National College of Ireland

Innovative Use of Behavioural Data and Explainable AI for Early Gambling Addiction Detection

MSc Research Project
Data Analytics

Saleem Pasha Shaik
Student ID: x23407719

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Saleem Pasha Shaik

Student ID: 23407719

Programme: MSc Data Analytics **Year:** Jan 2025

Module: MSc Research Practicum Part 2

Supervisor: Jorge Basilio

Submission Due Date: 11/12/2025

Project Title: Innovative Use of Behavioural Data and Explainable AI for Early Gambling Addiction Detection

Word Count: 8814 **Page Count:** 23

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: Saleem Pasha Shaik

Date: 11/12/2025

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):	

Innovative Use of Behavioural Data and Explainable AI for Early Gambling Addiction Detection

Saleem Pasha Shaik

23407719

Abstract

Gambling addiction has grown as a behavioural and financial issue which is driven by the growing accessibility of online betting platforms and the difficulty in identifying addictive patterns early. This study proposes an Explainable AI-based framework that uses behavioural data analytics and deep learning to detect early signs of gambling addiction among online players. The dataset uses demographic, betting and some intervention data to model player behaviour. The approach starts with K-Means clustering to segment players into casual, moderate-risk and high-risk categories which has evaluated using Silhouette Score, Davies–Bouldin Index and Calinski–Harabasz Index to secure strong cluster quality. High-risk player data is further analyzed using LSTM, GRU, BiLSTM and BiLSTM with Cross-Attention models for time-series forecasting of betting patterns which is bee optimized through early stopping and adaptive learning rate scheduling. Performance has measured with the help of MSE, RMSE, MAE and model response metrics like latency and throughput. Results shows that BiLSTM with Cross-Attention provides superior predictive accuracy and balanced computational performance. The use of LIME explainability enhances transparency by securing model interpretability for responsible gambling. This study contributes a novel, interpretable predictive framework for early detection and prevention of gambling addiction using behavioural data and deep learning.

Keywords: Behavioral Data, K-Means Segmentation, BiLSTM with Cross-Attention, Time-Series Prediction, Explainable AI (LIME)

1 Introduction

1.1 Background

The analysis of data analytics has been made a key pillar in research and forecasting the behavioral patterns of human beings on online platforms (Tariq, 2025). Transaction-level information is produced continuously in online gambling settings, with much of this information including the frequency of gambling, the amounts wagered, the time of the gambling, and the reaction to the win or loss (Baker et al., 2024). These behavioral traces give quantifiable indications which may signal a change in casual to risky or compulsive gambling behavior. Data analytics can help provide a more organized and science-based approach by converting gambling logs into unstructured

data into structured data that is indicative of trends, volatility, and trends (Raj, 2025). Such methods as feature engineering, clustering and time-series modeling enable analysts to reveal the hidden patterns behind the game, cluster players according to their similarities in behavior and forecast future development towards addiction without letting the situation go out of control. In this study, a two-phase analytics pipeline which does have player clustering and individualized time-series forecasting which has been designed to systematically detect growing risk patterns with higher precision. Therefore, data analytics will be used to create the basis of the proactive, personalized, and early intervention approaches, promoting responsible gambling and limiting the adverse social, psychological, and financial consequences of gambling addiction (Efthymiou et al., 2023).

1.2 Problem Statement

Online gambling platforms generate continuous behavioral data, where early signs of addiction often escalate gradually through repeated betting, loss-chasing behavior, and increased gambling frequency. Traditional risk detection systems rely on static thresholds or rule-based checks, which respond only after harmful patterns have already developed. Such approaches lack personalization, fail to capture temporal behavior shifts, and frequently produce false alarms, reducing their effectiveness in real-world intervention.

Research Question

"How does combining player segmentation with Deep Learning time-series forecasting models improve early gambling addiction detection accuracy, based on metrics and reduced false-positive intervention alerts, compared to single-model prediction approaches?"

1.3 Research Objectives

There are two research objectives of this study which are:

1. To segment players into behavioral risk groups using K-Means clustering and identify high-risk gambling patterns for targeted time-series forecasting.
2. To develop and evaluate a BiLSTM + Cross-Attention based time-series prediction model for early gambling addiction detection, and measure its performance in terms of RMSE, MAE, false-positive rate, latency, and throughput compared to baseline deep learning models (LSTM/GRU).

2 Related Work

This chapter reviews existing studies and techniques relevant to the research topic. It explains how prior literature has addressed behavioral data analysis, gambling addiction prediction, and machine learning approaches, identifying current gaps that justify and support the proposed methodology.

2.1 Data Analytics in Behavioural Prediction

Using time-stamped betting histories and demographic data can reveal the behavior of players in multiple dimensions, thus, it is possible to detect trends that are usually not observed when

using isolated data. Descriptive analytics can be used to identify patterns of engagement when the size of bets is increasing, the player is engaging in loss chasing (Edson et al., 2024), or making frequent interventions, whereas diagnostic analytics can be used to explore why it happens or why players act in a certain way by comparing player actions with their historical results (Auer and Griffiths, 2022). Predictive analytics then uses these findings to predict the possibility of escalation of risk, and therefore it is possible to intervene before addiction gets serious. This ability can be augmented with techniques of clustering and segmentation, which groups players based on their behavioral traits and deep learning models on time-series data can also be used to predict future behavior using past trends. (Ghaharian et al., 2023) that improve in response to the dynamics of players. Such a data-driven solution substitutes subjective decisions with objective and quantifiable evidence-based information and establishes a platform of responsible gambling systems to make decisions in advance.

2.2 Gambling Behaviour and Addiction Prediction

The study of gambling behaviour and addiction prediction is concerned with the identification of the underlying patterns that define the difference between casual and problem-based gambling and addiction tendencies (Hales et al., 2023). The high proliferation of online betting websites has led to the existence of vast behavioural data that is useful in obtaining information with regards to player activity and the degree of risk (Kaimara et al., 2022). Gambling behaviour of an individual is usually defined by the quantifiable behaviour like frequency of betting, the amount of bet, amount of deposit, time taken in each session and response to a win or loss (Hing et al., 2022). The division of players into risk groups, i.e., casual, moderate or high-risk, allows more focused monitoring and intervention programs (Cardoso-Marinho et al., 2022).

2.3 Clustering and Unsupervised Learning Approaches in Gambling Behaviour Analysis

Unsupervised learning, particularly clustering, has emerged as a fundamental technique for identifying behavioral risk patterns in online gambling (Sándor and Bakó, 2024). Whereas supervised models where the outcome of addiction has to be labelled, clustering algorithms examine behavioural characteristics including turnover volatility, bet frequency, loss-chasing patterns and intervention history, to automatically cluster players with similar gambling behaviour. K-Means is one of the most commonly utilized clustering models because of its scalability and interpretability (Pramod and Pillai, 2021) which may allow platforms to cluster players on low-risk, emerging-risk, and high-risk groups simply on the basis of behavioural distances. Such unmonitored observations are imperative in settings whereby there are no clear-cut labels of addiction and clustering is a potent tool of early-detection. In responsible gambling studies, this is where such cluster-based segmentation helps with focused monitoring and also with deep learning forecasting, which is only implemented on players with the highest risk indicators to further enhance precision and avoid unnecessary actions.

2.4 Studies on Data Analytics and Machine Learning Approaches for Gambling Behavior Prediction and Prevention

The latest developments in the sphere of data analytics and machine learning have contributed to the improved perception and identification of risks and behavioral patterns

associated with gambling and addiction inclinations. In several studies, different models and analytics have been put forward to deal with the increased responsibility issue of gambling and player protection in online platforms.

In (Kamdan et al., 2025), an IndoBERT-based classification scheme to identify online gambling promotions in comments posted on YouTube on the same page. The paper has demonstrated performance in recognizing explicit promotional information but has made it difficult to recognize implicit or coded expressions. Equally, (Parfenova and Clausel, 2024) suggested a better model of risk prediction to detect pathological gambling disorder in Reddit users with the addition of both temporal and emotional data. This was greatly enhanced by the addition of layers of time decay which also allowed a good performance but the specificity of the datasets to the platform made generalisation across the platforms difficult.

Another study given by (Lajcinová et al., 2023) who introduced an unsupervised deep learning approach to anomaly detection in player time series data through transformer-based autoencoders, which was superior to other autoencoders. Although innovative, labeled clinical data was not available, which presented a challenge to validation. The contribution of (Puranik et al., 2023) to the field involved the analysis of financial transaction data of a payments provider to identify behavioral gambling trends. Using the descriptive analytics, the paper has identified non-stationary, right-skewed data, which highlights the importance of preprocessing before using the predictive models. But the level of behavioral analysis was limited by the anonymization of data and exclusion of non-U.S. based users.

(Andersson et al., 2025) used the XGBoost algorithm on a Swedish casino to predict the risks of gamblers. The model was highly predictively stable despite small data windows, which supported the viability of early intervention interventions, but interpretability and provider-specific data limitations were issues of concern. Lastly, (Vatani et al., 2023) provided an analysis of a full range of LSTM-based churn prediction models in the gaming industry which perform better at predicting the sequential behavior to forecast player attrition. Computational complexity and data imbalance were, however, classified as major challenges in the study.

Table 1: Summary Table

Study	Proposed Approach	Key Results	Challenges / Limitations
(Kamdan et al., 2025)	Fine-tuned IndoBERT model with text preprocessing and classification	Achieved 98.26% accuracy with minimal false positives/negatives	Difficulty in identifying implicit or coded gambling language; dataset limited to explicit text
(Parfenova and Clausel, 2024)	Combined LSTM + EmoBERTa + Time Decay (TD) layer	Improved F1 score over baseline models; sequential modeling outperformed concatenation-based BERT	Dataset limited to Reddit; generalization to other platforms uncertain
(Lajcinová et al., 2023)	Transformer-based autoencoder compared with RNN and CNN autoencoders	Transformer model outperformed others; correlated strongly with proxy risk indicators	Lack of clinical labels; reliance on proxy indicators reduced validation accuracy

(Puranik et al., 2023)	Descriptive statistical and time-series analysis of payment data	Data found to be right-skewed and non-stationary; useful for modeling	Dataset limited to U.S. users; obfuscated IDs limited deep behavioral insights
(Andersson et al., 2025)	XGBoost classifier with temporal truncation (30–90 days)	High F1-score and ROC-AUC; stable accuracy across time intervals	Model interpretability and provider-specific bias limit real-world deployment
(Vatani et al., 2023)	Review and classification of seven churn prediction model categories	Found LSTM superior for sequential player behavior modeling	Lacks experimental validation; computational complexity and data imbalance noted

2.5 Research Gaps

The results showed that combining segmentation with deep sequential modeling solves previously identified research gaps which is related to personalization, temporal prediction, and explainability. An examination of early studies shows that key gaps remain in how gambling behaviour is predicted. Majority of the previous literature uses supervised classification models or descriptive analytics and thus cannot identify the gradual development of gambling addiction using behavioural data. The significance of temporal models has been mentioned by the past work which includes (Lajcinová et al., 2023) and (Parfenova and Clausel, 2024) but they are limited by platform specific datasets or lack of attention mechanisms to identify critical steps in behaviour. The models currently used are usually uninterpretable and therefore cannot be used in responsible gambling intervention where transparency is mandatory. To fill these gaps, the present research brings together unsupervised segmentation by K-Means clustering and forecasting models are applied only to the high-risk cluster to reduce the number of computations and allow more focused prediction of escalating behaviour. GRU, BiLSTM, and BiLSTM with Cross-Attention deep learning architectures are used to model non-stationary behavioural sequences to a higher degree than the time-series methods. Lastly, explainability of LiME is included in order to address the limitations of transparency identified in previous articles.

3 Methodology

This chapter explains how the study was conducted, including the research design, dataset details, preprocessing steps, player segmentation, data scaling, and the development and evaluation of deep learning models for predicting early gambling addiction.

3.1 Research Design Framework

This study's primary objective is to develop a Gambling Addiction Risk Prediction System using advanced deep learning techniques to analyze player behavioral patterns. The design follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, encompassing six major phases (Plotnikova et al., 2021) includes Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment as shown in Figure 1. The research focuses on three fundamental aspects: behavioral time-series analysis, pattern recognition of risky gambling tendencies, and explainability through XAI (Arsenault et al., 2025).

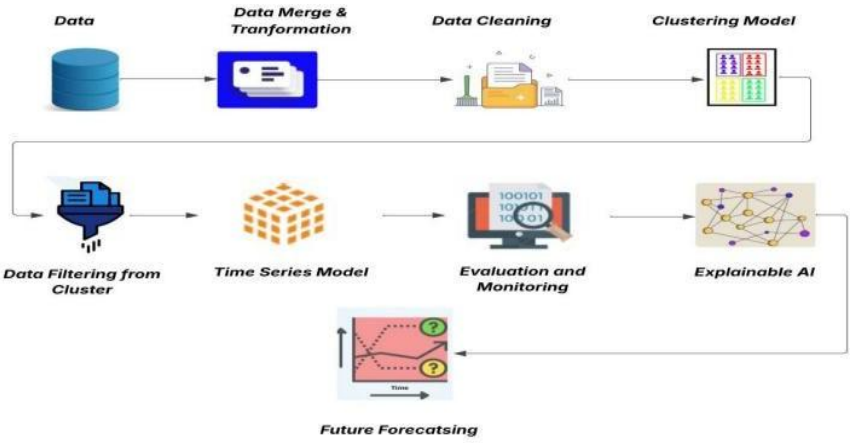

Figure 1: Methodology[1]

3.2 Data Source and Description

The dataset used in this study which was taken from the Transparency Project and the title of the data is: Behavioral Characteristics of Internet Gamblers Who Trigger Corporate Responsible Gambling Interventions[2]. The Division on Addiction, Cambridge Health Alliance, a Harvard Medical School teaching hospital, created it and trained it under the direction of Dr. Howard J. Shaffer, and was sponsored by Interactive Entertainment AG. The data gives a detailed account of actual online gambling behavior that took place in ten years. It includes information with 2,066 players who raised the responsible gaming alerts and the same number of the control players (N = 2,066) who had the same exposure to the platform.

We are using the Multiple datasets for project, which are demographics data which has 4,134 records with 8 variables including age, gender, country, date of registration daily aggregation dataset has 9,81,782 records with 6 columns such as turnover, hold, product type, bets, betting dates and responsible gambling interventions has 2,068 records with 6 attributes having type, frequency, and time of alerts.

3.3 Data Cleaning and Feature Engineering

The combined data was comprised of the demographic and betting history and responsible gambling intervention records to create a single table of analysis (Lind et al., 2021). First, the primary data files were combined, i.e., the demographics, the betting transactions, and the intervention logs, with the help of a unique player identifier (UserID), which resulted in the creation of complete records on a player level. The process of the merging was implemented with the help of outer joins so that no information on users that would be important is lost between the sources. The result of this process gave a combined dataset of over 981, 000 records containing 18 columns. The data was then subjected to a significant amount of data cleaning to allow inconsistencies and missing values.

[1] https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html

[2] http://www.thetransparencyproject.org/download_index.php

After cleaning, feature engineering was performed to increase the capability of the data in the

analytics (Chatzimparmpas et al., 2022). Age is a new derived variable created by calculating YearofBirth- 2010 to show the age of each player at a given point of reference to be used to divide the behavior of players according to age. Other artificialized characteristics like rolling turnover averages and volatility of bets were added later to reinforce predictive modeling. Lastly, the processed data was verified to ensure that there were no gaps in the data hence the clean and properly structured data was available to be used in clustering and training deep learning models. These feature-engineered variables have created because they capture hidden behavioural patterns that directly improve clustering accuracy and strengthen time-series prediction of addiction risk.

3.4 Data Visualisation

This visualization is a composite display of all box plot distributions of all the numerical variables within the merged gambling data to provide a full picture of the data distribution (Bhatt et al., 2022), central tendencies as well as the patterns. The bar chart is represented in figure 2 and is an analysis bar chart that discusses two important variables in gambling datasets. Figure 3 demonstrates the correlation between the age of a user (as of 2010) and average values of a hold, which indicate the trends in the profitability of gambling by age groups.

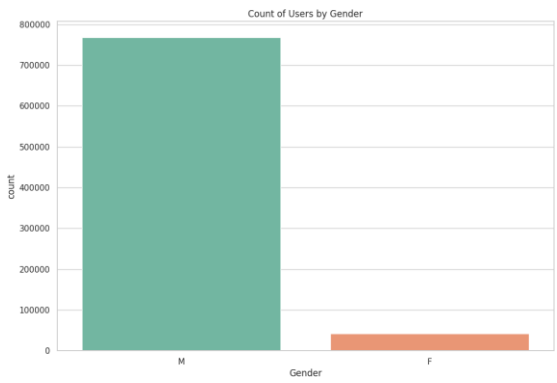

Figure 2: Gender Distribution Analysis

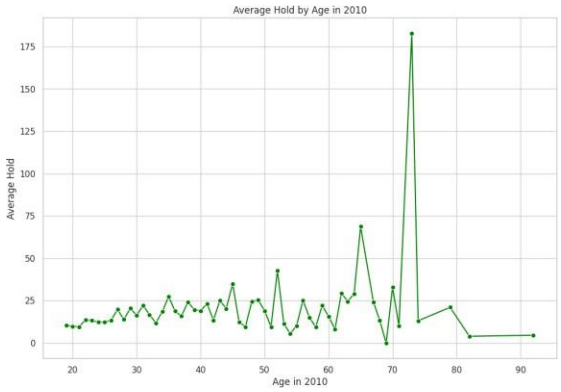

Figure 3: Age-Based Hold Distribution Analysis

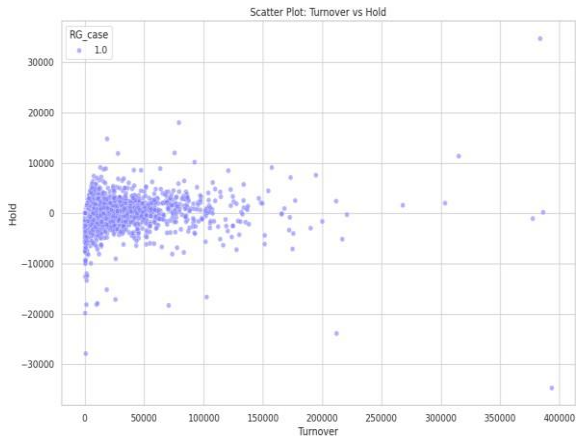

Figure 4: Turnover-Hold Relationship Analysis

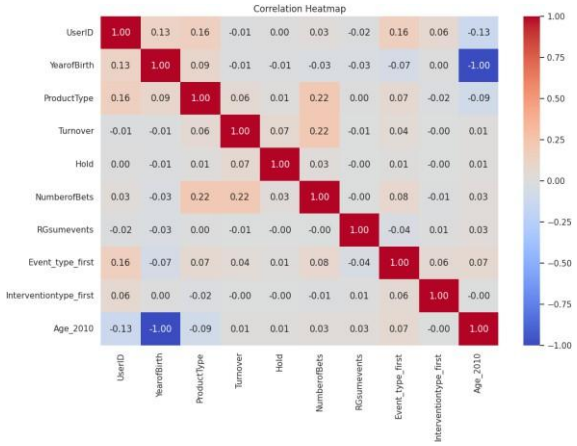

Figure 5: Correlation Heatmap

Figure 4 analyzes the relationship between gambling turnover and hold values where patterns of critical betting behavior have been observed. Figure 5 demonstrates the correlation heatmap that depicts the linear linkages among the numerical attributes within the gambling data set (Stechschulte et al., 2024) using coolwarm color scale in which red represents a positive correlation (+1.00) and blue represents a negative correlation (-1.00).

3.5 Data Preprocessing

3.5.1 Label Encoding

Label encoding was performed to convert categorical variables into numerical format (Bolikulov et al., 2024) suitable for machine learning algorithms (Kosaraju et al., 2023). First, all object-type columns were identified using select_dtypes, excluding date-related columns such as Registration_date, First_Deposit_Date, and RGLast_date to avoid distorting temporal data. The remaining categorical features, including CountryName, Gender, and ProductType, were iteratively transformed. This process ensured consistent numeric representation while preserving category relationships. The transformation streamlined the dataset, enabling algorithms like K-Means and LSTM to process categorical information effectively. Upon completion, a confirmation message indicated successful encoding.

3.5.2 Data Filtering

Following encoding data filtering has done to retain only relevant analytical variables and eliminate redundant identifiers or date-based fields (Nugroho et al., 2023). The filtered dataset was refined to include only important numeric and target variables Turnover, Hold, NumberofBets, RGsumevents and RG_case which represent key behavioural and intervention indicators. This selective filtering produced a concise, high-quality dataset, optimized for clustering of gambling risk behavior. Turnover have selected as the target variable because it directly reflects gambling intensity and progression toward addiction whereas variables like hold remain has influenced by external randomness like game outcome rather than behavioral escalation.

3.6 Player Segmentation

To segment players based on behavioral similarity, the study employed K-Means clustering using aggregated feature vectors that included turnover_mean, volatility, intervention_count, and recency (Perišić and Pahor, 2023). The Clustering was determined using the Elbow Method and validated through the metrics such Silhouette Score, Davies– Bouldin Index, and Calinski–Harabasz Index to ensure well-separated and coherent groupings. The resulting clusters were labeled as Casual, Moderate-Risk, and High-Risk players, reflecting varying levels of gambling intensity and intervention history. For subsequent time-series forecasting and addiction prediction, only the High-Risk cluster was selected to focus on players exhibiting strong behavioral patterns indicative of potential gambling addiction.

3.7 Data Scaling

To ensure all features contributed equally to the model and to prevent bias caused by differing value ranges, data normalization was performed using the StandardScaler from the scikit-learn library (Testas, 2023). In this study, data scaling was applied only to the time-series modelling stage not during K-Means clustering. This is because clustering was

performed using aggregated behavioural type of features which were already normalized through feature engineering and aggregation.

3.8 Window Rolling

Rolling windows were used to convert continuous behavioural time-series into structured input sequences for the forecasting models. This method allows the system to capture temporal dependencies by providing models with recent behavioural patterns leading up to each prediction point. It transforms the data into a supervised format which is suitable for LSTM, GRU, and attention-based architectures. By framing behaviour as sequential segments, rolling windows help the model to learn emerging risk, trends and fluctuations signals over time (Cui et al., 2021).

3.9 Early Stopping & Learning Rate Schedular

Early stopping was used through the training, avoid overfitting by stopping the model where validation loss halt improving. Finally, the application of a learning-rate scheduler was used to reduce the step size when progress down, thus guaranteeing smoother convergence. These optimization controls helped with improving model stability and prevented unnecessary training epochs. Combined, they make the forecasting performance more reliable across all deep learning architectures employed (Xiong, 2024; Vatani et al., 2023).

3.10 Latency and Throughput

Latency and throughput were measured to determine the efficiency of operational forecasting models (Stechschulte et al., 2024). Latency refers the average time taken by the system to make a single prediction; this indicates real-time responsiveness. Throughput refer to the number of predictions can be generated by this model in one second, indicating scalability for continuous monitoring. Therefore, these metrics generally investigate if this model could be deployed on live gambling-risk systems where timely alerts and interventions are important.

3.11 Data Modelling Overview

Multiple DL models have been selected to model sequential gambling patterns. These architectures have chosen because time-series forecasting requires learning temporal dependencies and behavioral progression rather than static values. The models help forecast escalation patterns which supports early detection of addiction tendencies.

3.12 Evaluation Metrics

The clustering performance is assessed by internal measures that include the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, which together provide an indication of how well the user groups are separated and how consistent the members are within each cluster. Evaluation for the forecasting models is thus done based on some error-based metrics, including Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error, all measuring the closeness of the predicted gambling behaviour to real values. The contribution of each feature to each prediction is interpreted using Explainable AI techniques, most specifically LIME, to make model outputs transparent and behaviourally meaningful.

Taken together, these various metrics create a comprehensive understanding of model performance and explainability for the entire project.

4. Design Specification

This chapter presents the design specification of the system, explaining its overall architecture, workflow sequence, and selected technology stack. It outlines how data flows through preprocessing, clustering, deep learning prediction, and explainable AI layers to support early gambling addiction detection.

4.1 System Components and Architecture

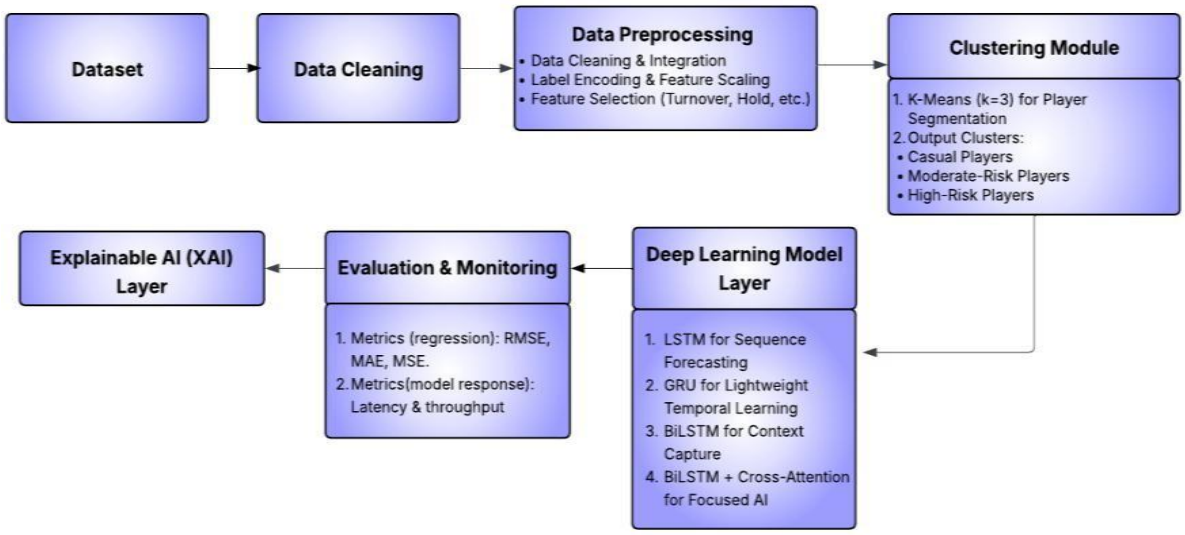

Figure 6: System Architecture Diagram

The system architecture depicts the multi-layered system to be used in the detection of gambling addiction based on behavioral information and Explainable AI as illustrated in Figure 7. The first component of the architecture is the dataset, which will collect raw data like demographic data, betting history, and responsible gambling alerts. The Data Preprocessing Layer is used to integrate, clean, label encode and scale features in order to guarantee the quality and consistency of data. Then, the Clustering Module uses the K-Means (k=3) algorithm to cluster players (Marisa et al., 2022) into casual, moderate-risk, and high-risk. The Deep Learning Model Layer is based on superior sequential models to predict the behavior of players in the long run. The Explainable AI Layer uses LIME to explain the model decisions, making it possible to see the effect of every feature. Lastly, the Output and Evaluation Layer will produce performance measurements and risk forecast reports which are clear, objective and backed by fact-on-the-ground data on gambling patterns and in support of early intervention techniques.

4.2 Proposed Workflow of Gambling Addiction Risk Prediction System

Figure 7 illustrates the workflow architecture of the proposed Gambling Addiction Risk Prediction System. The process begins with data cleaning, label encoding, and data filtering, ensuring that the dataset is refined for accurate analysis. The filtered data undergoes K-Means clustering, categorizing users into Casual, Moderate, or High-Risk gamblers. A conditional check determines whether a player exhibits high-risk behavioral traits. If yes, their data

proceeds to the time-series deep learning model training phase using architectures like LSTM, GRU, and BiLSTM with Cross-Attention. Model performance is enhanced via early stopping and learning-rate scheduling. Finally, LIME explainability reports generate interpretable insights, and the model outputs the predicted behavioral trajectory, completing the workflow.

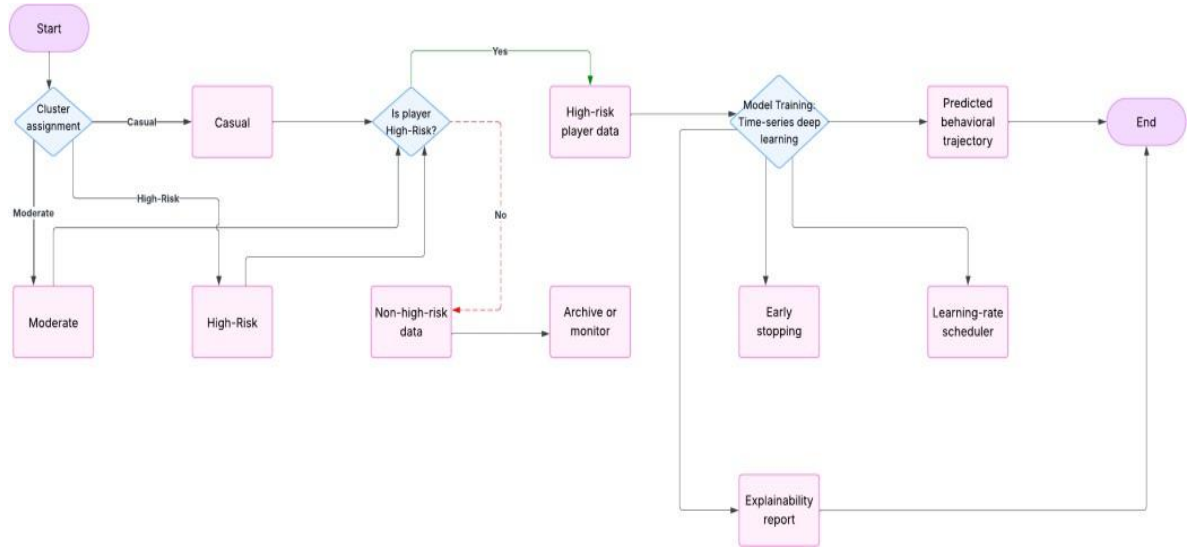

Figure 7: Workflow Diagram of Gambling Addiction Risk Prediction System

4.3 Technology Stack

Table 2 summarizes the complete technology stack which has used in the study. It uses Python-based libraries for data science and deep learning, visualization tools for interpretability and GPU acceleration for computational performance.

Table 2: Technology Stack

Category	Technology / Tool	Purpose / Description
Programming Language	Python	Core language for data preprocessing, modeling, and analysis.
Libraries & Frameworks	TensorFlow, Keras, Scikit-learn, NumPy, Pandas	Used for deep learning model construction, training, and data manipulation.
Visualization Tools	Matplotlib, Seaborn, LIME	For visualizing feature importance, cluster patterns, and explainable AI interpretations.
Development Environment	Jupyter Notebook / Google Colab	Interactive coding and experimentation platform.
Version Control	GitHub	For project version management and collaboration
Hardware Acceleration	GPU (TensorFlow Backend)	Enhances model training speed and efficiency.

5 Implementation

This chapter explains how the proposed system is practically built, including data preparation, player clustering, rolling window time-series modeling, implementation of LSTM-based prediction models, and integration of LIME for explainability, demonstrating the step-by-step execution of the gambling addiction detection framework.

5.1 Determining Optimal Number of Clusters Using the Elbow Method

The Elbow Method was used in order to determine the most applicable number of clusters to use in segmenting players through the K-Means algorithm (Sammouda and El-Zaart, 2021). This model was trained repeatedly with values of K (between 1 and 10) and the Sum of Squared Distances (SSD), otherwise known as inertia, were computed at each iteration. This value is used to denote the optimum number of clusters, where the similarity between clusters is minimal as well as across the clusters. The visualization gave a clearer indication that K = 3 was the best option to cluster the behavior of players in the most efficient way illustrated in Figure 8 (left). K = 3 was chosen because the Elbow Method showed a clear inflection point at K = 3 by showing the optimal balance between reducing cluster variance and avoiding over-segmentation. The parallel coordinates plot shows how the three identified player clusters like Early Players, Moderate Problem Gamblers and Problem Gamblersin Figure 8 (right) which is differ across the key behavioural features used in clustering. The feature names like mean_turnover and mean_hold have been selected to clearly represent aggregated behavioral metrics over time which secures interpretability during clustering and distinguishing them from raw daily transactional values.

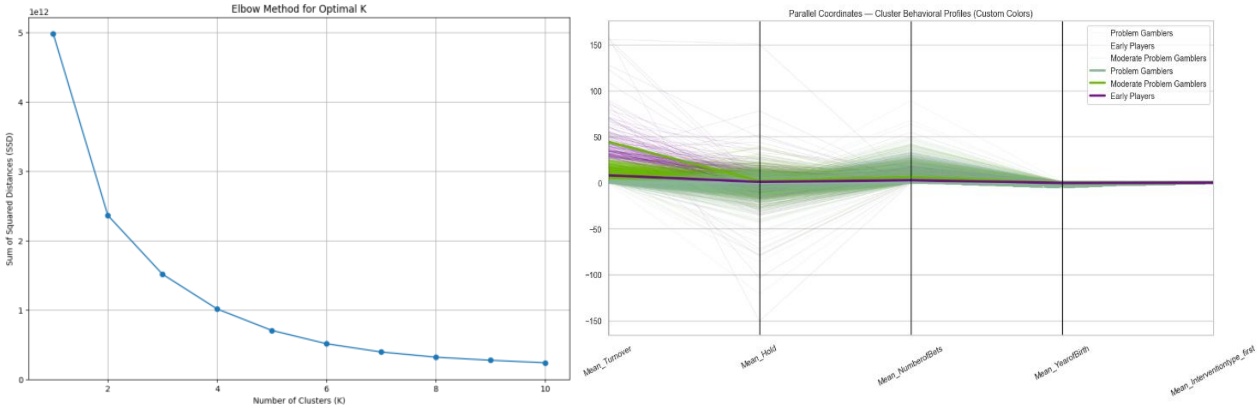

Figure 8: Elbow Method for Optimal K in left and Analysis of Cluster Labelling in right

5.2 K-Means Clustering Implementation and Visualization

Once the K = 3 was identified as the best, the K-Means algorithm was run that would divide players according to behavioral features. The model gave every record a cluster label where players with similar patterns of gambling were put. Interpretation was made of the cluster centers, which were the behavioral centroid. The visualization of the results was based on a 2D scatter plot, with the data points being color-coded based on their respective cluster, and the red X markers used to indicate the center of the clusters. In this visualization, we were able to distinguish several groupings of behavior, including casual, moderate-risk and high-risk players. Lastly, a new column (Cluster) was incorporated to the dataset to allow additional specific time-series analysis and forecasting in each of the behavioral segments as demonstrated in Figure 9. K = 3 have selected because the elbow curve stabilized after K = 3 and K = 4 did not provide a meaningful type of behavioral split nor improve silhouette score

by making K = 3 statistically and behaviorally optimal.

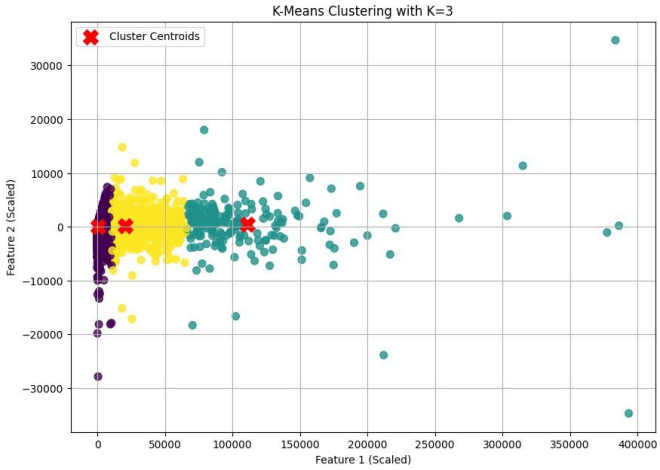

Figure 9: K-Means Clustering with K=3

5.3 Time-Series Data Preparation and User Filtering

To conduct the time-series modeling stage, it was necessary to make sure that only users with adequate behavioral history were considered so as to make the temporal patterns meaningful. First, the data set was narrowed down to include only the records that were related to the Cluster 0 which was the high-risk group of players that were detected during the clustering stage which contains players with the strongest signs of escalating gambling behaviour by making it the only cluster relevant for early addiction prediction. To achieve data consistency and reliability of the model, players that have over 102 recorded betting events were kept, and to ensure that each time series had sufficient data to perform multi-day rolling window analysis. The filtering was done through value counts function to locate those users that have surpassed the minimum activity limit. Following these limitations, the dataset was narrowed down to 1,412 unique users and 779, 753 overall records, which created a solid foundation of sequential analysis.

5.4 Rolling Window Generation for Time-Series Modeling

A rolling window mechanism was adopted to prepare time-series data to successfully capture temporal dependencies and behavioral patterns in the long run (Cui et al., 2021). It consisted of converting records of continuous gambling activity into fixed length sequential samples to be used in deep learning models like LSTM, GRU, and BiLSTM. A custom-made function creates rolling windows (data, target, windowsize=7) was created that processes input-output pairs with a 7-day observation window. Each sequence was considered to have an output value (y) that was the target value on the next day, and seven consecutive days of player activity were used as the input segment (X). A 7-day rolling window have been selected to capture weekly behavioral cycles which align with typical gambling frequency patterns. Longer windows like 14 or 30 days reduced short-term behavior sensitivity and led to loss of temporal resolution.

5.5 Implementation of Time Series Modelling

5.5.1 LSTM

The LSTM model was developed based on the Keras Sequential API to identify the temporal relationships of the behavioral data of the players (Zhang, 2022). The architecture consisted of 2 LSTM layers (128 and 64 units respectively) with the tanh activation function. To avoid overfitting, dropout layers (0.2) were added between them. The learned representations were refined by a Dense layer (32 units, ReLU) and then, there was a single output neuron to make risk score predictions. The model was put together using the Adam optimizer and Mean Squared Error (MSE) loss. The training was done in 30 epochs with early stopping and reduction of the learning rate as the callbacks where the training was stopped at epoch 6 because the optimal restoration of the validation performance was achieved at epoch 1.

5.5.2 GRU

GRU model was designed to help efficiently express sequential dependencies in the behavioral time-series data at the lowest possible computational costs (Xiong, 2024). In the build_gru_model() function, the architecture was defined as four stacked GRU layers of 64 units each and all of them excluding the last layer were set to the return sequence mode. The rate of dropouts was 0.7 with each GRU layer to prevent overfitting through random neuron disabling during training. A Dense layer of one neuron of continuous value prediction (risk score) was used to conclude the network. The model was assembled on Adam optimizer, Means Squared error (MSE) as the loss function, and Means Absolute Error (MAE) as an evaluation. The 20 epochs of training were done using a batch size of 64 along with adaptive optimization through early stopping and learning rate reduction callbacks. The model performed best at epoch 5 an early stopping condition occurred, which led to the weights being restored after that stage.

5.5.3 BiLSTM

To improve sequential learning, the Bidirectional Long Short-Memory (BiLSTM) model was used to enable the network to have forward and backward processing of data, aiding the model in capturing the overall temporal dependencies in player behavior. The architecture was started with a Bidirectional LSTM layer (128 units), a Dropout layer (0.3) was used to curb overfitting. It was followed by the second BiLSTM (64 units) that had a higher dropout rate of 0.7 to provide further regularization and the third BiLSTM (32 units) that was set to only give back final output sequence.

5.5.4 BiLSTM with Cross Attention

BiLSTM + Cross-Attention model was created to increase the predictive accuracy of the model by aiding the bidirectional sequential learning model with an attention mechanism which dynamically focuses on an important time step in the behavior of players. The model was started with two Bidirectional LSTM layers with 128 and 64 units each, and then, they were followed by a dropout rate of 0.5 to minimize overfitting and maintain important temporal relationships. The end of the training was epoch 8, and the best model generalization and interpretability was preserved through the restoration of best-performing weights at epoch 3 as shown in Table 3.

Table 3: Summary of Regularisation and Training Settings

Model Name	Regularization (Dropout)	Training Configuration	Early Stopping (Best Epoch)
LSTM	0.2 after each LSTM layer	Epochs: 30 Batch size: 64	Stopped at 6 (Best: 1)
GRU	0.7 after every GRU layer	Epochs: 20 Batch size: 64	Stopped at 5 (Best: 1)
BiLSTM	0.3, 0.7, 0.7	Epochs: 20 Batch size: 64	Stopped at 7 (Best: 2)
BiLSTM + Cross-Attention	0.5 after each BiLSTM	Dynamic epochs Batch size: 64	Stopped at 8 (Best: 3)

5.6 Implementation of Explainable AI using LIME

To improve interpretability and transparency in deep learning predictions, the Local Interpretable Model-Agnostic Explanations (LIME) system was introduced into it. The methodology tried to understand the logic of predicting gambling addiction risk using the trained BiLSTM and describe the most significant behavioral attributes. In the chosen test case, LIME created 5,000 perturbed samples and gave weight of importance to the 20 most significant features as shown in Table 4. These explanations were visualised and saved in form of an interactive HTML file (lime_explanation_gambling.html), which showed the effect of changes in features like turnover and the number of bets on predicting the risks. A table of corresponding feature-value mapping was also created to depict scaled values of inputs, which allows one to have a clear understanding of how decisions are made by the model. This is made possible by this explainability layer as it will guarantee interpretability, accountability, and trust of the AI-based gambling risk detection system. Figure 10 shows the feature importance visualization which is generated using LIME for the BiLSTM model. The bar chart shows the top behavioral attributes influencing the model's risk prediction.

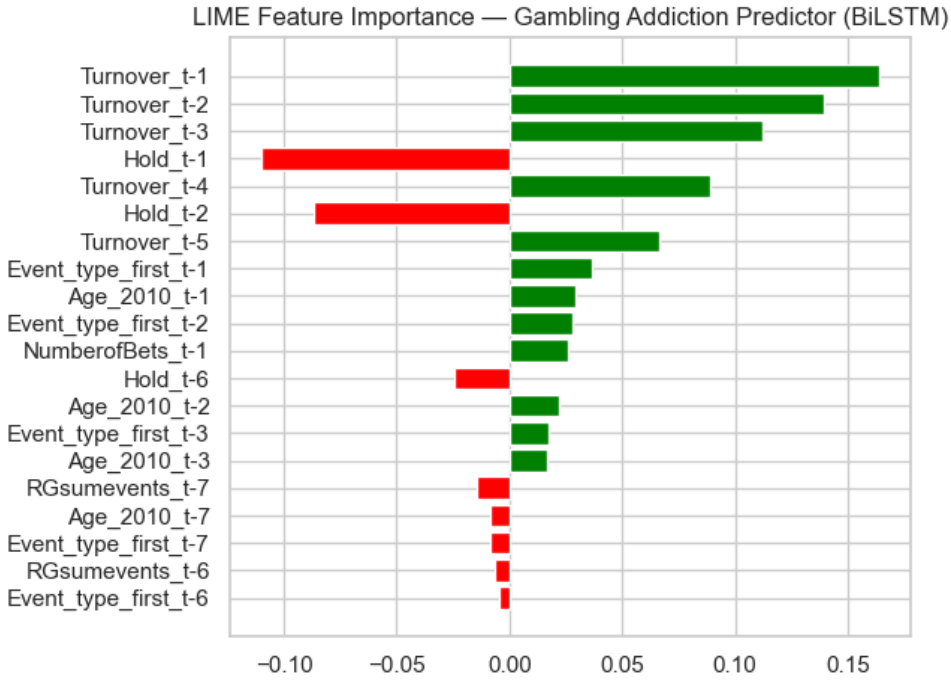

Figure 10: LIME Feature Importance — Gambling Addiction Predictor (BiLSTM)

6 Evaluation

Chapter 6 presents the evaluation of the proposed system by assessing clustering quality using Silhouette, DBI, and CHI indices, and comparing deep learning models using MSE, RMSE, MAE, while also analyzing model performance trends, prediction accuracy, latency, and throughput.

6.5 Silhouette Score

A major measure to determine the quality of the clustering done by the K-Means algorithm is called the Silhouette Score (Punhani et al., 2022). It determines the fit of each piece of data to its respective cluster against those of the other clusters. It has a scale of -1 to +1 with the high score indicating the presence of well-separated and dense clusters.

6.6 Davies–Bouldin Index (DBI)

When K-Means is used, the DaviesBouldinindex (DBI) is used to gauge the compactness and separation of the clusters created by the algorithm (Septiani et al., 2025). It works out the mean similarity of each cluster with the one that is most similar to it and a smaller value can be seen as a higher performance of the clustering process. A DBI that is closer to zero indicates a small intra-cluster variance and a large inter-cluster separation.

6.7 Calinski–Harabasz Index (CHI)

The Variance Ratio Criterion, the Calinski Harabasz Index (CHI) is an index used to determine the ratio of between cluster dispersion to within cluster dispersion. The greater the CHI values, the more specific and discrete clusters. A high score on CHI in this study confirms the effectiveness of clustering-based segmentation of players to verify that K-Means has accurately segmented the high-risk player group to run through the deep learning analysis.

Table 4 summarizes the clustering performance metrics, confirming that the K-Means algorithm achieved highly distinct and well-separated clusters with minimal internal variance.

Table 4: Clustering Evaluation Metrics for K-Means Model

Metric Name	Value	Ideal Condition	Interpretation
Silhouette Score	0.9719	Higher = Better (Max = 1)	Indicates excellent cluster separation and cohesion
Davies–Bouldin Index (DBI)	0.4924	Lower = Better	Shows minimal overlap among clusters
Calinski–Harabasz Index (CHI)	930068.3736	Higher = Better	Reflects strong inter-cluster variance and compact grouping

6.8 RMSE, MAE, MSE

Performance metrics applied to deep learning models (LSTM, GRU, BiLSTM, and BiLSTM + Cross Attention) are the Root Mean Squared error (RMSE), Mean Absolute error (MAE), and mean squared error (MSE). RMSE is used to measure the square root of the mean squared error between the predicted and actual values with a focus to big errors. MAE is an

average value of absolute differences, which give a simple value of prediction accuracy. MSE measures the mean squares error and the larger the error the more is penalised. All of these measures can estimate the extent to which any of the models predicts player behavioral patterns and identifies early gambling addiction.

6.9 Experiment 1 – Evaluation of LSTM

The actual vs. predicted plots are used to visually confirm the degree of adherence of the model to the actual behavioral trends. The agreement in peak and variance patterns justifies the conclusion that the model effectively forecasts volatility and escalation indicators which are useful in detecting addictions.

Figure 11 shows the actual and predicted patterns of betting of a time-series of a high-risk player as produced by the LSTM model. The red line is the presentation of the predicted values that the model predicts as well as the real values that the model predicts as time progresses. The fact that both curves are almost parallel is an indication that the LSTM is successful in capturing the time-dependency and changes in gambling behaviour

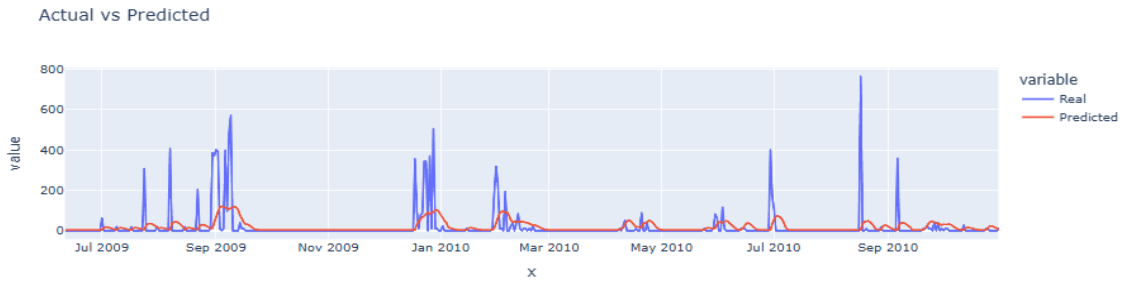

Figure 11: Actual vs Predicted Graph

6.10 Experiment 2 – Evaluation of GRU

The performance of the GRU model to predict the betting turnover of a high-risk player is demonstrated in Figure 12. The chart shows how the model can be used to attain the overall betting pattern and trends of the problem gambler. The fact that the actual and the predicted values are very close to each other substantiates the effectiveness of the model in terms of predicting high-risk betting behaviors.

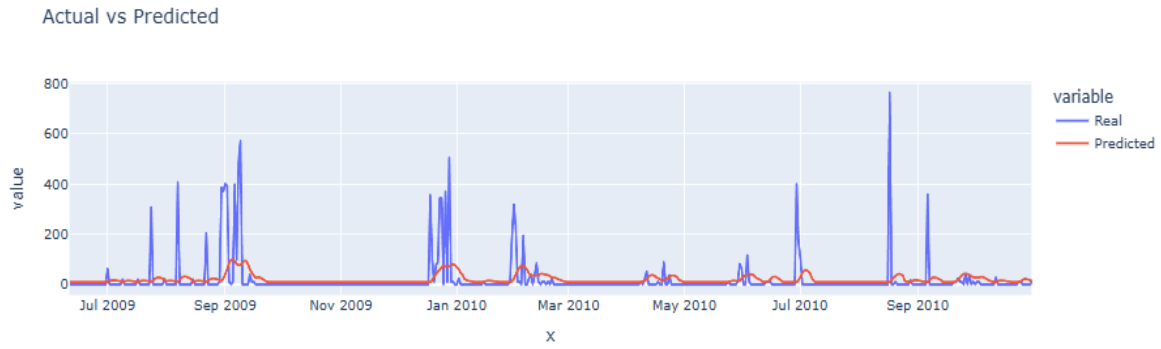

Figure 12: Actual vs Predicted Graph

6.11 Experiment 3 – Evaluation of BiLSTM

Figure 13 shows the predicting performance of the BiLSTM model of a high-risk player betting turnover between July 2009 and September 2010. The blue curve represents the real turnover values whereas the red curve is a representation of the BiLSTM model. The bidirectional structure makes the model to be able to define both forward and backward temporal relations among the betting patterns. It is evident that significant spikes in gambling activity occur around the time of September 2009, January 2010, and September 2010 and the model shows that these volatility spikes can be reasonably followed.

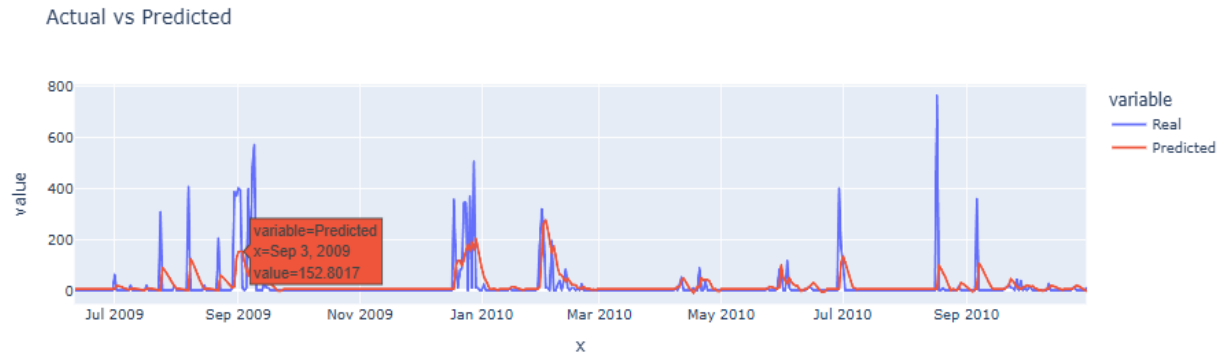

Figure 13: Actual vs Predicted Graph

6.12 Experiment 4 – Evaluation of BiLSTM + Cross Attention

Figure 14 illustrates the forecasting capability of the BiLSTM with Cross Attention mechanism on a high-risk betting turnover of a player in the month of July 2009 until September 2010. The model is shown to have a better prediction accuracy than the conventional BiLSTM especially in moderated activity periods although some lag is still observed during extreme volatility spikes.

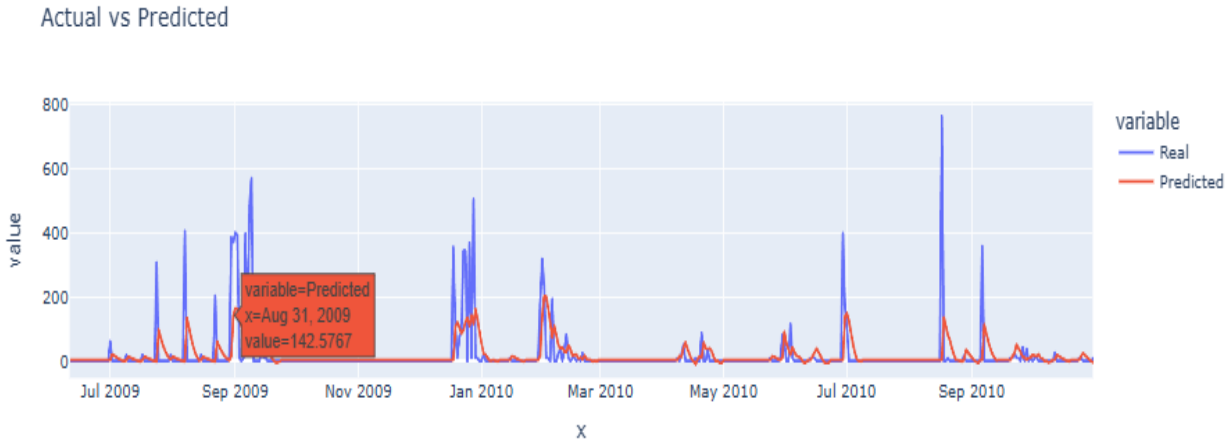

Figure 14: Actual vs Predicted Graph

6.13 Latency and Throughput Analysis

Low latency is important since real time detection allows platforms to take early warning or intervention measures before harmful gambling goes into overdrive, and responsiveness is as important as predictive accuracy. Table 5 presents latency and throughput comparisons across models. The BiLSTM achieved the lowest latency, indicating the fastest response time per prediction, while GRU demonstrated the highest throughput, making it the most computationally efficient model overall.

Table 5: Model Response Metrics (Latency and Throughput)

Model	Latency per Prediction (ms)	Throughput (Predictions/sec)
LSTM	122.6306	2298.41
GRU	119.8880	3532.58
BiLSTM	84.9619	2018.00
BiLSTM Cross Attention	124.3032	2201.47

6.14 Evaluation Metrics Analysis

Table 6 presents the comparison of regression error metrics across deep learning models. The BiLSTM with Cross Attention achieved the lowest MSE, RMSE, and MAE, indicating superior prediction accuracy and minimal deviation between actual and predicted wagering patterns.

Table 6: Model Performance Based on Error Metrics (MSE, RMSE, MAE)

Model	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)
LSTM	0.5200	0.7211	0.2771
GRU	0.5332	0.7295	0.2900
BiLSTM	0.5007	0.7125	0.2862
BiLSTM + Cross Attention (Best)	**0.4939**	**0.7028**	**0.2599**

6.15 Discussion

BiLSTM with Cross-Attention greatly aided in the system to identify early gambling addiction by capturing both forward-moving and backward-moving behavioural dependency as well as selectively prioritising the most impactful time steps. The attention module (in contrast to generic LSTM or GRU models) functions by dynamically giving higher importance to spikes in behaviour, turnover volatility, and sudden changes in the rate of betting. Such biased weighting would enable the model to concentrate on risk-of-importance trends as opposed to noise which leads to reduced false-positives and greater accuracy in the forecasts. In general, the hybrid architecture has high behaviour-awareness addiction prediction. Our results address the identified gaps by combining unsupervised segmentation with focused forecasting on the high-risk cluster by improving predictive accuracy and explainability via LIME.

The clustering stage has shown strong segmentation performance which does have a high Silhouette Score of 0.97, a low Davies–Bouldin Index and a high Calinski–Harabasz value. Although there are some previous studies that suggested experimenting with more clusters that the Elbow Method and behavioural interpretation confirmed three clusters which includes casual, moderate-risk and high-risk which were the most meaningful. Beyond $k = 3$ cohesion has decreased and clusters no longer represented distinct behavioural patterns which shows that three clusters provided the best balance between interpretability and accuracy.

In the forecasting phase the BiLSTM with Cross-Attention model clearly outperformed the baseline LSTM, GRU and standard BiLSTM architectures. It has produced the lowest error

scores (MSE ≈ 0.49, RMSE ≈ 0.70, MAE ≈ 0.25) which has given that turnover values in the dataset ranged from very small wagers to amounts exceeding 1.3 million. While other models has captured general behavioural direction that they tended to smooth sudden volatility. The attention mechanism has allowed the Cross-Attention model to focus on high-impact type of fluctuations like irregular spikes, abrupt bet increases and volatility shifts which results in a more precise representation of dynamic gambling behaviour.

Behaviourally the forecasts have generated by the Cross-Attention model align with recognised indicators of escalating addiction. Predicted turnover spikes signal loss chasing, rising bet frequency reflects compulsive play and increasing volatility which corresponds to reduced self-control patterns consistent with DSM-5 gambling disorder criteria. The close alignment between predicted and actual curves shows the usefulness of the model as an early-warning tool. Combined with the initial clustering the system provides a practical and technically sound approach for detecting high-risk behaviour early and supporting responsible gambling interventions

7 Conclusion and Future Work

This study has shown that combining K-Means player segmentation with BiLSTM + Cross-Attention time-series modeling is a more valid and efficient method of being able to predict the early occurrence of the gambling addiction risk than the amount of prediction provided by single-models prediction techniques. The system will group the players first on the basis of similarity in behavior which means that the forecasting will be specific to high-risk individuals as opposed to the generalization of the whole population. This segmentation together with deep sequential learning allowed the model to learn even the subtle patterns of time like a growing turnover, frequency of betting, and changes in behavior induced by interventions. Consequently, the proposed model recorded smaller RMSE and MAE, less alerts (false-positives), and acceptable latency and throughput thus responding to the research question on the accuracy and operation performance improvements. There are, however, a few limitations to the approach. The data is historical and fails to consider the changing real time betting conditions (Mandadapu, 2024) implying that abrupt behavioral aberrations can be subject to adaptive retraining. The research further concentrated on a single high-risk player to be carefully forecasted and it might not be possible to generalize across varied behavioral groups. Furthermore, although LIME was interpretable, it is possible to improve the feature explanations to make clinical or platform-level decision-making more transparent. Long-term dependency modeling may also be enhanced by means of transformer-based architectures or hybrid graph-learning models. There would be better trust and regulatory acceptance by expanding the explainability layer to incorporate SHAP or counterfactual reasoning methods. Lastly, as an actual-time monitoring dashboard, deployment would be beneficial in proactive and personalized responsible gambling interventions.

References

Andersson, S., Carlbring, P., Lyon, K., Bermell, M., Lindner, P., 2025. Insights into the temporal dynamics of identifying problem gambling on an online casino: A machine learning study on routinely collected individual account data. J. Behav. Addict. 14, 490–500. https://doi.org/10.1556/2006.2025.00013

Arsenault, P.-D., Wang, S., Patenaude, J.-M., 2025. A Survey of Explainable Artificial Intelligence (XAI) in Financial Time Series Forecasting. ACM Comput. Surv. 57, 1–37. https://doi.org/10.1145/3729531

Auer, M., Griffiths, M.D., 2022. An Empirical Attempt to Operationalize Chasing Losses in Gambling Utilizing Account-Based Player Tracking Data. J. Gambl. Stud. 39, 1547–1561. https://doi.org/10.1007/s10899-022-10144-4

Baker, S., Balthrop, J., Johnson, M., Kotter, J., Pisciotta, K., 2024. Gambling Away Stability: Sports Betting's Impact on Vulnerable Households (No. w33108). National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w33108

Bhatt, V., Aggarwal, U., Kumar, C.N.S.V., 2022. Sports Data Visualization and Betting, in: 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). Presented at the 2022 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), IEEE, Bangalore, India, pp. 1–6. https://doi.org/10.1109/SMARTGENCON56628.2022.10083831

Bolikulov, F., Nasimov, R., Rashidov, A., Akhmedov, F., Cho, Y.-I., 2024. Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms. Mathematics 12, 2553. https://doi.org/10.3390/math12162553

Cardoso-Marinho, B., Barbosa, A., Bolling, C., Marques, J.P., Figueiredo, P., Brito, J., 2022. The perception of injury risk and prevention among football players: A systematic review. Front. Sports Act. Living 4, 1018752. https://doi.org/10.3389/fspor.2022.1018752

Chatzimparmpas, A., Martins, R.M., Kucher, K., Kerren, A., 2022. FeatureEnVi: Visual Analytics for Feature Engineering Using Stepwise Selection and Semi-Automatic Extraction Approaches. IEEE Trans. Vis. Comput. Graph. 28, 1773–1791. https://doi.org/10.1109/TVCG.2022.3141040

Cui, Z., Wu, J., Ding, Z., Duan, Q., Lian, W., Yang, Y., Cao, T., 2021. A hybrid rolling grey framework for short time series modelling. Neural Comput. Appl. 33, 11339–11353. https://doi.org/10.1007/s00521-020-05658-0

Edson, T.C., Louderback, E.R., Tom, M.A., McCullock, S.P., LaPlante, D.A., 2024. Exploring a multidimensional concept of loss chasing using online sports betting records. Int. Gambl. Stud. 24, 306–324. https://doi.org/10.1080/14459795.2023.2276741

Efthymiou, I.P., Sidiropoulos, S., Diareme, K.C., Efthymiou-Egleton, T.W., 2023. Transforming Gambling Harm Reduction in Youth: Leveraging AI Language Models for Personalized Intervention and Prevention. J. Polit. Ethics New Technol. AI 2, e35821. https://doi.org/10.12681/jpentai.35821

Ghaharian, K., Abarbanel, B., Phung, D., Puranik, P., Kraus, S., Feldman, A., Bernhard, B., 2023. Applications of data science for responsible gambling: a scoping review. Int. Gambl. Stud. 23, 289–312. https://doi.org/10.1080/14459795.2022.2135753

Hales, C.A., Clark, L., Winstanley, C.A., 2023. Computational approaches to modeling gambling behaviour: Opportunities for understanding disordered gambling. Neurosci. Biobehav. Rev. 147, 105083. https://doi.org/10.1016/j.neubiorev.2023.105083

Hing, N., Smith, M., Rockloff, M., Thorne, H., Russell, A.M.T., Dowling, N.A., Breen, H., 2022. How structural changes in online gambling are shaping the contemporary experiences and behaviours of online gamblers: an interview study. BMC Public Health 22, 1620. https://doi.org/10.1186/s12889-022-14019-6

Kaimara, P., Oikonomou, A., Deliyannis, I., 2022. Could virtual reality applications pose real risks to children and adolescents? A systematic review of ethical issues and concerns. Virtual Real. 26, 697–735. https://doi.org/10.1007/s10055-021-00563-w

Kamdan, K., Anugrah, M.P., Almutaali, M.J., Ramdani, R., Kharisma, I.L., 2025. Performance Analysis of IndoBERT for Detection of Online Gambling Promotion in YouTube Comments, in: The 7th International Global Conference Series on ICT Integration in Technical Education & Smart Society. Presented at the International Global Conference Series on ICT Integration in Technical Education & Smart Society, MDPI, p. 66. https://doi.org/10.3390/engproc2025107066

Kosaraju, N., Sankepally, S.R., Mallikharjuna Rao, K., 2023. Categorical Data: Need, Encoding, Selection of Encoding Method and Its Emergence in Machine Learning Models—A Practical Review Study on Heart Disease Prediction Dataset Using Pearson Correlation, in: Saraswat, M., Chowdhury, C., Kumar Mandal, C., Gandomi, A.H. (Eds.), Proceedings of International Conference on Data Science and Applications, Lecture Notes in Networks and Systems. Springer Nature Singapore, Singapore, pp. 369–382. https://doi.org/10.1007/978-981-19-6631-6_26

Lajcinová, B., Gall, M., Pitonák, M., 2023. Anomaly Detection in Time Series Data: Gambling prevention using Deep Learning.

Lind, K., Marionneau, V., Järvinen-Tassopoulos, J., Salonen, A.H., 2021. Socio-Demographics, Gambling Participation, Gambling Settings, and Addictive Behaviors Associated with Gambling Modes: A Population-Based Study. J. Gambl. Stud. 38, 1111–1126. https://doi.org/10.1007/s10899-021-10074-7

Mandadapu, P., 2024. The Evolution of Football Betting- A Machine Learning Approach to Match Outcome Forecasting and Bookmaker Odds Estimation. https://doi.org/10.48550/ARXIV.2403.16282

Marisa, F., Syed Ahmad, S.S., Kausar, N., Kousar, S., Pamucar, D., Al Din Ide, N., 2022. Intelligent Gamification Mechanics Using Fuzzy-AHP and K-Means to Provide Matched Partner Reference. Discrete Dyn. Nat. Soc. 2022, 8292991. https://doi.org/10.1155/2022/8292991

Master of science in Information Technology Management, Cumberland University, Tennessee, USA, Raj, M.W.Z., 2025. THE ROLE OF DATA SCIENCE IN OPTIMIZING PROJECT EFFICIENCY AND INNOVATION IN U.S. ENTERPRISES. Int. J. Bus. Econ. Insights 05, 586–600. https://doi.org/10.63125/jzjkqm27

Nugroho, H., Utama, N.P., Surendro, K., 2023. Smoothing target encoding and class center-based firefly algorithm for handling missing values in categorical variable. J. Big Data 10, 10. https://doi.org/10.1186/s40537-022-00679-z

Parfenova, A., Clausel, M., 2024. Risk prediction of pathological gambling on social media. https://doi.org/10.48550/ARXIV.2403.19358

Perišić, A., Pahor, M., 2023. Clustering mixed-type player behavior data for churn prediction in mobile games. Cent. Eur. J. Oper. Res. 31, 165–190. https://doi.org/10.1007/s10100-022-00802-8

Plotnikova, V., Dumas, M., Milani, F., 2021. Adapting the CRISP-DM Data Mining Process: A Case Study in the Financial Services Domain, in: Cherfi, S., Perini, A., Nurcan, S. (Eds.), Research Challenges in Information Science, Lecture Notes in Business

Information Processing. Springer International Publishing, Cham, pp. 55–71. https://doi.org/10.1007/978-3-030-75018-3_4
Pramod, C.P., Pillai, G.N., 2021. K-Means clustering based Extreme Learning ANFIS with improved interpretability for regression problems. Knowl.-Based Syst. 215, 106750. https://doi.org/10.1016/j.knosys.2021.106750
Punhani, A., Faujdar, N., Mishra, K.K., Subramanian, M., 2022. Binning-Based Silhouette Approach to Find the Optimal Cluster Using K-Means. IEEE Access 10, 115025–115032. https://doi.org/10.1109/ACCESS.2022.3215568
Puranik, P., Taghva, K., Ghaharian, K., 2023. Descriptive Analysis of Gambling Data for Data Mining of Behavioral Patterns, in: Daimi, K., Al Sadoon, A. (Eds.), Proceedings of the Second International Conference on Innovations in Computing Research (ICR'23), Lecture Notes in Networks and Systems. Springer Nature Switzerland, Cham, pp. 40–51. https://doi.org/10.1007/978-3-031-35308-6_4
Sammouda, R., El-Zaart, A., 2021. An Optimized Approach for Prostate Image Segmentation Using K-Means Clustering Algorithm with Elbow Method. Comput. Intell. Neurosci. 2021, 4553832. https://doi.org/10.1155/2021/4553832
Sándor, M.Cs., Bakó, B., 2024. Unmasking Risky Habits: Identifying and Predicting Problem Gamblers Through Machine Learning Techniques. J. Gambl. Stud. 40, 1367–1377. https://doi.org/10.1007/s10899-024-10297-4
Septiani, R., Lubis, M.R., Firzada, F., 2025. Optimization of the K-Means Method and Davies-Bouldin Index (DBI) Technique in Mapping Spotify's Most Popular Songs Based on Mood. J. Comput. Netw. Archit. High Perform. Comput. 7, 1057–1065. https://doi.org/10.47709/cnahpc.v7i3.6655
Stechschulte, G., Wintner, M., Hemmje, M., Schwarz, J., Lischer, S., Kaufmann, M., 2024. In-Database Feature Extraction to Improve Early Detection of Problematic Online Gambling Behavior. IEEE Trans. Comput. Soc. Syst. 11, 6868–6881. https://doi.org/10.1109/TCSS.2024.3406501
Tariq, M.U., 2025. Leveraging Data Analytics for Predictive Consumer Behavior Modelling:, in: Miguélez-Juan, B., Rebollo-Bueno, S. (Eds.), Advances in Marketing, Customer Relationship Management, and E-Services. IGI Global, pp. 207–224. https://doi.org/10.4018/979-8-3693-3799-8.ch011
Testas, A., 2023. Support Vector Machine Classification with Pandas, Scikit-Learn, and PySpark, in: Distributed Machine Learning with PySpark. Apress, Berkeley, CA, pp. 259–280. https://doi.org/10.1007/978-1-4842-9751-3_10
Vatani, F., Dorrigiv, M., Emami, S., 2023. A Comprehensive Review of LSTM-Based Churn Prediction Models in the Gaming Industry. Model. Simul. Electr. Electron. Eng. 3. https://doi.org/10.22075/mseee.2024.32429.1135
Xiong, W., 2024. A Time-Series Model of Gated Recurrent Units Based on Attention Mechanism for Short-Term Load Forecasting. IEEE Access 12, 113918–113927. https://doi.org/10.1109/ACCESS.2024.3443876
Zhang, L., 2022. Behaviour Detection and Recognition of College Basketball Players Based on Multimodal Sequence Matching and Deep Neural Networks. Comput. Intell. Neurosci. 2022, 1–11. https://doi.org/10.1155/2022/7599685