

An mRMR-Based Feature Optimization Framework for Gaming Addiction Classification

Rizki Surya Permana*, Arif Rahman Hakim, Seshariana Rahma Melati, Muhammad Farhan Maulana

Bachelor of Telecommunications Engineering, School of Electrical Engineering, Telkom University, Jakarta 11710, INDONESIA

Article Info

Abstract

Article history:

Received
04-08-2026

Revised
01-09-2026

Accepted
12-09-2026

Available Online
19-10-2026

Keywords:

Feature Selection,
Gaming Addiction,
Machine Learning,
mRMR, Support Vector
Machine

Gaming addiction classification is challenging because problematic gaming behavior is influenced by multiple behavioral, psychological, lifestyle, and contextual factors, while redundant or weakly informative features may reduce the effectiveness of machine learning models. To address this issue, this study investigates minimum Redundancy Maximum Relevance (mRMR) feature selection to obtain a more compact and discriminative feature representation for gaming addiction classification. Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN) were evaluated using two configurations: a baseline using the complete predictor set and an mRMR-based configuration using the 15 highest-ranked features. All preprocessing, feature selection, hyperparameter tuning, and cross-validation were performed on the training data, while the independent test set was used only for final evaluation. The results show that the effect of mRMR varies across classifiers, indicating that feature reduction does not universally improve classification performance. The strongest improvement was obtained by SVM, whose accuracy increased from 0.940 to 0.960 and F1-score from 0.824 to 0.857 after applying mRMR. The mRMR-SVM model also achieved a precision of 1.000 and a recall of 0.750. In contrast, the baseline XGBoost model achieved a higher recall of 0.914, indicating a trade-off between false-positive reduction and positive-case detection. These findings demonstrate that mRMR can improve classification when the selected feature space is well aligned with the learning mechanism of the classifier, particularly for margin-based SVM. The selected features further indicate that gaming addiction classification is associated not only with gaming intensity but also with psychosocial and lifestyle characteristics. Therefore, the proposed mRMR-SVM configuration provides a compact and effective representation for gaming addiction classification, although validation on larger and more diverse datasets is required before broader generalization.



© 2026 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY 4.0) license <https://creativecommons.org/licenses/by/4.0/>.
Published by Universitas Katolik Indonesia Atma Jaya.

1. INTRODUCTION

The rapid development of digital technology has made gaming an important part of modern society. Digital games provide entertainment, social interaction, competition, skill development, and economic opportunities. However, excessive and poorly controlled gaming behavior may develop into problematic gaming and interfere with social relationships, academic achievement, occupational performance, physical health, and psychological well-being.

The World Health Organization (WHO) recognizes gaming disorder in the International Classification of Diseases, 11th Revision (ICD-11), as a pattern of impaired control over gaming, increasing priority given to gaming over other activities, and continuation or escalation despite negative consequences [1],[2]. The inclusion of gaming disorder in ICD-11 emphasizes clinically significant functional impairment rather than gaming time alone [3]-[5]. Consequently, gaming-addiction risk should be examined using multiple behavioral, psychological, social, and lifestyle-related indicators.

Problematic gaming has been associated with psychological distress, poor sleep, anxiety, depression, loneliness, reduced self-control, lower physical activity, and impaired academic or daily functioning [6]-

*Corresponding author : Rizki Surya Permana
Email address : rizkisuryapermana@telkomuniversity.ac.id

[12]. Because these factors may interact in complex and nonlinear ways, machine learning offers a data-driven approach for classifying gaming-addiction risk. Previous studies have evaluated Random Forest, Support Vector Machine, Extreme Gradient Boosting, K-Nearest Neighbors, and related algorithms using questionnaire, behavioral, and longitudinal data, with performance varying across datasets and validation designs [13]-[18].

A major challenge in multidimensional gaming-addiction data is the presence of irrelevant and redundant features. Variables such as gaming duration, screen time, late-night sessions, sleep patterns, psychological conditions, and productivity may provide overlapping information. Including all variables may increase model complexity, reduce interpretability, and negatively affect classification performance, particularly when the number of samples is limited.

The minimum Redundancy Maximum Relevance (mRMR) method provides a suitable feature-optimization strategy because it selects variables that are strongly related to the target while minimizing redundancy among selected features [19],[20]. Unlike feature-extraction methods that transform variables into latent components, mRMR retains the original meaning of behavioral and psychosocial indicators. This property is important because gaming-addiction research requires both accurate classification and interpretable identification of relevant contributing factors.

Although prior work has compared multiple machine-learning algorithms for behavioral addiction [13]-[18] and applied feature-selection methods to neurophysiological, radiomic, or multiclass internet-addiction data [21]-[26], controlled evidence on mRMR for survey-based gaming-addiction classification remains limited. Recent feature-reduction studies also indicate that the benefit of selection depends on the dataset, selected subset, and learning algorithm [27]-[29]. A controlled comparison is therefore required to determine whether mRMR consistently improves performance across different classifiers.

Accordingly, this study proposes an mRMR-based feature optimization framework for gaming-addiction classification. Random Forest, Support Vector Machine, Extreme Gradient Boosting, and K-Nearest Neighbors are evaluated under two scenarios: classification using the original valid feature set and classification using an optimized feature subset selected by mRMR. All models are compared using consistent preprocessing, validation, hyperparameter-optimization, and evaluation framework.

The main contribution of this study is the development of an interpretable feature-optimization framework that evaluates the effect of mRMR across four distinct machine-learning paradigms. The study also identifies a compact set of relevant and complementary behavioral and psychosocial features and determines the most effective combination of feature subset and classifier. Model performance is assessed using a confusion matrix, accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve. The resulting model is intended to support data-driven behavioral analysis and preliminary risk classification rather than replace professional or clinical assessment.

2. LITERATURE REVIEW

2.1. Gaming Addiction and Associated Factors

The rapid development of digital technology has made gaming an important part of modern society, providing entertainment, social interaction, competition, skill development, and economic opportunities. However, excessive and poorly controlled gaming behavior may develop into problematic gaming, interfering with social relationships, academic achievement, occupational performance, physical health, and psychological well-being.

The World Health Organization (WHO) recognizes gaming disorder in the International Classification of Diseases, 11th Revision (ICD-11), as a pattern characterized by impaired control over gaming, increasing priority given to gaming over other activities, and continued gaming despite negative consequences [1],[2]. The rationale and broader diagnostic implications of this inclusion have also been discussed in clinical and ICD-11 literature [3]-[5]. Gaming duration alone does not necessarily indicate addiction because functional impairment and loss of behavioral control are more decisive indicators [1],[3]. Gaming addiction should therefore be examined through multiple behavioral, psychological, social, and lifestyle-related factors rather than through duration alone.

Previous studies have associated problematic gaming with poor sleep quality, stress, anxiety, depression, loneliness, reduced self-control, limited social interaction, and decreased productivity [6]-[12]. Wong et al. reported that Internet Gaming Disorder (IGD) was associated with psychological distress and poor sleep, while self-compassion and physical activity were protective factors [7]. A systematic review and meta-analysis confirmed associations between problematic gaming and sleep disturbance and related psychosocial difficulties [10]. Among university students, loneliness and hopelessness were associated with IGD through the mediating role of self-control [11], whereas personal distress mediated relationships involving self-esteem, self-efficacy, loneliness, and problematic video gaming [12]. These factors may interact in complex and nonlinear ways, making conventional univariate analysis insufficient to capture the underlying patterns. This multidimensional and overlapping

characteristic renders gaming-addiction data high-dimensional when represented numerically for automated classification, thereby motivating the use of machine-learning approaches supported by an appropriate feature-optimization strategy.

2.2. Machine Learning for Gaming-Addiction Classification

Machine learning (ML) offers a data-driven approach for identifying complex patterns and classifying individuals according to gaming-addiction risk. Islam et al. compared nine classifiers using questionnaire-based online gaming-addiction data in Bangladesh [13]. Hasan et al. used machine learning to estimate mental-health effects associated with mobile gaming among urban children [14], while Güllü et al. examined digital-game addiction in relation to obesity and physical activity [15]. Gan et al. combined risk-factor analysis with six ML models for adolescent internet behavioral addiction [16], and Liu et al. developed an explainable longitudinal prediction framework for internet addiction [17]. Kornev et al. also proposed a machine-learning platform for predicting gaming behavior [18]. These studies demonstrate substantial variation in model performance across data sources and experimental designs.

Studies using neuroimaging and physiological signals further support the feasibility of ML-based IGD classification. Batbat et al. combined ERP and fNIRS features and achieved up to 88.17% accuracy using Random Forest after ReliefF feature selection [21]. Wang et al. evaluated radiomics-based feature selection and multiple classifiers using subcortical MRI data [22], while Yeh et al. applied mRMR to 564 EEG features to estimate IGD risk associated with social anxiety [23]. Multimodal EEG and neuropsychological features have also been used to differentiate IGD and alcohol-use-disorder characteristics [24], and MRI radiomics has been used to distinguish IGD participants from healthy controls [25]. The diversity of algorithms and datasets used across these studies confirms that gaming-addiction classification performance depends heavily on the quality of the feature representation used, making an appropriate feature-optimization strategy an important component for improving both accuracy and efficiency of classification models in this domain regardless of the classification algorithm chosen.

2.3. Machine-Learning Classification Algorithms

This study focuses on four classifiers that represent different learning paradigms and have been repeatedly evaluated in gaming- and internet-addiction research [13]-[18],[26]. Random Forest and XGBoost are tree-ensemble methods, SVM is a margin-based classifier, and KNN is an instance-based method. Their contrasting learning mechanisms enable a broader assessment of whether the same feature-optimization strategy produces consistent effects across classifiers. SVM has also been applied to multimodal neurophysiological and neuropsychological data in related IGD studies [24]. These four algorithms represent distinct learning paradigms ensemble tree-based (RF, XGBoost), margin-based (SVM), and instance-based (KNN) allowing a more comprehensive evaluation of whether feature optimization consistently affects performance across different types of classification models, rather than being limited to a single learning paradigm.

2.4. Feature Optimization in Machine Learning

A major challenge in multidimensional gaming-addiction data is the presence of irrelevant and redundant features. Variables such as gaming duration, screen time, late-night gaming sessions, sleep patterns, psychological conditions, and productivity often provide overlapping information. Including all such variables without selection may increase model complexity, reduce interpretability, and negatively affect classification performance, particularly when the number of samples is limited.

Feature optimization is an important preprocessing step for reducing irrelevant and redundant predictors. Feature-selection methods may be designed to identify all variables associated with a target, as in Boruta-based analysis [26], or to identify a compact minimal-optimal subset. mRMR belongs to the latter category and selects features using relevance and redundancy criteria [19],[20]. Unlike feature-extraction methods such as Principal Component Analysis, which transform original variables into latent components, mRMR retains the original meaning of behavioral and psychosocial indicators. This interpretability is important when the selected factors must remain substantively understandable.

Other feature-selection approaches used in related domains include ReliefF [21], Recursive Feature Elimination, ANOVA, Kruskal-Wallis testing, and model-based importance measures [22],[25]. Fast-mRMR was developed to improve scalability for high-dimensional data [27], while mRMR has also been implemented in large-scale industrial machine-learning platforms [28]. More recent work has integrated Boruta and mRMR to reduce feature count and training time while maintaining or improving classification performance [29].

2.5. Minimum Redundancy Maximum Relevance

Minimum Redundancy Maximum Relevance (mRMR) is a mutual-information-based feature-selection method introduced by Peng, Long, and Ding [19], with an early formulation also presented by Ding and Peng [20]. Its core principle is to select a feature subset that maximizes relevance to the target while minimizing redundancy among the selected features. Relevance is measured using the average mutual information between each feature and the target, whereas redundancy is measured using the average mutual information between pairs of selected features [19]. This formulation is expressed as:

$$\max_S \left[\left(\frac{1}{|S|} \right) \sum \text{MI}(f_i; c) - \left(\frac{1}{|S|^2} \right) \sum \sum \text{MI}(f_i; f_j) \right] \quad (1)$$

where S denotes the selected feature subset, f_i is an individual feature, c is the class variable, and MI denotes mutual information [19].

mRMR operates iteratively. The first feature is selected according to its relevance to the target. In subsequent iterations, each candidate is evaluated using its relevance to the target and its redundancy with the features already selected, and the highest scoring candidate is added to the subset [19]. The process continues until the required number of features is obtained. This minimal optimal characteristic is useful when repeated feature selection must be performed efficiently while retaining the original meaning of the selected variables.

mRMR has been adopted in domains ranging from microarray gene-expression analysis [20] to industrial machine-learning platforms [28]. In a domain closely related to gaming addiction, mRMR was applied to 564 EEG features to estimate IGD risk associated with social anxiety [23]. Further developments include Fast-mRMR for high-dimensional data [27] and the Boruta-mRMR combination, which integrates all relevant and minimal-optimal selection strategies [29].

2.6. Previous Studies

Previous studies have demonstrated that the performance of machine-learning algorithms for gaming-addiction classification varies considerably depending on the characteristics of the dataset, feature representation, and classification strategy. A comparative study involving KNN, Logistic Regression (LR), SVM, Gaussian Naive Bayes (GNB), Random Forest (RF), AdaBoost, Decision Tree (DT), XGBoost, and Gradient Boosting evaluated nine classifiers without applying feature optimization [22]. Because all available features were directly used for classification, the study primarily focused on differences in classifier performance rather than on the effect of reducing irrelevant or redundant information.

The influence of model characteristics can also be observed in subsequent studies. A comparison of DT, LR, SVM, and RF showed that RF achieved the highest accuracy of 90.87% [23], suggesting that ensemble-based learning may provide advantages when the classification problem involves nonlinear relationships among multiple predictors. Similarly, another study compared MLP, RF, KNN, SVM, LR, and XGBoost while incorporating regression-based feature selection [24]. The introduction of feature selection in this study indicates that predictive performance may depend not only on the choice of classifier but also on the representation of the input features.

Different results have also been reported across other gaming-related classification tasks. A comparison involving KNN, SVM, Multinomial Naive Bayes, and RF identified SVM as the best-performing classifier with an accuracy of 92.75% [22]. This result suggests that when the selected feature space provides sufficient class separation, margin-based classification may outperform distance-based and ensemble approaches. In another study, DT, RF, SVM, and LR were evaluated using a multi-class framework consisting of four addiction-severity levels [24]. Because the target was divided into multiple severity categories rather than a binary outcome, model evaluation relied on multi-class ROC analysis, making the classification problem more complex than conventional binary addiction detection.

The use of multimodal and neuroimaging data has further demonstrated that classification performance is strongly affected by the type of input information. A Random Forest-based approach using 80 multimodal EEG-fNIRS features achieved an accuracy of 88.17% [29], indicating that combining physiological modalities can provide discriminative information for gaming-disorder classification. However, when structural MRI features were used, RF achieved an accuracy of only 0.73 ± 0.08 . This difference suggests that predictive performance cannot be attributed solely to the classification algorithm because the quality and representation of the input features substantially influence the resulting model.

Similar observations were reported in MRI radiomics-based research, where ten classifiers and a clinical nomogram were evaluated. The best performance was obtained using an ANOVA-based feature-selection procedure combined with Logistic Regression on features extracted from the caudate nucleus

and amygdala regions. This finding further indicates that appropriate feature selection can substantially affect classification performance by focusing the model on the most informative variables.

Feature optimization has also been investigated more explicitly in gaming-disorder research. An automated machine-learning approach using MATLAB 'fitcauto' applied mRMR directly to EEG-based Internet Gaming Disorder data. Although this study demonstrated the potential of mRMR for reducing feature dimensionality, its application was based on physiological EEG features rather than multidimensional behavioral and psychosocial data. In another study, a hybrid Boruta-mRMR approach was combined with SVM, RF, XGBoost, and GBM [28]. The feature-selection process reduced the number of variables while maintaining or improving classification accuracy, indicating that eliminating redundant features may improve computational efficiency without necessarily sacrificing predictive performance.

Collectively, these studies show that classification performance is influenced by both the learning algorithm and the feature representation. Studies without feature optimization mainly demonstrate differences among classifiers, whereas studies incorporating feature-selection methods indicate that reducing irrelevant or redundant information can alter model performance substantially. However, existing studies have predominantly applied feature optimization to physiological, neuroimaging, or specialized datasets, while systematic evaluation of mRMR on multidimensional gaming-addiction data remains limited. Therefore, the present study compares Random Forest, SVM, XGBoost, and KNN under two configurations: a baseline using the complete feature set and an mRMR-based configuration using 15 selected features. This comparison is intended to determine whether feature reduction improves classification performance and whether its effect differs according to the learning mechanism of each classifier.

2.7. Research Gap and Study Positioning

Based on the literature reviewed above, the following research gaps can be identified. First, although several studies have compared multiple machine-learning algorithms for gaming and behavioral-addiction classification [13]-[18],[26], the effect of mRMR across ensemble tree-based, margin-based, and instance-based learning paradigms has not been sufficiently investigated using the same survey-based dataset. Second, model comparisons in previous studies differ in preprocessing, validation procedures, class distributions, feature-selection methods, and evaluation criteria [13]-[18],[21]-[26]. These differences make it difficult to determine whether an observed improvement results from the classifier, feature optimization, or the experimental design. A controlled comparison using consistent preprocessing, validation, hyperparameter optimization, and evaluation criteria is therefore needed. Third, the application of mRMR in the gaming-disorder context has mainly involved high dimensional EEG data [23]. Its use with moderately dimensional survey and behavioral indicators, including gaming duration, sleep quality, psychological distress, loneliness, self-control, physical activity, and social functioning [6]-[12], remains underexplored despite the lower cost and greater accessibility of such data. Fourth, mRMR retains the original meaning of selected variables while reducing redundancy [19],[20]. This interpretive advantage has not been widely leveraged in questionnaire-based gaming-addiction classification, where identifying meaningful behavioral and psychosocial factors is important in addition to maximizing predictive metrics.

Based on these gaps, this study is positioned to address the identified shortcomings by proposing an mRMR-based feature optimization framework for gaming-addiction classification. Four classification algorithms representing distinct learning paradigms Random Forest, Support Vector Machine, Extreme Gradient Boosting, and K-Nearest Neighbors are evaluated under two scenarios: (1) classification using the original valid feature set, and (2) classification using an optimized feature subset selected by mRMR. All models are compared using a consistent preprocessing, validation, hyperparameter-optimization, and evaluation framework (confusion matrix, accuracy, precision, recall, F1-score, and area under the ROC curve), allowing a controlled determination of whether and to what extent mRMR consistently improves classification performance across all tested learning paradigms, while also identifying the most relevant and complementary set of behavioral and psychosocial features for gaming-addiction classification.

3. METHODOLOGY

The experimental procedure was designed to ensure a fair and unbiased comparison among the classification models. The workflow begins with dataset preparation, preprocessing, and dataset splitting into training and test sets. All mRMR feature selection, cross-validation, and hyperparameter tuning were performed exclusively on the training set. The best configuration of each model was selected based on cross-validation performance, while the test set remained unseen during model development. The

independent test set was then used only for the final evaluation of Random Forest, SVM, XGBoost, and KNN. The overall process is illustrated in Figure 1.

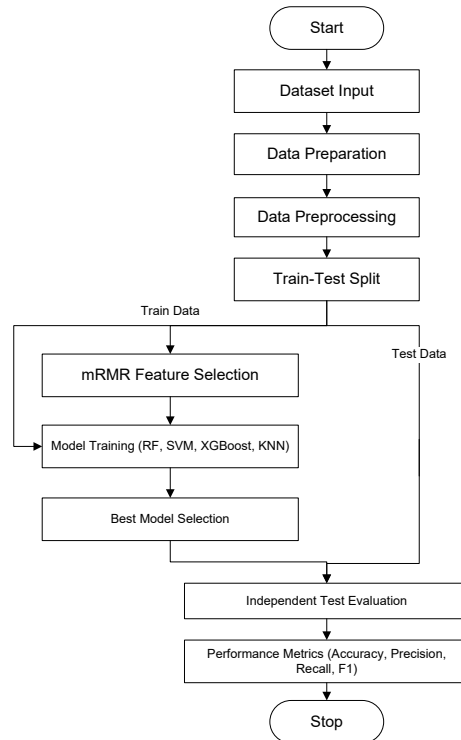


Figure 1. Flowchart of the Experimental Procedure, Including Data Preprocessing, mRMR Feature Selection, and Model Evaluation.

3.1. Research Design and Dataset

The dataset used in this study was obtained from kaggle. It contains 250 observations representing demographic, gaming behavior, psychological, lifestyle, and productivity related characteristics. The data were originally collected through survey. Further details regarding the dataset source and collection procedure are provided to ensure transparency and reproducibility. This study focuses on binary gaming-addiction classification. The target variable consists of two classes: non-addiction (0) and addiction (1), determined using an *addiction_score* threshold of 50. Although the dataset also provides severity information, the severity categories were not used as prediction targets because the primary objective of this study is to evaluate the capability of machine-learning algorithms to distinguish individuals with and without indications of gaming addiction.

The dataset consists of 250 observations containing behavioral, psychological, lifestyle, and productivity-related characteristics. The classification target, *addiction_binary*, is derived from *addiction_score*, where scores below 50 are assigned to the non-addiction class and scores of 50 or higher to the addiction class. The *addiction_severity* variable represents categorical levels derived from the same score. Therefore, both *addiction_score* and *addiction_severity* were excluded from the predictor variables to prevent target leakage.

This study employs a quantitative experimental design to develop an mRMR-based feature optimization framework for gaming addiction classification. Four supervised machine learning algorithms Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN) are evaluated under two scenarios. The first scenario uses the complete set of valid predictors as baseline models, while the second applies minimum Redundancy Maximum Relevance (mRMR) to select an optimized feature subset before classification.

3.2. Data Preprocessing and Validation

Missing numerical values are replaced using median imputation, while missing categorical values are replaced using the most frequent category or an Unknown category. Nominal categorical variables are transformed using one-hot encoding, whereas numerical features are standardized using:

$$\left[z_i = \frac{x_i - \mu}{\sigma} \right] \quad (2)$$

where (x_i) is the original value, (μ) is the training-set mean, and (σ) is the training set standard deviation.

The dataset is divided into development and independent test sets using a stratified 80:20 ratio. Stratified five-fold cross-validation is applied to the development set for feature selection and hyperparameter optimization. All preprocessing, feature selection, and model-training procedures are performed within each training fold to prevent data leakage. The same partitions and random seed are used for all model configurations. Because the addiction class is smaller than the non-addiction class, class-sensitive evaluation is prioritized. Class weighting is applied where supported. Any oversampling technique, when evaluated, is performed only on the training portion of each cross-validation fold.

3.3. mRMR-Based Feature Optimization

The mRMR method selects features by maximizing their relevance to the target while minimizing redundancy among the selected variables. Feature relevance is expressed as:

$$\left[D(S, c) = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \right] \quad (3)$$

while feature redundancy is defined as:

$$\left[R(S) = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i; x_j) \right] \quad (4)$$

where $(I(x_i; c))$ represents the mutual information between feature (x_i) and target class (c) , and $(I(x_i; x_j))$ represents the mutual information between two selected features. The mRMR objective is therefore:

$$[\max \Phi(S) = D(S, c) - R(S)] \quad (5)$$

Several subset sizes, such as the top 5, 10, 15, 20, and 25 features, are evaluated. The optimal number of features is selected according to mean cross-validation performance. Feature-selection frequency across folds is also recorded to identify stable predictors.

3.4. Classification and Experimental Scenarios

RF, SVM, XGBoost, and KNN are selected because they represent ensemble-based, margin-based, boosting-based, and instance-based learning paradigms. Their principal hyperparameters are optimized using grid search or randomized search within stratified cross-validation. Two experimental scenarios are evaluated, including:

- Baseline scenario: RF, SVM, XGBoost, and KNN are trained using all valid predictors.
- mRMR scenario: the same four classifiers are trained using the optimized mRMR feature subset.

This design produces eight model configurations Original features–RF, Original features–SVM, Original features–XGBoost, Original features–KNN, mRMR–RF, mRMR–SVM, mRMR–XGBoost, and mRMR–KNN. Each baseline classifier is directly compared with its mRMR-based counterpart under identical experimental conditions.

3.5. Performance Evaluation and Model Selection

Model performance is evaluated using a confusion matrix, accuracy, precision, recall, F1-score, specificity, balanced accuracy, ROC-AUC, and precision–recall AUC. The principal metrics are calculated as:

$$\left[Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \right] \quad (6)$$

$$\left[Precision = \frac{TP}{TP+FP} \right] \quad (7)$$

$$\left[Recall = \frac{TP}{TP+FN} \right] \quad (8)$$

$$\left[F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \right] \quad (9)$$

Accuracy is not used as the sole selection criterion because the dataset is imbalanced. Greater emphasis is placed on recall, F1-score, balanced accuracy, ROC-AUC, and PR-AUC to assess the ability of each model to identify the addiction class. The final proposed model is selected from the mRMR-based configurations by considering independent test performance, recall and F1-score for the addiction class, cross-validation stability, number of selected features, and absence of target leakage. The selected model is intended to support preliminary gaming-addiction risk classification rather than replace professional or clinical assessment.

4. RESULT AND DISCUSSION

4.1. Experimental Results

The experiments evaluated four machine-learning algorithms Support Vector Machine, Random Forest, K-Nearest Neighbors, and XGBoost under two scenarios. The baseline scenario used all valid predictors after removing identifiers and variables associated with target leakage, whereas the second scenario used 15 features selected through the minimum Redundancy Maximum Relevance method. Both scenarios were evaluated using the same training testing partition, preprocessing procedure, and evaluation metrics. Table 1 presents the classification performance of all evaluated configurations.

Table 1. Performance comparison of baseline and mRMR-based models

Scenario	Model	Accuracy	Precision	Recall	F1_Score	Roc_Auc	Pr_Auc
Baseline	SVM	0.940	0.778	0.875	0.824	0.991	0.959
mRMR	SVM	0.960	1.000	0.750	0.857	0.982	0.936
Baseline	Random Forest	0.860	0.556	0.625	0.588	0.929	0.676
mRMR	Random Forest	0.820	0.455	0.625	0.526	0.878	0.465
Baseline	KNN	0.880	0.667	0.500	0.571	0.939	0.767
mRMR	KNN	0.860	0.600	0.375	0.462	0.765	0.524
Baseline	XGBoost	0.920	0.667	0.914	0.800	0.961	0.802
mRMR	XGBoost	0.920	0.700	0.875	0.778	0.958	0.781

The results show that the influence of mRMR was different for each classifier. The most favorable result was obtained by the mRMR-SVM configuration, which achieved the highest accuracy of 0.960 and the highest F1-score of 0.857. The model also achieved a precision of 1.000, indicating that all observations predicted as gaming addiction were correctly classified. However, its recall was 0.750, showing that the model did not identify all positive cases. Comparison of model performance presented in Table 2.

Table 2. Comparison of Model Performance

Model	f1_score		Recall		pr_auc	
	baseline	mrmr	baseline	mrmr	baseline	Mrmr
Random Forest	0.5882	0.5263	0.625	0.625	0.6759	0.4651
SVM	0.8235	0.8571	0.875	0.750	0.9594	0.9363
XGBoost	0.8000	0.7778	0.914	0.875	0.8021	0.7810
KNN	0.5714	0.4615	0.500	0.375	0.7672	0.5240

Table 2 compares the classification performance obtained using the complete feature set and the 15 features selected by mRMR. This comparison is important for evaluating whether reducing irrelevant and redundant features improves model performance. The results also demonstrate that the effect of mRMR varies across classifiers, with the most significant improvement observed for SVM. The baseline SVM produced slightly lower accuracy and F1-score, at 0.940 and 0.824, respectively. Nevertheless, it achieved higher recall, ROC-AUC, and PR-AUC than mRMR-SVM. This indicates that the baseline SVM had stronger overall class-ranking capability and detected a larger proportion of positive observations, although it generated more false-positive predictions.

4.2. mRMR-Selected Features

The mRMR procedure selected 15 features from the preprocessed feature space. The selected features are presented in Table 3.

Table 3. Features selected by mRMR

No	Type	Feature
1	numeric	daily_playtime_hours
2	categorical	country_Other
3	categorical	gender_Male
4	numeric	late_night_sessions_hours
5	numeric	lootbox_openings
6	numeric	screen_time_total_hours
7	numeric	dopamine_dependency_index
8	categorical	relationship_status_Married
9	numeric	loneliness_score
10	categorical	rank_tier_Master
11	numeric	consecutive_hours_max
12	categorical	occupation_Employed
13	categorical	rank_tier_Diamond
14	numeric	exercise_frequency_per_week
15	categorical	platform_Console

The selected subset covers several dimensions of gaming-addiction behavior. The most prominent features were related to gaming intensity, including daily playtime, late-night gaming, total screen time, and maximum consecutive gaming hours. These variables indicate that the classification problem cannot be represented only by total playing duration, but also by the temporal pattern and continuity of gaming activity. The selection of *lootbox_openings* and *dopamine_dependency_index* suggests that reward seeking behavior and repeated engagement with in-game reinforcement mechanisms contribute useful information to the classification process. Meanwhile, *loneliness_score* and *exercise_frequency_per_week* indicate that psychosocial conditions and lifestyle balance also contribute to distinguishing the two classes.

Several one hot encoded categorical variables were also selected, including gender, marital status, occupation, gaming rank, country group, and gaming platform. These variables should be interpreted as dataset-specific predictive associations rather than causal determinants of gaming addiction. For example, the selection of the male category does not demonstrate that male users are inherently more likely to experience gaming addiction. It only indicates that this category contributed information to the classifier within the observed dataset.

The selected categories *rank_tier_Master* and *rank_tier_Diamond* originated from the same original variable. Their simultaneous selection indicates that different competitive rank categories provided distinct information after one-hot encoding. However, category frequency and feature stability should be examined before drawing substantive conclusions, particularly because the dataset contains a limited number of observations.

4.3. Selection of the Proposed Model

Based on the overall evaluation results presented in Table 2, mRMR-SVM was selected as the proposed model because it achieved the highest accuracy (0.96) and F1-score (0.857), while also obtaining perfect precision (1.00). These results indicate that mRMR-SVM provides a strong and relatively balanced classification performance, particularly in minimizing false-positive predictions. In addition, the model uses only 15 selected features, resulting in a more compact representation than the baseline configuration. This reduced feature space may improve computational efficiency and facilitate interpretation of the variables that contribute most strongly to the classification results.

However, the selection of the best performing model should also consider the intended application and the relative importance of different types of classification errors. Although mRMR-SVM achieved superior accuracy, precision, and F1-score, the baseline XGBoost model obtained a substantially higher recall of 0.914 compared with 0.750 for mRMR-SVM. This indicates that XGBoost is more effective at identifying actual gaming addiction cases, although its lower precision of 0.667 suggests a higher rate of false positive predictions. Therefore, if avoiding false positives is the primary objective, mRMR-SVM

is the more appropriate model. In contrast, if the main objective is to identify as many at risk individuals as possible and minimize false negatives, the baseline XGBoost model may be preferable.

This trade off is particularly important in preliminary screening applications, where false negative predictions may have greater consequences than false positives because an undetected at risk individual may not be referred for further assessment. Consequently, although mRMR-SVM was selected as the proposed model based on its superior overall performance and more compact feature representation, its lower recall should be explicitly acknowledged as a limitation. Before practical deployment, further optimization, such as decision threshold adjustment, class weighted learning, or cost sensitive optimization, can be explored to enhance sensitivity while preserving the strong precision and overall performance of the proposed model.

4.4. Discussion

The experimental results demonstrate that the effect of mRMR-based feature selection varies across classification algorithms. Among the evaluated configurations, the strongest improvement was obtained by SVM using the 15 selected features. Specifically, the accuracy increased from 0.940 to 0.960, corresponding to an improvement of 2.0 percentage points, while the F1-score increased from 0.824 to 0.857, representing an improvement of 3.3 percentage points. This result indicates that removing less relevant and potentially redundant features produced a more discriminative feature space with a clearer separation margin, enabling SVM to achieve better classification performance with a more compact feature representation.

However, mRMR did not universally improve the performance of all machine learning models. Its effectiveness depends on the relationship between the selected feature representation and the learning mechanism of each classifier. In contrast to SVM, tree-based classifiers such as Random Forest and XGBoost benefited from access to a larger feature space. These algorithms can capture conditional and nonlinear interactions that may not be fully represented by the relevance and redundancy criteria used in mRMR. Consequently, features considered redundant during feature selection may still provide complementary information within tree-based ensemble models.

A similar limitation was observed for mRMR-KNN. The reduction in performance indicates that feature selection alone does not necessarily produce a more discriminative distance space. Although the selected variables may have high individual relevance to the target, they can still generate overlapping local neighborhoods between classes. This limitation may become more pronounced when the dataset is relatively small and imbalanced.

Despite these model differences, the 15 selected features provide a compact representation of the multidimensional characteristics associated with gaming addiction. Gaming intensity variables constitute an important part of the selected subset, while psychosocial, behavioral, and lifestyle-related variables are also retained. This finding indicates that classification is not determined solely by gaming duration. Late-night gaming behavior, prolonged continuous sessions, reward-seeking activity, loneliness, exercise habits, and contextual characteristics collectively contribute to distinguishing the classes.

The interpretation of the selected features should be approached cautiously. Categorical variables such as `'gender_Male'`, `'relationship_status_Married'`, and `'rank_tier_Master'` were selected because they provided discriminative information within the observed dataset; however, their inclusion should not be interpreted as evidence of a causal relationship with gaming addiction. These variables may reflect sample-specific demographic or behavioral patterns and may also interact with other unobserved factors.

In addition, the relatively small dataset of 250 observations limits the robustness and generalizability of the findings. Feature rankings and classification performance may vary under different data partitions, particularly because the addiction class contains substantially fewer observations than the non-addiction class. Consequently, validation on larger, more diverse, and independently collected datasets is necessary before broader conclusions can be drawn.

The results should nevertheless be interpreted cautiously because the dataset contains only 250 observations, including 42 positive cases. With an 80:20 data split, the test set contains only approximately eight positive observations. Under these conditions, one or two classification errors can substantially affect precision, recall, and F1-score. For example, a precision of 1.000 may occur when the classifier predicts only a limited number of positive cases without producing false positive predictions. Furthermore, the relatively small sample size may affect the stability of feature rankings and limit the generalizability of the observed model performance to broader gaming populations.

Therefore, although the mRMR SVM configuration with 15 features demonstrated the most favorable overall performance and was adopted as the proposed configuration, its superiority should be interpreted in conjunction with cross validation results rather than relying solely on a single test-set evaluation. Cross validation means and standard deviations can provide a more robust assessment of model stability across different data partitions. In addition, repeated stratified cross-validation and feature selection stability

analysis may be considered in future work to further verify the robustness of the selected features and the generalizability of the model.

5. CONCLUSION

This study developed and evaluated an mRMR-based feature optimization framework for binary gaming-addiction classification using Support Vector Machine, Random Forest, K-Nearest Neighbors, and XGBoost. The models were evaluated under two configurations, which are a baseline using the complete predictor set and an mRMR-based configuration using the 15 highest-ranked features. The selected subset represented multiple dimensions of gaming behavior, including gaming intensity, late-night activity, prolonged gaming sessions, reward behavior, psychosocial characteristics, lifestyle factors, and contextual variables. These features were interpreted as predictive characteristics within the observed dataset rather than as causal determinants of gaming addiction.

The experimental results demonstrate that the effect of mRMR is classifier dependent. The strongest improvement was obtained by SVM, where accuracy increased from 0.940 to 0.960, corresponding to an improvement of 2.0 percentage points, while the F1-score increased from 0.824 to 0.857, representing an improvement of 3.3 percentage points. The mRMR–SVM model achieved an accuracy of 0.960, precision of 1.000, recall of 0.750, F1-score of 0.857, ROC-AUC of 0.982, and PR-AUC of 0.936. These results suggest that removing weakly informative and redundant features produced a more discriminative representation for the margin-based SVM classifier while reducing the dimensionality of the input space.

However, mRMR did not consistently improve the performance of Random Forest, KNN, and XGBoost. This finding confirms that feature reduction should not be assumed to benefit all learning algorithms equally. Tree-based models may continue to exploit conditional and nonlinear interactions among variables that appear redundant under relevance redundancy criteria, while distance-based classifiers may not necessarily obtain a more discriminative neighborhood structure after feature reduction.

Based on accuracy and F1-score, mRMR SVM was selected as the proposed configuration. Nevertheless, model selection should consider the intended application. The perfect precision obtained by mRMR–SVM indicates strong control of false positive predictions, whereas its lower recall shows that some positive cases remain undetected. In contrast, the baseline XGBoost model achieved a higher recall of 0.914 and may therefore be more suitable when maximizing the detection of potentially addicted individuals is the primary objective.

The findings should be interpreted cautiously because the dataset contains only 250 observations, including 42 positive cases, and the binary target is defined using a dataset specific addiction score threshold. The limited sample size and class imbalance may affect feature-selection stability, performance estimates, and the generalizability of the results. Therefore, the reported findings should be considered preliminary rather than definitive evidence of population level gaming addiction risk factors.

Future research should validate the framework using larger, more diverse, and independently collected datasets, together with repeated or nested stratified cross-validation, confidence intervals, external validation, threshold optimization, and feature-stability analysis. Further studies may also compare mRMR with alternative feature selection and feature extraction approaches, including recursive feature elimination, embedded tree-based methods, PCA, and hybrid mRMR–PCA strategies. These extensions would provide stronger evidence regarding the robustness, generalizability, and practical suitability of machine-learning models for preliminary gaming addiction risk classification.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the reviewers and editorial team for their constructive comments and valuable suggestions, which contributed to improving the quality and clarity of this manuscript.

AUTHOR CONTRIBUTIONS

R.S.P.: Conceptualization, Methodology, Software, Data Curation, Formal Analysis, Writing – Original Draft. A.R.H.: Methodology, Validation, Formal Analysis, Visualization, Writing – Review & Editing. S.R.M.: Investigation, Validation, Writing – Review & Editing. M.F.M.: Supervision, Project Administration, Validation, Writing – Review & Editing. All authors have read and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the research presented in this manuscript.

ETHICAL CLEARANCE STATEMENT

This study utilized a secondary dataset containing anonymized records and did not involve direct interaction with human participants or the collection of personally identifiable information. Accordingly, no additional human subject data collection was conducted as part of this research. The dataset was used exclusively for academic research purposes.

DECLARATION OF AI TOOL USAGE

The authors used ChatGPT (OpenAI) to assist with language refinement, grammatical correction, and improving the clarity and structure of the manuscript. All AI-assisted content was reviewed, verified, and revised by the authors. The authors take full responsibility for the accuracy, originality, scientific integrity, and final content of the manuscript.

REFERENCE

- [1] World Health Organization, "Addictive behaviours: Gaming disorder," Oct. 22, 2020. [Online]. Available: <https://www.who.int/news-room/questions-and-answers/item/addictive-behaviours-gaming-disorder>. [Accessed: Aug. 4, 2026].
- [2] World Health Organization, "6C51 Gaming disorder," ICD-11 for Mortality and Morbidity Statistics. [Online]. Available: <https://icd.who.int/browse/2025-01/mms/en>. [Accessed: Aug. 4, 2026].
- [3] J. Billieux, D. J. Stein, J. Castro-Calvo, S. Higuchi, and D. L. King, "Rationale for and usefulness of the inclusion of gaming disorder in the ICD-11," *World Psychiatry*, vol. 20, no. 2, pp. 198-199, 2021, doi: 10.1002/wps.20848.
- [4] A. Musetti, G. Floros, M. Chiappedi, and V. Stavropoulos, "Gaming disorder in the ICD-11: The state of the game," *BMC Psychiatry*, vol. 25, art. no. 1114, 2025, doi: 10.1186/s12888-025-07576-8.
- [5] J. B. Saunders, "Alcohol, drugs and addictive behaviors in ICD-11: Major changes have improved the diagnostic system," *European Psychiatry*, vol. 68, suppl. 1, p. S11, 2025, doi: 10.1192/j.eurpsy.2025.76.
- [6] Q. Wang, K. Mati, and Y. Cai, "The link between problematic internet use, problematic gaming, and psychological distress: Does sleep quality matter?" *BMC Psychiatry*, vol. 21, art. no. 103, 2021, doi: 10.1186/s12888-021-03105-5.
- [7] M. Y. C. Wong, G. F. Yuan, C. Liu, S. K. K. Lam, and H. W. Fung, "The relationship between internet gaming disorder, sleeping quality, self-compassion, physical activity participation and psychological distress: A path analysis," *Cambridge Prisms: Global Mental Health*, vol. 11, e67, 2024, doi: 10.1017/gmh.2024.36.
- [8] T. Alshammari, S. Alseraye, A. Rogowska, N. Alrasheed, and M. Alshammari, "Examining the indirect effect of online gaming on depression via sleep inequality and anxiety-A serial and parallel mediation analysis," *Journal of Clinical Medicine*, vol. 11, no. 24, art. no. 7293, 2022, doi: 10.3390/jcm11247293.
- [9] M. S. Mohamed, G. Rukh, H. B. Schiöth, S. Vadlin, S. Olofsdotter, C. Åslund, and K. W. Nilsson, "Worsened anxiety and loneliness influenced gaming and gambling during the COVID-19 pandemic," *Journal of Clinical Medicine*, vol. 12, no. 1, art. no. 249, 2023, doi: 10.3390/jcm12010249.
- [10] R. Kristensen, S. Pallesen, D. L. King, M. Hysing, and E. K. Erevik, "Problematic gaming and sleep: A systematic review and meta-analysis," *Frontiers in Psychiatry*, vol. 12, art. no. 675237, 2021, doi: 10.3389/fpsy.2021.675237.
- [11] Y. Yu, V. W. I. Fong, J. H.-Y. Ng, Z. Wang, X. Tian, and J. T. F. Lau, "The associations between loneliness, hopelessness, and self-control and internet gaming disorder among university students who were men who have sex with men: Cross-sectional mediation study," *Journal of Medical Internet Research*, vol. 25, e43532, 2023, doi: 10.2196/43532.
- [12] A. Cudo, N. Kapiś, and A. Zabielska-Mendyk, "Personal distress as a mediator between self-esteem, self-efficacy, loneliness and problematic video gaming in female and male emerging adult gamers," *PLOS ONE*, vol. 14, no. 12, e0226213, 2019, doi: 10.1371/journal.pone.0226213.
- [13] S. Islam, A. N. Tusher, M. S. Mia, and M. S. Rahman, "A machine learning based approach to predict online gaming addiction in the context of Bangladesh," in *Proc. 13th Int. Conf. Computing Communication and Networking Technologies (ICCCNT)*, 2022, pp. 1-7, doi: 10.1109/ICCCNT54827.2022.9984508.

- [14] T. Hasan, M. Z. Bonny, R. K. Das, S. Sultana, M. T. Islam, and A. Sattar, "Prediction of Bangladeshi urban children's mental health for the effect of mobile gaming using machine learning," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 14, no. 3, pp. 859-865, 2024, doi: 10.18517/ijaseit.14.3.18534.
- [15] Güllü M, Yagin FH, Gocer I, Yapici H, Ayyildiz E, Clemente FM, Ardigò LP, Zadeh AK, Prieto-González P and Nobari H, "Exploring obesity, physical activity, and digital game addiction levels among adolescents: A study on machine learning-based prediction of digital game addiction," *Front. Psychol.* 14:1097145, 2023, doi: 10.3389/fpsyg.2023.1097145
- [16] Y. Gan, L. Kuang, X.-M. Xu, M. Ai, J.-L. He, W. Wang, S. Hong, J. Chen, J. Cao, and Q. Zhang, "Application of machine learning in predicting adolescent Internet behavioral addiction," *Frontiers in Psychiatry*, vol. 15, art. no. 1521051, 2025, doi: 10.3389/fpsyt.2024.1521051.
- [17] Liu J, Chen L, Chen Y, Luo J, Yu K, Fan L, Yong C, He H, Liao S, Ge Z and Jiang L, "Explainable machine learning prediction of internet addiction among Chinese primary and middle school children and adolescents: a longitudinal study based on positive youth development data (2019–2022)," *Front. Public Health* 13:1590689, 2025, doi: 10.3389/fpubh.2025.1590689.
- [18] D. Kornev, R. Sadeghian, S. Nwoji, Q. He, A. Gandjbakhche, and S. Aram, "Machine learning-based gaming behavior prediction platform," in *Neuroergonomics and Cognitive Engineering*, vol. 42, pp. 122-128, 2022, doi: 10.54941/ahfe1001826.
- [19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005, doi: 10.1109/TPAMI.2005.159.
- [20] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185-205, 2005, doi: 10.1142/S0219720005001004.
- [21] T. Batbat, A. Güven, M. Altinkaynak, D. Yeşilbaş, E. Uğurgöl, E. Demirci, N. Dolu, and M. İzzetoğlu, "Multimodal neurocognitive assessment of internet gaming disorder using ERP and fNIRS during an oddball paradigm: A machine learning-based classification," *Behavioural Brain Research*, vol. 507, art. no. 116192, 2026, doi: 10.1016/j.bbr.2026.116192.
- [22] L. Wang, L. Zhou, S. Liu, Y. Zheng, Q. Liu, M. Yu, X. Lu, W. Lei, and G. Chen, "Identification of patients with internet gaming disorder via a radiomics-based machine learning model of subcortical structures in high-resolution T1-weighted MRI," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 133, art. no. 111026, 2024, doi: 10.1016/j.pnpbp.2024.111026.
- [23] P.-Y. Yeh, C.-L. Lin, C.-K. Sun, and S.-Y. Tung, "Application of EEG-based machine learning in time-frequency and brain connectivity analysis among individuals at high risk of internet gaming disorder with social anxiety," *International Journal of Mental Health and Addiction*, vol. 24, no. 2, pp. 1841-1864, 2026, doi: 10.1007/s11469-025-01560-9.
- [24] J. Y. Lee, M. S. Song, S. Y. Yoo, J. H. Jang, D. Lee, Y. C. Jung, W. Y. Ahn, and J. S. Choi, "Multimodal-based machine learning approach to classify features of internet gaming disorder and alcohol use disorder: A sensor-level and source-level resting-state electroencephalography activity and neuropsychological study," *Comprehensive Psychiatry*, vol. 130, art. no. 152460, 2024, doi: 10.1016/j.comppsy.2024.152460.
- [25] X. Han, L. Wei, Y. Sun, Y. Hu, Y. Wang, W. Ding, Z. Wang, W. Jiang, H. Wang, and Y. Zhou, "MRI-based radiomic machine-learning model may accurately distinguish between subjects with internet gaming disorder and healthy controls," *Brain Sciences*, vol. 12, no. 1, art. no. 44, 2022, doi: 10.3390/brainsci12010044.
- [26] N. T. Thity, A. Rahman, A. Dulmini, M. N. Yasmin, and R. Rois, "An illustration of multi-class ROC analysis for predicting internet addiction among university students," *PLOS ONE*, vol. 20, no. 7, e0325855, 2025, doi: 10.1371/journal.pone.0325855.
- [27] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, V. Bolón-Canedo, J. M. Benítez, F. Herrera, and A. Alonso-Betanzos, "Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data," *International Journal of Intelligent Systems*, vol. 32, no. 2, pp. 134-152, 2017, doi: 10.1002/int.21833.

- [28] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *Proc. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2019, pp. 442-452, doi: 10.1109/DSAA.2019.00059.
- [29] T. K. Phan, L. Ma, T. N. Nguyen, and S. Do, "BoMGene: Integrating Boruta-mRMR feature selection for enhanced gene expression classification," *arXiv preprint*, arXiv:2510.00907, 2025, doi: 10.48550/arXiv.2510.00907.