Enhancing Kernel-based Model Predictive Power Through Enhanced Relief-based Algorithm for the Early Detection of Alcohol Use Disorder Among Secondary Students

Myat Noe WIN, Sri Devi RAVANA*, Tutut HERAWAN, Liyana SHUIB

Department of Information Systems, Faculty of Computer Science & Information Technology, University of Malaya, Kuala Lumpur, 50603, Malaysia
myatnoewin786@gmail.com, sdevi@um.edu.my (**Corresponding author*),
liyanashuib@um.edu.my, tutut@um.edu.my

Abstract: Despite its efficacy, machine learning in health sciences faces limitations with regard to addiction prediction due to integrating diverse data sources, addressing biases, and interpreting complex models. This may reduce the effectiveness of predictive models in identifying at-risk individuals and informing intervention strategies. The current challenge lies in identifying the optimal number of features for model training and determining the influential factors for alcohol addiction. Therefore, this paper explores and proposes an enhanced feature engineering algorithm which not only ranks the feature importance, but also automatically extracts the optimal features for the prediction model, which in return improves the predictive power of kernel-based models. By using a feature aggregation approach, the features identified by different Relief-based algorithms (such as Relief, ReliefF and RReliefF) were merged into a unified set as a ranked feature list, and the Relief-based algorithms were integrated with the XGBoost boosting algorithm for the implementation of an automated feature selection process. The proposed method provided 11 influential features to be included as n_features in the predictive model. Three different families of classifiers, namely the Linear, Ensemble-based and Kernel-based classifiers were analysed in combination with the enhanced Relief-based algorithm to evaluate the response of the proposed model to the respective algorithms. In this context, the enhanced RReliefF algorithm improved the Kernel-based model by 7.47% in terms of the discriminative power and by 12.69% with regard to the predictive power, in comparison with the baseline model. These findings aided in resolving the limitations related to the manual optimal feature selection typical of the current feature engineering methods, thereby opening a new research avenue for automatic feature engineering in a low-code context. Overall, the proposed enhanced algorithm ensures technical correctness by leveraging ReliefF algorithm's feature rankings effectively for an improved performance in the context of kernel-based models like Support Vector Machine (SVM), making them more accessible and actionable for clinicians and healthcare professionals working in alcohol addiction-related prevention and intervention.

Keywords: Feature Aggregation, Automatic Feature Selection, Influential features, Kernel-based model, RReliefF.

1. Introduction

The high prevalence of Alcohol Use Disorder (AUD) and its associated risks among students underscores the need for early identification and intervention (Kast & Avery, 2019). Despite the significant global burden of alcoholism, there remains a lack of effective, evidence-based treatments, emphasizing the potential of novel, technology-supported approaches, including mechanistically distinct medications (Stokłosa et al., 2023).

Machine learning has shown promise in predicting alcohol addiction and facilitating the early detection of AUD among students (Stüben et al., 2023). However, unlike other domains, AUD prediction in clinical settings must balance accuracy with interpretability and computational efficiency. Model interpretability is essential for identifying factors contributing to addiction risk, enabling tailored clinical interventions.

Additionally, an efficient computational resource utilization is critical for processing large datasets and deploying predictive models in a timely manner. Thus, the central challenge lies in achieving interpretability and computational efficiency without compromising predictive performance.

Feature engineering plays a pivotal role in addressing these challenges. Without a robust strategy, predictive models could lack transparency or demand excessive computational resources, which would limit their practical applicability (Chatzimparmpas et al., 2022). The existing techniques for feature extraction face the ongoing challenge of selecting an optimal subset of features from the dataset's attributes (Uddin et al., 2018).

The existing methods often prioritize overall accuracy, which can mask the performance of the model in identifying true positives (Hung et al., 2023). Improving recall is crucial in medical diagnosis to ensure that individuals with a certain condition are correctly identified and receive an appropriate intervention. Many feature selection methods fail to optimize feature discriminative power, resulting in models that struggle to distinguish between classes (Chakroun et al., 2023). This results in poor model performance and unreliable predictions.

https://doi.org/10.24846/v33i4y202406

For AUD predictions, recall is critical for diagnosis and preventing false negatives. Similarly, discriminative power is important for accuracy of predictions and better decision making. Traditional feature selection methods may not fully leverage the potential of both kernel-based and linear models, leading to a suboptimal performance in classification tasks. Specifically, the key technical challenge is enhancing recall and discriminative power while preserving model interpretability and efficiency for real-world applicability.

Henceforth, this study aims to address these issues by employing an automatic feature engineering approach to identify the optimal subset of features and evaluate kernel-based models' predictive performance. This approach reduces subjectivity associated with manual feature selection, mitigates the risk of overlooking optimal solutions, and adapts dynamically to changes in datasets.

Therefore, unlike Traditional Relief and ReliefF algorithms, which focus on local, direct associations and do not account for minority class bias, making them less effective at improving recall in imbalanced datasets, the enhanced RReliefF with Boosting employs a comprehensive feature aggregation and selection strategy that identifies features relevant to positive cases. This enables the model to better capture challenging cases, significantly improving recall. Ultimately, it enhances predictive power and interpretability, improving AUD prediction in clinical applications.

This paper is organized as follows. Section 2 reviews the state-of-the-art approaches. Section 3 outlines the proposed methodology, while Section 4 presents the experimental setting. Section 5 discusses the obtained results and Section 6 refers to the application of the proposed model to real-world scenarios. Finally, Section 7 outlines the conclusion of this paper and future research directions.

2. Related Works

Multiple studies have been conducted to identify influential features in AUD with high dependency on statistical methods, including the works of Brunborg et al. (2021), Wang et al. (2018) and Arshad et al. (2015), as well as developing predictive models with different feature selection techniques (Win et al., 2022a; Sarić et al., 2019).

Notably, Win et al. (2022a) developed an AI-based virtual therapist to support alcohol relapse prevention. Authors utilized different machine learning techniques to personalize interventions. Feature engineering is achieved through manual selection to be implemented based on a traditional machine learning model.

Symons et al. (2020) compared clinical psychologists and "trained" machine learning models to explore the predictive accuracy for alcohol dependence treatment outcomes. Sarić et al. (2019) used a decision tree-based algorithm to identify alcohol addicts among high school students. Singh et al. (2022) explored machine learning algorithms for predicting alcohol consumption rates and the Random Forest and Decision Tree algorithms as the highest performing models.

Ruiz-España et al. (2023) adopted the filter method and wrapper method for feature selection to select the optimal number of features from the alcoholic patients' dataset. The findings showed that Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel coupled with a wrapper feature selection method achieved an Area Under the Curve (AUC) value of 0.906, but required nearly 19 minutes of computational time on a high-performance Intel processor.

Palaniappan et al. (2017) employed deep learning techniques, including Multilayer Perceptron (MLP) and AutoMLP, to classify alcohol addiction among secondary students with a dataset consisting of 15 manually selected features. By contrast, Ismail et al. (2018) trained various machine learning models directly on the chosen dataset without employing any feature selection techniques. Kinreich et al. (2021) used a multi-feature machine learning approach to predict the alcohol use disorder.

Although the above-mentioned studies have successfully reached their objectives, the common gap across the studies can be interpreted as the absence of balance between performance and the resource consumption ratio. Deep learning-based feature extraction has disadvantages, including environmental limitations, the number of data sources required for training, and the performance of the algorithm (Zhao et al., 2021). Likewise, wrapper methods feature a high computational complexity and take a significant amount of time to complete (Verhaeghe et al., 2022). Balancing the demand for top-notch features with the associated expenses underscores the necessity for a more efficient and scalable feature engineering (Zhang et al, 2020; Mamikonyan et al., 2020).

On the other hand, Support Vector Machine (SVM), a kernel-based model, is typically considered the most complex among the three approaches (i.e. Linear, Kernel-based and Ensemble-based), due to its ability to create non-linear decision boundaries by mapping input features into a higher-dimensional space using kernel functions (Olorisade et al., 2022). Likewise, Nikhitha et al. (2021) reported kernel-based methods guarantee structural risk minimization and global optimal solutions, but they may struggle with large-scale datasets due to their computational complexity.

Therefore, exhaustive search and deep-learning based feature extractions are not the ideal candidates to combine with complex models like the kernel-based approach due to the potential added complexity and increased resource consumption. Given the inherent complexity of kernel-based models, there is a need to explore underrated feature engineering approaches that complement the complexity of kernel-based models and leverage AUD prediction. In short, these facts underscore the need for a simple yet effective feature engineering approach for enabling kernel-based AUD prediction in clinical settings.

3. The Proposed Method

Unlike the previous studies which focused on feature ranking based on their importance or statistical representations, the proposed method integrates the Relief algorithm with the XGBoost boosting algorithm for the implementation of the automated feature selection process to identify the optimal number of n features to be used in alcohol addiction predictive models. Through iterative feature subset evaluation, the proposed method systematically evaluates different subsets of features and selects the subset that maximizes predictive performance, thereby leveraging ReliefF's feature rankings effectively.

The proposed method includes five major steps, namely aggregated feature ranking, automated feature selection, optimal feature subset formulation, prediction model implementation and evaluation. Firstly, the original dataset consists of multiple data points, which makes it challenging to identify the influential features for alcohol addiction prediction. Therefore, the highest feature importance scores across different Relief algorithms (Relief, ReliefF and RReliefF) are extracted and aggregated as a ranked list.

Secondly, the XGBoost boosting algorithm is equipped on top of the ranked list with cross-validation to measure the model`s predictive performance for a different subset of ranked features iteratively. During cross-validation, XGBoost provides a reliable estimate of the feature subset`s performance, leveraging its ability to capture complex patterns and interactions, and assisting it to identify the most informative features. Thirdly, optimal features are identified as the best combination of the highest predictive accuracy with the input set of features from the ranked list.

Fourthly, the optimal number of features is used as input to train different classification models i.e. Linear, Kernel-based and Ensemble-based models to observe the performance of diverse classification families on the enhanced relief algorithm. Lastly, the balanced accuracy and discriminative power across the selected models are compared and evaluated to find the best predictive model that complements enhanced feature engineering. With this approach, the most influential factors for alcohol addiction are identified, leading to more accurate predictions. The proposed method is depicted in Figure 1.

Regarding the selection of methods, the Relief-based algorithm family is chosen due to its efficiency and scalability. Relief is computationally

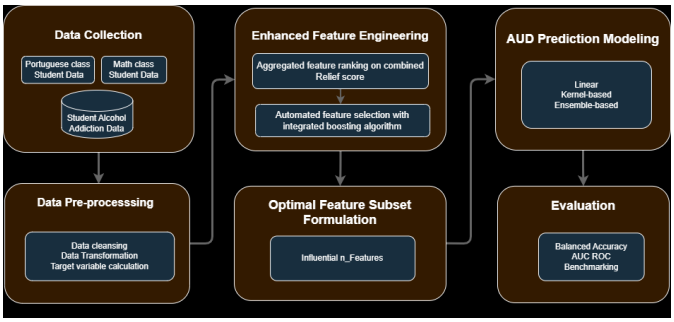

Figure 1. Overview of the proposed method

efficient compared to wrapper methods, as it does not require a repeated training of the model. This efficiency is valuable for datasets with a number of attributes and records which is moderate to large. Morevoer, Relief algorithms scale well with the number of instances, making it suitable for AUD datasets.

On the other hand, with regard to the classification algorithms, different types of classifiers from each classification family are chosen as Linear (LR), Kernel-based (SVM) and Ensemble-based (RF) to evaluate the reaction of enhanced feature engineering approach on different classification types. In terms of complexity management, the chosen algorithms (LR, SVM, RF) are scalable and can be parallelized, which makes them suitable for handling the dataset efficiently. Besides, systematic hyperparameter tuning (i.e. grid search, random search) is employed to find the optimal settings for each model, ensuring a balanced complexity and performance.

3.1 Enhanced Relief Algorithm for Feature Engineering

The major technical correction proposed for the above-mentioned method lies in overcoming the limitation of the Relief algorithm based on instance-level feature ranking and leveraging ReliefF algorithm's rankings into actionable subsets for machine learning model training. Traditional Relief algorithm includes three technical steps, namely Initialization, Feature Importance Calculation and Feature Ranking. The drawback is that it does not suggest the relevant feature subset for the model training.

On the other hand, the proposed feature engineering enhancement includes four steps, that is Initialization, Aggregated Feature Importance Calculation, Feature Ranking, and Automated Feature Selection. The proposed method enhances technical correctness by empirically validating feature subsets i.e., instead of manually choosing n_features, the algorithm evaluates multiple feature subsets and selects the one yielding the highest cross-validated score, thus improving predictive performance. The significant improvement consists in addressing the arbitrary feature selection issue related to the traditional Relief algorithm. The high-level algorithm for the proposed model is illustrated below in Figure 2.

```
Algorithm EnhancedReliefFeatureEngineering
Input: Dataset D, number of neighbors k, range of features to evaluate F_range
Output: Performance metrics for the Linear model, Kernel model and Ensemble model

1. Initialize empty lists for Relief, ReliefF, and RReliefF rankings
2. For each algorithm in {Relief, ReliefF, RReliefF}:
   a. Initialize weight vector W for feature importance
   b. For each instance i in D:
      i. Find k nearest hits and k nearest misses
      ii. Update weights W based on differences between i and its neighbors
   c. Rank features based on weights W
   d. Store rankings in the corresponding list

3. Combine top-ranked features from Relief, ReliefF, and RReliefF
4. Create a consolidated rank list of features

5. Initialize best_score to 0
6. Initialize best_n_features to 0

7. For n in F_range:
   a. Select top n features from consolidated rank list
   b. Train an XGBoost model using these features
   c. Perform cross-validation and compute performance metric (F1 score)
   d. If current performance metric > best_score:
      i. Update best_score to current performance metric
      ii. Update best_n_features to n

8. Select top best_n_features from consolidated rank list
9. Train Linear model, Kernel model, Ensemble model using these features
10. Evaluate the respective models on validation/test set
11. Compute and return performance metrics

End Algorithm
```

Figure 2. The enhanced feature engineering algorithm

3.2. Complexity Analysis for the Enhanced Feature Engineering Algorithm

For feature ranking, the importance of each feature is computed for each data point (instance). If m is the number of instances and f is the number of features, the complexity is denoted as $O(3.m.f) = O(m.f)$. Next, for the consolidated ranking, it involves sorting and merging, which has a complexity of $O(flogf)$.

On the other hand, for automatic feature selection with cross-validation, each number of features n in the range F_{range} is trained iteratively. Suppose the maximum number of features is N and the number of cross-validations is k. For each number of features n, XGBoost takes an amount of computational time which is proportional to $m.nlogn$. Since this exercise is executed for each subset of features and for each validation, the total cost is $O(k.m.f^2logf)$.

Next, different classifiers are used for training on the best number of features found in the previous step, which interprets the complexity as $O(m.(n^*)^2)$. Since $n^* \leq f$ the total cost is $O(m.f^2)$. Lastly, as evaluation is concerned, the total cost of evaluating takes an amount of computational time which is proportional to $m.n^*$ which is $O(m.f)$ since $n^* \leq f$. The dominant term here is $O(k.m.f^2logf)$ from the automated feature selection step, because it grows the fastest as f and m increase. The total computational time for the algorithm is the sum of computational times for all the steps. However, the step with the highest time complexity dominates the total time. Thus, the total time complexity of the algorithm is primarily $O(k.m.f^2logf)$.

4. Experimental Setting

The experimental setting of the study includes data acquisition, data preprocessing, modeling, model evaluation, and deployment. The experimental testbed is Intel(R) Core (TM) i5-6500 CPU @ 3.20GHz, 16GB RAM and 6GB GPU.

4.1 Datasets

This experiment utilized two datasets. The first, the Student Dataset by Cortez & Silva (2008), comprises 33 attributes and 1,044 records. The second, the Drug Dataset by Fehrman et al. (2017) contains 32 attributes and 1,885 records. The Student Dataset serves as the primary dataset for end-to-end experimentation, while the Drug Dataset is used to replicate selected experiments, which demonstrates the robustness and performance of the proposed approach.

4.2 Data Preprocessing

Due to the absence of the target variable alcohol addiction in the analysed data, a new variable called addiction was created with existing attributes using the recommended method in equation (1), proposed by Shukla et al. (2018):

$$addiction = \frac{[(Walc \times 2) + (Dalc \times 5)]}{7} \quad (1)$$

The new target variable, addiction, was then mapped to a binary outcome as it is shown in equation (2), where 1 means having an alcohol addiction and 0 means not having an alcohol addiction:

$$\begin{aligned} addiction > 3 &= 1 \\ addiction \leq 3 &= 0 \end{aligned} \quad (2)$$

The dataset was then divided into 2 sets, as training and a testing set, with a ratio of 8:2.

4.3 Aggregated Feature Ranking

Relief-based algorithms, namely Relief, ReliefF, and RReliefF, are employed to extract features that have the highest feature importance scores. Among all feature engineering methods, Relief algorithms are recommended for exploration due to their ability to balance interpretability and computational efficiency effectively. Relief algorithms have been shown to prioritize relevant features while handling noise and imbalanced data, which makes them suitable for tasks where model transparency and computational resources are paramount, such as alcohol addiction prediction in clinical settings. The generic formula of Relief-based feature selection methods can be represented as in equation (3), where x_i is the instance being evaluated, $x_i^{(r)}$ is the nearest instance of the same class and $x_i^{(\delta)}$ is a different class, d is the distance between instances and N represents the number of features as n_features.

$$Relief-based(x_i) = \sum_{j=1}^{k} \frac{-d(x_i, x_i^{(r)}) + d(x_i, x_i^{(\delta)})}{N} \quad (3)$$

After computing feature importance scores using Relief-based algorithms, an aggregated feature selection approach is employed to extract influential features. The top features with the highest scores from each algorithm are selected independently, ensuring the identification of a subset of the most influential features.

Overlapping features across all three sets are then identified, as they indicate a consistent importance across algorithms, and are considered influential regardless of the specific Relief-based algorithm used. These influential features, extracted from the overlapping sets, are presented as a unified set in the dataset, which emphasizes their robust importance across different algorithms and enhances feature selection reliability.

4.4 Automated Feature Selection

Following the completion of feature ranking, the XGBoost boosting algorithm is implemented within a cross-validation framework to evaluate the performance of different subsets of features. This step uses a ranked list based on aggregated feature engineering as the input and iteratively selects top n-ranked features to determine the subset of features that yields the best cross-validated performance, ensuring robustness in feature selection. The XGBoost algorithm is known for its strong performance and ability to handle complex datasets. When used in the feature selection process, it provides a robust measure of how well a subset of features performs. XGBoost can capture the non-linear relationships and interactions between features, which makes it an excellent choice for evaluating the true importance of features. This capability ensures that the selected features are genuinely useful for prediction, even if their relationships with the target variable are complex.

Using cross-validation to evaluate different feature subsets ensures that the feature selection process is robust and generalizes well for new data. This systematic approach is technically correct as it mitigates the risk of overfitting and ensures that the selected features are truly relevant for the prediction task.

4.5 Optimal Feature Subset Formulation

The outcome of automatic feature extraction determines the optimal set that maximizes the employed model`s predictive performance. The selection is based on criteria such as predictive accuracy, model interpretability, or computational efficiency The resulting new feature list, refined through rigorous validation, is ready in this stage for integration into the proposed classification model. This approach not only enhances model interpretability but it also ensures that the selected features contribute significantly to the classification task at hand.

The selected optimal feature subsets are validated against the existing literature in the public health domain to ensure their relevance and consistency with regard to the known factors influencing the studied condition. This cross-verification enhances the credibility of the feature selection process and aligns the model with the established public health knowledge.

This phase confirms a subset of features from a larger set of available features as the subset of optimal features that maximize the performance of a machine learning model for alcohol addiction prediction.

4.6 Modelling

For modelling purposes, three distinct families of classifiers were selected. Logistic Regression (LR) was chosen for the Linear family, Random Forest (RF) for the ensemble-based family, and Support Vector Machine (SVM) for the kernel-based family.

The LR is a type of generalized linear model (GLM) which is used for modelling the relationship between a dependent variable and one or more independent variables (Arshad et al., 2023). RF is an ensemble learning method, belonging to the family of decision tree-based models. RF builds multiple decision trees during training and outputs the mode of the classes (classification) or average prediction (regression) for the individual trees (Zhu, 2020).

The SVM was chosen for its capability to handle non-linear classification tasks by mapping data into a higher-dimensional space using kernel functions. They are a class of algorithms for classification and regression, representing the current state of the art (Mathar et al., 2020), also claimed to be facing challenges with complexity and a higher consumption time (Ruiz-España et al., 2023).

By comparing LR, RF, and SVM, the performance of different types of classifiers can be evaluated across various aspects such as predictive accuracy, interpretability, and the ability to handle different types of data distributions. This comprehensive comparison helps in selecting the most appropriate classifier for alcohol addiction prediction based on the specific characteristics of the analysed dataset.

For the model training process, the dataset was split according to a 80:20 ratio into a training and a testing set, and the selected classifier was trained on the training set and evaluated with the testing set. Rounds of experiments were conducted with each classifier for each Relief-based algorithm, which helped to assess the stability and consistency of the models' performance across different feature selection techniques. By repeating the experiments, any variability or fluctuations in model performance was detected, ensuring the reliability of the results and providing insights into the robustness of the chosen classifiers.

4.7 Evaluation

The evaluation utilized Precision, Recall, the F-score, AUC, Log Loss, Matthews Correlation Coefficient (MCC), and Training Time (T-time) to comprehensively assess model performance and align with the research objectives. Precision evaluates the proportion of true positives among the predicted positives, while Recall measures the model's ability to identify all true positives, both critical for minimizing errors in AUD

detection. F-score balances Precision and Recall, which is particularly valuable for imbalanced datasets. AUC assesses discriminative power across classification thresholds, offering an aggregate performance view, while Log Loss penalizes overly confident incorrect predictions, encouraging reliable probabilistic outputs. MCC provides a balanced evaluation of the binary classification performance, even with imbalanced classes. Finally, T-time ensures computational efficiency, supporting real-world deployment. These metrics collectively ensure the robust evaluation of predictive power, interpretability, and efficiency.

5. Results and Discussion

5.1 Identification of Optimal Features with Enhanced Feature Engineering

Feature importance scores were calculated for each Relief algorithm, and the consolidated rankings were derived after eliminating highly correlated features, resulting in the selection of 26 out of 33 original features. Relief and RReliefF showed the highest similarity, sharing 63.64% of features, with seven common attributes: Age, Mother Education, Free Time, Higher Education, Family Relation, Father Education, and Gender. Comparatively, ReliefF and RReliefF overlapped on 54.55% of features with six shared features: Gender, Higher Education, Address, Failures, Family Size, and Family Relation. Relief and ReliefF exhibited the lowest similarity at 9%, with Family Relation as the only shared feature.

These findings highlight varying degrees of consistency for feature importance across the Relief-based algorithms. Some features are consistently highly ranked, while others show inconsistent rankings due to differences in how the algorithms handle feature interactions, data noise, and methodological assumptions. Consolidated rankings were generated by aggregating importance scores across all three methods, identifying features with the highest combined importance, as it is illustrated in Table 1.

Table 1. Combined Rank Score for the Ranked Features

Ranked Features	Combined Rank Score
age	1.066
goout	0.888
freetime	0.836
Fedu	0.793
G3	0.768
sex	0.768
famsup	0.755
traveltime	0.744
studytime	0.646
health	0.619
failures	0.611
address	0.596
Medu	0.519
guardian	0.511
famrel	0.466
Fjob	0.328
romantic	0.294
famsize	0.27
activities	0.255
internet	0.223
paid	0.203
schoolsup	0.178
higher	0.155
Mjob	0.074
Pstatus	0.072

Next, the automatic feature selection process was initiated for the ranked features for Relief algorithm, which returned 11 feature subsets as the optimal feature size.

The enhanced feature engineering approach identified 11 influential features for predicting alcohol use disorder among students: Age, Family Relation, Academic Failures, Family Size, Mother Education, Father Education, Free Time,Higher Education, Address/Neighborhood, Frequency of Going Out, and Gender. These features reflect key dimensions of socio-economic background, academic performance, family dynamics, and personal characteristics, which have a significant influence on alcohol addiction risk (Haardörfer et al., 2020).

The relevance of these features is consistently supported by public health research. Age is a critical demographic factor, with younger individuals exhibiting higher susceptibility to alcohol misuse (Calvo et al., 2020). Family-related factors,

including family relationships, parental education, and family support, highlight the influence of familial environments on substance use behaviours (Chi et al., 2022). Academic factors such as performance and aspirations for higher education further underscore the connection between educational outcomes and health behaviours (Kendler et al., 2020). Additionally, gender and residential location are well-documented predictors of alcohol use patterns (Edmonds et al., 2022).

The convergence of evidence across various studies validates these features as reliable predictors in alcohol addiction modelling, strengthening their inclusion in predictive frameworks to improve risk assessment and intervention strategies.

5.2 Enhancing the Predictive Power of Kernel-based Models

The experiments were conducted across three scenarios: (1) classification models utilizing Traditional Relief, (2) classification models employing Traditional ReliefF, and (3) models leveraging the proposed Enhanced RReliefF for feature selection. To benchmark performance, the proposed automated feature selection method was compared against a manually configured model with a predefined number of features on the same dataset, validating the improvements achieved. The results for different metrics for each classification model is illustrated in Table 2.

The results demonstrate that the Enhanced RReliefF model consistently outperforms the Traditional Relief and ReliefF models across all metrics and classifiers. For LR, Enhanced RReliefF achieves the highest recall (0.82), F1-score (0.783), and AUC (0.839) with a balanced MCC (0.65), demonstrating its ability to identify positive instances while maintaining robustness. This finding is consistent with the previous study of Arshad et al. (2023) which stated that LR outperformed complex models such as Decision Tree and Support Vector Machines.

Similarly, for SVM, Enhanced RReliefF improves recall (0.83) and F1-score (0.817), surpassing RF in recall, and closing the gap in performance between SVM and RF. This highlights its capability to boost the effectiveness of kernel-based models to that of ensemble-based ones.

For RF, Enhanced RReliefF improves recall (0.8) and AUC (0.848) while reducing log loss, validating its predictive power. The reduction in log loss and comparable T-time across the three classifiers further supports the adaptability and computational efficiency of the Enhanced RReliefF.

Overall, Enhanced RReliefF refines feature selection, while enhancing predictive performance, particularly in SVM, which matches RF with regard to recall (0.83 vs. 0.8) and surpasses RF for the F1-score (0.817 vs. 0.794). In comparison with the baseline model (Ismail et al., 2018), SVM with Enhanced RReliefF shows a 12.69% improvement for the F-score from 0.725 to 0.817. These findings establish Enhanced RReliefF as the superior feature engineering strategy for achieving a high prediction accuracy, interpretability, and computational efficiency across diverse classifiers.

Table 2. Model Evaluation with different metrics

LR							
Algorithm/Type	**Precision**	**Recall**	**F-score**	**AUC**	**Log Loss**	**MCC**	**T-time**
Traditional Refief	0.736	0.59	0.655	0.653	0.583	0.45	0.5
Traditional ReliefF	0.857	0.711	0.777	0.828	0.441	0.66	0.7
Enhanced RReliefF	0.75	0.82	0.783	0.839	0.457	0.65	0.8
SVM							
Algorithm/Type	**Precision**	**Recall**	**F-score**	**AUC**	**Log Loss**	**MCC**	**T-time**
Traditional Refief	0.61	0.614	0.612	0.692	0.607	0.34	2.3
Traditional ReliefF	0.828	9.77	0.798	0.788	0.462	0.64	2.8
Enhanced RReliefF	0.804	0.83	0.817	0.849	0.439	0.67	3.23
RF							
Algorithm/Type	**Precision**	**Recall**	**F-score**	**AUC**	**Log Loss**	**MCC**	**T-time**
Traditional Refief	0.648	0.78	0.708	0.787	0.512	0.38	1.9
Traditional ReliefF	0.687	0.69	0.688	0.845	0.498	0.56	2.15
Enhanced RReliefF	0.788	0.8	0.794	0.848	0.446	0.62	2.5

On the other hand, ROC curves and PR curves are discussed to evaluate the performance of the employed models based on the proposed approach. Figure 3 illustrates the ROC and PR curves for Traditional Relief.

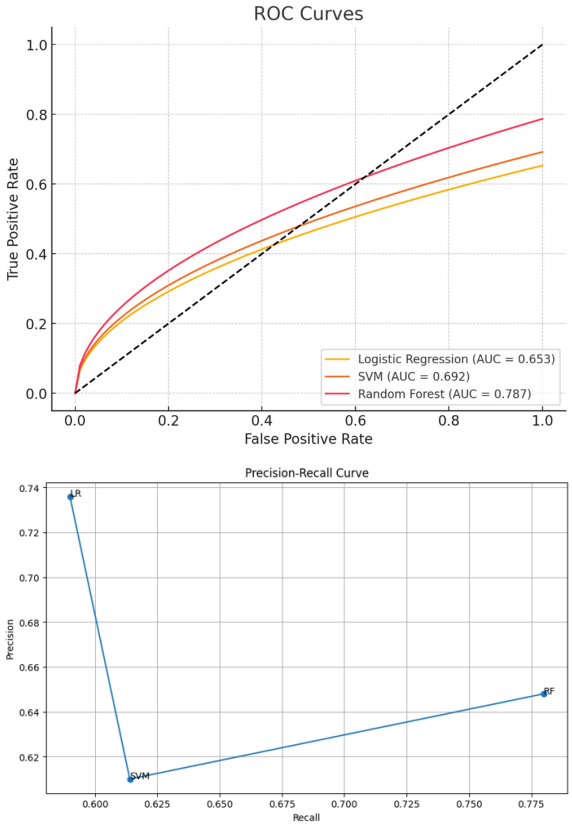

Figure 3. ROC and PR curve for the Traditional Relief algorithm

The ROC curves show that RF outperforms both LR and SVM in terms of overall classification ability. The gap between RF and the other models suggests that RF's ensemble approach provides a significant advantage in handling the dataset with Relief features. LR and SVM have moderate AUC values, the value of AUC for SVM being slightly better than that for LR. This suggests that SVM's capability to handle the feature space provided by Relief is marginally better than that of LR but not as effective as that of RF.

The PR curves reveal different strengths for each model. RF's high recall indicates that it captures most of the positive instances, which is crucial for applications where missing a positive instance is costly such as AUD prediction. LR's higher precision in comparison with SVM suggests it is better at avoiding false positives, but its lower recall indicates it misses more actual positives. SVM's balanced PR metrics show it provides a middle ground between precision and recall but does not excel in either, indicating a consistent but not outstanding performance.

Likewise, for the evaluation with ReliefF, both LR and SVM see a significant improvement in AUC with ReliefF, indicating that this feature selection method has greatly enhanced their ability to distinguish between classes. RF remains the highest-ranking samong the models.

For PR curve, the increase in precision and recall suggests that LR benefits significantly from the improved features, balancing between capturing positives and avoiding false positives. SVM's performance boost with ReliefF shows its dependency on feature selection for an optimal performance and, while RF shows a high AUC, its precision and recall are balanced but lower than expected. This suggests that while RF can leverage feature interactions, this may not always translate into a higher precision and recall, especially if the selected features introduce some redundancy or noise.

All in all, the performance gap between RF and the other models is narrower with ReliefF, indicating an improved performance for all, as it is shown in Figure 4.

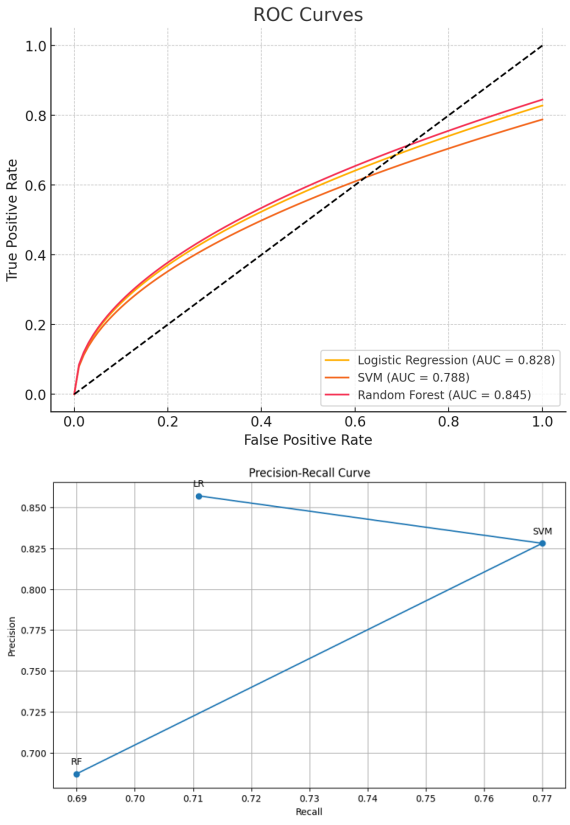

Figure 4. ROC and PR curve for the ReliefF algorithm

Last but not least, for the experiment with the enhanced RReliefF as it is shown in Figure 5, the ROC curve reveals that SVM shows the highest AUC among the three employed models, which indicates an excellent performance in distinguishing between classes. This suggests that SVM was able to capitalize on the features selected through the proposed automatic feature selection algorithm with the enhanced RReliefF, leading to a substantial increase in accuracy.

The PR curves reveal that while RF has a more balanced precision and recall, SVM offers the highest precision and recall, indicating fewer false positives and a better capture of true positives. LR also shows a strong performance but it is slightly outperformed by SVM for both metrics. Notably, the enhanced RReliefF algorithm consistently improves the accuracy of all three models i.e. the Linear, kernel-based and ensemble-based models. In comparison with the baseline model (Ismail et al., 2018), SVM with Enhanced RReliefF shows a 7.47% improvement for AUC from 0.790 to 0.849.

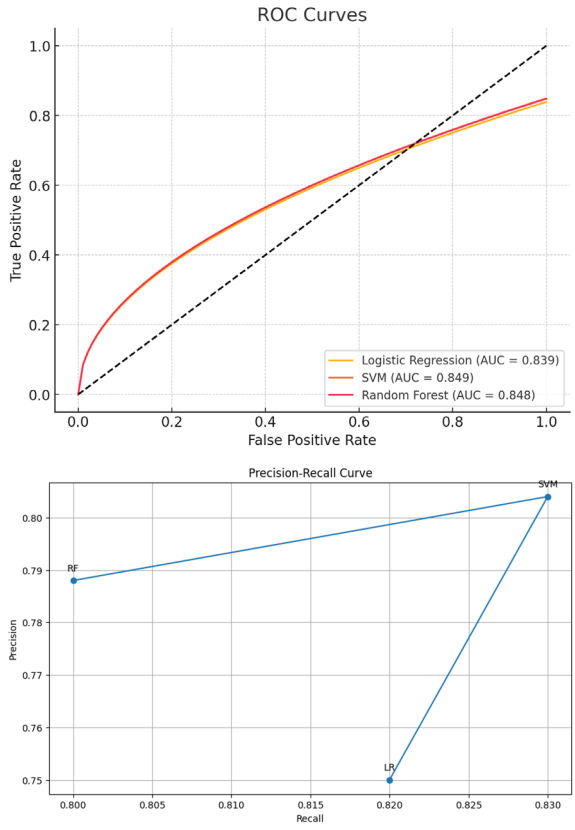

Figure 5. ROC and PR curve for the Enhanced RReliefF algorithm

Additionally, a similar experiment was carried out with the second dataset to evaluate the discriminative power of the employed models with regard to the enhanced RReliefF approach. The findings are illustrated in Figure 6.

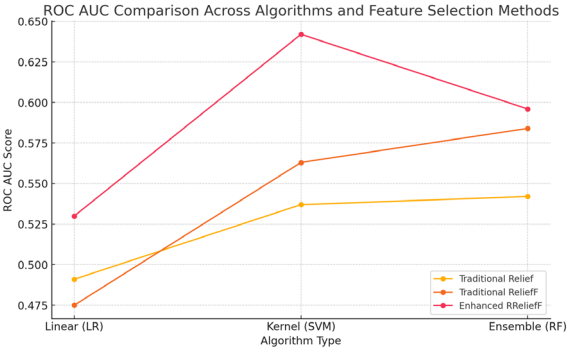

Figure 6. ROC comparasion for Drug Dataset

As a similar observation in relation to dataset 1, for traditional Relief, RF performs slightly better than SVM and LR. For traditional RReliefF, RF still leads, but the gap between RF and SVM narrows. On the other hand, Enhanced RReliefF significantly boosts the performance of SVM (0.642), bringing it closer to the performance level of RF (0.596). This demonstrates that Enhanced RReliefF is particularly effective for kernel-based models like SVM, while SVM relies heavily on the quality of the feature space to find the optimal hyperplane for classification. The improved feature selection enhances SVM's ability to separate classes more effectively, leading to higher performance metrics.

In short, the findings from the round of experiments which were carried out are consistent across different datasets. It can be concluded that traditional Relief algorithms are not sufficient for improving Linear, Ensemble- and Kernel-based models. Relief's weight changes are based solely on the nearest hit and nearest miss. This makes Relief more vulnerable to noisy data, potentially compromising the reliability of its calculations (Cai et al., 2018). The rationale behind the enhanced RReliefF being a helpful feature engineering method lies in its ability to selectively extract the most influential qualities, resulting in a reduced amount of noisy data. This, in turn, enables the machine learning models to make more accurate predictions.

Enhanced RReliefF may work comparatively better for SVM than for LR and RF due to SVM's reliance on identifying optimal decision boundaries, which benefits greatly from having a more discriminative feature set provided by the enhanced RReliefF. Additionally, SVM's

sensitivity to noisy and irrelevant features makes enhanced RReliefF's ability to handle such issues through resampling particularly advantageous, resulting in an improved classification performance for SVM in comparison with LR and RF in certain scenarios.

Overall, the findings highlight the importance of not only selecting the appropriate classifiers but also of leveraging effective feature engineering techniques to enhance model performance. The significant improvement observed in SVM's accuracy with RReliefF underscores the impact of feature engineering on model effectiveness, showcasing its potential to elevate the predictive power of machine learning models, especially in scenarios with balanced accuracy improvement in complex models like kernel-based ones. Henceforth, based on the findings presented, it's evident that SVM performance is significantly improved through the proposed automatic feature engineering approach without resorting to complex feature engineering approaches.

On the other hand, the potential limitations of the proposed approach lie in the area of generalizability, class imbalance and computational requirements. The proposed enhanced RReliefF algorithm is tailored to the specific dataset and problem at hand. While it significantly enhances model performance for AUD prediction, its generalizability to other medical conditions or datasets with different characteristics may require further validation and adjustment. Although Enhanced RReliefF addresses the bias toward minority classes by focusing on features which are relevant to positive cases, extreme class imbalances may still pose challenges. In comparison with deep learning models or wrapper methods, Enhanced RReliefF with SVM is relatively faster but still more resource consuming. This represents an area for potential improvement, by focusing on optimizing computational efficiency without compromising performance.

As a future direction, the application of the Enhanced RReliefF algorithm can be extended to various healthcare domains where accurate and early prediction is critical. For instance, it could be adapted to predict other mental health disorders, such as depression and anxiety, by identifying relevant biomarkers and clinical features. Additionally, it can be applied to chronic disease management, like diabetes or cardiovascular diseases, to improve the early detection and monitoring of at-risk individuals. Enhanced RReliefF's robust feature selection can be leveraged in precision medicine to tailor treatments based on individual patient profiles. By addressing the computational challenges and being integrated with electronic health records, Enhanced RReliefF can significantly contribute to advancing personalized medicine and improving patient outcomes across various healthcare applications.

5.3 Improvement in Efficiency and Sustainability

By leveraging RReliefF with automatic feature selection, SVM achieves an improved accuracy and discriminative power while maintaining interpretability, thereby reducing the need for resource-intensive feature engineering methods such as wrapper methods which need most 19 minutes for the classification task, while SVM on enhanced RReliefF takes 3 minutes and 23 seconds instead. Hence, this streamlined approach not only enhances SVM's predictive performance but it also mitigates computational resource consumption, which makes it a practical and efficient choice for classification tasks in real-world applications.

Additionally, the Big O Notation related to the proposed method demonstrates that for datasets with moderate to large sizes instances and features), $O(k.m.f^2 logf)$ is often manageable. It indicates that the algorithm can process datasets of a reasonable complexity within a reasonable time frame, especially considering the complexity introduced by cross-validation and XGBoost training.

By comparison, deep learning models, such as neural networks, often require significant computational resources (both in terms of time and hardware) due to their complexity and the need for extensive training iterations. Those models can have higher computational demands which can be expressed by $O(N.E)$ where N is the number of instances and E is the number of epochs. In this case, the cost for the proposed model, $O(k.m.f^2 logf)$ can be seen as efficient, particularly for tasks where feature interpretability and model transparency are valued.

In brief, the proposed algorithm strikes a balance between computational efficiency and interpretability, making it suitable for scenarios where understanding feature importance and maintaining manageable computational costs are priorities.

6. The Proposed Model's Application to Real-World Problems

In Malaysia, there is a threat related to binge drinking among the youngsters of the Orang Asli Community. Research has been conducted by the University of Malaya Centre for Addiction Sciences (UMCAS) in collaboration with Orang Asli Gombak Hospital (HOAG) and it was noticed that there are few limitations with regard to the current alcohol-related brief interventions (Win et al., 2022b). Currently, there are no available technology-based interventions to support alcohol use disorder prevention within this community.

Henceforth, through the proposed solution which automates the feature selection process, by efficiently identifying the most relevant predictors of alcohol addiction, to make data-driven decisions, healthcare experts can ensure the chosen number of features is backed by empirical evidence rather than intuition or arbitrary choices. The proposed algorithm is designed as an automatic selection process; therefore, it can dynamically adapt to different datasets or changes in a given dataset, providing a tailored solution. This ensures that the model trained with the existing student dataset can also be implemented for the real-life dataset of HOAG when available.

With regard to the expected results, healthcare specialists no longer need to manually sift through vast amounts of data to determine which features are important. This saves time and resources, allowing them to quickly develop and implement intervention strategies without extensive manual data analysis. By leveraging robust machine learning techniques to identify the most relevant features, healthcare workers can trust that the models they use are based on accurate and comprehensive data. This confidence in the model's reliability ensures that interventions are both evidence-based and effective, leading to better patient outcomes. All in all, the proposed algorithm can be embedded in the hospital management system of HOAG to assist in the brief intervention for the Orang Asli Community.

7. Conclusion

In summary, this paper achieved the goal of successfully identifying the optimal feature subsets for alcohol addiction classification and also the goal of improving the predictive power of a kernel-based classifier (SVM) in addiction prediction. The proposed feature engineering method leverages RReliefF which marks a shift from traditional feature importance ranking to automatic feature selection, due to which 11 influential factors for alcohol addiction were discovered from the given dataset. The proposed method addresses the arbitrary feature selection issue related to the traditional Relief algorithm, enhancing technical correctness by empirically validating feature subsets. Furthermore, the research involved training three different machine learning algorithms (Linear, Kernel-based and Ensemble-based) to predict alcohol addiction among adolescents, with the SVM coupled with the enhanced RReliefF algorithm proving to be the best-performing model. These insights have significant implications for the development of effective interventions and support systems aimed at addressing alcohol addiction among adolescents.

The practical advantages typical of the enhanced RReliefF algorithm combined with SVM, including interpretability, computational efficiency, and dynamic adaptation, make it a better choice for addiction prediction tasks, especially in real-world scenarios where model interpretability and computational resources are important considerations.

In terms of future work, a more balanced dataset could be included to improve the quality of the employed models and to better identify the influential factors with a view to further improving its generalisation ability.

Acknowledgements

This research was supported by the Impact-Oriented Interdisciplinary Research Grant Programme (IIRG) of Universiti Malaya (IIRG005B-2020SAH).

REFERENCES

Arshad, M. R. M., Omar, M., & Shahdan, N. A. (2015) Alcoholism among youth: A case study in Kuala Lumpur, Malaysia. *International Journal of Culture and History*, 1(1), 21-28.

Arshad, M., Shahriar, S. & Anjum, K. (2023) The Power Of Simplicity: Why Simple Linear Models Outperform Complex Machine Learning Techniques - Case Of Breast Cancer Diagnosis. To be published in *Journal of Machine Learning Research*. [Preprint] https://doi.org/10.48550/arXiv.2306.02449 [Accessed: 1st December 2024].

Brunborg, G. S., von Soest, T., & Burdzovic Andreas, J. (2021) Adolescent income and binge drinking initiation: prospective evidence from the MyLife study. *Addiction*. 116(6), 1389-1398.

Cai, J., Luo, J., Wang, S. & Yang, S. (2018) Feature selection in machine learning: A new perspective. *Neurocomputing*. 300, 70-79. doi: 10.1016/j.neucom.2017.11d077.

Calvo, E., Allel, K., Staudinger, U., Castillo-Carniglia, Á., Medina, J. & Keyes, K. (2020) Cross-country differences in age trends in alcohol consumption among older adults: a cross-sectional study of individuals aged 50 years and older in 22 countries.. *Addiction*. 116(6), 1399-1412. doi: 10.1111/add.15292.

Chatzimparmpas, A., Martins, R., Kucher, K. & Kerren, A. (2022) FeatureEnVi: Visual Analytics for Feature Engineering Using Stepwise Selection and Semi-Automatic Extraction Approaches. *IEEE Transactions on Visualization and Computer Graphics*. 28(4), 1773-1791. doi: 10.1109/TVCG.2022.3141040.

Chakroun, M., Bouhamed, S. A., Kallel, I. K., Solaiman, B. & Derbel, H. (2022) Feature selection based on discriminative power under uncertainty for computer vision applications. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*. 21(1), 111-120.

Chi, R., Lu, S., Zhang, N., Zhang, M., Guo, K., Du, S., Guo, J., Hu, X. & Ma, G. (2022) The Association Between Family Environment and Adolescent Alcohol Drinking Behavior: A Cross-Sectional Study of Six Chinese Cities. *Frontiers in Nutrition*. 9, 903216. doi: 10.3389/fnut.2022.903216.

Cortez, P. & Silva, A. M. G. (2008) Using data mining to predict secondary school student performance. In: Brito, A. and Teixeira, J. (eds.) *Proceedings of 5th Annual Future Business Technology Conference (FUBUTEC 2008)*, *9-11 April 2008, Porto, Portugal*. Oostende, EUROSIS-ETI. pp. 5-12.

Edmonds, A., Rhew, I., Jones-Smith, J., Chan, K., Castro, A., Rubinsky, A., Blosnich, J. & Williams, E. (2022) Neighborhood Disadvantage, Patterns of Unhealthy Alcohol Use, and Differential Associations by Gender, Race/Ethnicity, and Rurality: A Study of Veterans Health Administration Patients. *Journal of studies on alcohol and drugs*. 83(6), 867-878. doi: 10.15288/jsad.21-00110.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V. & Gorban, A. N. (2017) The five factor model of personality and evaluation of drug consumption risk. In: Palumbo, F., Montanari, A. and Vichi, M. *Data science: innovative developments in data analysis and clustering.* Cham, Switzerland, Springer International Publishing, pp. 231-242.

Haardörfer, R., Windle, M., Fairman, R. & Berg, C. (2020). Longitudinal changes in alcohol use and binge-drinking among young-adult college students: Analyses of predictors across system levels. *Addictive Behaviors*. 112, 106619. doi: 10.1016/j.addbeh.2020.106619.

Ismail, S., Azlan, N. I. A. N. & Mustapha, A. (2018) Prediction of alcohol consumption among Portuguese secondary school students: A data mining approach. In: *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), 28-29 April 2018, Penang, Malaysia*. IEEE. pp. 383-387.

Kast, K. A. & Avery, J. D. (2019) Medical students and substance use disorders. In: Zappetti, D. & Avery, J. D. (eds.) *Medical Student Well-Being: An Essential Guide*. Berlin, Germany, Springer, 59-71.

Kendler, K., Ohlsson, H., Fagan, A., Lichtenstein, P., Sundquist, J. & Sundquist, K. (2020) Nature of the Causal Relationship Between Academic Achievement and the Risk for Alcohol Use Disorder. *Journal of Studies on Alcohol and Drugs*. 81(4), 446-453. doi: 10.15288/jsad.2020.81.446.

Kinreich, S., McCutcheon, V., Aliev, F., Meyers, J., Kamarajan, C., Pandey, A., Chorlian, D., Zhang, J., Kuang, W., Pandey, G., Viteri, S., Francis, M., Chan, G., Bourdon, J., Dick, D., Anokhin, A., Bauer, L., Hesselbrock, V., Schuckit, M., Nurnberger, J., Foroud, T., Salvatore, J., Bucholz, K. & Porjesz, B. (2021) Predicting alcohol use disorder remission: a longitudinal multimodal multi-featured machine learning approach. *Translational Psychiatry*. 11(1), 166. doi: 10.1038/s41398-021-01281-2.

Mamikonyan, N., Abazyan, S. & Janpoladov, V. (2020) Multi-Memory Grouping Wrapper with Top Level BIST Algorithm. *Open Access Library Journal. 7, e6294.* doi: 10.4236/oalib.1106294.

Mathar, R., Alirezaei, G., Balda, E. & Behboodi, A. (2020) Support Vector Machines. *In: Fundamentals of Data Analytics with a View to Machine Learning.* Springer Nature Switzerland AG, pp. 83-105.

Nikhitha, N. K., Afzal, A. L. & Asharaf, S. (2021) Deep Kernel machines: a survey. *Pattern Analysis and Applications*. 24, 537-556. doi: 10.1007/s10044-020-00933-1.

Olorisade, B., Brereton, P. & András, P. (2022) Structural Complexity and Performance of Support Vector Machines. In: *2022 International Joint Conference on Neural Networks (IJCNN), 18-23 July 2022, Padua, Italy*. IEEE. pp. 1-8.

Palaniappan, S., Hameed, N. A., Mustapha, A. & Samsudin, N. A. (2017) Classification of alcohol consumption among secondary school students. *JOIV: International Journal on Informatics Visualization*. 1(4-2), 224-226. doi: 10.30630/joiv.1.4-2.64.

Ruiz-España, S., Ortiz-Ramón, R., Pérez-Ramírez, Ú., Díaz-Parra, A., Ciccocioppo, R., Bach, P., Vollstädt-Klein, S., Kiefer, F., Sommer, W. H., Canals, S. & Moratal, D. (2023) MRI texture-based radiomics analysis for the identification of altered functional networks in alcoholic patients and animal models. *Computerized Medical Imaging and Graphics*. 104, 102187.

Sarić, R., Jokić, D., & Čustović, E. (2019) Identification of Alcohol Addicts Among High School Students Using Decision Tree Based Algorithm. In: Badnjevic, A., Škrbić, R. & Gurbeta Pokvić, L. (eds.) *CMBEBIH 2019* (*IFMBE Proceedings*, vol. 73). Cham, Switzerland, Springer. doi: 10.1007/978-3-030-17971-7_69.

Shukla, A. K., Singh, P. & Vardhan, M. (2018) Predicting alcohol consumption behaviors of the secondary level students. In: *Proceedings of the 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), 26-27 March 2018, Jaipur, India.* pp. 26-27.

Singh, A., Singh, V., Gourisaria, M. & Sharma, A. (2022) Alcohol Consumption Rate Prediction using Machine Learning Algorithms. In: *2022 OITS International Conference on Information Technology (OCIT)*, *14-16 December 2022, Bhubaneswar, India.* IEEE. pp. 85-90.

Stokłosa, I., Więckiewicz, G., Stokłosa, M., Piegza, M., Pudlo, R., & Gorczyca, P. (2023) Medications for the Treatment of Alcohol Dependence - Current State of Knowledge and Future Perspectives from a Public Health Perspective. *International Journal of Environmental Research and Public Health*. 20(3), 1870. doi: 10.3390/ijerph20031870.

Stüben, N., Franke, A. G. & Soyka, M. (2023) Evaluation of a primary E-health intervention for people with alcohol use disorder: clinical characteristics of users and efficacy. *International Journal of Environmental Research and Public Health*. 20(15), 6514. doi: 10.3390/ijerph20156514.

Symons, M., Feeney, G., Gallagher, M., Young, R. & Connor, J. (2020) Predicting Alcohol Dependence Treatment Outcomes: A Prospective Comparative Study of Clinical Psychologists vs 'Trained' Machine Learning Models. Addiction. 115(11), 2164-2175. https://doi.org/10.1111/add.15038.

Uddin, M. F., Lee, J., Rizvi, S., & Hamada, S. (2018). Proposing enhanced feature engineering and a selection model for machine learning processes. *Applied Sciences*. 8(4), 646. doi: 10.3390/app8040646.

Verhaeghe, J., Donckt, J., Ongenae, F. & Van Hoecke, S. (2022) Powershap: A Power-full Shapley Feature Selection Method. To be published in *Journal of Machine Learning Research*. [Preprint] https://doi.org/10.48550/arXiv.2206.08394 [Accessed: 1st December 2024].

Wang, H., Hu, R., Zhong, J., Du, H., Fiona, B., Wang, M. & Yu, M. (2018) Binge drinking and associated factors among school students: a cross-sectional study in Zhejiang Province, China. *BMJ open*. 8(4), e021077.

Win, M. N., Han, L. W., Kumar, E. K. K. S. C., Keat, T. Y. & Ravana, S. D. (2022a) AI-based personalized virtual therapist for alcohol relapse. *Enthusiastic: International Journal of Applied Statistics and Data Science*. 2(2), 82-96.

Win, M., Ravana, S., Nasiruddin F., & Mohammad, A. (2022b). Computerized Clinical Decision Support System: Adaptive Intervention in Controlling Alcohol Addiction among Indigenous in Malaysia. In Proceedings of the Seminar on Information Retrieval and Knowledge Management 2022 (SIRKM'22), 2-3 March 2022, Virtual venue. Bangi, Universiti Kebangsaan Malaysia. pp. 48-51.

Zhao, D., Yang, J., Liu, H., Huang, K. & Li, X. (2021) A study of individual identification of radiation source based on feature extraction and deep learning. *Journal of Physics: Conference Series* (*2nd International Conference on Computer Vision and Data Mining (ICCVDM 2021)*, *20-22 August 2021, Changsha, China*). 2024, 012072. doi: 10.1088/1742-6596/2024/1/012072.

Zhang, Y., Tiňo, P., Leonardis, A. & Tang, K. (2020) A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 5(5), 726-742. doi: 10.1109/TETCI.2021.3100641.

Zhu, T. (2020) Analysis on the Applicability of the Random Forest. *Journal of Physics: Conference Series*. 1607(1), 012123. doi: 10.1088/1742-6596/1607/1/012123.