Applying eXplainable AI Techniques to Interpret Machine Learning Predictive Models for the Analysis of Problematic Internet Use among Adolescents

Aleksandar S. Stanimirovic[1,*], Mina S. Nikolic[1], Jelena J. Jovic[2], Dragana I. Ignjatovic Ristic[3,4], Aleksandar M. Corac[2], Leonid V. Stoimenov[1], Zoran H. Peric[1]

[1]Faculty of Electronic Engineering, University of Nis, Aleksandra Medvedeva 14, 18000 Nis, Serbia

[2]Faculty of Medicine, University of Pristina in Kosovska Mitrovica, Kosovska Mitrovica, Serbia

[3]Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia

[4]Psychiatric Clinic, Clinical Center "Kragujevac", Kragujevac, Serbia

[*]aleksandar.stanimirovic@elfak.ni.ac.rs; mina.nikolic@elfak.ni.ac.rs; jelena.jovic@med.pr.ac.rs; draganaristic4@medf.kg.ac.rs; aleksandar.corac@med.pr.ac.rs; leonid.stoimenov.elfak.ni.ac.rs; zoran.peric@elfak.ni.ac.rs

***Abstract*—This research focusses on the potential application of artificial intelligence (AI) techniques in the analysis of behavioural addictions, specifically addressing problematic Internet use among adolescents. Using tabular data from a representative sample from Serbian high schools, the authors investigated the feasibility of employing eXplainable AI (XAI) techniques, placing special emphasis on feature selection and feature importance methods. The results indicate a successful application to tabular data, with global interpretations that effectively describe predictive models. These findings align with previous research, which confirms both relevance and accuracy. Interpretations of individual predictions reveal the impact of features, especially in cases of misclassified instances, underscoring the significance of XAI techniques in error analysis and resolution. Although AI's influence on the medical domain is substantial, the current state of XAI techniques, although useful, is not yet advanced enough for the reliable interpretation of predictions. Nevertheless, XAI techniques play a crucial role in problem identification and the validation of AI models.**

***Index Terms*—Artificial intelligence; Machine learning; Medical services; Addiction.**

I. INTRODUCTION

Artificial intelligence (AI) has become an inevitable factor in many areas of human society. AI offers solutions that have led to accelerated transformation in areas such as finance, transport, education, and especially medicine. In the field of medicine, artificial intelligence has gained importance through various applications, such as applications for image analysis, natural language processing, predictive analysis, and others [1]–[5]. AI helps medical professionals in better diagnosis, treatment planning, drug discovery, and better planning and resource allocation.

However, with more frequent use and increased complexity of the models used, such as machine learning (ML) models and deep learning models, numerous questions and doubts have arisen related to understanding decision-making process of these models [6]. The "black box" paradigm is often used to describe most AI models. The metrics used to assess the quality of AI models lack the ability to describe the internal functioning of these models. This has led to the development of a distinct field of artificial intelligence known as eXplainable AI (XAI). XAI is tasked with providing explainability and interpretability of AI models.

In recent years, XAI has gained increasing importance, especially in key areas such as medicine. The application of AI in medicine usually involves the use of complex models whose decision-making process is difficult to understand and interpret. In addition, the "black box" nature of these models is extremely critical to their application in clinical conditions, as they cannot provide the necessary explanations to medical experts [7]. In such cases, the existence of XAI is crucially important. XAI should provide medical professionals with the ability to understand the logic that exists behind the decisions proposed by artificial intelligence models, which would significantly increase confidence in those models and ensure that patient care would remain the shared responsibility of AI and humans [8].

XAI methods are highly dependent on the type of data [9]. For images, techniques are usually used to highlight parts of the image or select pixels that participate in the decision-making process. For textual data, techniques for extracting

Manuscript received 17 December, 2024; accepted 19 March, 2024.

This work has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia.

relevant features from semistructured text are used. For tabular data, XAI techniques are usually based on determining and comparing the relevance of different features.

Due to the rapid development of deep neural networks, many applications based on these models are being developed in the medical field for processing medical images or unstructured textual data [10]. However, the importance of tabular data in medicine and the need for its analysis should not be ignored. Patient records, laboratory results, and clinical data are often organised and stored in tabular format [5]. Therefore, AI models used for tabular data require appropriate XAI techniques. These methods help healthcare providers understand the variables that influence AI predictions, facilitating better-informed decisions and personalised patient care plans [11], [12].

The focus of this research is the possibility of applying different machine learning techniques to address behavioural addictions. The group of authors reported a limited number of articles (less than 0.25 %) dealing with the interdisciplinary field of machine learning and addiction [2]. Only a small subset of identified articles addresses the issue of behavioural addictions. Specifically, the authors of this paper address explainable artificial intelligence (XAI) approaches in the context of behavioural addictions. Given that data in this domain, mostly originated from electronic health records and surveys, is tabular, our focus is on XAI techniques to interpret and enhance the analysis of tabular data in behavioural addiction applications.

II. BACKGROUND

Explainable AI is not a new field since, in expert systems of the 1980s, there were reasoning architectures to support an explanation function for complex AI systems [13]. Trust is intricately linked with explainability, since the level of trust in an AI system depends on the visibility a human has in its operations. Consequently, AI algorithms should provide understandable justifications for their outputs, offering insights into the inner workings of the AI system.

In the literature, various terms such as "understandability" ("intelligibility"), "comprehensibility", "interpretability", and "transparency" are interchangeably used in the context of eXplainable AI (XAI) [14]. Transparency in an AI model implies inherent self-understandability, while interpretability involves explaining the meaning of the model in understandable terms to a human. Understandability, a most essential concept within XAI, denotes the characteristic of a model to make a human understand its function - how the model works - without the need to explain its internal structure or the algorithmic means by which the model processes data internally [15]. Both transparency and interpretability are strongly related to understandability: While transparency refers to the self-understandability of the model, understandability measures the degree to which a human can understand a decision made by a model [14]. Comprehensibility, associated with an algorithm's ability to represent learnt knowledge in a human-understandable manner [16], is interconnected with understandability, relying on the audience's ability to grasp the model's knowledge. One possible definition of XAI, emphasising the perspective of the end user [13], is provided in [17] as: "to create new or modified ML techniques that produce explainable models, enabling end users to understand, trust, and manage emerging AI systems effectively when combined with explanation techniques".

Based on the literature, the concepts of XAI within different application domains are categorised as [18]–[20]:

– *Stage of Explainability*: The stage of explainability refers to the phase of the AI process when a model generates the explanation for the decision it provides. According to [19], [20], the stages are as follows:

  – Ante-hoc methods: Involves generating the explanation for the decision from the very beginning of the training phase [18]. It can also be devided into premodelling and during modelling explainability [13], [21]. The goal of premodelling explainability is to understand and describe data used to develop models, whereas the goal for during modelling explainability is to develop inherently more explainable models. This stage is applicable to transparent (intrinsic) models like linear regression, logistic regression, k-nearest neighbour, rule-based learners, general additive models, Bayesian models, and decision trees [7].

  – Post-hoc methods: Comprise the external or surrogate models and the base model [18]. The base model remains unchanged, and the external model mimics its behaviour to generate explanations. Post-hoc methods are further classified as model-agnostic (applying to any AI/ML model) and model-specific (confined to models). These methods are associated with models that are perceived as "black box" models from the user's perspective, e.g., support vector machines and neural networks.

– *Scope of Explainability*: Defines the extent of an explanation produced by some explainable methods. Recent studies [19]–[21] distinguish between global and local scopes. The global scope makes the whole inferential technique of a model transparent or comprehensible to the user. On the other hand, local scope refers to explicitly explaining a single instance of inference to the user.

– *Input and Output formats*: Alongside core concepts, stages, and scopes, input and output formats are significant in the development of XAI methods [19], [20]. The mechanisms of explainable models unquestionably differ when learning different input data types, such as images, numbers, texts, etc. The most common forms of explanations are numeric, rules, textual, visual, and mixed.

As mentioned previously, the focus of this paper is on XAI techniques for interpreting and enhancing the analysis of tabular data derived from digital health records and surveys.

III. PROBLEMATIC INTERNET USE

The authors have already highlighted the limited research in the multidisciplinary area of medicine and AI on behavioural addictions. This paper focusses specifically on the application of AI and XAI methods in addressing behavioural addictions, with a focus on problematic Internet use. The term "Internet addiction" and "problematic Internet use" (PIU) are used to describe patterns of Internet use marked by excessiveness, loss of control, neglect of other essential activities, and continued engagement despite adverse consequences, including distress and functional

impairment [22].

Adolescents and young adults, as a particularly sensitive demographic, are increasingly facing very real and sometimes severe consequences in their daily lives arising from inadequate Internet use [23], [24]. This demographic group, including adolescents and young adults, represents a significant fraction of Internet users and is the group with the highest risk of developing PIU within the general population. Adolescents, the most exposed population, are particularly prone to various types of addiction, and constitute the demographic that engages the most with the Internet [25], [26].

The exploration of how specific Internet activities and content may both decrease and increase the risk of developing PIU is particularly interesting. Monitoring the development of the PIU by analysing the impact of various online activities, the content that adolescents engage in, their daily habits, and the potential influence of temperament poses a considerable challenge. To address this, a nationally representative sample of adolescents was formed to monitor the development of PIU. Various analytical methods were used to fully analyse the database of samples collected. Conventional data analysis approaches sometimes proved insufficient. Consequently, the application of ML methods emerges as a solution to overcome these limitations, facilitating a more robust identification of target groups for interventions.

To create a nationally representative sample of adolescents, the research focussed on high school students aged 16 and 17. Stratification occurred first by regions, then by cities within regions, and finally by schools within cities. Within schools, a random sample, determined by the number of classes, decided the participating departments. The final stratified proportional sample comprised 48 high schools, representing approximately 10 % of all high schools in the Republic of Serbia. The final sample included 2113 adolescents, out of the 2239 initially surveyed, with 56 % being female and 44 % male, and an average age (mean) of 16.73.

A four-part questionnaire was created for the research participants, covering the following aspects [23], [27], [28]:

– Sociodemographic characteristics: This section collected information on gender, age, satisfaction with socioeconomic status, and academic achievement.

– Intensity of Internet use: The frequency and duration of Internet use.

– Internet content and types of online activities: Participants provided details on the content they were interested in (e.g., politics, sports, music, pop culture, pornography) and the types of online activities they engaged with (e.g., email, chat, social networks, online games).

– Habits: This section explored the habits of the participants, including engagement in sports, fast food consumption, alcohol consumption, intake of energetic drinks, coffee consumption, and smoking.

The questionnaires were used to rate participants according to the Internet use disorder scale (IUDS) [27]. The IUDS included 18 items that participants rate according to a five-level Likert scale (from {1 - minimally} to {5 - completely}), including the questions related to compulsive Internet use, the symptoms of abstinence, and increased tolerance, as well as the questions related to the problems at work and school (Kronbah's alpha coefficient $\alpha = 0.815$). According to the scores of the subjects on the scale (cut-off 39/40), they were divided into two groups: those with and without PIU.

Furthermore, TEMPS-A for adolescents (A TEMPS-A) was used to determine the temperament of the participants. TEMPS-A is a self-evaluation questionnaire that determines which type of affective temperament is the following: depressive, cyclothymic, hyperthymic, irritable, and anxious [28]. According to the definition of affective temperament, the hyperthymic temperament is released from the depressive characteristics, and vice versa, the depressive temperament does not contain any of the hyperthymic components. A TEMPS-A is a completely new version of the scale adjusted to the age of adolescents [27].

IV. Data Preprocessing

The analysis of the adolescent Internet use sample (PIU data set) involved descriptive statistical methods. Descriptive statistics, such as measures of central tendency (arithmetic mean), measures of variability (standard deviation), and structural indicators expressed as percentages, were used. The PIU data set was exported as a text file containing comma-separated values.

The next step focussed on preparing the PIU data set for the ML algorithms. All data preprocessing is performed with Python 3.10 using the scikit-learn library [29]. Initially, features with missing values (reflecting participants who avoided or omitted responses to certain parts of the questionnaire) were identified. To address this, records with missing class-label attributes (approximately 5 % of the data set) were removed. The remaining missing values were treated as missing at random (MAR) [30] and imputed using the imputation of k-nearest neighbours (kNN). KNN imputation involves finding the k most similar records and imputing the missing value with a summary metric from those k records. In this case, Euclidean distance was used to measure, with a value of k set at 10. The mean value of the feature was used for the imputation.

The data set was normalised using the standard scaler, ensuring that each feature had a mean of zero and a standard deviation of one [30]. During preprocessing, the isolation forest approach was used to identify outliers. Identified outliers were removed, leaving 1908 records in the PIU data set.

For the remaining records, it was determined that the distribution of the class label feature was skewed. Clearly, in such scenarios, the metrics of the ML model may become biased. To address the imbalanced data set, the resampling approach was used. There are two types of methods that can be used: undersampling and oversampling. In most scenarios, oversampling is a preferred technique because undersampling may result in the removal of records carrying valuable information. For the resampling of the PIU data set, the synthetic minority oversampling technique (SMOTE) [31], [32], implemented in the imbalance-learn library [33], was chosen. SMOTE is an oversampling technique in which synthetic samples are generated for the minority class, achieving a binary class distribution of 1:1. These synthetic samples are created using the kNN algorithm to identify k

neighbours, which are then used to interpolate new synthetic instances.

V. METHODOLOGY

After completing the data preparation process, various feature selection and feature importance techniques described in this paper were applied to the prepared PIU data set. The entire implementation is based on the use of the Python 3.10 programming language and the scikit-learn library, and it is available on the public GitHub repository [34]. Due to the random nature of different machine learning algorithms, results may vary from case to case. Therefore, the results of the execution of the code in this paper were recorded in the form of an HTML file, which is an integral part of the GitHub repository.

An overview of the XAI methods used is presented in Table I. In this research, eight different XAI techniques were considered. These techniques are divided into two main groups based on their approach: feature selection techniques and feature importance techniques. While feature selection techniques are categorised as ante-hoc methods, all feature importance techniques are considered post-hoc methods. Feature selection techniques offer an interpretation of the entire model (global scope), whereas certain feature importance techniques also provide a local interpretation of individual predictions (local scope). Most techniques, excluding tree-based classifiers, are model-agnostic, implying that their model interpretations are not dependent on the specific model itself. As mentioned earlier, given the nature of the PIU data set, the emphasis is placed on techniques that facilitate working with tabular data.

TABLE I. THE RESULTS OF THE ANALYSIS OF GLOBAL FEATURE IMPORTANCE VALUES.

	Method	Selection type	Stage	Interaction	Scope
Feature selection	Univariate feature selection	filter	ante-hoc	model agnostic	global
	Lasso	embedded	ante-hoc	model agnostic	global
	ElasticNET	embedded	ante-hoc	model agnostic	global
	Genetic selection	wrapper	ante-hoc	model agnostic	global
Feature importance	Tree classifiers	N/A	post-hoc	model specific	global
	LIME	N/A	post-hoc	model agnostic	local
	SHAP	N/A	post-hoc	model agnostic	global and local
	Permutation feature importance	N/A	post-hoc	model agnostic	global and local

The selection of univariate features involves selecting the most relevant features based on univariate statistical tests. It is a variant of the correlation-based feature selection method. In the case of the PIU data set, the analysis of variance (ANOVA) F-test was used [35]. This test computes the F-value for each feature in relation to class-label, trying to identify the most relevant features for predicting the class-label feature. Features that are highly dependent on the class-label will have high scores.

Both Lasso (L1 regularisation) and ElasticNET (a combination of L1 and L2 regularisation) regressions from the scikit-learn library were used for feature selection in the PIU data set. The coefficients of both regressions represent the linear relationship between the features and the class-label. A larger absolute value of a coefficient indicates a stronger effect of the corresponding feature on the class-label. The sign of a coefficient indicates the direction of the effect: positive for positive correlation, negative for negative correlation. Coefficients with a value of zero indicate that the corresponding features are not relevant to the model.

The sklearn-genetic-opt library [36] offers an implementation of wrapper techniques for feature selection based on an evolutionary approach. These techniques enable partition of data set features into two sets: one comprising features influencing class-label prediction and the other containing features with no impact on class-label prediction. These techniques were applied to train the RandomForest model for classification on the PIU data set.

Decision tree-based estimators already have embedded mechanisms to compute the importance of features. The relative rank (i.e., depth) of a feature used as a decision node in a tree can be used to assess the relative importance of that feature with respect to the predictability of the target variable. The features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. Therefore, the expected fraction of the samples they contribute to can be used as an estimate of the relative importance of the features.

For the PIU data set, various classifiers from the scikit-learn library were used to create predictive models: DecisionTreeClassifier, RandomForestClassifier, and GradientBoostingClassifier as ensemble methods, along with AdaBoostClassifier and BaggingClassifier as ensemble metaestimators. Additionally, the XGBoost implementation of the gradient boosting tree model was used [37]. To create prediction models, the PIU data set was divided into training and test sets in a 90:10 ratio (stratified split with a shuffle). Hyperparameter optimisation was performed for all models using the grid search with the cross-validation method, and the validation of the obtained models was performed using the k-fold cross-validation with the value of k set to 10 and stratified shuffle splits. Folds are made by preserving the percentage of samples for every class. The trained models were evaluated using standard classification metrics, including accuracy, precision, recall, and F1-measure (Fig. 1).

All models, except the BaggingClassifier model, support the computation of feature importance. For the BaggingClassifier, the computation of feature importance is implemented based on the mean of feature importance in each individual estimator that comprises the model. Due to the random split of cross-validation folds, results from validation tend to vary. However, given the minimal margin in results across various tree models, the feature importance of the PIU data set is calculated as the mean of the importance they have in each of the individual models (Fig. 2).

For the analysis of global and local interpretation results, we used a RandomForest classifier. The classifier was trained using the PIU data set training set and accuracy assessment was performed on the test set. The classifier achieved an estimated accuracy of around 76 %, and the complete

classification metrics are displayed in Fig. 3

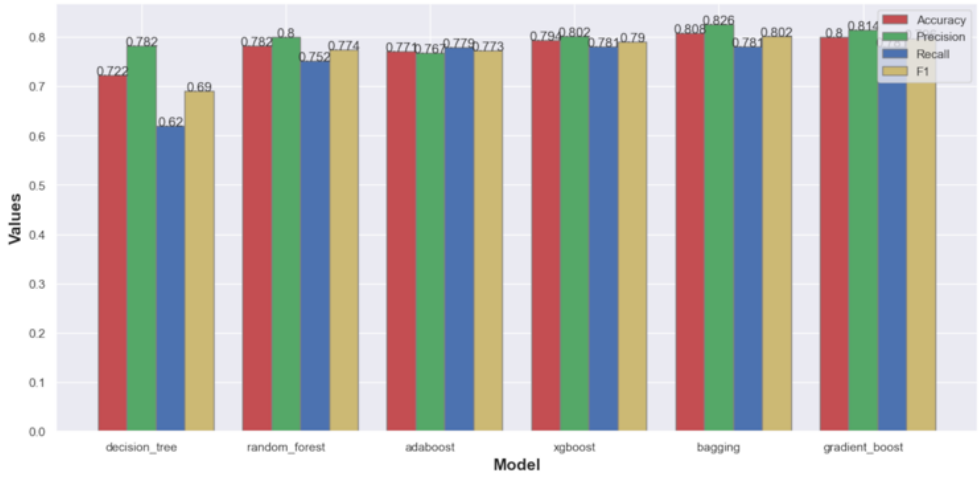

Fig. 1. Comparison of classification metrics for different tree classifiers.

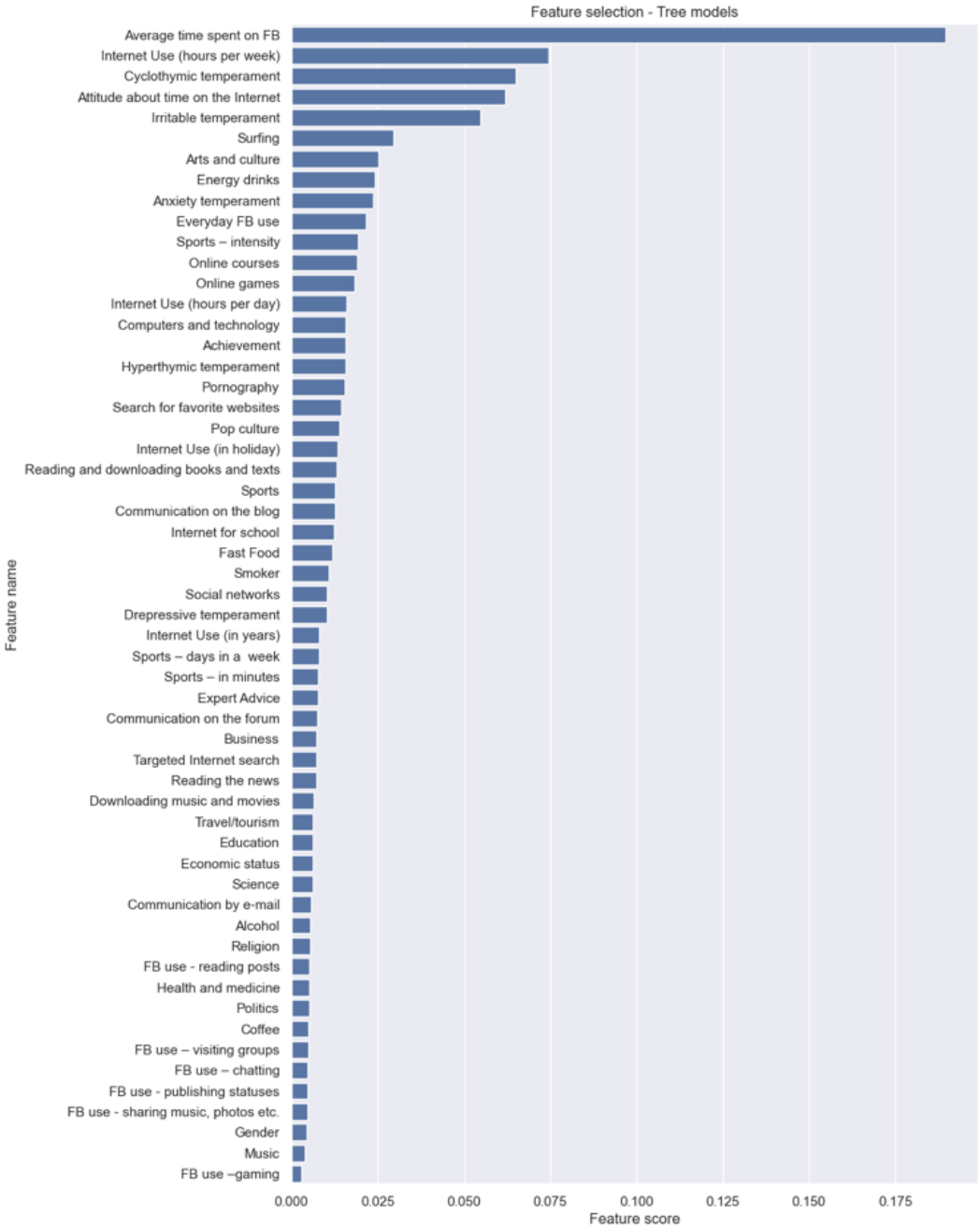

Fig. 2. Feature importance results using tree classifiers.

```
Test accuracy: 0.74
Confusion Matrix:
[[46  9]
 [16 25]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.74      0.84      0.79        55
         1.0       0.74      0.61      0.67        41

    accuracy                           0.74        96
   macro avg       0.74      0.72      0.73        96
weighted avg       0.74      0.74      0.74        96
```

Fig. 3. Classification metrics for a RandomForest classifier.

Local interpretable model-agnostic explanations (LIME) [38] offers local interpretability for any supervised ML algorithm. We decided to evaluate LIME against the predictions generated by the RandomForest classifier model. LIME is designed to attribute a model's prediction to human-understandable features. It is important to note that LIME does not provide global interpretability for ML models. If we require global interpretability using LIME, we need to run the explanation model on a diverse but representative set of instances from the data set to return a nonredundant explanation set that is a global representation of the model.

The shapley additive explanations (SHAP) method [39] works by computing the Shapley values [40] for each feature in the input space. The model can be applied to any supervised ML algorithm. Once the Shapley values are computed, various visualisation techniques can be used to gain insight into the decision-making process of the model. We decided to evaluate SHAP against predictions generated by the RandomForest classifier model.

Eli5 [41] is a Python library to inspect machine learning classifiers and explain their predictions. The library implements various XAI techniques, including permutation feature importance. In this paper, we leverage Eli5 to assess the effectiveness of the permutation feature importance method in the context of RandomForest classifier predictions.

VI. RESULTS

During this research, the authors of the paper treated feature selection techniques and feature importance techniques equally as eXplainable Artificial Intelligence (XAI) methods that can be utilised to explain and interpret AI models. However, some authors make a clear distinction between these two approaches, claiming that only feature importance techniques can be considered XAI techniques [42]. The literature often does not mention a direct connection between feature selection techniques and the interpretability of AI models. This connection is somewhat implicitly assumed, since a smaller number of features leads to the creation of simpler AI models, and there are studies that directly establish a link between the interpretability of AI models and their complexity [8]. The authors of this paper believe that this fact establishes a direct connection between feature selection techniques and XAI and that feature selection techniques can play a significant role in XAI.

The results of the analysis of the importance values of global features are presented in Table II. For each feature, we counted the number of occurrences in the top 10 features identified by each XAI method used, except for the LIME method, which does not support global feature importance. The top_n All methods column represents the number of occurrences of features in the results of all methods, top_n Feature selection the number of occurrences of features in the results of feature selection methods (Univariate feature selection, Lasso, ElasticNET, and Genetic algorithm), and top_n Feature importance column the number of occurrences of features in the results of feature importance methods (Tree classifiers, SHAP, and Permutation feature importance). Due to the random nature of data splits during model training and the random nature of the model itself, the obtained results vary between different runs. However, in more than 80 % of instances, the same features are consistently chosen, indicated by bold features in Table II. Although the order of features (based on the frequency of their selection) differs, the set of selected features remains consistent. This is significant because it shows that different XAI methods consistently choose a set of the most relevant features.

TABLE II. THE RESULTS OF THE ANALYSIS OF GLOBAL FEATURE IMPORTANCE VALUES.

Feature selection methods		**All methods**		**Feature importance methods**	
Feature	**top_n**	**Feature**	**top_n**	**Feature**	**top_n**
Internet Use (hours per week)	3	Internet Use (hours per week)	6	Internet Use (hours per week)	3
Average time spent on FB	3	Average time spent on FB	6	Average time spent on FB	3
Cyclothymic temperament	3	Cyclothymic temperament	6	Cyclothymic temperament	3
Attitude about time on the Internet	3	Attitude about time on the Internet	6	Attitude about time on the Internet	3
Everyday FB use	3	Everyday FB use	6	Everyday FB use	3
Surfing	3	Surfing	4	Surfing	3
Smoker	3	Irritable temperament	4	Irritable temperament	3
Internet Use (hours per day)	2	Internet Use (hours per day)	4	Internet Use (hours per day)	2
Anxiety temperament	2	Arts and culture	4	Social networks	2
Hyperthymic temperament	2	Online games	3	Online games	2
Achievement	2	Social networks	3	Politics	1
		Anxiety temperament	3	Anxiety temperament	1
		Smoker	3	Arts and culture	1
		Hyperthymic temperament	2	Energy drinks	1
		Achievement	2		

In addition to the significant overlap in the results obtained

by different methods, the results closely align with the expectations derived from previous research. It is evident that all features related to time spent on the Internet (Internet Use (hours per week), Internet Use (hours per day), Average time spent on FB), and the use of Internet services (social networks and Internet surfing) significantly influence the occurrence of PIU addiction. Interestingly, there is no overlap between the time spent on the Internet and the time spent on FB. Participants in the study may perceive the use of social networks as a separate activity from Internet use, not considering social networks as part of the Internet. This poses an interesting question for future research.

Another group of features that significantly influence PIU disorder are different types of affective temperament. The common neurobiological basis for the development of behavioural addictions and substance addiction is well-explained today. Specifically, in numerous studies, the cyclothymic temperament consistently exhibits the strongest association with substance addiction [43]. Additionally, our results show a strong influence of irritable temperament on the development of PIU addiction.

It is important to note that, unlike other methods, both Lasso and ElasticNET identified features with a negative impact on PIU addiction, suggesting their potential role as protective factors. Both methods have identified the same group of these features, including Sports and Hyperthymic temperament, for example. This opens an interesting question for future research regarding a possible preventive protective model.

For analysis, local interpretations of individual predictions were generated using LIME [38], SHAP [39], and permutation feature importance methods [41]. Figure 4 is an example of a local interpretation of a permutation feature importance method for a single instance from the PIU data set. The interpretation includes the contribution of each feature to the final prediction, distinguishing between features with a positive impact and those with a negative impact on a prediction.

The results of the local interpretation analysis are presented in Table III and are based on data from local interpretations generated using the permutation feature importance method. Similar results can be obtained using the LIME and SHAP methods as well. Table III contains information about each type of result from the confusion matrix, the number of instances belonging to that type of result, and the average number of features from the set of relevant features determined by global interpretations of the model, which were used during the prediction of these instances.

Based on the data from Table III, it is evident that the model generates predictions, whether accurate or inaccurate, based on features determined as relevant according to global interpretations of the model. In the case of misclassified instances, XAI models have proven crucial for further analysis. Figures 5 and 6 represent local interpretations generated by the LIME and SHAP methods, respectively, for an instance that was misclassified as a negative instance (false negative instance).

Figure 5 illustrates the graphical representation of the LIME interpretation for a misclassified (false negative) individual instance from the PIU data set. On the left side of the image, the prediction probabilities are displayed. Each possible prediction value is assigned a colour consistent throughout the image. In the central part of the graphical representation, the most relevant features for the obtained prediction are highlighted. For each feature, it indicates the prediction for which it is relevant and the numerical value of its relevance. On the right side of the image, a table displays the actual values of the most relevant features for the observed instance.

y=PIU yes (probability **0.735**) top features

Contribution[?]	Feature
+0.502	<BIAS>
+0.081	Internet Use (hours per week)
+0.072	Average time spent on FB
+0.026	Internet Use (hours per day)
+0.022	Health and medicine
+0.018	Surfing
+0.017	Search for favorite websites
+0.016	Anxiety temperament
+0.015	Pornography
+0.015	Social networks
	... 23 more positive ...
	... 9 more negative ...
-0.005	FB use – visiting groups
-0.006	Alcohol
-0.007	Sports
-0.008	Gender
-0.014	Irritable temperament
-0.014	Drepressive temperament
-0.016	Religion
-0.016	Business
-0.024	Attitude about time on the Internet
-0.033	Arts and culture

Fig. 4. Local interpretation of the permutation feature importance method for a single instance from the PIU data set.

TABLE III. THE RESULTS OF THE ANALYSIS OF LOCAL (SINGLE PREDICTION) INTERPRETATIONS.

Result type	No. of instances	Average usage of relevant features
True positive	25	5.04
True negative	46	4.80
False positive	9	4.89
False negative	16	3.94

Figure 6 illustrates the graphical representation of SHAP interpretation for the same misclassified instance from the PIU data set. The Shapley force plot is used to illustrate the contribution of each feature towards individual prediction. The features are ranked by importance, and the arrows illustrate how each feature contributes to the model's output for a specific prediction.

Based on these interpretations, it is evident that, although the model used relevant features for the prediction, their specific values were such that they led to an inaccurate prediction. This highlights the need for additional analysis of the collected data to identify potential shortcomings, such as processing errors, inaccurate information provided by participants during the original data collection, or the need to improve the questionnaire to achieve finer differences between research participants. XAI methods provide valuable information on misclassified results, providing a solid foundation for future research on PIU addiction.

As mentioned previously, there are a few studies that explore the potential of using machine learning (ML) techniques, in the field of behavioural addictions especially when it comes to employing XAI techniques for interpreting predictive ML models in this domain. The authors suggest that the findings of this study can be compared to similar

research conducted in the field of AI and ML models in medicine [7], [12], [21], [44], [45]. Since there is currently no universally accepted method for evaluating XAI techniques and their outcomes, the evaluation primarily relies on human interpretation.

Fig. 5. Local interpretation of LIME for a single false negative prediction.

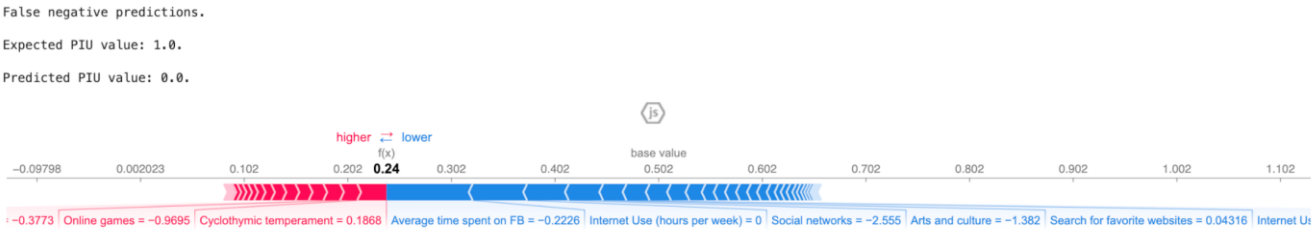

Fig. 6. SHAP local interpretation for a single false negative prediction.

One notable advantage of this research is that the ML and XAI techniques were applied to an aspect related to Internet use (PIU) among adolescents. Consequently, the study utilised raw data, which brings it closer to real-world scenarios where XAI techniques can be applied. The data set used to create ML models consisted of 56 features, with interpretations provided by XAI techniques. This number of features exceeds those used in other comparable studies, making the interpretation of obtained ML models more challenging, yet yielding more substantial results. Furthermore, the medical experts involved in the original study interpreted these results extensively, thus largely validating their findings. The lack of interpretation of the results of the XAI technique by experts in the domain is one of the limiting factors in research in this area.

VII. CONCLUSIONS

The main goal of this research is to investigate the possibility of applying AI techniques to the analysis of behavioural addictions. The paper focusses on the analysis of problematic Internet use among adolescents. The data set used was created as a result of research conducted among high school students in the Republic of Serbia.

As part of the research, various XAI techniques, with special considerations for techniques for tabular data, were used to interpret predictive AI models and their results. The results showed that feature selection and feature importance techniques were successfully applied for the interpretation of predictive AI models created based on tabular data. Specifically, a detailed analysis of global interpretations of predictive AI models showed that the results obtained well describe the models used. Furthermore, the results showed a large overlap with the results of the original PIU research, confirming their relevance and accuracy.

Additionally, the analysis of interpretations of individual predictions confirmed that the results coincide with the results of global interpretations, features that were identified as relevant in global interpretations were also found to be relevant in local interpretations. Furthermore, a detailed analysis of misclassified instances was performed with a special focus on false negative instances, which are of critical importance in the medical domain. XAI techniques offer interpretations of these results and offer a good starting point for analysing these errors and effectively solving them.

Future work aims to apply the described XAI methods with more complex AI models. Of particular interest is the possibility of reusing the described XAI techniques, or other similar techniques, to artificial neural network models and other types of data (text, image, etc.).

In the future, a great impact of AI on the field of medicine is expected. As for XAI techniques, regardless of their usefulness, at this moment they have not yet reached a sufficient level of maturity to guarantee the interpretation of AI models with complete certainty. XAI models themselves are not yet ready for direct application in clinical practice [46]. However, XAI techniques remain crucial for AI experts, aiding in the identification of typical problems and overall validation of AI models.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Z. Obermeyer and E. J. Emanuel, “Predicting the future — Big data, machine learning, and clinical medicine”, *The New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, 2016. DOI: 10.1056/NEJMp1606181.

[2] P. Cresta Morgado, M. Carusso, L. A. Alemany, and L. Acion, "Practical foundations of machine learning for addiction research. Part I. Methods and techniques", *The American Journal of Drug and Alcohol Abuse*, vol. 48, no. 3, pp. 260–271, 2022. DOI: 10.1080/00952990.2021.1995739.

[3] S. Rameshbabu and S. Ramakrishnan, "Machine learning approach for diagnosis and prognosis of cardiac arrhythmia condition using a minimum feature set and auto-segmentation-based window optimisation", *Elektronika ir Elektrotechnika*, vol. 29, no. 5, pp. 51–61, 2023. DOI: 10.5755/j02.eie.34357.

[4] T. Proscevičius, V. Raudonis, A. Kairys, A. Lipnickas, and R. Simutis, "Autoassociative gaze tracking system based on artificial intelligence", *Elektronika ir Elektrotechnika*, vol. 101, no. 5, pp. 67–72, 2010.

[5] H. Ozkan, G. Tulum, O. Osman, and S. Sahin, "Automatic detection of pulmonary embolism in CTA images using machine learning", *Elektronika ir Elektrotechnika*, vol. 23, no. 1, pp. 63–67, 2017. DOI: 10.5755/j01.eie.23.1.17585.

[6] E. Bernardo and R. Seva, "Affective design analysis of Explainable Artificial Intelligence (XAI): A user-centric perspective", *Informatics*, vol. 10, no. 1, p. 32, 2023. DOI: 10.3390/informatics10010032.

[7] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery", *Diagnostics*, vol. 12, no. 2, p. 237, 2022. DOI: 10.3390/diagnostics12020237.

[8] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission", in *Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730. DOI: 10.1145/2783258.2788613.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", in *Proc. of 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

[10] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)", *Computer Methods and Programs in Biomedicine*, vol. 226, art. 107161, 2022. DOI: 10.1016/j.cmpb.2022.107161.

[11] J. Chen, L. Song, M. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation", in *Proc. of the 35th International Conference on Machine Learning*, 2018, pp. 883–892.

[12] J. Duell, X. Fan, B. Burnett, G. Aarts, and S.-M. Zhou, "A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records", in *Proc. of 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021, pp. 1–4. DOI: 10.1109/BHI50953.2021.9508618.

[13] S. Chakrobartty and O. El-Gayar, "Explainable Artificial Intelligence in the medical domain: A systematic review", *AMCIS 2021 Proceedings*, 2021, pp. 1–10.

[14] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.

[15] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks", *Digital Signal Processing*, vol. 73, pp. 1–15, 2018. DOI: 10.1016/j.dsp.2017.10.011.

[16] A. Fernandez, F. Herrera, O. Cordon, M. J. del Jesus, and F. Marcelloni, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?", *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 69–81, 2019. DOI: 10.1109/MCI.2018.2881645.

[17] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program", *AI magazine*, vol. 40, no. 2, pp. 44–58, 2019. DOI: 10.1609/aimag.v40i2.2850.

[18] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks", *Applied Sciences*, vol. 12, no. 3, p. 1353, 2022. DOI: 10.3390/app12031353.

[19] G. Vilone and L. Longo, "Explainable Artificial Intelligence: A systematic review", 2020. arXiv: 2006.00093.

[20] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats", *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, 2021. DOI: 10.3390/make3030032.

[21] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension", *BMC Medical Informatics and Decision Making*, vol. 19, art. no. 146, 2019. DOI: 10.1186/s12911-019-0874-0.

[22] V. Starcevic *et al.*, "Problematic online behaviors and psychopathology in Australia", *Psychiatry Research*, vol. 327, art. 115405, 2023. DOI: 10.1016/j.psychres.2023.115405.

[23] J. Jovic *et al.*, "Internet use during coronavirus disease of 2019 pandemic: Psychiatric history and sociodemographics as predictors", *Indian Journal of Psychiatry*, vol. 62, suppl. 3, pp. S383–S390, 2020. DOI: 10.4103/psychiatry.indianJPsychiatry_1036_20.

[24] K. S. Young and M. Brand, "Merging theoretical models and therapy approaches in the context of Internet gaming disorder: A personal perspective", *Frontiers in Psychology*, vol. 8, p. 1853, 2017. DOI: 10.3389/fpsyg.2017.01853.

[25] D. Backović, M. Maksimović, and D. Stevanović, "Psychosocial risk factors and substance abuse in adolescents", *Vojnosanitetski Pregled*, vol. 64, no. 5, pp. 331–336, 2007. DOI: 10.2298/vsp0705331b.

[26] J. Jović *et al.*, "Problematic internet use among adolescents - Gender differences", *European Neuropsychopharmacology*, vol. 27, suppl. 4, p. S1080, 2017. DOI: 10.1016/S0924-977X(17)31879-5.

[27] J. Jović *et al.*, "The development of temperament evaluation of Memphis, Pisa, Paris, and San Diego-auto-questionnaire for adolescents (A-TEMPS-A) in a Serbian sample", *Psychiatria Danubina*, vol. 31, no. 3, pp. 308–315, 2019. DOI: 10.24869/psyd.2019.308.

[28] D. Hinić, S. H. Akiskal, K. K. Akiskal, J. Jović, and D. Ignjatović-Ristić, "Validation of the Temps-A in university student population in Serbia", *Journal of Affective Disorders*, vol. 149, nos. 1–3, pp. 146–151, 2013. DOI: 10.1016/j.jad.2013.01.015.

[29] F. Pedregosa *et al*., "Scikit-learn: Machine learning in Python", *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] S. García, J. Luengo, and F. Herrera, "Dealing with missing values", in *Data Preprocessing in Data Mining. Intelligent Systems Reference Library*, vol. 72. Springer, Cham, 2015, pp. 59–105. DOI: 10.1007/978-3-319-10247-4_4.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. DOI: 10.1613/jair.953.

[32] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data", *BMC Bioinformatics*, vol. 14, art. no. 106, pp. 1–16, 2013. DOI: 10.1186/1471-2105-14-106.

[33] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning", *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[34] Source code of XAI experiment for PIU addiction database. [Online]. Available: https://github.com/acast975/xai_addiction

[35] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed., John Wiley & Sons, 2012.

[36] sklearn-genetic-opt library documentation. [Online]. Available: https://sklearn-genetic-opt.readthedocs.io/en/stable/index.html#

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[38] M. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier", in *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[39] SHAP library: A game theoretic approach to explain the output of any machine learning model. [Online]. Available: https://github.com/shap/shap

[40] L. S. Shapley, "A value for n-Person Games", in *Contributions to the Theory of Games II*. Princeton University Press, Princeton, 1953, pp. 307–317. DOI: 10.1515/9781400881970-018.

[41] ELI5 library: A library for debugging/inspecting machine learning classifiers and explaining their predictions. [Online]. Available: https://github.com/TeamHG-Memex/eli5

[42] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models", *SN Applied Sciences*, vol. 3, art. no. 272, pp. 1–12, 2021. DOI: 10.1007/s42452-021-04148-9.

[43] L. Rovai *et al.*, "Opposed effects of hyperthymic and cyclothymic temperament in substance use disorder (heroin- or alcohol-dependent patients)", *Journal of Affective Disorders*, vol. 218, pp. 339–345, 2017. DOI: 10.1016/j.jad.2017.04.041.

[44] O. Daramola *et al.*, "Towards AI-enabled multimodal diagnostics and management of COVID-19 and comorbidities in resource-limited settings", *Informatics*, vol. 8, no. 4, p. 63, 2021. DOI: 10.3390/informatics8040063.

[45] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective", *BMC Medical Informatics and Decision Making*, vol. 20, art. no. 310, pp. 1–9, 2020. DOI: 10.1186/s12911-020-01332-6.

[46] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health-care", *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021. DOI: 10.1016/S2589-7500(21)00208-9.