Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2954

2025, 44(2) 93-116

Crowdsourced data leaking user's privacy while using anonymization technique

Naadiya Mirbahar [a, *], Kamlesh Kumar [b], Asif Ali Laghari [a], Mansoor Ahmed Khuhro [c]

[a] *Department of Computer Science, Sindh Madressatul Islam University, Karachi, Pakistan*

[b] *Department of Software Engineering, Sindh Madressatul Islam University, Karachi, Pakistan*

[c] *Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi, Pakistan*

[*] Corresponding author: Naadiya Mirbahar, Email: naadiya.khudabux@yahoo.com

Received: 04 August 2023, Accepted: 27 March 2025, Published: 01 April 2025

KEYWORDS

Big educational data

Classification

Privacy leakage

Anonymization

Machine learning

ABSTRACT

Due to the tremendous value embedded in big educational data, numerous research institutes have collected large volumes of student behavioral data. To fully utilize the underlying values, the collected data may be shared with third parties, such as worldwide intelligent data experts. However, this may pose privacy risks to data owners, even though the data collectors usually anonymize the data before crowdsourcing. To demonstrate that anonymization alone is insufficient to protect user privacy, we conducted an experimental study using offline and online behavioral traces collected through campus cards and smartphones. Our study demonstrates that a student's identity can be identified with high probability based on anonymized behavior payment traces. The analysis of results demonstrates that only ten features, i.e., Transmission Control Protocol (TCP), synchronization attempts, content length, downlink traffic, last acknowledgement packet delay, uplink traffic, cell ID, base station ID, day, hour (offline payment, time) day, hour, minute (online payment time), and point of sale ID (POS_ID) are sufficient to uniquely identify an individual. Five supervised standard learning algorithm classifiers have been utilized to predict the user identity i.e., Extra Tree, Bagging, Decision Tree, Nearest Neighbor (KNN), and Random Forest Tree classifiers. The evaluation results showed that the achieved accuracy reached 99.99%, 99.95%, 99.02%, 98.84%, and 99.56%, respectively.

1. Introduction

With the rapid growth of big data, educational data has garnered significant attention from research communities [1, 2]. The existing literature has revealed that the utilization of big educational data can help student management [3], student behavior monitoring [4], student performance improvement [5], providing mobile App recommendations [6], student mental health promotion [7], teaching skill enhancement [8], public health improvement [9], and friend circle identification [10].

In the area of big educational data, there are a few institutes that collect various educational big data, such as reality mining [11] and device analyzer [12]. Although data collectors have extracted valuable insights, such as location tracking [13] and behavioral diversity over time [14], their findings rely on sharing these datasets with third parties to maximize their utility. To fully utilize these big educational data, those data collectors usually share the collected datasets with third parties [15]. Although these datasets are anonymized before sharing, they may still pose privacy risks. One common method of data collection in this domain is crowdsourcing, which involves collecting large volumes of user-generated data from multiple individuals, often through online or institutional systems, for research and analysis [16]. In the context of this study, educational institutions collect behavioral data from students via campus cards

and smartphones, then anonymize and share these datasets with third parties, such as data analysts or researchers. Machine learning models can analyze behavioral patterns and successfully identify individuals, raising concerns about data security and the effectiveness of current anonymization techniques.

To illustrate that anonymized big educational data can still result in privacy leakage, we conducted experiments on an anonymized dataset. Furthermore, the risk of privacy leakage increases when different types of educational data are integrated. Privacy leakage occurs when personal or sensitive information is unintentionally exposed, even after anonymization. Anonymization is achieved by removing or modifying Personal Identifiable Information (PII) to prevent direct identification while maintaining the dataset's usability. Studies suggest that even in anonymized financial and behavioral datasets, identification remains a significant risk, particularly when attackers exploit a minimal set of distinguishing features. Identification occurs when attackers use external datasets and advanced machine learning techniques to cross-reference anonymized records with identifiable information, ultimately breaching user privacy. This demonstrates that anonymization alone is often insufficient to fully protect sensitive data from being traced back to its original owners. [17]. Furthermore, the integration of multiple data sources significantly increases the risk of privacy leakage [18]. For example, some universities in China have implemented new campus card systems that integrate traditional campus card functions with WeChat, the country's most widely used smartphone-based social network application. The data records generated through campus cards usually reflect the offline behaviors of a student, such as payments at the university cafeteria or campus grocery store. The data records from WeChat Pay reflect both offline and online behaviors.

Our experimental study demonstrates that these two anonymized educational datasets leak privacy to varying degrees and are complementary in terms of identity privacy leakage.

The contributions of this work are as follows:

- Illustrating privacy leakage issues associated with both offline and online data traces;
- Identifying key features that significantly contribute to privacy leakage, and;
- Conducting extensive experiments using various classification models to demonstrate the impact of privacy leakage.

The remainder of this paper is organized as follows. Section 2 provides a review of related work, outlining existing research on privacy risks in educational data. Section 3 presents the system architecture, detailing the framework used in this study. Sections 4 and 5 describe the dataset and feature selection process, respectively, highlighting the key attributes used for analysis. Section 6 examines the increased privacy leakage resulting from integrating multiple datasets. Section 7 discusses the experimental results, demonstrating the impact of machine learning models on re-identification risks. Section 8 provides an in-depth analysis and discussion, exploring their implications and the broader context of data privacy. Section 9 explores ethical considerations related to data privacy and responsible data sharing. Section 10 outlines the study's limitations and proposes directions for future research. Finally, Section 11 presents the conclusion, summarizing the key findings and their implications.

2. Related Work

Big educational data, as an active branch of big data [19], introduced new opportunities to monitor student activities, identify the deviant ideologies of students, and foster a safe learning environment. Besides, it helps improve the efficiency and effectiveness of student learning, as well as their knowledge retention [20]. It also provides new opportunities for the well-being of communities by improving public health and medicine [21]. Bienkowski et al. [22] proposed to apply data mining to analyze big educational data, improve the online learning system, and support management decisions.

Wang proposed to mine campus card data, refine campus services and business operations, and identify potentially poor students through payment history [23]. Chai et al. [24] analyzed the campus card data to support the school logistics management. Furthermore, Lu et al. [25] proposed an encounter-based model to discover offline social relationships through campus card data. Based on campus data evaluation, Wan [26] suggested that serious health problems exist in Chinese college students.

As machine learning (ML) models increasingly rely on sensitive data, privacy-preserving techniques have become essential to mitigate risks of data exposure. Various modern approaches have been developed to ensure confidentiality, security, and compliance while maintaining model performance. Modern techniques like differential privacy, federated learning, homomorphic encryption, secure multi-party computation (SMPC), and blockchain-based ML offer varying levels of security. Differential privacy adds

mathematical noise to training data or model outputs to prevent the re-identification of individual records [27], while federated learning Instead of sending raw data to a central server, ML models are trained locally on user devices, and only aggregated updates are shared for the strong privacy [28]. Homomorphic encryption enables computations on encrypted data, ensuring confidentiality, while SMPC allows multiple parties to collaborate without sharing raw data [29]. Despite advancements in privacy-preserving techniques, real-world threats persist, including membership inference attacks [30], adversarial poisoning in federated learning [31], and vulnerabilities that allow data reconstruction [32]. Additionally, A hybrid approach combining multiple privacy techniques is necessary to balance privacy, data utility, and computational feasibility in practical applications. Despite these advancements, real-world risks persist. Membership inference attacks can reveal whether specific data points were used in model training, potentially compromising privacy [33]. Adversarial attacks in federated learning can introduce malicious updates, leading to biased or manipulated models [34]. Data reconstruction risks remain a challenge, when attackers attempt to reverse-engineer anonymized information [35]. To achieve both privacy and utility, a multi-layered privacy-preserving approach is crucial for safeguarding data confidentiality while maintaining the effectiveness of machine learning models. However, there is an inherent trade-off between privacy protection and data utility in research and analytics. Higher data utility enhances model performance by providing more granular and detailed data, but this also increases the risk of re-identification and privacy breaches. On the other hand, implementing stronger privacy measures—such as differential privacy, k-anonymity, and data perturbation—can effectively mitigate data leakage risks but may lead to reduced data quality, lower model accuracy, and limited analytical insights. Striking a balance between these competing priorities is essential to ensure that data-driven innovations continue to deliver valuable insights without compromising user privacy. Sensitive information can be obtained through mining the pattern by the recipient [36]. A previous study has shown that anonymized financial data is not safe to release to third parties [37]. It has been suggested that only a few bits of an anonymized smartphone dataset are enough to identify individuals when it is linked with an external Netflix movie dataset [38].

The integration of different types of educational data may increase the risk of privacy leakage. For example, a few Chinese universities are deploying new campus card systems, which integrate the functionality of their original campus cards into WeChat, the most popular smartphone-based social application in China. The data records generated through campus cards usually reflect the offline behaviors of a student, such as payments at the university cafeteria or campus grocery store, while the data records from WeChat Pay reflect both offline and online behaviors of students. Our experimental study suggests that these two anonymized educational data can leak privacy to different extents and these two data are complementary in terms of identity leakage.

3. System Architecture

To evaluate whether anonymized data can still leak privacy, numerous standard machine learning methods have been utilized to infer identity privacy from offline and online student behavior data. The underlying design philosophy of this process is that privacy leakage can be confirmed if any of the machine learning methods can infer privacy information from the anonymized data. The whole evaluation process consists of four phases, i.e., a preprocessing phase, a training phase, a model validation phase, and a testing phase (Fig. 1).

The preprocessing phase aims to clean the data records with abnormal or missing values by removing them from the dataset. The identity information from each user is removed from both datasets, such as the International Mobile Subscriber Identity (IMSI), the International Mobile Equipment Identity (IMEI), and the account number associated with the campus card. Since it is high dimensional data various feature selection techniques were employed, including Embedded methods (Linear Regression, Random Forest, Ridge, LASSO, Stability Selection), the Wrapper method (Recursive Feature Elimination - RFE), and the Filter method (Univariate Selection), to optimize classification performance while minimizing redundancy and computational complexity, where 10 out of 63 are optimal features are selected. The feature selection results are presented in section 5. We also removed the rows and columns containing missing values by eliminating all of them. Furthermore, to maintain the statistical properties, the data associated with the students with less than 30 records per month are also removed from the dataset. The records from 250 students are randomly selected for evaluation.

In our work, we defined student behavior as daily activities. In addition, we converted the machine time format to the human-readable format, because humans usually do an activity in terms of the day, hour, and minute. Thus, we adopted day, hour, and minute extracted from the data as features The features are

converted to the numeric vector, each dimension of which is a numerical representation of the corresponding attribute because the chosen algorithm requires a numerical representation to facilitate processing. Usually, a data set is anonymized by eliminating the identity information. To verify whether campus card data and smartphone data are complementary, these two data sets are merged through the common data field, i.e. the phone number.

In the second phase, each data set and the integrated dataset were partitioned into two independent subsets, the training subset, and the test subset, with the ratio being 3:1. Numerous supervised machine learning classifiers were applied to anonymized datasets to assess their ability to predict user identities based on behavioral data. The classifiers were trained on selected features extracted from smartphone usage and campus card transaction data. To evaluate their performance, the models underwent rigorous testing, including cross-validation and accuracy assessment. The experimental results, presented in Section 7, demonstrate the effectiveness of these classifiers in identity prediction, with high classification accuracy confirming the potential risks of privacy leakage despite data anonymization.

Among the ten selected features, TCP Synchronize Attempt Delay, content length, base ID, cell ID, upstream traffic, and downstream traffic, day, hour, and minute are selected from the smartphone usage data, while POS_ID, day, and minute are selected from the campus card dataset. The ten features are selected based on the average score of various feature ranking methods, which can be classified into three categories, namely, the embedded methods, the wrapper method, and the filtering method, which will be explained in detail in Section 5. Both the selected features and the ground-truth labels are fed into the privacy inference model to generate a trained classifier.

In the third phase, the hold out-validation was used to evaluate the performance of the learned model, where 75% of the data was used for training and 25% was used for testing purposes. The task requires predicting the labels to compare with the model output and validation set. The performance of the model is illustrated through the learning curve, which plots the prediction accuracy variation along with the change in the training set size (See Section 6).

Finally, in the testing phase, each learned model is evaluated with the unused test set. The goal of this phase is to evaluate the generalization of the trained classifiers. The performance evaluation results are graphically presented in Precision, Recall, and F1-score (Figures 14, 15, 16).

In the whole process, two data sets, namely campus card data and smartphone usage data, have been used. The underlying reason to integrate these two datasets is that these two data sets represent two different types of daily activities, respectively. The campus-card data contains rich information about students' offline activities, such as offline behavior at the campus cafeteria, while the smartphone usage data contains rich information on students' online activities, such as the usage of applications, as well as offline activities, such as offline Ali-Pay and Wechat-Pay.

This study successfully demonstrated that a student's online and offline behavior patterns are highly unique, which means that the user's identity can be easily recognized through a state-of-the-art machine learning algorithm. The experimental results show the cafeteria and smartphone data alone are enough to disclose the identity privacy of the student about 25% and 89%, respectively. Furthermore, we quantify the integration of these two anonymous data to extend the potential risk of student identification up to 99%.

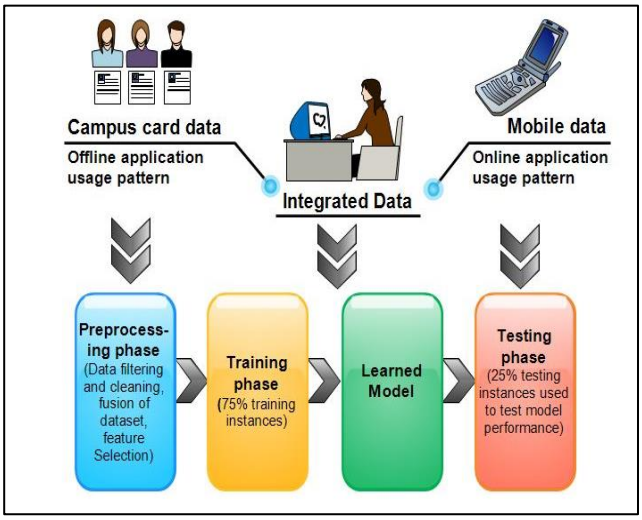

Fig. 1. The Privacy Disclosure Evaluation Process

4. Dataset Description

The datasets used in this work contain both online and offline behavior traces of 250 students. The offline behavior traces are collected from both campus card data and smartphone usage data, while the online behavior traces are collected from the smartphone usage data alone. We considered both online and offline behaviors, such as online shopping behavior and swapping campus cards in a cafeteria, respectively. In the following section, the two data sets will be described in detail.

4.1 Smartphone Data

The smartphone dataset considered in this study includes 57 attributes, which can be classified into 16 categories, as illustrated in Table 1.

4.2 Campus Card

DatasetThe campus card data includes identification information certifying the status of a student, as well as payment information for campus services, such as food, grocery, and shower. We selected the cafeteria payment transaction location and time information given in Table 2 as features.

Table 1

Attributes of the Smartphone Dataset

Domain	Features	Description
Mobile network information	Radio access technology type	RAT technology used by the device, namely; 1: UTRAN; 2: GERAN; 3: WLAN; 4: GAN; 5: HSPA 6: EUTRAN
	Machine IP Address Type	Internet Protocol address is used to identify any device connected to the network This feature contains the data about different IP classes.
	Serving gateway IP address	Gateway GPRS support node IP Address
	Access point name	Name of the gateway between the mobile network and another computer network
Application information	Application Type	Application grouped depending on the category
	Application subtype	Subcategory of applications
	Application Content	Content information of an application, such as 0 represents heartbeat, 1. Text, 3. Audio, 4. Video, and 5, other files.
	Portal Application Collection	Portal application set
Location information	Cell ID	Area code
	Base station ID	Base station Id
Time	Day	Used the app during the day
	Hour	Apps have been used within 24 hours.
	Minute	Apps have been used within the 60 minutes.
Data traffic	Upstream traffic	Data sent from the device
	Downstream traffic	Data received from the device
	Upstream IP Packet	Upstream IP packet size
	Downstream IP Packet	Downstream IP packet size
	Upstream TCP Outbound Packet	Number of upstream TCP outbound packets.
	Downstream TCP Outbound Packet	Number of downstream TCP retransmissions.
	Upstream TCP Retransmission Packet	Data of packets which have been either damaged or lost in upstreaming
	Downstream TCP Retransmission Packet	Packets which have been either damaged or lost in upstreaming.
	Upstream IP Fragmentation Packets	Fragmented datagrams pass through a link with a smaller maximum transmission unit.

Port Information	Downstream IP Fragmentation Packets	Fragmented datagrams pass through a link with a smaller maximum transmission unit.
	User Port number	Through the port channel, devices communicate
	Application Server Port	Server port number
Protocol information	Protocol Type	Protocol type code (IP, IPx, HTTP, FTP)
	L4 Protocol	Type of L4 protocol either TCP or UDP.
Network Performance	TCP Synchronize Attempt Delay	TCP link response delay
	TCP Synchronize Confirm Delay	Delay in sending the acknowledgement of the data received by the TCP sender
	TCP Synchronization	Delay in the First successful transmission
	First request To first request Delay	First transaction request to its first response packet delay
	First HTTP request To first request Delay	First HTTP response packet relative to the first HTTP request packet.
	Last HTTP packet delay	The last HTTP content packet is relative to the first HTTP request packet.
	The last acknowledgement packet delay	Last HTTP packet ACK relative to the first HTTP request packet delay.
	TCP Synchronization Attempts	Number of times TCP attempts to establish a connection
APP /HTTP content	HTTP content	HTTP content is text, pictures, video and other application
	Content length	Content length field in the protocol
	Cookie	Cookie field in HTTP packet header
	Event type	HTTP / WAP2.0 transaction type
	Application content	Data representing 0: Heartbeat; 1: Text; 2: Picture; 3: Audio 4: Video. And 5: other files.
Connection status	TCP Connection Status	TCP connection indication is either 0: success or 1: failure
	Application Status	Status of application
	HTTP WAP status	HTTP / WAP2.0-layer response code
Packet size	Window Size	data in bytes received by the device at a time
	Maximum Segment Size	maximum segment size of the TCP layer
Software information	Browser	Browser information
	User Agent	Terminal to the site to provide the terminal information
Protocol behavior	Destination Behavior	Target behavior, 0: session is the user to click the page; 1: site target generated by the page
	Operation behavior Identity	0: business login; 1: refresh; 3: unrecognized
	Operation Finish Identity	Finished state of the operation, 1: success; 2: failure; 3: unrecognized.
	Operation Delay	Service latency
	End session	Established session end
URI	URI	unambiguously identified physical or logical resources on a network
	Reference URI	It is either a URI or a relative path reference.

IP address	User IP address	User IP address
	Server IP	Application's server IP address
Host	Host	Host
	X online host	Access applications outside of the corporate network

Table 2

Attributes of Campus Card Dataset

Domain	Features	Description
Location	Place	Campus card transactions performed
	POS-ID	A device number.
	Cost	The amount that has been paid in the cafeteria /supermarket
Time	Day	The day of the month when payment is made
	Hour	Payment made on the period of the day
	Minute	Payment made on the period of a minute

4.3 Integrated Dataset

The purpose of integrating the two collected datasets is to illustrate that the two data sets are complementary in terms of privacy leakage. To verify this, besides the experiment conducted for each dataset to identify the students' typical user online and offline behavioral patterns, additional experiments have been conducted on the integrated data set to illustrate the increases in privacy leakage.

5. Feature Selection

Feature selection intends to select the minimal subset of features that can maintain the learning performance as much as possible. Our objective is to find out the optimal subset of a feature that provides predictive accuracy. In this study, seven feature selection techniques are considered to select the optimal feature subset to build the models. The feature selection techniques were driven by the need to optimize classification performance while minimizing redundancy and computational complexity. Embedded methods (Linear Regression, Random Forest, Ridge, LASSO, and Stability Selection) were chosen as they integrate feature selection into model training, improving efficiency and interpretability. Random Forest evaluates feature importance using impurity reduction, while LASSO and Ridge perform regularization to prevent overfitting, with LASSO shrinking some coefficients to zero and Ridge ranking features by contribution. Stability Selection enhances robustness by introducing noise to identify key features. The wrapper method, specifically Recursive Feature Elimination (RFE), was used to iteratively remove less relevant features, refining the subset for higher classification accuracy. Additionally, the filter method (Univariate Selection) was applied to evaluate individual feature significance based on statistical correlation with the target variable, ensuring that only the most relevant features were retained without relying on a specific model. Each feature selection technique estimates the significance of each feature in terms of its contribution to the classification performance. The optimal subset is formed based on the score of each feature, which reflects the importance of the corresponding feature. By convention, a high score implies a more valuable feature. For each feature, each of the seven feature selection algorithms will compute a score, and the average score will be calculated as the score for that feature. Then, all the features will be sorted in descending order of their scores, as shown in Tables 3 and 4 Based on that, the top features with the higher scores are chosen as input to the classifiers. A subset of high-scoring features F = {,…,} is selected using a certain threshold. If any feature's score is equal to or greater than the threshold, it will be selected. To

ensure robust and accurate user identity prediction, five supervised learning classifiers were selected based on their effectiveness in handling high-dimensional data and their ability to generalize well. The Extra Trees Classifier was chosen for its efficiency in managing complex datasets while reducing variance through randomized decision trees. The Bagging Classifier enhances robustness by aggregating predictions from multiple models, mitigating overfitting. The Decision Tree Classifier was included for its interpretability and capability to model non-linear relationships effectively. The k-Nearest Neighbors (KNN) Classifier was employed as a non-parametric baseline, particularly useful for smaller datasets and comparisons with tree-based models. Finally, the Random Forest Classifier, an ensemble learning method, was utilized to further mitigate overfitting and improve classification accuracy. This diverse selection of classifiers ensures a comprehensive evaluation of model performance and reliability in identifying users based on anonymized behavioral data.

5.1 Embedded Method

This method combines the construction of the model and the feature selection task. It is implemented by features that have built-in feature selection methods. We have selected five classifiers, which perform feature selection as a part of the model construction, i.e., a Linear Regression classifier, a Random Forest classifier, a Ridge, a LASSO, and stability.

5.1.1 Linear regression feature selection

Regression is a technique of modelling an output variable (y) based on some input variables (x, i.e., features). It is generally used to characterize the correlation between y and x. Linear regression is a linear model that assumes a linear relationship between y and x. More specifically, y can be calculated from a linear combination of the input variables, as shown in Fig. 2, where the red line represents the relation between x and y, which can be learned from the associated data points. The linear relation can be modelled by Eq. (1).

$$Y = \beta_0 + \beta_1 * x \quad (1)$$

Where β0 represents the intercept and β1 denotes the coefficient for x.

5.1.2 The random forest feature selection

Random forest (RF) [39] is a method that generates the forest of decision trees, where each tree is randomly sampled from the original data. Every node in the decision trees represents a feature, which will partition the data into multiple sub-spaces based on the possible feature values. Since the different ordering of the features (from the root to the leaves) may incur different classification accuracies, the relative ordering reflects the importance of features, which can be used to rank and select the features. In this work, we adopted Gini impurity [40], to characterize the classification accuracy. Gini Impurity is a measure used in decision tree algorithms to determine how "pure" a split is when classifying data. It quantifies the likelihood of an incorrect classification if a randomly chosen element is labelled according to the distribution of classes in a given node. It is calculated as given in Eq. (2).

$$Gini(split) = 1 - \sum_{i=1}^{n} p_i^2 \quad (2)$$

Where pi is the rate of data items that belong to class i. A lower Gini Impurity value indicates a purer node, meaning better classification accuracy.

5.1.3 The LASSO regularization method

The Least Absolute Shrinkage and Selection Operator (LASSO) is a powerful method to perform two main tasks: regularization and feature selection [41]. Regularization is a technique used in machine learning to prevent overfitting by adding a penalty to the model's complexity. Overfitting occurs when a model learns not only the patterns in the training data but also noise, making it perform well on the training data but poorly on new, unseen data. It is a particular case of L1 regularization that adds the sum of the absolute values of model coefficients as a penalty. It promotes sparsity by shrinking some coefficients to zero, effectively performing feature selection, as shown in Eq. (3).

$$w = argmin_w \sum_{i=1}^{N} (y_i - \sum x_{ij} w_{ij})^2 + \lambda \sum_{j=1}^{p} |w_{ij}| \quad (3)$$

We assume that the data xij and label yi for, where N is the total number of data points, i = 1,2, …, N represents the i-th data point, x_ij denotes the j-th feature of data point i, coefficient w_ij means the weight of feature j of data point i, and the Langanger $\lambda \geq 0$ tradeoffs the minimal square error and regularization term. The larger the Langanger λ, the more zero coefficients [42]. Model has 57 scores but only 12 of them are non-zero. The features having zero scores are useless in predicting the target value.

5.1.4 The regularized linear method (Ridge)

Ridge is similar to LASSO. The key difference between the two is that Ridge adopts L2 regularization, which penalizes large coefficients by adding the sum of their squares to the loss function. L2 regularization helps prevent overfitting by ensuring that all feature weights remain small rather than

eliminating them entirely. Unlike LASSO, which forces some coefficients to zero (performing feature selection), Ridge regression shrinks all coefficients towards zero but retains all features in the model.

Ridge regression regularizes the minimal square error by adding the sum of squares of the coefficients in the optimization function, as shown in Eq. (4). The effect of L2 regularization is that it forces the model to choose small values for the coefficients (wj), preventing any single feature from dominating the prediction. Therefore, the coefficients of those features with relatively less impact will be closer to zero, which can be used to rank the features based on their importance.

$$w = argmin_w \sum_{i=1}^{N}(y_i \ - \sum x_{ij}\, w_j)^2 + \lambda \sum_{j=1}^{p} | w_j|^2), \quad (4)$$

The effect of L2 regularization is that it forces the model to choose small values for the coefficients (wj), preventing any single feature from dominating the prediction. Therefore, the coefficients of features with relatively less impact will be closer to zero, which can be used to rank feature importance.

5.1.5 The stability selection

In stability selection [43], certain noise is added to the original data by creating bootstrap samples—randomly drawn subsets of the data using the replacement approach. Bootstrap sampling is a statistical technique where multiple random samples of the same size as the original dataset are created by drawing data points with replacement. This means that some data points may appear multiple times in a sample, while others may be excluded. This method helps estimate the stability and importance of selected features.

In this study, LASSO has been chosen as the base feature selection algorithm to identify the most relevant features in each bootstrap sample. The fundamental idea is straightforward: irrelevant features will have little influence on classification performance when perturbed by noise. In contrast, the model's performance should be significantly impacted when important features are perturbed. Consequently, features that contribute minimally to classification accuracy under noise are removed, and only the most robust and essential features are retained for the final classification model. If we have an original dataset X={x1,x2,...,xN},}, we generate B bootstrap samples X_b^* (where b=1,2,..., B). The bootstrap estimate of the mean is given in Eq. (5).

$$\hat{\mu}^* = \frac{1}{B} \sum_{b=1}^{B} \tilde{X}_b^* \quad (5)$$

5.2 The Wrapper Method

The wrapper methods [44] evaluate the usefulness of each feature by iterative training and testing a model based on different subsets of features. Instead of ranking features independently, the wrapper method selects features based on their direct impact on model performance. A model is trained using all available features, and then features are systematically removed or added based on their contribution to classification accuracy. Each iteration involves making predictions, and the change in prediction accuracy is used to assess feature importance—the greater the accuracy drop when a feature is removed, the more significant that feature is. Inza [45] suggested that the goal of the wrapper method is to identify the optimal subset of features that maximizes model performance by selecting those that create the largest margin of class separation. This process follows a sequential feature selection strategy, where features are either eliminated or added one by one based on classifier performance until an optimal feature set is determined. Common wrapper techniques include Recursive Feature Elimination (RFE), which recursively removes less important features, and Forward/Backward Feature Selection, which iteratively evaluates feature subsets to find the best-performing combination. By focusing on model-driven feature selection, wrapper methods enhance predictive accuracy but are computationally expensive, as they require multiple iterations of training and evaluation. The wrapper method can be mathematically expressed as given in Eq. (6)

$$S^* = \sum_{j=1}^{p} w_j x_j \ \ where\ \omega_j = \arg max\, f(S_k) \quad (6)$$

In the above Eq. $S^*$is the optimal subset of selected features, where x_j represents individual features, w_j is the weight or importance score of feature j. $f(S_k)$ is the performance function (e.g., accuracy, F1-score), which we aim to maximize.

5.2.1 Recursive feature elimination (RFE)

In this work, we adopt a special type of wrapper method, called Recursive Feature Elimination (RFE), which repeatedly removes features with smaller scores (i.e., less prediction reduction). RFE starts by ranking all features in the dataset and stops when all the features have been evaluated. Kumari [46] illustrated that RFE is a greedy optimization for finding the best-performing subset of features.

$$\omega_j = \sum_{i=1}^{N} \left|\beta_j\right| \ where\ \forall_j \in \ S_k \quad (7)$$

In this Eq. (7) , ω_j represents the importance score of feature j, which determines its contribution to the

model's predictive performance. The term β_j denotes the coefficient of feature j in a regression model or its importance score in a tree-based model, indicating how significantly it influences the output. Additionally, N refers to the total number of training samples, which affects the computation of feature importance by considering the overall dataset size during the selection process.

5.3 The Filter Methods

Filter methods are feature selection techniques that evaluate the relevance of each feature by measuring its statistical relationship with the target variable. Unlike wrapper methods, which rely on a machine learning model for feature selection, filter methods assess features independently of any specific algorithm. This makes them computationally efficient and suitable for handling high-dimensional datasets.

Several statistical techniques are used to determine the correlation between features and the target variable, including the Chi-squared test [47], mutual information [48], and correlation coefficient scores.

In this study, we have used Chi-Squared Test to rank features and measures the dependence between categorical features and the target variable. It is calculated as Eq. (8).

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (8)$$

Where $O_i$is the observed frequency of a category. and E_i is the expected frequency under the assumption of independence are individual data points of the feature and target variable, respectively and $\bar{X}$ and $\bar{Y}$are the means of X and Y.

5.3.1 Univariate Feature Selection

The univariate feature selection technique evaluates each feature individually to measure its correlation with the output variable. Unlike multivariate techniques, it does not account for dependencies between different features; instead, it ranks features based on their independent relationships with the target variable.

A common approach for univariate feature selection is Pearson's correlation coefficient, which is used in this study. It is calculated using Eq. (9). If 〖|r〗 _i | is close to 1, the feature is X_j strongly correlated with the target variable and is likely to be useful for prediction. If the feature is close to 0, it has little to no correlation with the target and may be irrelevant for model training.

$$r_i = \frac{\sum(X_j - \overline{X_j})(Y - \bar{Y})}{\sqrt{\sum(X_j - \overline{X_j})^2 \sum(Y - \bar{Y})^2}} \qquad (9)$$

In the above Eq. r_i is the correlation coefficient for the feature X_j. X_j represents the $j - th$ values of the feature across all samples. Y is the target variable. However, $\bar{X}_j$ and $\bar{Y}$are the means of feature X_j and the target Y, respectively.

5.4 The Integrated Feature Selection

To improve the model performance, we integrate the above feature selection methods by averaging their feature scores, as shown in Eq. (10).

$$\mathrm{X} = \frac{1}{l}\sum_{i=1}^{l} b_i = \frac{b_1 + b_2 + \cdots + b_l}{l} \qquad (10)$$

where fi represents the feature score computed by the i-th feature selection method and l denotes the total number of feature selection methods.

The features, the average scores of which are larger than a pre-defined threshold, 0.3, are selected as the input features for the subsequent processes. The selected features associated with the smartphone usage data and campus card data are ranked in descending order in terms of their scores, as shown in Tables 3 and 4, respectively. The corresponding graphical representation is shown in Fig. 3. The selected features from the smartphone data include TCP Synchronization Attempts, content length, downlink traffic, last acknowledgement packet delay, uplink traffic, cell ID, base station ID, day, hour, and minute. The selected features from the campus card dataset include minute, POS_ID, and day.

Table 3

Mobile dataset

Attributes	Lasso	Linear regression	Random Forest	Recursive feature selection	Ridge	Stability	Univariate	Mean
TCP synchronize Attempt	1	0	0	0.09	1	1	0	0.44
Content-Length	0	0	0.17	0.74	0	1	1	0.42
Downstream Traffic	0	0	0.26	0.75	0	0.96	0.61	0.37

Cell ID	0	0	1	0.53	0	1	0	0.36
Last ACK Packet Delay	0	0	0.3	0.77	0	0.96	0.48	0.36
Upstream Traffic	0	0	0.37	0.79	0	0.99	0.32	0.35
Base ID	0	0	0.83	0.6	0	0.96	0	0.34
Day	0	0	0.66	0.51	0	0.99	0	0.31
Hour	0.01	0	0.7	0.44	0.01	1	0	0.31
Minute	0.01	0	0.71	0.42	0.01	1	0	0.31
First request to first Response Delay	0	0	0.25	0.7	0	1	0.05	0.29
Browser	0.44	0	0.01	0.11	0.45	1	0	0.29
TCP synchronize Attempt Delay	0	0	0.33	0.68	0	0.94	0.02	0.28
TCP synchronize confirm delay	0	0	0.35	0.67	0	0.96	0.01	0.28
SGWG ID	0.03	0	0.37	0.37	0.03	1	0	0.26
Last HTTP Packet Delay	0	0	0.23	0	0	0.74	0.82	0.26
App server IP	0	0	0.23	0.56	0	1	0	0.26
Host ID	0	0	0.17	0.65	0	0.96	0	0.25
App server port	0	0	0.02	0.58	0	1	0.11	0.24
Event type	0.24	0	0.05	0.12	0.24	1	0	0.24
Upstream IP packet ID	0	0	0.16	0.54	0	1	0	0.24
HTTP WAP status	0	0	0.01	0.61	0	1	0	0.23
Downstream IP packet ID	0.01	0	0	0.49	0.01	0.93	0	0.21
Portal App collection	0.11	0	0.01	0.25	0.11	1	0	0.21
HTTP content ID	0.01	0	0.08	0.4	0.01	1	0	0.21
Upstream TCP retransmission Packet	0.01	0	0.01	0.39	0.01	1	0	0.2
Radio access technology	0	1	0	0.35	0	0	0	0.19
App subtype	0.04	0	0.06	0.19	0.05	1	0	0.19
Upstream TCP outbound packet	0.05	0	0	0.23	0.05	1	0	0.19
Operation delay	0	0	0.23	0.02	0	0.26	0.82	0.19
Downlink TCP outbound Packet	0.03	0	0	0.21	0.03	0.96	0	0.18
URI	0	0	0	0.98	0	0	0	0.14
X Online host	0	0	0	1	0	0	0	0.14
User-agent	0	0	0	0.96	0	0	0	0.14
Refer URI	0	0	0	0.95	0	0	0	0.14
Window size	0	0	0	0.93	0	0	0	0.13
Maximum segment size	0	0	0	0.89	0	0	0	0.13
Cookie	0	0	0	0.91	0	0	0	0.13
Destination behavior	0	0	0	0.88	0	0	0	0.13
Machine IP address	0	0.51	0	0.33	0	0	0	0.12
TCP connection status	0	0	0	0.84	0	0	0	0.12
Session is end	0	0	0	0.82	0	0	0	0.12
Operation behavior Identity	0	0	0	0.81	0	0	0	0.12
Operation finish identity	0	0	0	0.86	0	0	0	0.12
TCP synchronize success first Request delay	0	0	0.23	0.04	0	0.48	0	0.11

Upstream IP fragmentation packets	0	0	0	0.72	0	0	0	0.1
Downstream TCP retransmission packet	0	0	0	0.63	0	0	0	0.09
Downlink IP fragmentation packets	0	0	0.22	0.05	0	0.36	0	0.09
User IPv4	0	0	0	0.46	0	0	0	0.07
User port	0	0	0	0.47	0	0	0	0.07
Access point name	0	0.13	0	0.32	0	0	0	0.06
App type code	0	0.14	0	0.3	0	0	0	0.06
First HTTP Request packet delay	0	0	0.22	0.07	0	0.16	0	0.06
Protocol type	0	0.01	0	0.28	0	0	0	0.04
App Type	0	0	0	0.26	0	0	0	0.04
App content	0	0	0	0.18	0	0	0	0.03
App status	0	0	0	0.16	0	0	0	0.02
L4Protocal	0	0	0	0.14	0	0	0	0.02

Table 4

Campus card dataset

Attributes	Lasso	Linear Regression	Random Forest	Recursive Feature Selection	Ridge	Stability	Univariate	Mean
Minute	1	1	1	0	1	0	0.09	0.58
POS_ID	0.58	0.58	0.46	0.4	0.58	0	1	0.51
Day	0.65	0.65	0.68	0.2	0.65	0	0	0.4
Hour	0.33	0.33	0.16	0.6	0.33	0	0.01	0.25
Cost	0	0	0.5	1	0	0	0.17	0.24
Place	0.16	0.16	0	0.8	0.16	0	0.05	0.19

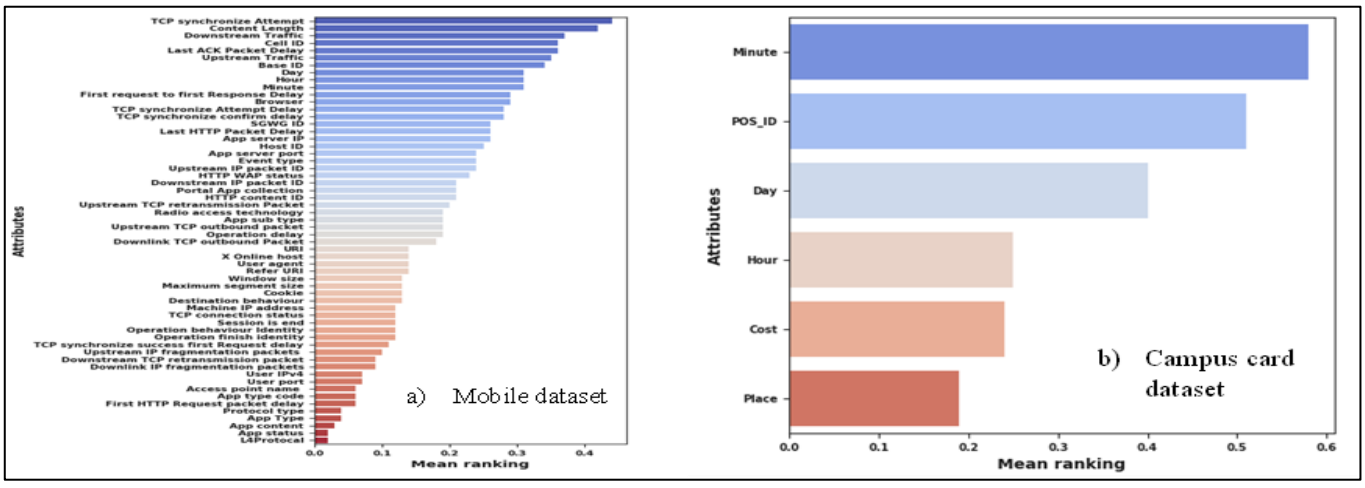

Fig 3. The Mean Feature Ranking Of Each Dataset Is Shown

6. The Increased Privacy Leakage Of Integrated Data

To illustrate that the integrated data can improve the model performance, we adopt a learning curve to visualize the model performance associated with the individual datasets and the integrated dataset. A learning curve is usually used to compare the validation and training performance along with the increase in the data size for a particular model. It refers to a graphical representation of the prediction accuracy on the y-axis and the training set size on the x-axis, which demonstrates the performance of the model along with the increasing number of data instances used to train the model. In supervised machine

learning, the learning curve is used to evaluate the effect of increasing training data on model performance and to detect overfitting by analyzing the gap between the training and validation curves.

6.1 Privacy Preservation Techniques

Anonymization techniques play a crucial role in preserving user privacy, but their effectiveness varies based on the method used and the type of data being protected. Generalization and suppression help reduce identifiability by modifying or removing sensitive attributes, yet they often compromise data utility and remain vulnerable to linkage attacks. K-anonymity ensures that each record is indistinguishable from at least k-1 others, but it does not prevent attackers from inferring missing details. Differential privacy, which introduces mathematical noise to data queries, provides stronger privacy guarantees while maintaining statistical accuracy, though improper tuning can reduce its effectiveness. Homomorphic encryption, allowing computations on encrypted data, offers maximum security but remains computationally expensive. While no single method provides complete protection, a hybrid approach combining multiple techniques is often necessary to balance privacy and data utility, reducing re-identification risks while ensuring meaningful data analysis.

6.2 The Data Set Comparisons Through Multiple Classifiers

To verify whether the integration of the smartphone usage data and campus card data may increase the risk of privacy leakage, we conducted experiments on both individual datasets and the integrated dataset. For comparison, the five classifiers (k-nearest Neighbor, Bagging, Extra Tree, Decision Tree, and Random Forest Tree classifiers) are applied to compare the classification accuracy of each classifier. Each data set (the smartphone, the campus card, and the integrated datasets) is trained with the five classifiers and the corresponding experiment results are shown in Figures 7, 8, and 9, respectively, in the form of the learning curve, which shows the relationship between the training set size and accuracy score on the training set and the validation set.

Firstly, we consider the campus card dataset. The learning curves of those five classifiers applied to the campus dataset are shown in Fig. 7, where the y-axis is the accuracy score, and the x-axis is the training set size. The higher the score, the better the performance of the corresponding model. From Fig. 7, we can see the existence of the overfitting problem, which can be mitigated by the introduction of more training samples. For all the classifiers, the gap between the training score and the validation score is large. Moreover, except for the k-nearest neighbor, the gap between all the other classifiers could be decreased by adding more training samples.

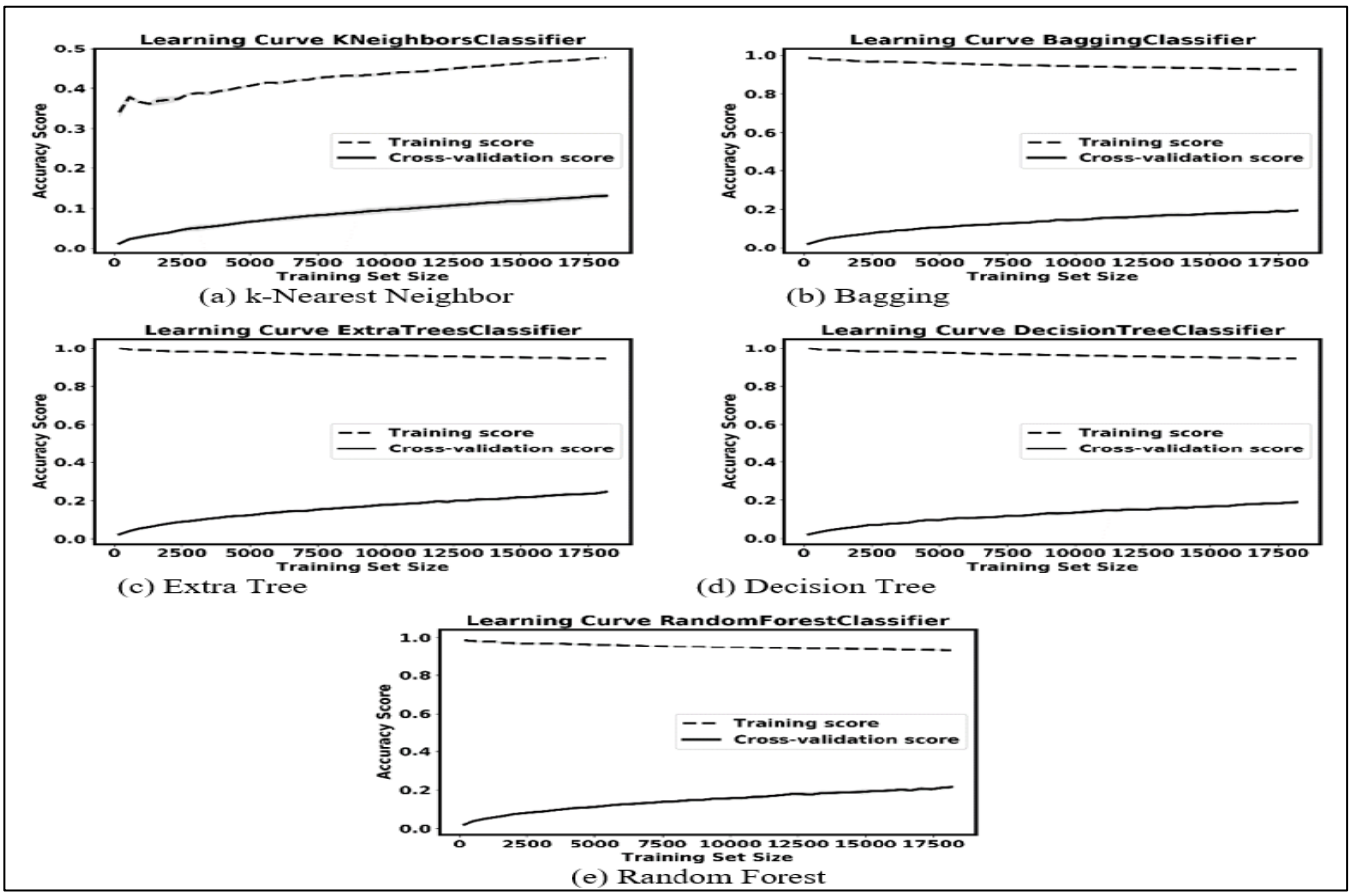

Fig 7. The Learning Curve Of The Classifiers Trained On The-Campus Dataset

Secondly, we apply the same five models to the mobile usage dataset, the experiment results of which are presented in Fig. 8. From Fig. 8, it can be observed that all the classifiers achieve better prediction accuracy on the mobile data set than the campus data but all the classifiers still suffer from the overfitting problem. The training score of all the classifiers is a higher score than the cross-validation score. Except for the k-nearest neighbor classifier, for all the other classifiers, the gap between the training score and validation score is significantly less than the corresponding one from the campus card data set.

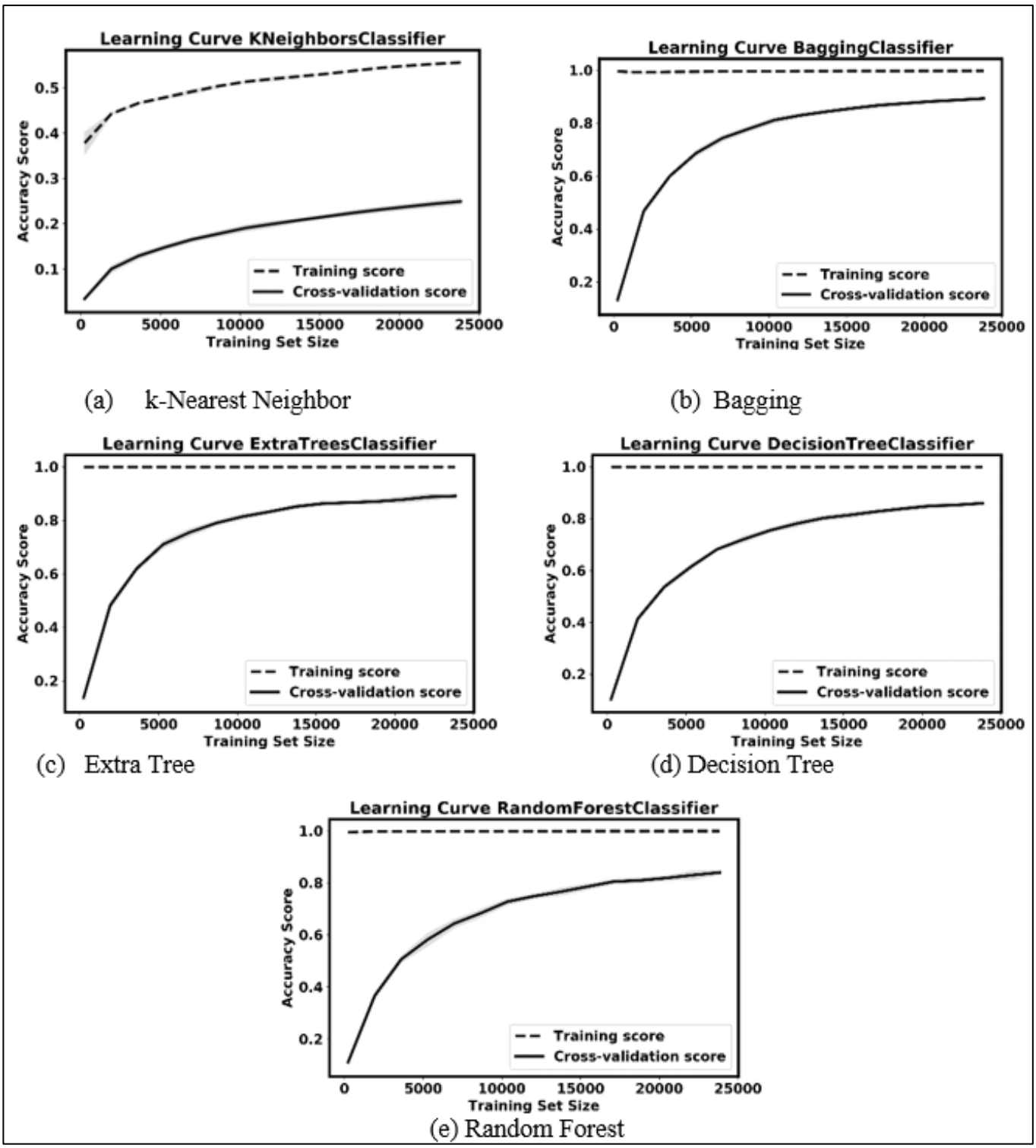

Fig 8. The Learning Curve Of The Classifiers Trained On The Smartphone Usage Dataset

Finally, we apply the same five models to the integrated dataset, the experiment results of which are presented in Fig. 9. From Fig. 9, it can be observed that both the training accuracy and validation accuracy for all five classifiers have been significantly improved. Moreover, the overfitting problem disappears due to the information complementarity between the campus card data set and the smartphone usage data set. This illustrates that the campus card data and smartphone usage data can be complementary to each other in terms of leaking user privacy.

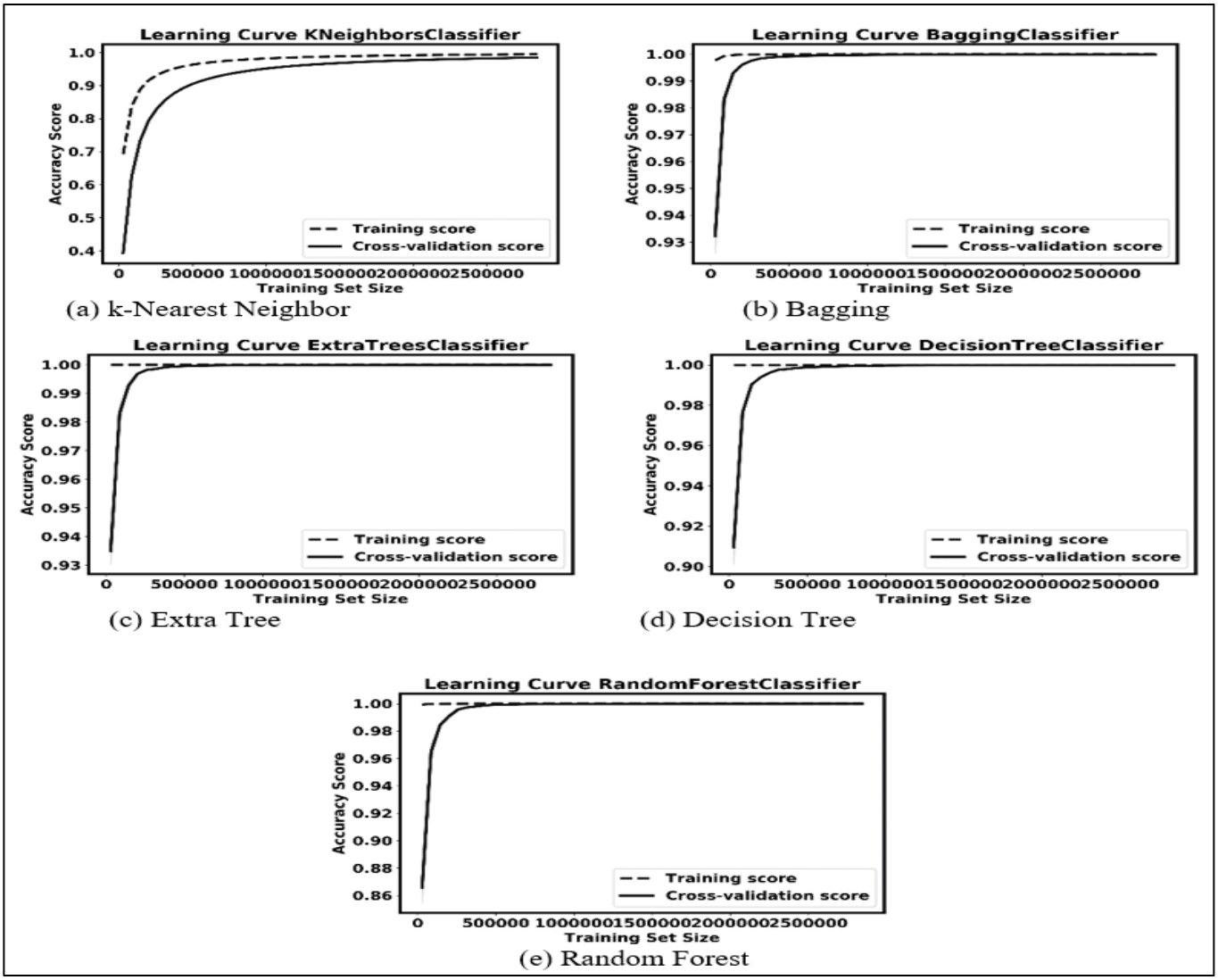

Fig. 9. The Learning Curve Of The Classifiers Trained On The Integrated Dataset

7. Experiment Evaluation

To assess whether anonymized integrated data can still lead to privacy leakage, we trained multiple standard machine learning classifiers. The significance of selected features was analyzed by high-score features applied to different feature selection techniques. A high score corresponds to the most likely to be selected to construct the optimal subset. The subset of features of each dataset is chosen to find the efficient training model. Through experimental analysis, it can be validated that the learning models can be trained with higher accuracy based on the selected features. This ensures that the selected features provide discriminative information. The results demonstrated that the linkage between smartphone data with campus card data has a potential implication of privacy loss. The k -Nearest Neighbor, Bagging, Random Forest Tree, Extra Tree, and Decision Tree classifiers have been applied after the feature selection.

7.1 Experiment Environment

We conducted three groups of experiments, on the three different real datasets. The powerful package of the Python Scikit-learn library [49] has been used for implementing the existing classifiers. Experiments were performed on a computer (R) Xeon (R) with CPU E5-2620 v4 @ 2.10GHz x 16 with 12 TB main memory and 512 GB RAM size. All the experiments were done under Ubuntu 16.04 LTS 64-bit.To verify that the anonymized integrated data may still leak privacy, we trained several standard machine learning classifiers. The significance of selected features was analyzed by high-score features applied to different feature selection techniques. A high score corresponds to the most likely to be selected to construct the optimal subset. The subset of features of each dataset is chosen to find the efficient training model. Through experimental analysis, it can be validated that the learning models can be trained with higher accuracy based on the selected features. This ensures that the selected features provide discriminative information. The results demonstrated that the linkage between smartphone data with campus card data has a potential implication of privacy loss. The k -Nearest Neighbor, Bagging, Random Forest Tree, Extra Tree, and

Decision Tree classifiers have been applied after the feature selection.

7.2 Model Evaluation

The experimental results demonstrate the power of integrating two datasets of student behavior disclosing the student's privacy. The integrated dataset refers to the set formed by joining two datasets smartphone and campus card through the user's phone number, which contains features of each dataset chosen with the mean ranking technique individually. The features selected with distinct behavioral characteristics are used to construct the classification model with better classification accuracy. For comparison, we also demonstrated the result of each dataset, campus card data, smartphone data, and integrated data respectively with the same sample size. The results show each dataset leaks enough privacy, while we can also observe from the results the risk of disclosing personal data.

The measures used in this study such as accuracy, precision, recall, and f1-score are derived from the confusion matrix. The annotations used are given in Table 5.

Table 5

Other notations.

Term	Meaning
y	Set of predicted labels
$\hat{y}$	Set of true labels
L	Set of labels
S	Set of samples
y_s	Subset of y with sample s, i.e $y_s =$ $:\{(s', l) \in y \mid s' = s\}$
y_l	Subset of y with the label l
$\hat{y}_s$	Subset of $\hat{y}$
$\hat{y}_l$	Subset of $\hat{y}$ with the label l

7.2.2 Average accuracy

To measure the accuracy of the model, in multilabel classification, the function returns the subset accuracy. If the entire set of predicted labels for a sample strictly matches the true set of labels, then the subset accuracy is 1.0; otherwise, it is 0.0. The best accuracy is 1.0 whereas the worst is 0.0. It is the predicted value of the i-th sample and is the corresponding true value, then the fraction of correct predictions over is defined as Eq. (11).

$$Accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (11)$$

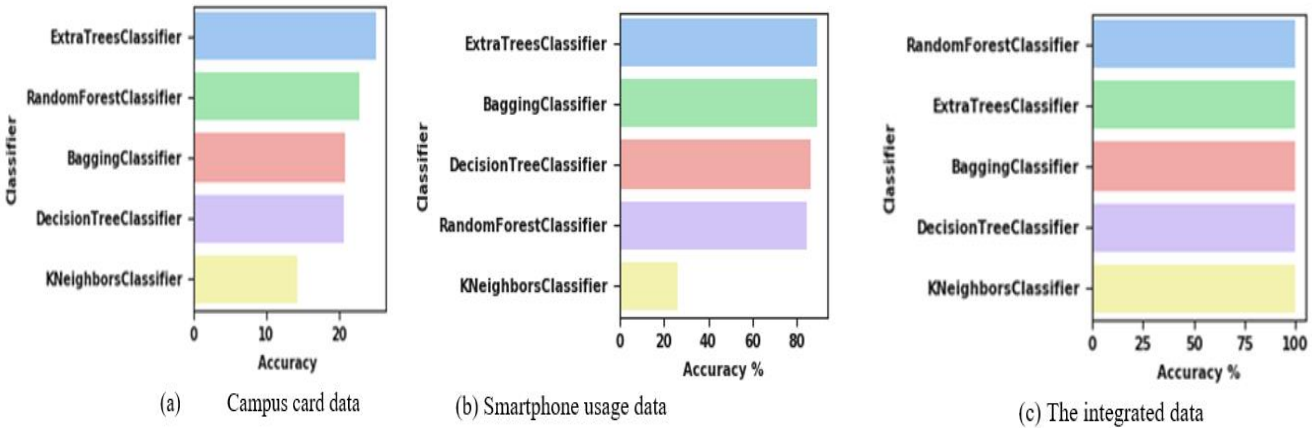

Fig 9. Classification Accuracy Comparison of Classifiers

The accuracy results of smartphone usage, campus card, and integrated datasets are plotted in Fig. 9, which illustrates that data integration can improve the accuracy of all classifiers. This implies that smartphone usage and campus card data are complementary in terms of privacy leakage.

7.2.1 Log Loss

The log loss function quantifies the accuracy of a classifier by penalizing false classifications. Minimizing the log loss is equivalent to maximizing the accuracy of the classifier. The formal definition of the log loss function is shown in Eq. (12).

$$L_{log}(Y, P) = -logPr(Y|P) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\, logp_{i,j} \quad (12)$$

where N represents the number of data samples, M denotes the number of different labels, is a binary variable indicating whether label j is correct for sample i, and is the probability, on which the model assigns label j to sample i.

The experiment results are shown in Fig. 10, from which it can be observed that the smartphone usage data and campus card are complementary in terms of privacy leakage because the integrated data significantly reduces the log loss of all the classifiers except the K-neighbors classifier.

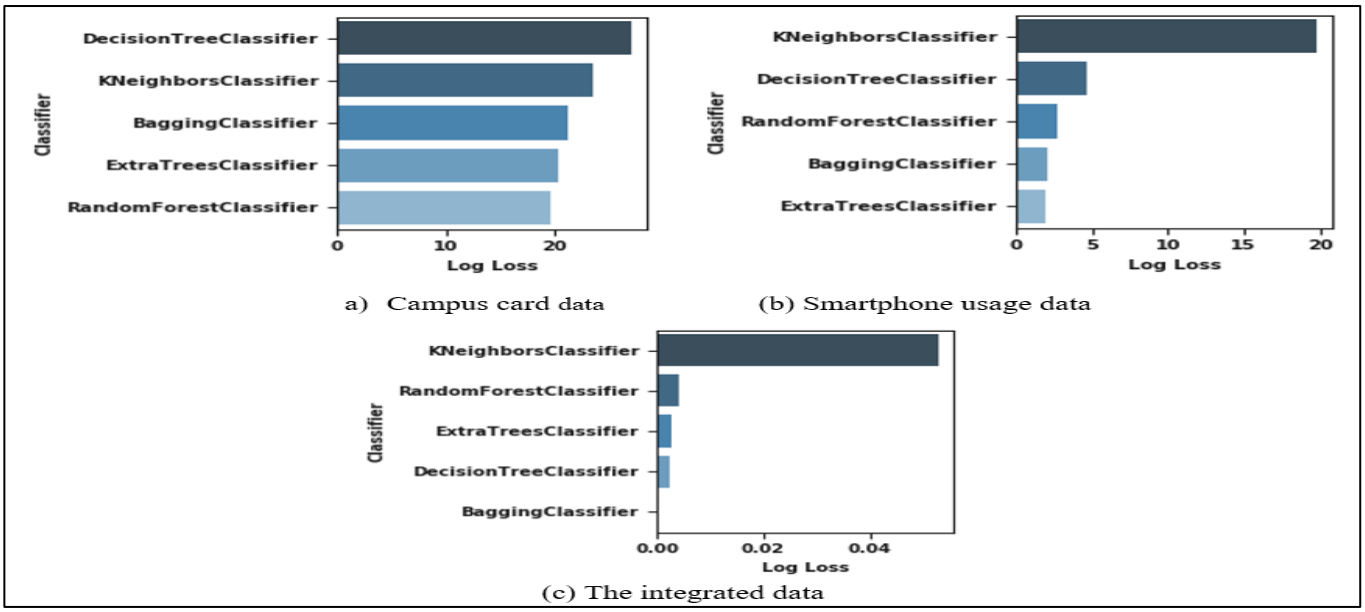

Fig 10. Log Loss Comparison of different Classifiers

7.2.3 Recall

The recall was used to know how much relevant information was extracted by the system. The recall is calculated as the number of correct positive predictions by the total number of positive tuples available in the dataset defined as the ratio of the total number of correctly classified samples to the total number of samples, as shown in Eq. (13).

$$Recall = \frac{T_p}{T_p + F_n}, \quad (13)$$

The experiment result is plotted in Fig. 14, which shows the recall of the five classifiers on the three datasets, respectively. It can be observed that is evident that the campus dataset provides a very low recall value for the classification of students' behaviors. For the campus dataset, the best precision is 25% and is achieved by Extra Tree Classifier. For the mobile dataset, the students' behaviors are classified with better recall than in the campus dataset. However, it is clear from the results that the integrated dataset yields much higher recall throughout all classifiers. It is obtained when the additional step of integration of datasets is performed to achieve a balanced recall score.

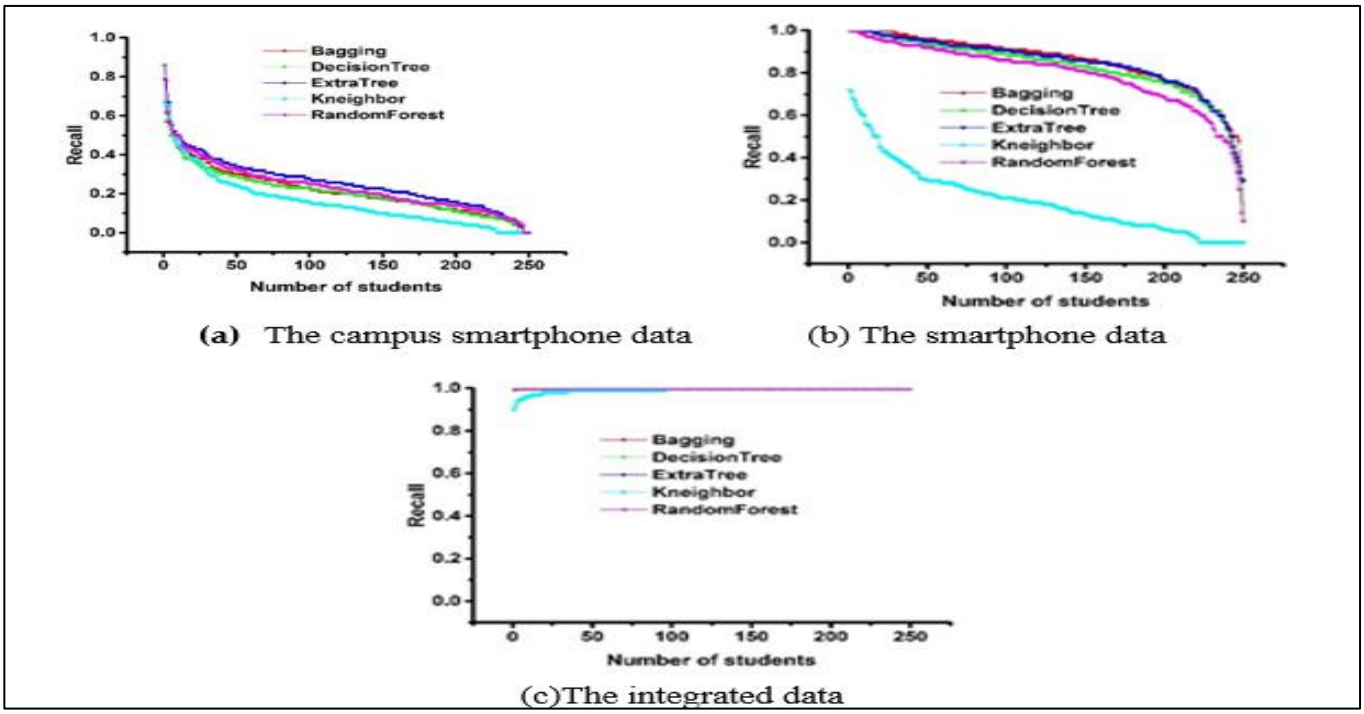

Fig 14. Recall Comparison Curves of 5 discrete Classifiers

7.2.4 Precision

The precision measure is known to correctly identify the number of relevant instances among the retrieved instances which is the fraction of Tp among the number of Tp plus the number of Fp given by the model. The ability of precision is not to label a positive sample as negative. The best precision value is 1 and the worst value is 0. Precision is considered as one label versus all other labels as if it had been reduced to a binary 'Label X' vs 'not Label X' problem. Precision can be calculated through Eq. (14).

$$Precision = \frac{tp_i}{tp_i + fp_i}, \quad (14)$$

The variation of the precision of the five classifiers along with the change in the number of students used in the training data for the individual data sets and the integrated dataset is visualized in Fig.15, from which it can be observed that the precision of the campus dataset decreases rapidly along with the increment of the number of students and all classifiers show similar precision with a marginal difference. The precision of the smartphone dataset decreases much less than that of the campus dataset, except for that of the KNN classifier. The precision of the integrated dataset shows a superior precision due to the complementation between the two individual datasets.

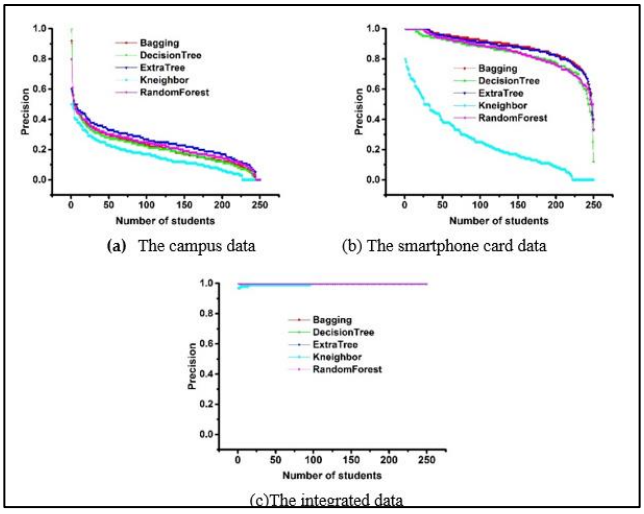

Fig 15. Precision Comparison Curves of the 5 discrete Classifiers

7.2.5 F1-score

The F1 score can be interpreted as a weighted average of precision and recall, where an F1 score reaches its best value at 1 and the worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The graphical representation of the F1 score is shown in Fig. 16 and can be computed through Eq. (15), and the comparison results of different classifiers are given in Table 6.

$$F1 - score = \frac{2 * (precision \ . \ recall)}{precision + recall} \quad (15)$$

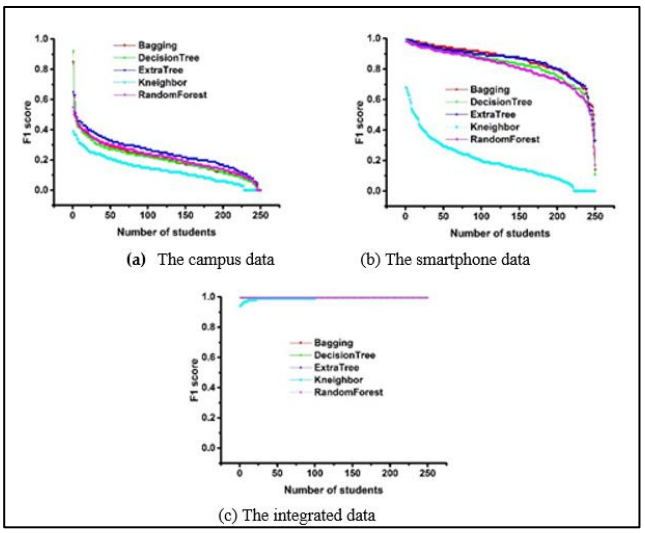

Fig 16. F1 Score Comparison Curves of the 5 discrete Classifiers

The F1 scores of the five classifiers applied to the individual dataset and the integrated dataset along with the increment of the number of students are visualized in Fig. 10. The visualizations suggest that classifiers have produced a better F1 score when additional information sources are used.

The classification algorithm used in this experiment performed well with integrated data with high precision and high recall, The F1 score approximately reached 99% with the increasing size of the sample. The classification results of all the applied to the integrated dataset were nearly identical. Our results demonstrate that even anonymized behavioral data (e.g., smartphone usage and campus card transactions) can still be used to re-identify individuals with over 99% accuracy when combined with machine learning techniques. This finding raises serious privacy concerns, as it highlights the potential for de-anonymization attacks, where adversaries can reconstruct personal identities from supposedly anonymized datasets. The risk is further amplified when multiple datasets are integrated, as seen in our study, where privacy leakage increases significantly after combining different data sources.

Our findings emphasize the urgent need for stronger privacy-preserving measures, such as differential privacy, secure multi-party computation (SMPC), and federated learning, to minimize the risk of data misuse while maintaining analytical value. We have incorporated this discussion in the revised manuscript to better reflect the broader implications of our results in the context of privacy and security.

8. Results and Discussion

8.1 Results

The study presents significant findings regarding identity disclosure, emphasizing the impact of data

sharing on users' privacy. The data in Table 6 illustrates a clear pattern where linked data increases the chances of privacy leakage, reinforcing concerns about personal information exposure in digital environments. The analysis reveals that when different datasets are combined, the risk of re-identification grows, making it easier for adversaries to infer sensitive user attributes.

Furthermore, the statistical analysis confirms that all five classifiers K-Nearest Neighbors (KNN), Extra Trees, Bagging, Decision Tree, and Random Forest consistently demonstrate the vulnerability of behavioral data to privacy breaches. The classification models indicate that existing privacy measures are insufficient in preventing unauthorized access to sensitive information. Notably, almost all classifiers show the highest accuracy in detecting potential privacy threats, suggesting that machine learning techniques can effectively identify privacy risks but may also be used by attackers to exploit weaknesses in data security.

The results also show variations between offline and online data, indicating potential underlying causes such as differences in data collection mechanisms, user behavior patterns, and exposure levels. Offline datasets generally exhibit lower privacy risks due to controlled access, whereas online datasets are more susceptible to leakage due to third-party tracking, metadata linkage, and user profiling.

8.2 Discussion

These findings align with previous studies on privacy risks in data-sharing environments, supporting the argument that behavioral data privacy requires stronger protection mechanisms. The high classification accuracy of the applied machine learning models highlights the predictability of user behavior, which adversaries can exploit for re-identification and unauthorized profiling. This confirms that privacy vulnerabilities persist even when data is anonymized, as patterns in user activity can still be inferred through statistical techniques.

A key takeaway from this study is that traditional anonymization techniques, such as data masking and differential privacy, may not be sufficient to mitigate risks in linked datasets. The strong performance of decision-tree-based classifiers suggests that advanced privacy-preserving techniques, such as homomorphic encryption, federated learning, and synthetic data generation, should be explored to enhance user privacy.

Additionally, the discrepancies observed between offline and online data suggest that context plays a crucial role in privacy exposure. Online data, being more dynamic and subject to tracking mechanisms, faces greater risks, whereas offline data benefits from controlled access and limited exposure. This reinforces the need for context-aware privacy frameworks that adapt protection levels based on the environment in which data is shared.

Despite these insights, certain limitations exist in the study. The results are highly dependent on the dataset characteristics and classification models used, and future research should explore additional classifiers, deep learning approaches, and real-world privacy attack simulations to further validate these findings. Moreover, user consent mechanisms and regulatory frameworks should be integrated into future studies to assess the role of legal and ethical considerations in privacy protection.

Overall, this study contributes to the ongoing discourse on data privacy and identity disclosure, emphasizing the urgent need for robust, adaptive, and scalable privacy-preserving techniques to safeguard behavioral data in both online and offline environments.

Table 6

Comparison of results of different classifiers

Canteen dataset					
Classifiers	KNN	Extra tree	Bagging	Decision tree	Random forest
Accuracy	0.125	0.234	0.202	0.197	0.212
Log loss	0.254	0.213	0.224	0.290	0.211
Precision	0.153	0.253	0.209	0.209	0.223
Recall	0.155	0.255	0.213	0.213	0.225
F-score	0.137	0.250	0.207	0.206	0.219
Mobile					
Accuracy	0.281	0.834	0.833	0.825	0.812
Log loss	0.200	0.020	0.025	0.050	0.030
Precision	0.253	0.882	0.878	0.844	0.833

Recall	0.208	0.857	0.859	0.833	0.796
F-score	0.220	0.859	0.871	0.838	0.818
Integrated dataset					
Accuracy	0.996	0.999	0.999	0.999	0.999
Log loss	0.0005	0.0025	0.0025	0.0001	0.00035
Precision	0.986	0.999	0.999	0.999	0.999
Recall	0.981	0.999	0.999	0.999	0.999
F-score	0.985	0.999	0.999	0.999	0.999

9. Ethical Considerations

Privacy leakage raises serious ethical concerns, particularly regarding user consent, autonomy, data security, and potential misuse of sensitive information. As this study demonstrates that anonymized behavioral data can still be used to re-identify individuals, it is crucial to address the ethical implications of data collection, sharing, and analysis.

9.1 User Consent And Transparency

One of the fundamental ethical concerns in data privacy is the issue of informed consent. Users often remain unaware of how their behavioral data such as campus card transactions and smartphone usage patterns are collected, stored, and shared with third parties. Many institutions anonymize data before sharing, but as shown in this study, anonymization alone is insufficient to prevent re-identification. It is ethically imperative for institutions to inform users about how their data is collected and used. Obtain explicit consent before collecting behavioral data and provide opt-out mechanisms for users who do not wish to participate in data collection.

9.2 Risk Of Re-Identification And Data Misuse

Although anonymization techniques are designed to protect user identities, this study demonstrates that machine learning models can re-identify individuals with high accuracy based on behavioral patterns. This creates a risk of privacy breaches, which could lead to unauthorized tracking and profiling of individuals, Discrimination or bias in decision-making processes if sensitive behavioral data is misused and financial or reputational harm if personal data is exploited by malicious entities. To mitigate these risks, organizations handling behavioral data must enforce stronger privacy-preserving techniques, such as differential privacy, secure multi-party computation (SMPC), and federated learning, to ensure that re-identification risks are minimized while preserving data utility.

9.3 Ethical Responsibilities of Data Collectors and Third Parties

Institutions and organizations collecting user data hold an ethical responsibility to ensure that third parties (such as researchers or data analysts) adhere to strict data governance policies. This includes: Implementing data access control measures to prevent unauthorized use, conducting regular privacy impact assessments to evaluate potential risks associated with data sharing and enforcing data retention policies to limit how long sensitive data is stored.

9.4 Compliance with Legal and Ethical Standards

Data privacy laws such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) establish strict guidelines for data protection. However, the findings of this study suggest that even anonymized datasets pose risks, meaning compliance alone is not always sufficient. Ethical data handling requires proactive privacy measures beyond legal compliance. The institutes ensure that data collection aligns with fundamental human rights regarding privacy and autonomy and regular audits and policy updates to address emerging threats in data security.

9.5 Balancing Data Utility and Privacy Protection

While large-scale data analysis can provide valuable insights into student behavior, academic performance, and institutional decision-making, it must be balanced with strong privacy safeguards. Institutions should strive for a privacy-aware data-sharing model that integrates privacy-preserving machine learning techniques without compromising data utility.

The ethical implications of privacy leakage from behavioral data go beyond technical concerns, touching on fundamental human rights and data protection ethics. Addressing these concerns requires transparent data collection policies, robust privacy-preserving techniques, strict governance of third-party access, and adherence to evolving legal standards. By adopting ethical data practices, institutions can ensure

that advancements in behavioral analysis do not compromise individual privacy and security.

10. Limitations and Future Work

The limitations of this study primarily revolve around dataset constraints, privacy threat assumptions, and the scope of anonymization techniques applied. The dataset consists of 250 students, which limits its generalizability to larger and more diverse educational institutions. Future research should incorporate datasets from multiple universities or regions to validate the findings on a broader scale. Additionally, this study primarily focuses on passive privacy risks arising from data anonymization techniques but does not assess adversarial attacks such as model inversion or membership inference attacks. Evaluating these threats in future studies would provide a more comprehensive risk assessment of anonymized datasets.

Another limitation of this study is the reliance on conventional anonymization techniques, such as removing personally identifiable information, without evaluating advanced privacy-preserving approaches like differential privacy, federated learning, or homomorphic encryption. Assessing these techniques in future research could help mitigate re-identification risks while maintaining data utility. Furthermore, the selected behavioral features used for identity prediction were optimized for accuracy, but potential biases in feature selection may impact real-world applications. A more systematic evaluation of feature biases should be conducted in future studies to ensure fairness and robustness. Finally, while this research highlights privacy risks, it does not fully explore the ethical and legal frameworks necessary for responsible data sharing, particularly concerning compliance with the General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA), and other privacy regulations. A more detailed regulatory analysis should be incorporated to ensure ethical data handling.

Future research should focus on expanding the dataset scope to analyze larger, multi-institutional datasets, thereby improving the generalizability of the findings. Incorporating attack models such as attribute inference, data reconstruction, and membership inference attacks would allow for a more in-depth examination of the vulnerabilities of anonymized datasets. Further studies should explore the implementation of advanced privacy-preserving techniques, including differential privacy, federated learning, and secure multi-party computation, to determine their effectiveness in preventing re-identification risks. Additionally, investigating hybrid privacy models that combine anonymization with cryptographic techniques could help balance privacy protection and data usability.

Developing policy and regulatory guidelines tailored to ethical data handling, particularly in crowdsourced educational data environments, should be prioritized in future studies. A structured framework could ensure that data collection and sharing practices align with privacy laws while minimizing risks to data subjects. Finally, future research should explore the practical implementation challenges of privacy-preserving methods, including computational overhead, scalability, and regulatory barriers, to facilitate their adoption in real-world applications.

11. Conclusion

The study highlights significant privacy risks associated with anonymized educational data, demonstrating that machine learning models can re-identify individuals. The growing popularity of big educational data has elevated the intrinsic risk of leakage of sensitive information in terms of security threats. Data sharing has privacy issues, which need to be solved. The objective of the data collector is to release useful data to data miners without disclosing data providers' identities and sensitive information about them. To achieve this goal, a proper privacy model must be developed to properly quantify the loss of privacy under different attacks.

This research paper discusses only data anonymization, which is not sufficient to guarantee the privacy protection of a student. To prove that we compared online and offline user behavior patterns from a different perspective a method has been introduced for a person’s privacy leakage by integrating both online and offline card datasets. The results suggested that the data anonymization technique alone is not sufficient for big data processing. Making the security system easy and adaptable to each smart card holder requires strong privacy protection. The modern systems i.e., m-commerce and campus card management store personal data with financial information. The ability to control what type of data to reveal and who can access it; is a growing concern currently in big data.

The analysis of the data security and privacy concern is expected to have an integrative and comprehensive security solution to meet the requirement for the preservation of the user’s privacy.

For privacy protection, privacy data identification and isolation are the intrinsic tasks that should be considered during the sharing of anonymized data.

12. Declarations

12.1 Ethical Approval

There is no human and/ or animal involved in this research.

12.2 Conflicts of Interest

The authors declare no conflict of interest.

13. Authors' Contributions

Nadia wrote the paper and completed the draft, Kamlesh Kumar supervised the revision for language correction, and Asif Ali and Mansoor Ahmed provided technical support.

14. Funding

This research was supported by the Higher Education Commission (HEC) of Pakistan through the Indigenous Scholarship Phase-II, Batch-VI. The study was conducted at Dr. Ibrahim Mulla High-Performance Computing Lab, Sindh Madressatul Islam University, Karachi.

15. Availability Of Data and Materials

The data used in this research is not publicly available online. It consists of mobile usage and campus card transaction data collected from students at Central South University, Changsha, Hunan, China and is restricted due to privacy and confidentiality considerations.

16. References

[1] D. Gupta and R. Rani, "A study of big data evolution and research challenges", Journal of information science, vol. 45, no. 3, pp. 322-340, 2019.

[2] M. Hilbert, "Big data for development: A review of promises and challenges", Development Policy Review, vol. 34, no. 1, pp. 135-174, 2016.

[3] J. Huang, "A big data based education information system for university student management", Journal of System and Management Sciences, vol. 13, no. 2, pp. 428-436, 2023.

[4] A. R. Baig and H. Jabeen, "Big data analytics for behavior monitoring of students", Procedia Computer Science, vol. 82, pp. 43-48, 2016.

[5] X. Yang and J. Ge, "Predicting student learning effectiveness in higher education based on big data analysis", Mobile Information Systems, vol. 2022, no. 1, p. 8409780, 2022.

[6] N. Mirbahar, A. A. Laghari, and K. Kumar, "Enhancing Mobile App Recommendations with Crowdsourced Educational Data Using Machine Learning and Deep Learning", IEEE Access, 2025.

[7] M. Li and F. Wang, "Analysis of college students' physical health test data based on big data and health promotion countermeasures", Advances in Multimedia, vol. 2022, no. 1, p. 6879597, 2022.

[8] S. Ma, "Enhancing sports education through big data analytics: Leveraging models for improved teaching strategies", Applied and Computational Engineering, vol. 57, pp. 184-189, 2024.

[9] R. Wyber, S. Vaillancourt, W. Perry, P. Mannava, T. Folaranmi, and L. A. Celi, "Big data in global health: improving health in low- and middle-income countries", Bulletin of the World Health Organization, vol. 93, pp. 203-208, 2015.

[10] L. Shugang and Z. Yuning, "Research on intelligent link prediction model of friend influence based on big data and complex network", 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), 2021: IEEE, pp. 738-743.

[11] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems", Personal and ubiquitous computing, vol. 10, pp. 255-268, 2006.

[12] D. T. Wagner, A. Rice, and A. R. Beresford, "Device Analyzer: Large-scale mobile data collection", ACM SIGMETRICS Performance Evaluation Review, vol. 41, no. 4, pp. 53-56, 2014.

[13] D. T. Wagner, A. Rice, and A. R. Beresford, "Device analyzer: Understanding smartphone usage", Mobile and Ubiquitous Systems: Computing, Networking, and Services: 10th International Conference, MOBIQUITOUS 2013, Tokyo, Japan, December 2-4, 2013, Revised Selected Papers 10, 2014: Springer, pp. 195-208.

[14] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users", Personal and Ubiquitous computing, vol. 7, pp. 275-286, 2003.

[15] A. Pentland, D. Lazer, D. Brewer, and T. Heibeck, “Using reality mining to improve public health and medicine”, Strategy for the Future of Health: IOS Press, 2009, pp. 93-102.

[16] J. K. Laurila et al., “The mobile data challenge: Big data for mobile computing research”, Pervasive computing, 2012.

[17] S. Li, S. Zhao, P. Gope, and L. Da Xu, “Data Privacy Enhancing in the IoT User/Device Behavior Analytics”, ACM Transactions on Sensor Networks, vol. 19, no. 2, pp. 1-13, 2022.

[18] K. Wang, C.-M. Chen, M. Shojafar, Z. Tie, M. Alazab, and S. Kumari, “AFFIRM: Provably forward privacy for searchable encryption in cooperative intelligent transportation system”, IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 22607-22618, 2022.

[19] M. Chen, S. Mao, and Y. Liu, “Big data: A survey”, Mobile networks and applications, vol. 19, pp. 171-209, 2014.

[20] N. Cele, “Big data-driven early alert systems as means of enhancing university student retention and success”, South African Journal of Higher Education, vol. 35, no. 2, pp. 56-72, 2021.

[21] D. Erickson and N. Andrews, “Partnerships among community development, public health, and health care could improve the well-being of low-income people”, Health Affairs, vol. 30, no. 11, pp. 2056-2063, 2011.

[22] M. Bienkowski, M. Feng, and B. Means, “Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief”, Office of Educational Technology, US Department of Education, 2012.

[23] J. Wang and P. Wang, “Innovation research on big data-driven student management work in universities”, 2021 International Wireless Communications and Mobile Computing (IWCMC), 2021: IEEE, pp. 2007-2012.

[24] Z. Chai, Z. Chai, M. Wang, and G. Quan, “How Will the Smart Logistics Service of Universities Based on Big Data Be Transformed and Developed?: Taking Jiangnan University as an Example”, 2023 5th International Conference on Computer Science and Technologies in Education (CSTE), 2023: IEEE, pp. 265-269.

[25] Y. Liu, M. Hu, and X. Lu, “Social frequency analysis of university students via digital campus cards”, 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2016, vol. 1: IEEE, pp. 369-372.

[26] L. WAN, “A research on the correlation of loneliness and social anxiety in college students”, Advances in Psychology, vol. 6, no. 4, pp. 391-397, 2016.

[27] M. Saifuzzaman, T. N. Ananna, M. J. M. Chowdhury, M. S. Ferdous, and F. Chowdhury, “A systematic literature review on wearable health data publishing under differential privacy”, International Journal of Information Security, vol. 21, no. 4, pp. 847-872, 2022.

[28] S. Shen, T. Zhu, D. Wu, W. Wang, and W. Zhou, “From distributed machine learning to federated learning: In the view of data privacy and security”, Concurrency and Computation: Practice and Experience, vol. 34, no. 16, p. e6002, 2022.

[29] L. Zhang, J. Xu, P. Vijayakumar, P. K. Sharma, and U. Ghosh, “Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system”, IEEE Transactions on Network Science and Engineering, vol. 10, no. 5, pp. 2864-2880, 2022.

[30] M. Xue et al., “Use the spear as a shield: An adversarial example based privacy-preserving technique against membership inference attacks”, IEEE Transactions on Emerging Topics in Computing, vol. 11, no. 1, pp. 153-169, 2022.

[31] Y. Zhao, J. Chen, J. Zhang, D. Wu, M. Blumenstein, and S. Yu, “Detecting and mitigating poisoning attacks in federated learning using generative adversarial networks”, Concurrency and Computation: Practice and Experience, vol. 34, no. 7, p. e5906, 2022.

[32] B. Balle, G. Cherubin, and J. Hayes, “Reconstructing training data with informed adversaries”, 2022 IEEE Symposium on Security and Privacy (SP), 2022: IEEE, pp. 1138-1156.

[33] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, “Membership inference attacks on machine learning: A survey”, ACM Computing Surveys (CSUR), vol. 54, no. 11s, pp. 1-37, 2022.

[34] Y. Zhou, J. Wu, H. Wang, and J. He, “Adversarial robustness through bias variance decomposition: A new perspective for federated

learning", Proceedings of the 31st ACM international conference on information & knowledge management, 2022, pp. 2753-2762.

[35] P. Papadopoulos, "Privacy-preserving systems around security, trust and identity", 2022.

[36] V. Avdiienko, "Mining patterns of sensitive data usage", 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, 2015, vol. 2: IEEE, pp. 891-894.

[37] I. S. Rubinstein and W. Hartzog, "Anonymization and risk", Wash. L. Rev., vol. 91, p. 703, 2016.

[38] F. Kargl, R. W. van der Heijden, B. Erb, and C. Bösch, "Privacy in mobile sensing", Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics: Springer, 2022, pp. 13-23.

[39] M. B. Kursa, "Robustness of Random Forest-based gene selection methods", BMC bioinformatics, vol. 15, pp. 1-8, 2014.

[40] X. Xie, M.-J. Yuan, X. Bai, W. Gao, and Z.-H. Zhou, "On the Gini-impurity preservation for privacy random forests", Advances in Neural Information Processing Systems, vol. 36, pp. 45055-45082, 2023.

[41] Z. Yan and Y. Yao, "Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO)", Chemometrics and Intelligent Laboratory Systems, vol. 146, pp. 136-146, 2015.

[42] M. Raab, "Data correction of the TOI system by statistical methods and machine learning/submitted by Michaela Raab", 2021.

[43] B. Yu, "Stability", 2013.

[44] G. Chen and J. Chen, "A novel wrapper method for feature selection and its applications", Neurocomputing, vol. 159, pp. 219-226, 2015.

[45] I. Inza, P. Larrañaga, R. Etxeberria, and B. Sierra, "Feature subset selection by Bayesian network-based optimization", Artificial intelligence, vol. 123, no. 1-2, pp. 157-184, 2000.

[46] P. Kumari, "Different Approaches of Quantitative Structure Retention Relationship of Small Molecules in Liquid Chromatography (QSRR)", Universite de Liege (Belgium), 2024.

[47] N. Balakrishnan, V. Voinov, and M. S. Nikulin, Chi-squared goodness of fit tests with applications. Academic Press, 2013.

[48] H. Zhou, X. Wang, and R. Zhu, "Feature selection based on mutual information with correlation coefficient", Applied intelligence, vol. 52, no. 5, pp. 5457-5474, 2022.

[49] L. Pappalardo, F. Simini, G. Barlacchi, and R. Pellungrini, "Scikit-mobility: A Python library for the analysis, generation, and risk assessment of mobility data", Journal of Statistical Software, vol. 103, pp. 1-38, 2022.

