How to disappear online

JOÃO VASCO ALMEIDA SOBRAL SIBORRO REIS, Instituto Superior Técnico, Portugal

In the digital age, personal data privacy has become a critical concern, driven by the proliferation of online platforms and the increasing volume of data storage. This thesis explores the challenges and methodologies surrounding digital erasure, with a particular focus on compliance with privacy regulations such as the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the Artificial Intelligence Act (AIA). The work investigates machine learning techniques, especially Random Forest and Decision Tree classifiers, to track, match, and prepare personal data for deletion across complex systems. By adapting intrusion recovery methodologies, this research proposes a system for accurately identifying and managing data marked for deletion, ensuring organizations can meet regulatory obligations. The analysis demonstrates how machine learning models are particularly effective for matching data, with this thesis focusing specifically on the identification and matching of logs as a crucial step in the broader process of data deletion. This thesis contributes to the broader discourse on data privacy by addressing the practical challenges of data management in distributed environments, paving the way for an initial approach for what would be one of the steps in order to obtain a more secure and compliant data deletion.

Additional Key Words and Phrases: Data Privacy, Digital Erasure, Swedish Act, GDPR, CCPA, AIA, Intrusion Recovery, Security, Machine Learning

1 Introduction

In the contemporary digital era, the significance of information management has escalated more than ever. As elucidated in the work by Gates and Matthews, personal data has emerged as a vital asset in realms such as targeted marketing and advertising, with its value extending even to illicit black markets [13]. This monetization of personal data raises profound concerns regarding data ownership and privacy. The ambiguity surrounding legal ownership of data and the existing legal frameworks pose significant challenges, especially in the context of digital erasure [13]. Organizations are faced with the increasingly difficult challenge of managing and protecting vast amounts of personal data. Regulations grant individuals the right to request the deletion of their personal information, making data erasure a crucial component of compliance. However, ensuring the complete and secure deletion of data across complex and distributed systems is far from straightforward [1]. Large organizations, with vast databases and intricate data flows, struggle to implement effective solutions for tracking, identifying, and securely erasing data [24]. This thesis addresses this problem by exploring innovative approaches to digital erasure, focusing on the evaluation of techniques that can help organizations prepare for and manage the process of securely removing data in compliance with privacy regulations. To combat data corruption and respect the privacy of individuals, several regulatory approaches have emerged. With the appearance of the GDPR [11], the AIA [7], and the CCPA[5], there has been a significant shift in the landscape of data privacy and digital ethics.

These frameworks collectively signify a global movement towards enhanced data privacy, ethical deployment of AI, and digital society strengthening consumer rights in the digital domain, setting critical standards for organizations globally. On the other hand, is the absence of practical solutions to erase the data. The implementation of the GDPR has posed significant challenges for many technology companies, requiring a comprehensive reevaluation of their data handling practices. Notably, the financial and operational burdens associated with achieving compliance have been substantial, often culminating in heavy sanctions for non-compliance. This situation is exemplified in a study, which delves into the experiences of nine major tech companies revealing the multifaceted nature of GDPR compliance issues [22]. One notable response to the stringent requirements of GDPR has been the decision by some American companies, including prominent news publishers, to restrict access to their platforms for EU users. This measure, while mitigating the risk of penalties, underscores the broader compliance challenges faced by these corporations. Furthermore, the complexities inherent in GDPR compliance are highlighted by a survey cited in the paper, which found that a majority of companies grapple with both financial constraints and the intricacies of deploying automated privacy rights management solutions [22].

The risk of data loss, whether through hardware failure, human error, or cyber threats, poses significant challenges for both individuals and organizations. As highlighted in [10], data loss prevention is a critical aspect of safeguarding confidential information within an organization's boundaries. The paper emphasizes the importance of understanding and mitigating the risks associated with data loss in the digital age, particularly as businesses navigate the complexities of digital risk management and GDPR compliance. Furthermore, the potential for irreparable consequences due to failures in personal data protection systems underscores the necessity of robust digital erasure methodologies. This thesis, therefore, aims to address these challenges by proposing an effective framework for digital erasure, ensuring the secure and compliant deletion of personal data in line with GDPR requirements. The framework will incorporate an innovative approach for risk mitigation, data protection, and compliance, contributing to the evolving discourse on digital privacy and data management in a world increasingly driven by data.

Intrusion recovery is a process in information security that focuses on identifying, mitigating, and recovering from unauthorized intrusions or attacks on a system. The goal is to restore the system to a safe, secure state by detecting and eliminating any compromised data or corrupted files. This process typically involves identifying the point of intrusion, analyzing the damage, isolating affected components, and restoring the system without the lingering effects of the attack. In the context of this thesis, intrusion recovery techniques are adapted to address the challenges of digital erasure. The similarities between handling corrupted data in intrusion scenarios and tracking data that must be deleted under privacy regulations, like the GDPR, allow for the application of similar methodologies. Specifically, the focus is on identifying personal data that has been flagged for deletion, tracing it across complex systems, and ensuring that it can be effectively removed in the future. By leveraging the principles of intrusion recovery, the thesis aims to improve the

Author's Contact Information: João Vasco Almeida Sobral Siborro Reis, joao.vasco.sobral@tecnico.ulisboa.pt, Instituto Superior Técnico, Lisbon, Portugal.

accuracy and efficacy of tracking data for eventual deletion. Intrusion recovery methods, such as identifying digital footprints and isolating compromised components, provide a valuable framework for analyzing how data can be securely marked and prepared for deletion, minimizing the risk of incomplete removal or unauthorized recovery. This approach ensures that data management processes align with the stringent requirements of digital privacy laws.

The primary objective of this thesis is to develop and evaluate a framework for analyzing the effectiveness of digital erasure techniques, particularly in the context of GDPR-compliant data management. Instead of directly implementing data deletion protocols, this work focuses on studying how accurately data can be identified and tracked for eventual deletion, and assessing the performance of various techniques used to facilitate this process. The thesis is structured around the following key steps:

Intrusion Identification and Data Tracking: The thesis proposes the adoption of methods from intrusion recovery to detect and track personal data marked for deletion. By treating such data as analogous to compromised or corrupted data in intrusion scenarios, this approach can enable the precise identification of digital footprints that should be erased.

Data Matching Using Machine Learning: A critical part of the work involves using machine learning models, such as Random Forest and decision trees, to evaluate how well they can identify and match data flagged for deletion. This step assesses the capability of these algorithms to accurately trace data points in large, complex datasets, simulating scenarios where deletion is necessary.

Analysis of Results: Instead of carrying out the deletion itself, the thesis evaluates the results of data identification processes. Metrics such as F1 Score, accuracy, and precision are used to determine how effectively data has been matched and isolated, and to assess the feasibility of eventual deletion. The analysis provides insights into the strengths and limitations of the methods in ensuring comprehensive data tracking and deletion preparedness. We additionaly discuss whether the techniques used for data matching and isolation are suitable for real-world digital erasure tasks. This involves analyzing the deletion potential based on the results obtained from the models. By analyzing these aspects, the thesis contributes to the field of digital privacy and data management by providing insights into the feasibility and effectiveness of machine learning approaches for preparing data for deletion.

The rest of the paper is structured as follows: Chapter 2 provides an overview of background and related work. Reviews the relevant literature on data deletion techniques and their application in machine learning. Chapter 3 describes the proposed solution, particularly the methods for matching logs and detecting unauthorized data. Chapter 4 introduces the first experimental scenario involving one form and one database, presenting initial results from matching attempts. Chapter 5 expands the complexity with a scenario involving database-to-database communication using numerical IDs, exploring the system's ability to integrate and match data across databases. Chapter 6 investigates matching in scenarios where textual identifiers are used, comparing the efficacy of string-matching algorithms such as Hamming and Levenshtein distances. Finally, Chapter 7 concludes the thesis by summarizing the results and proposing areas for future research, particularly focusing on real-world data application and enhanced machine learning models.

2 Related Work

User privacy in the digital world is an increasingly critical issue. A 2016 survey titled "The State of Privacy in Post-Snowden America" revealed that 65% of respondents consider it very important to control data collected about them, and 86% had taken steps to protect their privacy. This reflects growing awareness of digital privacy concerns. However, studies show significant gaps in data deletion practices on websites. For example, an analysis of 150 sites found that 26% didn't offer any option for users to delete their data. Furthermore, 83% of these websites lacked clear information about account deletion, highlighting the challenges users face in managing their personal data. The complexity of data deletion policies, requiring a university-level reading ability to understand, only adds to these issues. Historically, privacy laws have evolved, starting with the 1973 Swedish Data Act, the first national data protection law. This law set the stage for future legislation, including the EU's GDPR and CCPA, both of which emphasize user rights and transparency in data management. The ongoing evolution of privacy regulations reflects the growing complexity of digital data and the need for stronger user control and protection.

Given the growing significance of data in today's world, one might expect to find extensive research in this field. Surprisingly, this is not the case. Many existing approaches are either in their infancy or simply inadequate. A fundamental step in addressing data-related issues is to first consider whether the data should be deleted, followed by the actual process of deletion. This section delves into the latest methodologies for "Unlearning," which involves techniques for eliminating data that has been used in model training. This process is crucial for compliance with regulations like the GDPR. Additionally, we explore some methods in the field of Intrusion Recovery, where efforts are primarily focused on rectifying database intrusions.

2.1 Machine Unlearning

The study by Bourtoule et al. [2] delves into the critical need for machine unlearning in the context of digital privacy. This concept is especially relevant given the rise of privacy regulations like GDPR and CCPA, which mandate the right to be forgotten. Given dataset (D), one can train various models (like Deep Neural Networks (DNNs)) that effectively learn from this dataset. When a new data point (du) is added, creating a new dataset (D'), there are multiple ways to train a new model on D'. One approach is to use the parameters from a previously trained model (MA) as a starting point for the new model (MB), rather than beginning from scratch. This approach, however, faces challenges in measuring the influence of the new data point (du) on MB's parameters, making it difficult to reverse the process without having saved a copy of MA beforehand. The paper further discusses a strategy called "slicing" and the concept of plausible deniability in privacy, which involves retraining the model from scratch without the specific data point to ensure it doesn't influence the model's training. The 'SISA (Sharded Isolated Sliced and Aggregated) training' framework is presented by the authors as a novel way to accelerate the unlearning process in machine learning models. By dividing the training data into separate shards, this technique restricts the impact of each data point to the appropriate shard. Computational overhead is decreased because only the model linked to the impacted shard needs to be retrained in response to an unlearning request. The aggregation method mentioned is a label-based majority vote, where each model contributes equally to the final decision. This method is effective but might lose information in scenarios where models assign high scores to multiple classes. To address this, the paper evaluates a refined strategy where the entire prediction vectors (post-softmax vectors indicating a model's confidence in predicting each class) are averaged, and the label with the highest value is selected. Bourtoule et al. contribute significantly to the field by formulating a definition of unlearning and demonstrating the practicality of SISA training. It was definitely a step forward on a work using Machine learning that was primarily motivated by privacy. However, the approach faces challenges like the creation of weak learners and the need for comprehensive hyperparameter tuning.

2.2 Recovery from malicious transactions

In the context of the current data-driven business environment, managing large volumes of data effectively is a significant challenge. Traditional tracking methods become impractical in large companies, where the sheer volume of data and the complexity of adherence to specific protocols by numerous employees are overwhelming. This scenario underscores the importance of efficient data management strategies, particularly in complex organizational settings. In the field of intrusion recovery, which focuses on restoring corrupted or compromised data, there are several approaches that can be adapted for digital erasure. Sanare, with its ability to accurately trace and eliminate specific data alterations, presents an adaptable solution that extends beyond intrusion recovery to GDPR-compliant data management. Sanare [20],is an intrusion recovery system specifically designed for web applications that maintain shared state information. The primary functionality of Sanare involves two key processes: first, identifying all changes made by an intrusion and any subsequent changes that depend on the initial intrusion; second, erasing the data altered by the intrusion. This system, leveraging its deep learning scheme Matchare, effectively links HTTP (Hypertext Transfer Protocol) requests to database operations, which is crucial for tracking and reversing unauthorized data changes. Matos, Pardal, and Correia [19] propose Rectify, a service for recovering PaaS (Platform as a Service) web applications from intrusions, in their paper. Rectify operates by logging HTTP requests and database statements, using machine learning to correlate these logs for identifying malicious activities. Unlike the previous approach, which focuses on database transactions, Rectify addresses intrusions at the application level, recovering the system to a state as if the intrusion never occurred, using compensation operations. While this approach aim to recover from malicious activities, their methodologies focus on the application level, leveraging machine learning to associate HTTP requests with database modifications in a PaaS environment.

3 Solution

This work proposes a method for tracking unauthorized data by correlating logs from distinct sources, such as HTTP requests and database logs. The initial phase focuses on log matching to identify the location and nature of unauthorized data. We adapt the Sanare [20] framework to detect and mitigate unauthorized data creation or modification, which is conceptualized as an intrusion that must be removed. This method is particularly crucial for achieving GDPR compliance, as identifying the last interaction with such data is the first step in the deletion process [14].

3.1 System Proposal

In the context of GDPR, an *intrusion* is defined as the presence of personal data in a system beyond a user's request for its deletion. This process starts by identifying the HTTP requests that initiated data creation, allowing us to track the personal data and ensure its removal. According to Article 5 of GDPR, the principle of data minimization emphasizes limiting data collection to what is necessary, reinforcing the importance of this approach [11].

For the development of the proof of concept, we consider a system that consists of a web application connected to multiple data repositories, such as a relational database and a file system. The creation of personal data is initiated through HTTP requests, which serve as the mechanism for user interactions with the web application. The personal data itself is stored in a relational database, where it can be tracked and processed according to GDPR requirements. The system logs user activities and stores corresponding HTTP requests, allowing for the identification and potential deletion of data in compliance with privacy regulations. This setup reflects typical web-based systems where data flows between the user interface and the back-end storage repositories, creating a robust framework for evaluating digital erasure methods and data deletion strategies within complex data architectures.

The goal at this stage is to establish a clear connection between HTTP requests and the corresponding database operations to accurately map data dependencies for GDPR compliance. We propose using machine learning, particularly ensemble methods like Random Forests [3], which are effective for pattern recognition and high-dimensional data tasks [9]. These models reduce overfitting and enhance accuracy by aggregating predictions from multiple decision trees [16]. This method allows for precise log-data matching, improving the detection of unauthorized data [20].

Once the relevant data is identified, the next step is isolating it to ensure that the deletion process only affects the targeted data. The deletion process must be exhaustive and secure, ensuring that all traces of the data are removed in compliance with GDPR requirements [15]. This includes removing records from databases, logs, and backups, with post-deletion verification to confirm irreversibility.

3.2 Building the Dataset

We used a WordPress web form, as it powers a significant portion of websites globally. The form was designed to collect personal data while adhering to GDPR guidelines. A WordPress-based web form was created to gather demographic data (e.g., name, country, height, gender, etc.). The fields were designed to comply with

GDPR regulations, ensuring transparent data collection and processing. The form was tested locally using the MAMP (Macintosh, Apache, MySQL, PHP) server environment. The application was deployed using MAMP, allowing the simulation of a production environment for development and testing. MAMP supports technologies like PHP (Hypertext Preprocessor), MySQL, and Apache, ensuring smooth integration of web components. HTTP and database logs were captured through custom server configurations. We configured the Apache server to log all HTTP requests and enabled dumpio logging to capture request/response bodies. MySQL was set to log every executed statement, creating a comprehensive record of all user interactions and database operations. To collect large datasets, we automated form filling using Python's Selenium [1] and Pandas [2] libraries, simulating user interactions with randomized inputs. This generated a dataset of HTTP requests and corresponding database logs for machine learning analysis. The automation script used names from a CSV file, simulating diverse user data. Selenium WebDriver automated form submissions, while random sleep times mimicked human-like behavior. Apache and MySQL logs were parsed with regular expressions for detailed analysis. The automated script logged and parsed form submission data, storing it in a CSV file for future analysis. Challenges included handling dynamic web elements and complex log file formats, which were managed using explicit waits and optimized log parsing methods. The form included three main field types: text fields (e.g., name), numeric fields (e.g., height, weight), and dropdown menus (e.g., country, gender, etc.). This combination ensured accurate, diverse data collection while maintaining consistency for further analysis. The initial dataset consisted of SQL (Structured Query Language) logs aligned with corresponding HTTP request logs, producing 823 matched records. We also added 2,177 unmatched records to simulate more complex scenarios. This balanced dataset provided a strong basis for model training. The dataset was cleaned by removing irrelevant columns and duplicate entries. Categorical variables were encoded into numeric values, ensuring proper representation for machine learning models. For example, ethnicity values were mapped to numbers as follows:

```
ethnicity-mapping = {
    'Black or African American': 10,
    'Indian': 7,
    'Latino or Hispanic': 5,
    'Asian': 2,
    'White (Caucasian)': 0,
    'Other': -1
}
```

Missing values were consistently encoded as −2, while special cases like "Other" were mapped to −1. This ensured uniform treatment of missing and non-standard data, allowing for coherent analysis during model training. The dataset, now prepared with 36 variables, was ready for training. We used a Random Forest classifier for its robustness in handling multiple features and its ability to avoid overfitting [3]. This choice was motivated by the need to classify matched and unmatched records effectively. As we continued refining the model, we explored additional approaches to improve performance and interpretability.

4 Scenario 1: One Form - One Database

In this chapter, we explore the simplest approach to data deletion: attempting to delete a user's data from a single DB (Database) after a form submission. A form is filled out, generating HTTP requests that populate the database, as described in the previous chapter. The data is then stored in one database, and our goal is to match this data with the dataset built for this purpose.

4.1 Original Logs with no Correspondence Knowledge

This section analyzes the first attempt at training a Random Forest classifier to match the first 18 columns with the subsequent 18 columns of the dataset. The model uses an additional column labeled as "Matched" or "Unmatched" as the target variable. While the initial results are promising, both false positives (FP) and false negatives (FN) present challenges in the context of data deletion.

- **False Positives (FP)**: Incorrectly classifying data as a match could result in deleting valid data.
- **False Negatives (FN)**: Failing to identify data that should be deleted could lead to privacy or compliance issues.

Our goal is to achieve high precision and high recall. High precision minimizes the deletion of necessary data, and high recall ensures that all data flagged for deletion is correctly identified. The confusion matrix and performance metrics for the training and testing datasets are summarized below. The confusion matrix in Figure 1 shows that there are 159 true positives (TP), 414 true negatives (TN), 2 FP, and 25 FN, which are critical in understanding the overall performance of the model in this scenario.

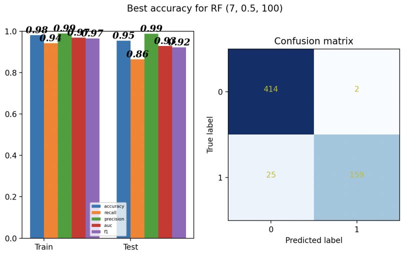


Fig. 1. Confusion matrix analysis showing 159 TP, 414 TN, 2 FP, and 25 FN.

The accuracy, recall, and precision metrics are shown in Figure 1, and are important for analyzing how well the model performed on both the training and test sets.

4.2 Impact of Max Depth, Number of Estimators, and Feature Proportion on Accuracy

Figure 2 illustrates how Random Forest classifier accuracy varies with the number of estimators and "max-depth" of trees, across different proportions of features (10%, 30%, 50%, 70%, 90%). With a "max-depth" of 2, accuracy remains flat, suggesting that shallow trees cannot capture complex patterns. At "max-depth" of 5, accuracy improves slightly with more estimators. The best accuracy

[1] Selenium: https://www.selenium.dev/

[2] Pandas: https://pandas.pydata.org/

is achieved with "max-depth" of 7, especially when using a larger proportion of features.

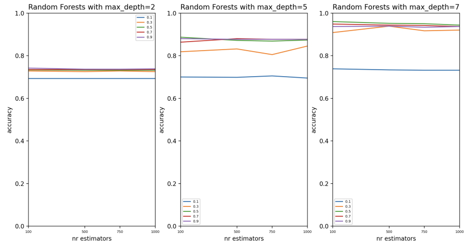


Fig. 2. Accuracy of the Random Forest classifier with varying "max-depth" (2, 5, 7) and feature proportions. Accuracy improves with increasing max-depth, estimators, and feature proportion.

4.3 Conclusion from the First Results

The Random Forest model shows strong performance, achieving high accuracy, recall, precision, AUC (Area Under the Curve), and F1 scores on both the training and testing sets. The best results are achieved with a max-depth of 7, 100 estimators, and feature proportion of 50%. However, the presence of false negatives (FN = 25) and false positives (FP = 2) indicates room for improvement in data deletion tasks [3].

4.4 Comparative Results

In a second set of experiments, the max_depth was increased to values of 7, 15, 20, and 35, allowing the model to capture more intricate patterns. This led to near-perfect accuracy, with the best model configuration achieving a **99% accuracy**, **97% recall**, **100% precision**, and **99% F1 score**. False negatives were reduced to zero, as show in the Figures (3, 4) and Table 1.

Metric	Train	Test		
Accuracy	1.00	0.99	True Positives (TP)	179
Recall	1.00	0.97	True Negatives (TN)	416
Precision	1.00	1.00	False Positives (FP)	0
AUC	1.00	0.99	False Negatives (FN)	5
F1 Score	1.00	0.99		

Table 1. Random Forest model with max_depth=15, showing near-perfect classification.

The second attempt at refining the Random Forest model addressed earlier issues, yielding near-perfect performance. These results demonstrate the model's effectiveness for data deletion tasks, provided that appropriate hyperparameter tuning is performed.

This chapter demonstrates that by carefully tuning parameters such as max-depth and feature proportions, the Random Forest model can achieve high accuracy in identifying matches and non-matches, ensuring reliable data deletion.

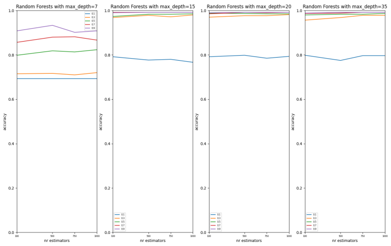


Fig. 3. Accuracy of Random Forest models with varying max_depth values, showing improvement with deeper trees.

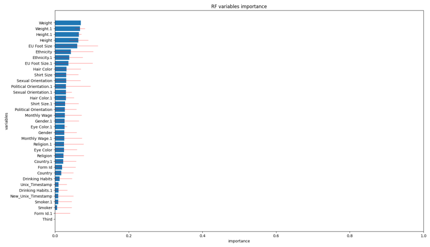


Fig. 4. Feature importance as determined by the Random Forest model.

5 Scenario 2: DB - DB with numerical IDs

Given the excellent results achieved in the initial phase of this research, where the correlation between HTTP requests and database logs was successfully identified, we extend the investigation to connections between logs from two different databases. This scenario mirrors real-world conditions where information is often dispersed across multiple databases.

5.1 Scenario Description

To simulate this environment, a second form was developed, incorporating user authentication. Users authenticate themselves before submitting data to a database. They then submit data to another form using the same credentials, with data stored in a second database. The forms were filled out randomly, ensuring varied results in both databases. In the initial phase, all variables were selected from dropdown menus, with up to four options. The two datasets represent the user's actions across different services. Since the authentication key is consistent across both services, all variables, including the key, are displayed in the dataframe. The goal is to assess how the Random Forest model performs when a consistent key links datasets across multiple databases. We store 42 variables per user, including 20 from each form submission and 2 for the authentication keys used during each login process. The dataset comprises 2177 rows of non-matches and 823 rows of correct matches, with 27.43% representing matched correspondences. The 43 features include 21 corresponding to HTTP request logs, 21 from the database logs, and 1 for labeling rows as matched or unmatched. These keys link the data from the two separate instances, providing a consistent identifier across the datasets. The objective is to determine the Random Forest model's performance under these conditions and how the presence of a key affects the dataset when integrating data across multiple databases.

5.2 With Different IDs

We analyzed the impact of varying the number of input variables on the performance of the Random Forest model. Three configurations were tested: one with 42 variables (including 2 IDs), one with 12 variables, and one with 4 variables. The goal was to understand how the number of variables influences the model's ability to correctly identify patterns, particularly focusing on the authentication keys present across both datasets.

With only 4 variables, the model demonstrated the best performance due to reduced complexity, allowing it to focus on the most relevant features (ID keys). As the number of variables increased to 12 and 42, performance decreased as the introduction of irrelevant features made it more difficult for the model to consistently test the ID variables. This resulted in worse generalization.

5.3 With the Same ID

In the next experiment, we computed the difference between corresponding columns when the authentication key was known. The model achieved perfect accuracy, precision, recall, and F1 scores due to the key's dominance. The confusion matrix confirmed that all instances were correctly classified, and the variable importance chart highlighted the ovewhelming influence of the key. The attached image (Figures 5) illustrates this point. This result underscores the importance of identifying keys for matching datasets across databases and highlights the potential for simplification when such keys are known.

5.4 Experiment with Text Variables

We also explored a scenario using text-based inputs instead of dropdowns. Text inputs introduced increased complexity and variability, leading to poorer performance. The confusion matrices and accuracy charts reflect this decline, which is attributed to the variability and noise from the text inputs.

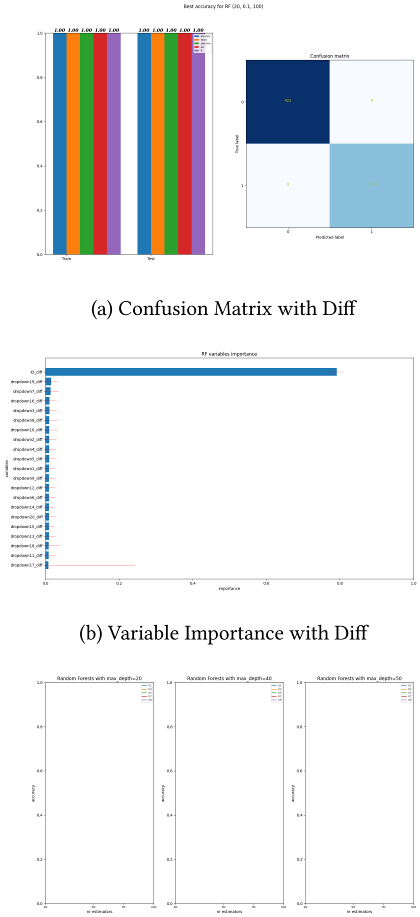
These findings emphasize the challenges of working with diverse data types in machine learning and underscore the importance of feature engineering and preprocessing. As seen in Figure 6.

5.5 Analysis of SHAP Values in Three Experimental Scenarios

We used SHAP(SHapley Additive exPlanations) values to analyze the contribution of different features to model predictions across three experiments. In the first experiment, the most influential features were height and weight differences (Figure 7) . In the second experiment, the IDkey dominated the predictions. In the third experiment, the IDkey also revealed dominance in terms of relevance.

5.6 Conclusion

In this chapter, we explored various scenarios involving multiple databases with differing and consistent IDs. The experiments revealed that: Reducing the number of irrelevant variables improved model performance by allowing it to focus on critical features (e.g., IDs); When the key is known, calculating differences between corresponding columns simplifies the classification task, leading to perfect performance; Text-based inputs introduce significant variability, reducing model performance compared to structured dropdown inputs. Overall, careful consideration of the number and type of input variables is essential for optimizing model performance when matching data across multiple databases.



(a) Confusion Matrix with Diff

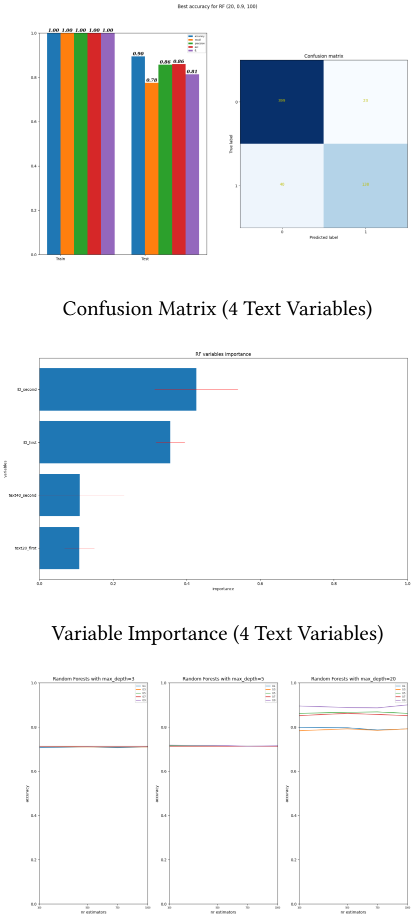
(b) Variable Importance with Diff

(c) Accuracy vs Estimators with Diff

Fig. 5. Random Forest performance after calculating the difference between corresponding columns when the key is known. The model achieves perfect accuracy due to the significant impact of the key.

6 Scenario 3: DB - DB with textual IDs

In many data integration scenarios, especially involving personal data, a key challenge is ensuring the integrity and consistency of identifiers, such as person names, across multiple databases. This issue is particularly relevant in contexts where a person's name may appear with slight variations, abbreviations, or even missing components in one database compared to another [23]. In our study, we explored this scenario by simulating a situation where names



Confusion Matrix (4 Text Variables)

Variable Importance (4 Text Variables)

Accuracy vs Estimators (4 Text Variables)

Fig. 6. Random Forest performance with 4 text variables (including 2 IDs).

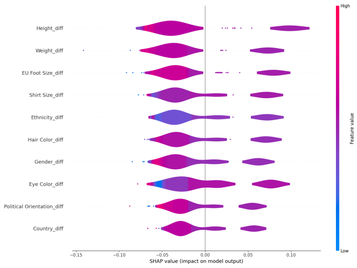


Fig. 7. SHAP summary plot for the first experiment analyzing feature differences from the first form.

from forms had slight differences, including variations in abbreviations, missing letters, additional characters, or differing numbers of names. To evaluate the effectiveness of different string comparison methods, we considered longer names with an average of five components (first name, middle names, and surnames). This reflects realistic complexities encountered in multi-source data environments. To measure how similar or different two names were, we tested both Hamming and Levenshtein distances. The Hamming distance, which counts the number of differing characters at the same positions in strings of equal length, quickly proved inadequate due to its limitations: it only works well when comparing strings of the same length and fails to account for insertions or deletions [26].

6.1 Levenshtein Distance: A Superior Metric

The Levenshtein distance [17] (Edit Distance), on the other hand, measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into another [21]. This approach is particularly suited for our scenario because it accurately captures real-world variations in names. For instance, it can handle cases where one name includes additional middle names or where abbreviations are used. Figures 8 illustrate how the Levenshtein distance provided significantly more reliable classification results compared to Hamming distance, as observed in the confusion matrix in Figure 10. The Random Forest model using Levenshtein distance achieved perfect classification with zero misclassifications in the confusion matrix, as seen in Figure 8, indicating a robust and accurate identification of name matches.

6.2 Performance Analysis

Figure 9 shows the feature importance derived from the model using Levenshtein distance. Levenshtein distance consistently emerged as the most influential feature, underscoring its superior role in distinguishing matching and non-matching names. This further highlights why Levenshtein distance is preferable in data linkage tasks involving person names. Our results demonstrate that Levenshtein distance offers a more practical and effective solution for comparing names across databases, particularly in scenarios involving varied name formats and inconsistencies. This approach allows for more accurate data integration, reducing the risk of mismatches and ensuring greater integrity in person identification [8].

6.3 Testing Name Matching with Shorter Names

In this experiment, we aimed to investigate the performance of name matching algorithms when dealing with shorter names, as some individuals, particularly in certain cultural contexts, may have only two names (e.g., "John Doe"). The goal was to understand how the effectiveness of the matching process would change when the names used as identifiers are shorter and more likely to overlap across different individuals [21]. We simulated a scenario with shorter names to test the effectiveness of both Hamming and Levenshtein distances in distinguishing between different individuals. In contrast to the previous experiment with longer names, this setup presented a more challenging situation due to the higher probability of identical name sequences appearing for distinct individuals, such as "John Doe" appearing for two different people [26]. As expected, the results were considerably worse compared to scenarios involving longer names. This is primarily because shorter names offer less uniqueness as identifiers. In some cases, names of non-matching individuals were identical, leading to false positives in the classification task. For instance, if two distinct individuals were recorded as "John Doe," the

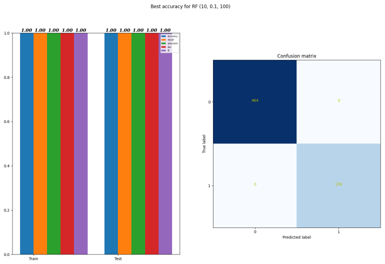


Fig. 8. Confusion matrix showing zero misclassifications when using Levenshtein distance, demonstrating its superior accuracy in classifying name matches.

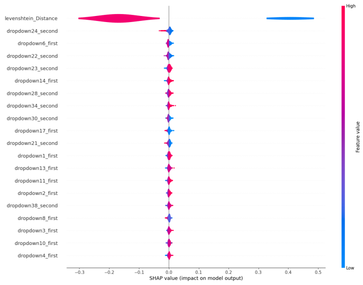


Fig. 9. SHAP values for the Levenshtein distance model, showing its high impact on predictions.

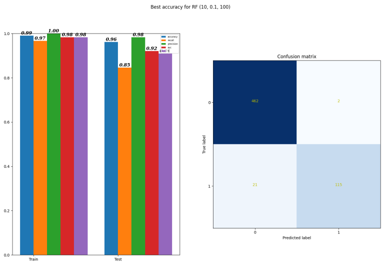


Fig. 10. Confusion matrix showing the results when using Hamming distance, highlighting its limitations compared to Levenshtein distance.

algorithm would struggle to differentiate them based on name alone, significantly undermining the integrity of the matching process [6].

Figure, 11, show the performance of the Random Forest model using Hamming distance for shorter names. Notably, the confusion matrix in Figure 11 demonstrates an increase in misclassifications, confirming the difficulty in distinguishing between distinct records with similar names. On the other hand, 12 illustrate that while Levenshtein distance still performed better than Hamming, the overall accuracy decreased compared to the scenario with longer names. These findings highlight the challenges of using shorter names as keys in data integration tasks and reinforce the importance of using more robust identifiers. While Levenshtein distance remains a preferable option over Hamming distance, its effectiveness is diminished when dealing with less distinctive name sets. This underscores the critical need for comprehensive approaches in managing person identifiers, especially in contexts where data integrity is essential [6, 21, 26].

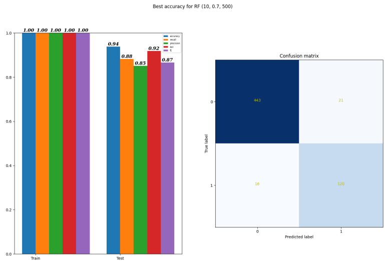


Fig. 11. Confusion matrix using Hamming distance with shorter names, showing increased misclassifications.

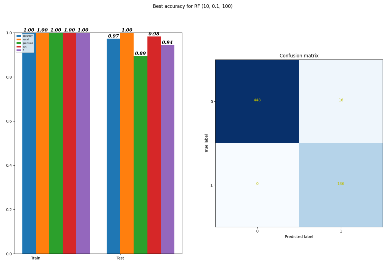


Fig. 12. Confusion matrix using Levenshtein distance with shorter names.

6.4 Encrypted Identifiers and Text Data

In this experiment, we aimed to evaluate the model's performance using six variables, including two unknown identifiers (IDs) and four text data columns. This setup simulates a real-world scenario where each participant is required to write their name as an identifier and provide two texts describing themselves. To train a Random Forest model under these conditions, we encoded the data by calculating the Hamming distances between all possible combinations of columns. This approach was chosen because the specific columns representing the IDs were unknown, making it necessary to consider all column pairs. Despite the inherent challenges, the model demonstrated the ability to find matches with reasonable accuracy, as seen in the results. This scenario reflects real-life situations where certain information in databases, such as personal identifiers, is encrypted along with associated data. However, users may not know which column specifically corresponds to the encrypted identifier. In our simulation, the individual's name was considered as the key, representing a common case in database matching tasks. The model was still able to identify most of the corrected matched cases even with the ID being unknown. Figures 13 and 14 illustrate the model's performance under these conditions. Figure 13 presents the confusion matrix, highlighting the model's ability to correctly identify matches even with encrypted identifiers, and Figure 14 demonstrates the importance of each variable in the Random Forest model.

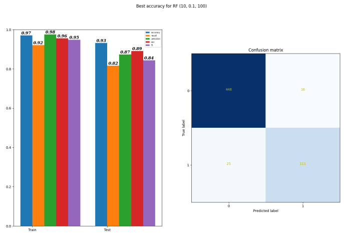


Fig. 13. Confusion matrix showing the model's classification performance with encrypted IDs and text data.

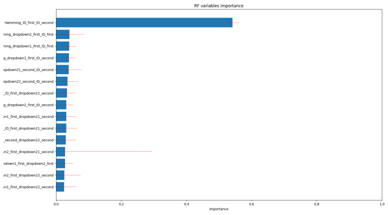


Fig. 14. Variable importance plot highlighting the significance of each column in the Random Forest model.

6.5 Results with Levenshtein Distance and Conclusion

When the same experiment was conducted using Levenshtein distance, the results significantly improved. Levenshtein distance, which accounts for insertions, deletions, and substitutions, provided a more comprehensive measure of similarity between strings, enhancing the model's ability to identify correct matches even when identifiers were encrypted or varied in structure. Figure 15 illustrate the superior performance of Levenshtein distance in this context. The Random Forest model achieved higher accuracy, fewer misclassifications, and demonstrated the critical role of Levenshtein distance features in distinguishing between matching and non-matching records.

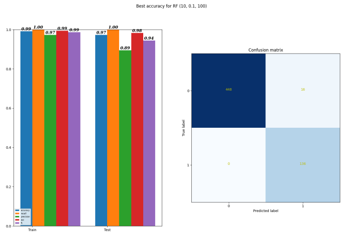


Fig. 15. Confusion matrix for the Levenshtein distance, with significantly fewer misclassifications.

The comparison shows that Levenshtein distance is a more effective metric for matching encrypted or anonymized data across databases, especially when the structure of identifiers is not predetermined. The enhanced ability to handle varied and complex edits makes Levenshtein distance a preferred choice over Hamming distance, providing a more reliable solution for data integration tasks involving person identifiers.

7 Conclusion and Future Work

The motivation behind this thesis comes from the increasing need for organizations to comply with data privacy regulations, particularly the GDPR [11]. As organizations collect data across various distributed systems, it becomes critical to trace where unauthorized data is located, especially in cases of breaches or unauthorized modifications. This work addresses the challenge of improving data tracking and deletion processes in compliance with GDPR. One key step toward achieving this goal is the matching of logs from different systems (such as HTTP requests and database logs). Through the proposed machine learning techniques, we aimed to develop methods that enhance this matching process, allowing for more accurate identification of unauthorized data and thus enabling timely and effective responses to data breaches or privacy violations[21, 23]. This thesis proposes a method that focuses on correlating logs from HTTP requests with database logs, aiming to detect unauthorized data actions and facilitate its potential removal. The proposed solution, which adapts the SANARE framework, emphasizes the data matching phase as an essential first step toward identifying where unauthorized data exists, before considering further actions such as deletion. This approach is highly relevant in the context of GDPR, where organizations must quickly locate and mitigate data breaches [18]. By linking logs from different data sources, we can more effectively trace data origins and identify security breaches. This thesis specifically focuses on the initial task of matching logs from different systems as a precursor to later data deletion or containment measures. To test the proposed models, synthetic datasets were generated, representing realistic user actions and submissions across multiple systems[23]. These datasets included HTTP request logs and database entries, simulating real-world interactions where users provide information that is stored in different databases. The experiments involved varying the number of input features (from 4 to 42 variables) to evaluate the model's ability to recognize patterns under different conditions. The focus was on understanding the performance of Decision Trees [4] and Random Forests [3] when handling both well-aligned and noisy datasets, providing insights into how these models behave in scenarios where data is incomplete or contains irrelevant information. The experiments yielded several important findings regarding the performance of the models under different conditions:

If IDs are the same: When the values between the HTTP requests and database logs were perfectly aligned, the models—both Decision Tree and Random Forest—performed with exceptional accuracy. This confirms that when the data identifiers are consistent across different sources, the models are highly effective in matching the logs.

If IDs are different: When dealing with misaligned data from different databases, the models' performance suffered. The ID key (key shared between both databases) remained the most important feature in both models. In this scenario, Random Forest demonstrated a clear advantage due to its ability to handle variability better than a single Decision Tree, which can fall more easily to overfitting. The ensemble learning approach of Random Forest, combined with its feature randomization, allowed it to perform better in cases where the IDs did not perfectly match, as it could rely on multiple trees to generate a consensus decision, mitigating over-reliance on specific ID values. [3]

DB x DB with textual identifiers: In scenarios where textual identifiers were used across databases, it became crucial to apply effective string-matching techniques. The Levenshtein distance proved to be particularly advantageous over the Hamming distance for handling variations in textual data, such as person names. Levenshtein distance, which accounts for insertions, deletions, and substitutions, allowed for a more flexible and accurate comparison of strings with minor differences, like abbreviations, missing letters, or additional characters. [17] This flexibility was critical when matching similar but not identical identifiers across databases, ensuring greater accuracy in identifying correct matches. The Hamming distance, limited to fixed-length string comparisons, failed to accommodate such variations, making Levenshtein the superior choice for scenarios with inconsistent or partially aligned textual identifiers.

7.1 Future Work

Building on the findings of this thesis, there are several important directions for future research: **Real-world data:** Future experiments should incorporate real-world data to provide more comprehensive and practical insights. While synthetic data allowed for controlled experiments, testing the models on real-world datasets would reveal their behavior under true operational conditions, where unpredictability, noise, and incomplete data are more prevalent. **Exploration of other classifiers:** Additional machine learning models, such as Gradient Boosting [12] and like Multilayer Perceptrons (MLP) [25], should be explored. Gradient Boosting offers high sensitivity to outliers, making it suitable for detecting rare, unauthorized data changes, but it is also prone to overfitting if not carefully managed. On the other hand, MLP models, which handle continuous data well, might struggle with discrete data like dropdown options, as used in this study. However, both Gradient Boosting and MLP remain viable alternatives for future research, potentially offering improvements in accuracy and generalization over the models tested in this thesis. This thesis laid the groundwork for understanding how machine learning models can be applied to the task of log matching across different systems. While the proposed methods performed well in controlled simulations, further work is needed to refine these models for real-world applications. Future efforts should focus on expanding the dataset to include real-world information and exploring additional classifiers that may offer improvements. This will enhance the models' effectiveness in real-world scenarios, ensuring better log matching and bringing us one step closer to full compliance with data privacy regulations.

References

[1] Mahmoud Alharbi and Ibrahim Yahya. 2021. Challenges and Issues of GDPR Compliance for Developing Countries. *Journal of Information Security* 12, 1 (2021), 1–12.

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*. IEEE, 141–159. https://doi.org/10.1109/SP40001.2021.00019

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.

[5] California Office of the Attorney General. 2020. California Consumer Privacy Act (CCPA). State of California - Department of Justice - Office of the Attorney General. https://oag.ca.gov/privacy/ccpa Accessed: 2024-01-10.

[6] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*. 73–78.

[7] European Commission. 2021. Artificial Intelligence Act. European Commission proposal for a regulatory framework. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

[8] Alfredo Cuzzocrea, Ladjel Bellatreche, and Il-Yeol Song. 2013. Data warehousing and OLAP over big data: current challenges and future research directions. In *Proceedings of the 16th International Workshop on Data Warehousing and OLAP*.

[9] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings (Lecture Notes in Computer Science, Vol. 1857)*, Josef Kittler and Fabio Roli (Eds.). Springer, 1–15. https://doi.org/10.1007/3-540-45014-9_1

[10] Menachem Domb, Sujata Joshi, and Pandey Roshan. 2019. Risk Mitigation Model for Data Loss: A Case Study Approach. *Journal of Advanced Research in Dynamical and Control Systems* 11 (09 2019). https://doi.org/10.5373/JARDCS/V11/20192590

[11] European Commission. 2018. General Data Protection Regulation. https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en. Accessed: 2024-01-10.

[12] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[13] Carrie Gates and Peter Matthews. 2014. Data Is the New Currency. In *Proceedings of the 2014 workshop on New Security Paradigms Workshop, Victoria, BC, Canada, September 15-18, 2014*. ACM, 105–116. https://doi.org/10.1145/2683467.2683477

[14] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi, and Meiko Jensen. 2018. Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. In *IEEE International Conference on Big Data (IEEE BigData 2018), Seattle, WA, USA, December 10-13, 2018*. IEEE, 5027–5033. https://doi.org/10.1109/BIGDATA.2018.8622621

[15] Kalle Hjerppe, Jukka Ruohonen, and Ville Leppänen. 2019. The General Data Protection Regulation: Requirements, Architectures, and Constraints. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*. 265–275. https://doi.org/10.1109/RE.2019.00036

[16] Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1. IEEE, 278–282.

[17] Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 8 (1966), 707–710.

[18] Qianmu Liu, Zhuo Xiao, and Chi Wang. 2017. Privacy-preserving data sharing in cloud computing. *Security and Communication Networks* 2017 (2017), 1–9.

[19] David R. Matos, Miguel L. Pardal, and Miguel Correia. 2017. Rectify: black-box intrusion recovery in PaaS clouds. In *Proceedings of the 18th ACM/IFIP/USENIX Middleware Conference, Las Vegas, NV, USA, December 11 - 15, 2017*, K. R. Jayaram, Anshul Gandhi, Bettina Kemme, and Peter R. Pietzuch (Eds.). ACM, 209–221. https://doi.org/10.1145/3135974.3135978

[20] David R. Matos, Miguel L. Pardal, and Miguel Correia. 2023. Sanare: Pluggable Intrusion Recovery for Web Applications. *IEEE Trans. Dependable Secur. Comput.* 20, 1 (2023), 590–605. https://doi.org/10.1109/TDSC.2021.3139472

[21] Gonzalo Navarro. 2001. A guided tour to approximate string matching. *Comput. Surveys* 33, 1 (2001), 31–88.

[22] Patrick Oyakhare. 2022. Complying with GDPR: The Difficulties American Big Techs Face. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4275535.

[23] Erhard Rahm and Hong-Hai Do. 2000. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin* 23, 4 (2000), 3–13.

[24] Stephan Rosenbaum, Christoph Müller-Bloch, and Helmut-Tilo Wagner. 2020. GDPR compliance implementation issues and solutions: The perspective of German SMEs. *Journal of Business Research* 122 (2020), 903–914.

[25] Frank Rosenblatt. 1961. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms*. Spartan Books.

[26] Li Yujian and Liu Bo. 2007. A normalized Levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 6 (2007), 1091–1095.