noise or irrelevant patterns in data.

The model's evaluation on the trained data and test data was analysed on the accuracy metric, achieving training accuracy of 91.23% and testing accuracy of 90.48%, with an accuracy of

other metrics and confusion matrix generated to highlight the precision, recall, and andF-1 score of each class to ensure a comprehensive performance evaluation.

The final outputs from the implementation were the trained model and the pre-processing artefacts, the hybrid cnn-rnn model is saved using the Keras native format, which includes model architecture, weights and optimiser configuration to ensure its compatibility with future Keras developments for effortless reuse or need for further training. The essential pre-processing tools, scaler and encoders, are saved as serialised files using the Pythons pickle library, which ensures the same transformations are applied during training and consistent during inference to maintain data integrity and prediction accuracy.

5.3 Tools and Technologies

- **Programming Environment:** Python
- **Frameworks and Libraries:**
    - TensorFlow/Keras for deep learning model development and training.
    - Scikit-learn for data preprocessing and evaluation metrics.
    - NumPy and Pandas for efficient data handling and transformation.
    - Matplotlib and Seaborn for generating visualizations.
- **Hardware:** The model was trained on Apple's M1 Silicon processor with 8GB RAM.

6 Evaluation

The evaluations for the experiments carried out for the prediction of customer lifetime value in online gaming, the analysis of model's performance, comparing the findings with previous research and implications for the academic and industry perspective.

6.1 Experiment 1: Baseline Model

To establish the baseline model for the prediction of the CLV based on the literature, the traditional machine learning models Random Forest and gradient boosting are applied. The dataset produced in the earlier steps is used for the model building, and using the same dataset and the default parameters, the model is trained for both random forest and gradient boosting. The model is evaluated on the metrics of accuracy, precision, recall and F-1 score. The model's performance, as shown in (Fig 8), is recorded as an accuracy of 89% and an f-1 score of 88% across all the classes. The baseline model performs well with the data but struggles with imbalanced class, particularly in the high clv class where it underpredicts it, this shows the need for better-sophisticated models capable of capturing complex relationships. Whereas the gradient boosting model performs slightly better with an accuracy of 91% and an f-1 score of 91%, as shown in (Fig 8). The model shows better accuracy and f-1 score but underpredicts some categories and the need for fine-tuning the techniques, refining its performance, and ensuring robust predictions of all classes.

Random Forest:

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.84	0.87	1045
1	0.87	0.93	0.90	1912
2	0.92	0.85	0.89	1047
accuracy			0.89	4004
macro avg	0.89	0.88	0.88	4004
weighted avg	0.89	0.89	0.89	4004

Gradient Boosting Model Evaluation:
Accuracy: 0.9118381618381618

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.88	0.90	1045
1	0.90	0.95	0.92	1912
2	0.93	0.88	0.90	1047
accuracy			0.91	4004
macro avg	0.92	0.90	0.91	4004
weighted avg	0.91	0.91	0.91	4004

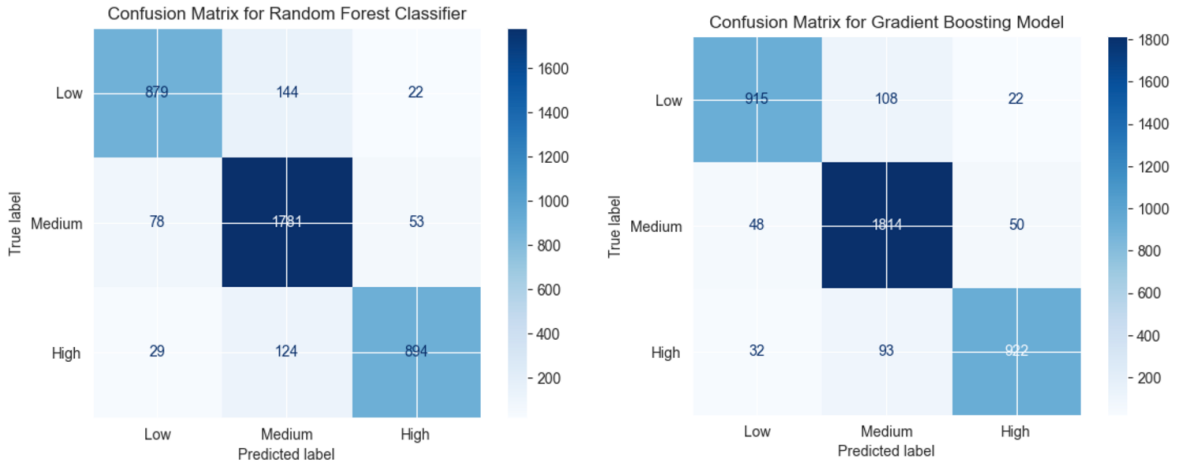


Fig 8: Classification Report

6.2 Experiment 2: CNN Model

To evaluate the capability of CNN to identify the spatial patterns in player data, the processed data is passed through a basic convolutional layer with 25 epochs and a batch size of 64, and the model outperforms the previous traditional methods with similar accuracy and better capturing of spatial dependencies. The CNN branch is trained independently with features playtime, purchases and sessions and evaluated on the metrics and the confusion matrix. The (Fig 9) accuracy of 89% and a loss of 0.36 signify the improvement in the performance of capturing the dependencies in features. However, it lacks the ability to consider the sequential dependencies which affect the temporal patterns.

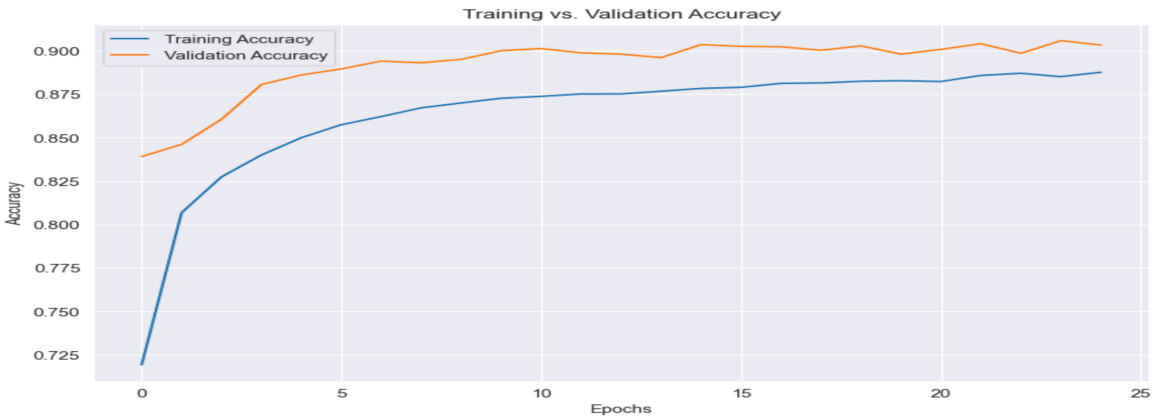


Fig 9: CNN Training vs Validation accuracy

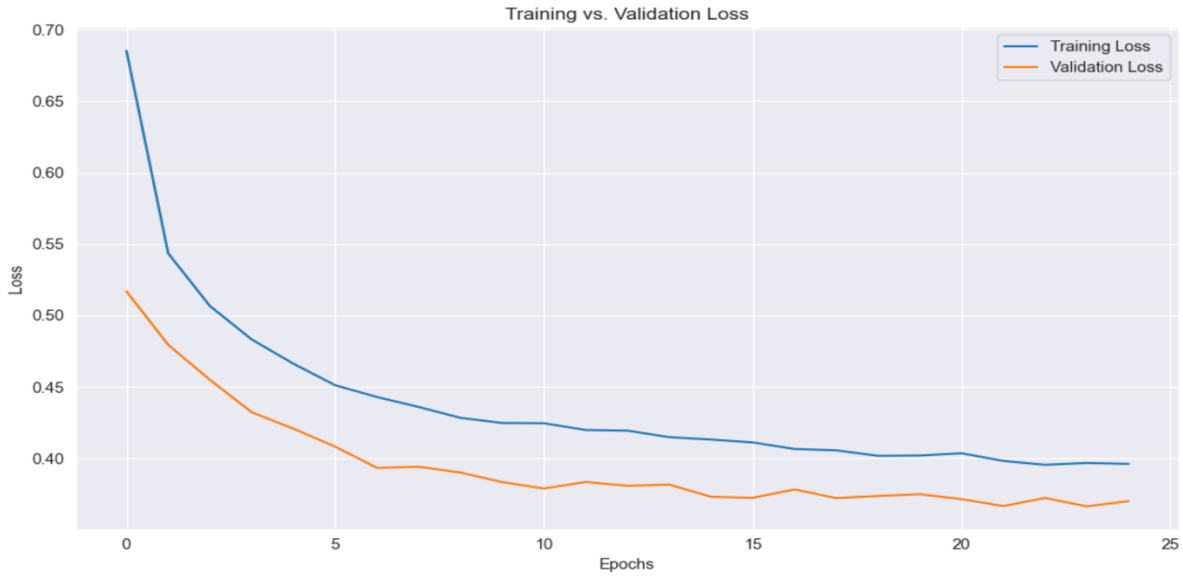


Fig 10: CNN training vs validation loss

6.3 Experiment 3: RNN Model

The RNN model works best for the textual data, and to assess its performance in capturing the sequential dependencies in the player's behaviour, the same dataset pre-processed is used to feed the model. The RNN model is trained independently using the LSTM layers, focusing on sequential data like the session history and purchase frequency. The RNN model effectively captures the temporal patterns and improves the classification of the categories in medium and high. The model achieves an accuracy of 91% and a loss of 0.32 (Fig 11). however, there is a very limited improvement based on the previous models, which suggests the need for a complementary spatial feature extraction process.

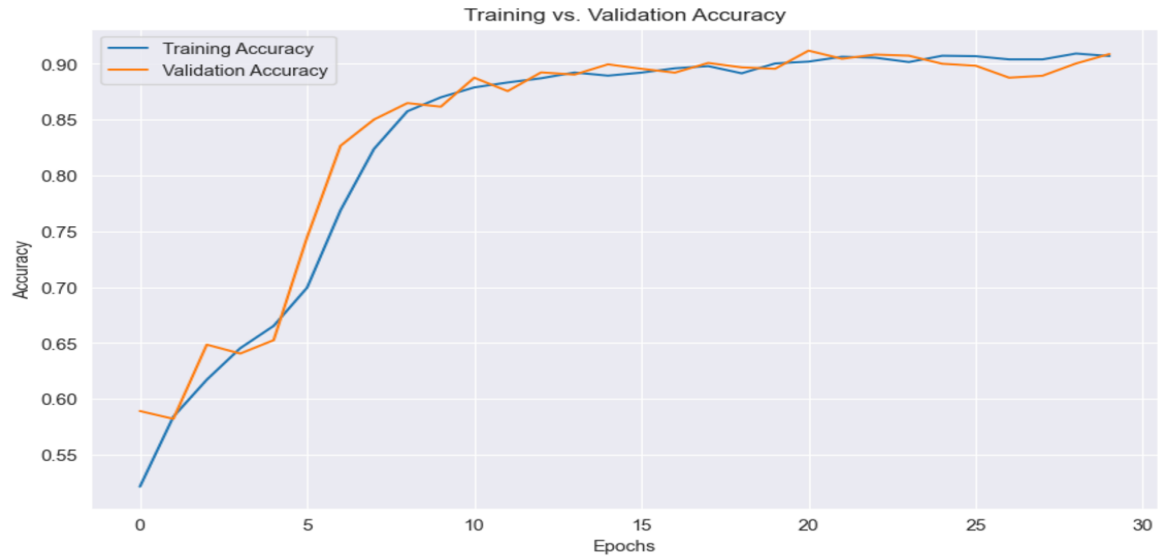


Fig 11: RNN Accuracy Plot

6.4 Hybrid CNN-RNN Model

The (Table 1) describes the previous models and their accuracies where the models performed very similarly to the hybrid model, but the traditional models struggled with imbalanced classes and couldn't predict the high engagement level category, whereas the CNN showed similar accuracy but captured the spatial patterns better but lacked in capturing the temporal patterns. The RNN outperforms the traditional and CNN models to capture the temporal features, The hybrid model developed uses both the CNN and RNN models to leverage their strengths and gives a consistent accuracy and f-1 with closely aligned validation and training scores (Fig 12), which show minimal overfitting in the model.

Table 1: Model Results Comparison

Model	Accuracy	F1-Score	Precision	Training Loss	Validation Loss
Random Forest	89%	0.89	0.89	N/A	N/A
Gradient Boosting	91%	0.91	0.91	N/A	N/A
CNN	89%	0.89	0.89	0.36	0.36
RNN (LSTM)	91%	0.91	0.91	0.32	0.32
Hybrid (CNN-RNN)	**90.4%**	**0.90**	**0.91**	**0.38**	**0.35**

The previous models lacked in some ways to capture the dependencies and extraction process thus, an ensemble CNN-RNN model to integrate spatial and temporal feature extraction for holistic CLV prediction. The hybrid model was trained with a learning rate of 10^{-3} using the Adam optimiser and categorical cross entropy loss function. The model achieved a test accuracy of 90.4% (Fig 12) and a loss of 0.36 (Fig 13) on the evaluation dataset, this shows the models robustness in predicting customer engagement levels. Observing the Precision and recall close to 0.90 as in (Fig 14) backs the models robustness. The accuracy and the loss curve over 30 epochs show constant convergence, with the validation metrics tracing closely to the training metrics which reflects the minimal overfitting.

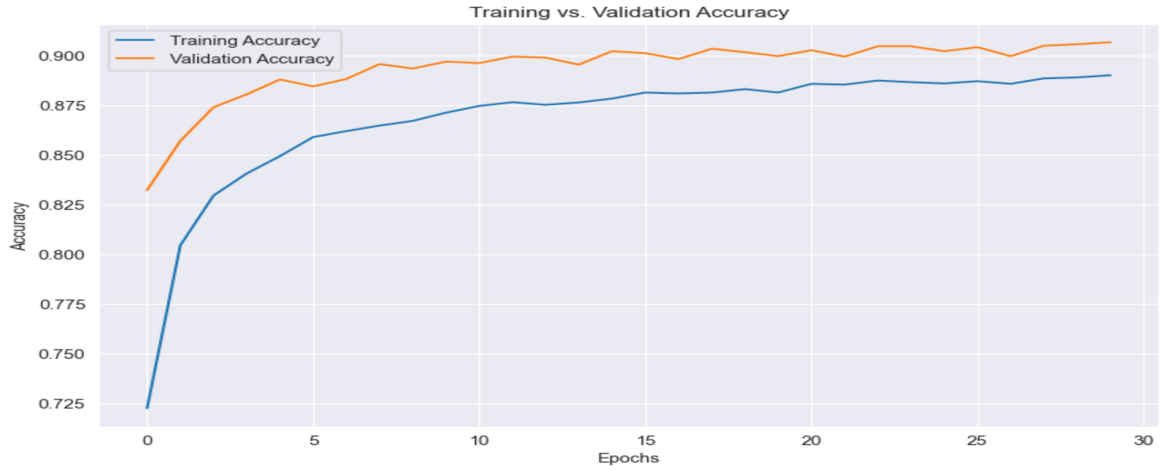
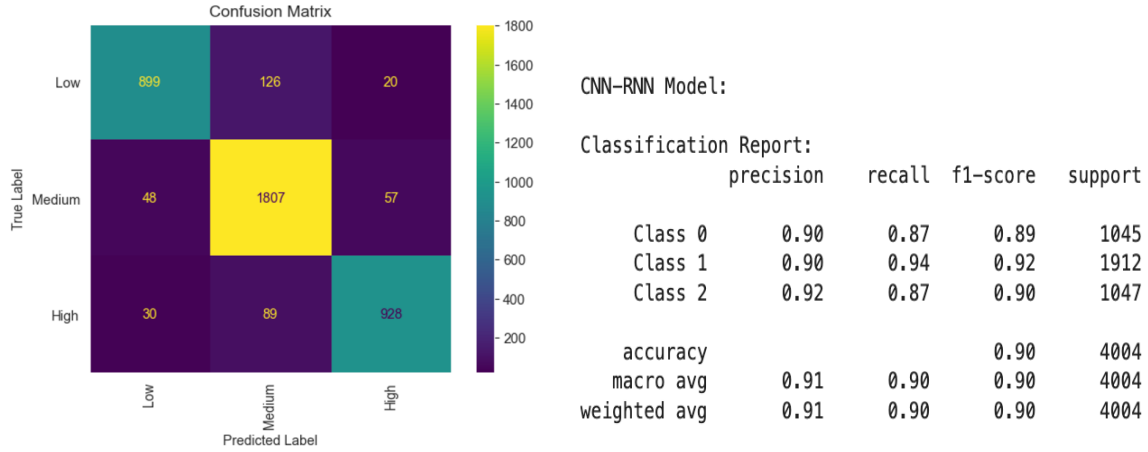


Fig 12: CNN-RNN Accuracy Curve

Fig 13: CNN-RNN Loss Curve

The confusion matrix (Fig 14) shows the precession across the engagement levels, with a majority of the classes correctly in over 90% of cases. This hybrid model outperforms the traditional classifiers and single-stream architectures like CNN and RNN on a similar dataset, showcasing the model's ability to extract the temporal and spatial features effectively.



CNN-RNN Model:

Classification Report:

	precision	recall	f1-score	support
Class 0	0.90	0.87	0.89	1045
Class 1	0.90	0.94	0.92	1912
Class 2	0.92	0.87	0.90	1047
accuracy			0.90	4004
macro avg	0.91	0.90	0.90	4004
weighted avg	0.91	0.90	0.90	4004

Fig 14: CNN-RNN Model Confusion Matrix

The model provides a high-accuracy solution that can be generalised for predicting customer engagement and sets a benchmark for predictive models in the gaming analytics domain. This model's performance aligns with the findings from (Zhao et al.(2020)) and (Ascarza et al.(2024)), which show the effectiveness of hybrid architectures in personalised retention strategies and lifetime value prediction, respectively.

6.5 Discussion

The results from the hybrid model show the integration of convolutional layers for feature extraction and the recurrent layer for sequential pattern recognition and have achieved a test accuracy of 90% and a loss of 0.36. this is well suited for the player engagement data because of its ability to capture the spatial and temporal patterns. Here, the convolutional layer processes the numerical features while the recurrent LSTM layer has sequential dependencies between the player behaviours over time.

The results show that the combining of CNN and RNN enhances the prediction performance compared to the standalone models. These outcomes validate the methodology for accurate prediction of the CLV using the advanced machine learning models.

To analyse the factors that affect the customer lifetime value in online gaming, the correlation matrix and the feature important analysis have shown the key factors that influence the CLV. The important features are the variables, playtime hours, sessions per week and in-game purchases, which show the strong predictors of the engagement levels. Based on the observations from the correlation matrix, sessions per week have a direct positive correlation with engagement. Similarly, in-game purchases show the monetary commitments of the user, which are directly linked to longer clv as of financial commitments. Demographic factors like age have shown that the interaction of different users where the mature players have shown steadier engagement patterns compared to the younger age group. This demonstrates how the behavioural and demographic variables affect the CLV and provide actionable insights on the metrics to prioritise in the retention strategies.
Based on the model's predictions, personalised retention strategies can be derived from the players. The high-value players predicted from the model based on their high engagement scores can be targeted specifically with loyalty benefits and exclusive rewards to sustain their activity; the lower engagement group, the players at risk, can be re-engaged using targeted offers based on their play patterns and game updates/ benefits tailored to their preferences. The features engineered can be used for a better understanding of the user and the creation of customised intervention strategies, with these strategies, player retention can be improved and enhance lifetime value.

While the model performs exceptionally, the synthetic data and the publicly available data do not contain all the real-world gaming behaviours. The incorporation of live operational data for better representative modelling would help understand player behaviour much better. Currently, the model architecture effectively balances the accuracy and complexity, experimenting with alternative deep-learning models as autoencoders and transformers could have a better insight into the prediction.

7 Conclusion and Future Work

This research aims at finding how the advanced machine learning model can be developed to predict the customer lifetime value accurately, and what factors affect the CLV in online gaming and the relation to player retention and engagement, and how can customer retention be improved using the predicted clv. The implementation of the hybrid CNN-RNN model, the interplay of the temporal and spatial features for accurate prediction of the clv in online gaming. The hybrid approach combines the strengths of the convolutional layer for spatial feature extraction with the recurrent layers for temporal pattern analysis. This model achieved an accuracy of 90.4% and strong generalisation across the diverse player database. The model performs better than the baseline models and uncovers complex relationships. The factors sessions per week and total playtime are the key predictors and the player level as well as in-game purchases are critical determinants of CLV. By segmenting the players into groups based on their CLV, actionable strategies from the insights can be used to tailor the promotions, session-based incentives and targeted player engagement plans.

This research advances the field of customer analytics by integrating deep learning into CLV prediction, which aligns with state-of-the-art studies and gives a boost to the academic field. This will also provide practical insights for gaming companies to optimise player retention

strategies and enhance monetisation tactics. While the model has its advantages and cons, the publicly available dataset may not comprehensively represent global gaming behaviour and can limit the model's generalisation. However, certain computational resources could have made the processing and real-time deployment smoother output and less time-consuming.

Future research can enhance the current prediction models by exploring several enhancements and by developing a modular architecture specific to different player demographics, which would refine the prediction accuracy. Testing the model in the live environment is essential to evaluate real-world performance and adaptability. Incorporating new additional features for model prediction, such as behavioural and social features and sentiment analysis, could enhance the model's predictive capabilities. Finally, the model's framework can be extended to different domains, such as e-commerce or media streaming, for a much broader analysis of customer behaviour.

References

A. Kumar, Singh, K. U., Kumar, G., Choudhury, T. and Kotecha, K., 2023. Customer lifetime value prediction: Using machine learning to forecast CLV and enhance customer relationship management. In: *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkiye, pp. 1-7. doi: 10.1109/ISMSIT58785.2023.10304958.

A. Perišić, and M. Pahor, 2022. RFM-LIR feature framework for churn prediction in the mobile games market. *IEEE Transactions on Games*, 14(2), pp. 126-137. doi: 10.1109/TG.2021.3067114.

A. Valdivia, 2021. Customer lifetime value in mobile games: A note on stylized facts and statistical challenges. In: *2021 IEEE Conference on Games (CoG)*, Copenhagen, Denmark, pp. 1-5. doi: 10.1109/CoG52621.2021.9619092.

Anders Drachen, Pastor, M., Liu, A., Fontaine, D. J., Chang, Y., Runge, J., Sifa, R. and Klabjan, D., 2018. To be or not to be... social: Incorporating simple social features in mobile game customer lifetime value predictions. In: *Proceedings of the Australasian Computer Science Week Multiconference (ACSW '18)*. Association for Computing Machinery, New York, NY, USA, Article 40, pp. 1–10. doi: 10.1145/3167918.3167925.

Arık, K., Gezer, M. and Tayali, S., 2022. The study of indicators affecting customer churn in MMORPG games with machine learning models. *Upravlenets*, 13, pp. 70-85. doi: 10.29141/2218-5003-2022-13-6-6.

Ascarza, E., Netzer, O. and Runge, J., 2024. Personalized game design for improved user retention and monetization in freemium mobile games. *Columbia Business School Research Paper No. 4653319*.

Bauer, J., Linzmajer, M., Nagengast, L., Rudolph, T. and D'Cruz, E., 2020. Gamifying the digital shopping experience: Games without monetary participation incentives increase customer satisfaction and loyalty. *Journal of Service Management*, ahead-of-print. doi: 10.1108/JOSM-10-2018-0347.

Datta, H., Foubert, B. and Van Heerde, H.J., 2015. The challenge of retaining customers acquired with free trials. *Journal of Marketing Research*, 52(2), pp. 217–234. doi: 10.1509/jmr.12.0160.

E. Lee, Kim, B., Kang, S., Kang, B., Jang, Y. and Kim, H.K., 2020. Profit optimizing churn prediction for long-term loyal customers in online games. *IEEE Transactions on Games*, 12(1), pp. 41-53. doi: 10.1109/TG.2018.2871215.

Farris, P., Bendle, N., Pfeifer, P., & Reibstein, D. (2010). Marketing Metrics: The Definitive Guide to Measuring Marketing Performance.

Flunger, R., Mladenow, A. and Strauss, C., 2017. The free-to-play business model. In: *Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services (iiWAS '17)*. Association for Computing Machinery, New York, NY, USA, pp. 373–379. doi: 10.1145/3151759.3151802.

Flunger, R., Mladenow, A. and Strauss, C., 2019. Game analytics on free to play. In: Younas, M., Awan, I. and Benbernou, S. (eds) *Big Data Innovations and Applications. Innovate-Data 2019*. Communications in Computer and Information Science, vol 1054. Springer, Cham. doi: 10.1007/978-3-030-27355-2_10.

Flunger, R., Mladenow, A. and Strauss, C., 2022. Game analytics—Business impact, methods and tools. In: Kryvinska, N. and Poniszewska-Marańda, A. (eds) *Developments in Information & Knowledge Management for Business Applications*. Studies in Systems, Decision and Control, vol 377. Springer, Cham. doi: 10.1007/978-3-030-77916-0_19.

J. T. Kristensen and P. Burelli, 2019. Combining sequential and aggregated data for churn prediction in casual freemium games. In: *2019 IEEE Conference on Games (CoG)*, London, UK, pp. 1-8. doi: 10.1109/CIG.2019.8848106.

Khajvand, M. and Tarokh, M.J., 2010. Recommendation rules for an online game site based on customer lifetime value. In: *2010 7th International Conference on Service Systems and Service Management*, Tokyo, Japan, pp. 1-6. doi: 10.1109/ICSSSM.2010.5530103.

K. Ke and Liren, X., 2010. Lifetime value management of network game customers. In: *2010 3rd International Conference on Information Management, Innovation Management and Industrial Engineering*, Kunming, China, pp. 192-195. doi: 10.1109/ICIII.2010.52.

Khajvand, M. and Tarokh, M. J., 2010. Recommendation rules for an online game site based on customer lifetime value. In: *2010 7th International Conference on Service Systems and Service Management*, Tokyo, Japan, pp. 1-6. doi: 10.1109/ICSSSM.2010.5530103.

M. Zhao, Wu, R., Tao, J., Qu, M., Li, H. and Fan, C., 2020. Multi-source data multi-task learning for profiling players in online games. In: *2020 IEEE Conference on Games (CoG)*, Osaka, Japan, pp. 104-111. doi: 10.1109/CoG47356.2020.9231585.

Ming, Y., Chen, J.E. and Li, C., 2021. The impacts of acquisition modes on achieving customer behavioral loyalty: An empirical analysis of the credit card industry from China. *International Journal of Bank Marketing*, 39(1), pp. 147-166. doi: 10.1108/IJBM-07-2020-0382.

Shiwei Zhao, Wu, R., Tao, J., Qu, M., Zhao, H. and Fan, C., 2023. PerCLTV: A general system for personalized customer lifetime value prediction in online games. *ACM Transactions on Information Systems*, 41(1), Article 23, pp. 1-29. doi: 10.1145/3530012.

Su, H., Du, Z., Li, J., Zhu, L. and Lu, K., 2023. Cross-domain adaptative learning for online advertisement customer lifetime value prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4), pp. 4605-4613. doi: 10.1609/aaai.v37i4.25583.

Tuguinay, J., Prentice, C., Moyle, B., Vada, S. and Weaven, S., 2024. A journey from customer acquisition to retention: An integrative model for guiding future gaming marketing research. *Cornell Hospitality Quarterly*, 65(3), pp. 335-353. doi: 10.1177/19389655231214718.

Voigt, S. and Hinz, O., 2016. Making digital freemium business models a success: Predicting customers' lifetime value via initial purchase information. *Business & Information Systems Engineering*, 58(2), pp. 107-118. Available at: https://aisel.aisnet.org/bise/vol58/iss2/2.

Wang, G.Y., 2022. Churn prediction for high-value players in freemium mobile games: Using random under-sampling. *Statistika: Statistics and Economy Journal*, 102, pp. 443-453. doi: 10.54694/stat.2022.18.

Xuanze Zhao, Sam, T., Zhang, X. and Liu, Y., 2024. The influencing factors of game brand loyalty. *Heliyon*, 10(10), e31324. doi: 10.1016/j.heliyon.2024.e31324.

Zhao, S., Wu, R., Tao, J., Qu, M., Zhao, H. and Fan, C., 2020. Multi-source data multi-task learning for profiling players in online games. In: *2020 IEEE Conference on Games (CoG)*, Osaka, Japan, pp. 104-111. doi: 10.1109/CoG47356.2020.9231585.

Zhao, X., Lin, J., Zhu, W., Wang, Y. and Zhang, X., 2022. Personalization strategies in mobile games: From monetization to retention. *Journal of Interactive Marketing*, 60, pp. 28-41. doi: 10.1016/j.intmar.2022.08.007.

Zhao, Y., Hu, S. and Wu, R., 2019. Customer lifetime value prediction for mobile freemium games. *IEEE Transactions on Computational Social Systems*, 7(3), pp. 792-800. doi: 10.1109/TCSS.2019.2936547.

Zhao, Y., Zhang, Y., Liu, Y., Lu, L. and Wu, R., 2021. Evaluating player loyalty with probabilistic retention and monetization models in mobile games. *ACM Transactions on Social Computing*, 4(2), Article 22, pp. 1-22. doi: 10.1145/3466894.

Zhao, Y., Zhang, Z. and Wu, R., 2020. Player behavior modeling and customer lifetime value prediction using a multi-relational approach. In: *2020 IEEE Conference on Games (CoG)*, Osaka, Japan, pp. 166-173. doi: 10.1109/CoG47356.2020.9231559.

Zhu, M., Kang, S., Jang, Y. and Kim, H.K., 2023. Multi-faceted feature extraction for mobile game customer lifetime value prediction. *Expert Systems with Applications*, 211, p. 118576. doi: 10.1016/j.eswa.2022.118576.

Zhu, X., Li, Y. and Zhang, Y., 2023. Predicting customer lifetime value using machine learning: A case study in online gaming. *International Journal of Forecasting*, 39, pp. 873-885. doi: 10.1016/j.ijforecast.2022.10.016.