PRIVACY-PRESERVING RECORD LINKAGE METHODS FOR HOMELESSNESS DATA

Felipe Castaño Gonzalez[1]

[1]Department of Electrical and Software Engineering, University of Calgary, Calgary, Canada

ABSTRACT

This study evaluated threshold-based classification and XGBoost for privacy-preserving record linkage (PPRL) of homelessness data. The threshold method, combined with Bloom filters and the Dice coefficient, achieved precision up to 85% and accuracy of 82.6% but required significant computational resources, making full-scale implementation challenging. XGBoost, enhanced with feature engineering and ADASYN for class balancing, achieved precision and recall above 0.80, with 72.8% overall accuracy, while being more efficient for larger datasets. Threshold methods are suitable for resource-limited settings, while XGBoost provides robust performance where computational capacity allows. These approaches demonstrate the potential for unifying fragmented homelessness data, improving policy-making and resource allocation while maintaining privacy.

***Index Terms*—** Privacy-Preserving, Record Linkage, Homelessness, Bloom Filter, Machine Learning

1. INTRODUCTION

Homelessness is a global issue that affects millions of people and causes significant personal and social costs [1], [2]. This issue is particularly significant in developed nation, where homelessness is recognized as a serious public health problem affecting a considerable proportion of the population. Urban homelessness is a persistent issue in Western countries, despite efforts to alleviate it [3], [4]. In response to this growing concern, several high-income nations have moved to develop standardized definitions of homelessness to assess progress and determine eligibility for different social assistance. However, these efforts have not been without substantial problems, since disagreements and discrepancies in the implementation of these criteria continue to create barriers in the fight against homelessness [5]. Homelessness is now recognized as a complex social and public health problem that goes beyond lack of housing [6].

Homelessness has a serious and far-reaching impact on health, with those who are homeless enduring major health inequalities across a wide range of physical and mental health disorders. Homelessness has continuously been linked to a disproportionate burden of illnesses, including chronic diseases such as asthma, chronic obstructive pulmonary disease (COPD), epilepsy, and different heart conditions. High rates of drug misuse and mental illness worsen the already perilous condition for homeless persons [1]. Many studies have revealed that the frequency of mental illness or intellectual and cognitive disability is significantly higher among homeless persons compared to the general population in Western nations [7]. The death rate among homeless populations is frighteningly elevated with homeless males experiencing a mortality rate roughly eight times higher and women 12 times higher than their housed counterparts in the general population [1]. Furthermore, a substantial percentage of homeless individuals suffer from cognitive impairment, contributing to chronic homelessness and frequent use of acute care services, which leads to an elevated risk of early death [2].

Understanding the root causes of homelessness requires a thorough evaluation of both human and systemic issues. According to current research, homelessness is caused by a complex interaction of these elements, with the presence or absence of a safety net playing an important role in defining an individual's susceptibility to homelessness. Individual factors such as poverty, adverse childhood experiences, mental health issues, substance misuse, a history of violence, and involvement with the criminal justice system all contribute significantly to an increased risk of becoming homeless [5]. However, systemic forces play an important role in sustaining homelessness. Homelessness persists in society due to a lack of affordable housing alternatives, restricted employment prospects for low-skilled individuals, and inadequate income support systems [5]. Income inequality is an important structural element that correlates with increased rates of homelessness [5]. Addressing these structural concerns is critical to the worldwide endeavor to minimize homelessness and lessen its terrible impact on people and society as whole.

Homelessness in Canada is a complicated and diverse problem that affects a significant percentage of the population. The Canadian Observatory on Homelessness defines homelessness as "the situation of an individual, family, or community without stable, secure, permanent, and adequate housing, or the immediate prospect, means, and ability to acquire it." [8]. This concept emphasizes the wide scope of homelessness and the crucial need for comprehensive solutions. The situation is particularly severe

among Indigenous peoples, who are disproportionately affected. Homelessness among Indigenous populations is eight times more prevalent than among other groups, and this inequality has worsened over time, with Indigenous peoples now being about 11 times more likely to be homeless than non-Indigenous individuals [9].

In Canada, recent data provides insight into the scale of homelessness across the country [10]. In 2016, it was estimated that, on average 235,000 people were experiencing homelessness and over 35,000 homeless people on any given night. Additionally, another 50,000 people each night were believed to be experiencing hidden homelessness, living in a temporary situations not typically captured in traditional homelessness statistics [11]. To address this issue, the homeless shelter capacity in Canada for 2022 was 18,467 [12] with an average occupancy rate of 88.6%, reflecting a high demand for shelter services [13].

Focusing on Calgary, the 2022 Point-in-Time Count reported that 2,782 individuals were experiencing some form of homelessness in the city. Of these individuals, 71% were sheltered, while 29% were unsheltered, indicating ongoing challenges in providing adequate housing solutions [14]. Additionally, the Calgary Drop-In Centre reported that between 2022 and 2023, 6,839 unique individuals accessed their shelter services, highlighting the continued need for resources and support for the homeless population in Calgary [15].

Data plays a critical role in shaping effective public policies, particularly in addressing complex issues such as homelessness. Reliable and comprehensive information allows policymakers to identify trends, allocate resources, and design targeted interventions that maximize the impact of limited resources. For instance, research highlights the importance of understanding homelessness trends and population dynamics to guide public education campaigns, resource distribution, and policy evaluations [16]. However, the limitations of existing data cannot be overlooked. Challenges such as incomplete coverage, infrequent collection, and potential inaccuracies undermine the ability of governments and organizations to make informed, timely decisions [3]. These gaps can limit the ability of policies to adapt effectively to changing social dynamics, highlighting the critical need for robust and systematic data collection strategies. In the absence of reliable data, initiatives aimed at addressing homelessness may be poorly targeted or ineffective, potentially worsening the challenges faced by this vulnerable population.

To address the critical issue of unreliable data on individuals experiencing homelessness, this research focuses on evaluating a privacy-preserving record linkage (PPRL) method to enable accurate account of individuals accessing shelters in Calgary. Current data collection practices often rely on approximate counts, which lack precision and hinder effective policymaking. This study proposes an assesses of using Bloom Filters to ensure privacy while facilitating data integration across shelters, which currently operate with unlinked records. By unifying shelter reports, it becomes possible to produce a feasible and accurate count of individuals experiencing homelessness annually. The research compares the performance of threshold-based techniques and machine learning models, including decision trees, random forests, and logistic regression, for record linkage. This provides a framework to improve data quality, which impacts policy-making and resource allocation decisions.

2. RELATED WORK

Previous work has explored the integration of machine learning models and threshold-based methods for privacy-preserving record linkage. Lee and Jun [17] introduced a hybrid approach combining distance-based record linkage and micro-aggregation, enabling the linkage of de-identified heterogeneous datasets while balancing privacy and utility through adjustable thresholds. Similarly, advancements such as Bloom filter encodings have enhanced machine learning applications by preserving privacy in domains like domain generation algorithm detection [18]. These foundational studies highlight the potential of threshold-based and machine learning methods to address the challenges of balancing privacy and utility in data linkage.

Building on these concepts, current research demonstrates progress in Privacy-Preserving Record Linkage (PPRL) using methods that leverage Bloom filters and hashing techniques to secure sensitive information across large datasets. Schnell et al. [19] and Randall et al. [20] both explore Bloom filter applications for encrypted identifiers, showcasing high accuracy and efficiency in linking while maintaining privacy. However, these approaches underscore the importance of optimizing threshold settings and enhancing security measures against re-identification, especially in large-scale implementations. Addressing these concerns, Durham et al. [21] propose Composite Bloom Filters (CBFs) to counter frequency-based attacks, while Schnell et al. [22] further strengthen these methods with random hashing and balanced Bloom filters to mitigate cryptographic vulnerabilities, paving the way for more robust PPRL systems..

Alternative approaches, such as Karapiperis and Verykios [23] LSH-based framework, demonstrate efficient matching with homomorphic encryption, yet reveal a gap in applying high-accuracy PPRL at multi-source scales. Vatsalan and Christen [24] address multi-party scalability by combining Bloom filters and secure summation, proving feasible for real-world multi-database integration but pointing to persistent challenges in computational efficiency and privacy

fidelity. Kuzu et al. [25] further emphasize the vulnerabilities in Bloom Filter Encodings (BFEs) against re-identification attacks and propose frequency-hiding as a defense, though limitations in real-world viability remain evident.

Vatsalan et al.'s [26] comprehensive review highlights the diversity and shortcomings in PPRL methods, noting the lack of scalable, secure protocols capable of handling cross-institutional, multi-party linkages without compromising data privacy or linkage accuracy. Koneru et al. [27] and Lee and Jun [17] explore hybrid phonetic encoding and PPDM for specific linkage contexts like open government data, showing improvements in accuracy but limited scalability and cross-domain adaptability.

Our proposal focuses on unifying shelter records to enable accurate annual counts of homeless individuals, leveraging the privacy-preserving capabilities of Bloom filters and comparing threshold-based methods with machine learning models. This approach aims to improve data quality, thereby supporting better policymaking and resource allocation to address homelessness effectively.

3. METHODS

This project aimed to evaluate privacy-preserving record linkage of homelessness data using threshold-based classification techniques and machine learning model. The record linkage process was enhanced by incorporating thresholds for similarity scoring alongside a XG boost machine learning model. The effectiveness of these approaches was assessed using metrics such as accuracy, precision, recall, and confusion matrix.

3.1. Dataset

The dataset used in this study is a synthetic dataset created based on record structure from the Calgary Drop-In Centre. The process began with organizing a dataset containing personal identifiers, including names and dates of birth. The data consisting of 15,991 records with identifiers and demographic information was cleaned by filtering relevant columns, converting date values into numerical formats. Records were categorized into unique and duplicate groups based on the relationship between Id and IdLinked fields. Unique records are those where the Id matches the IdLinked, while duplicates are identified when these values differ. The dataset was further refined by organizing unique records alongside their corresponding duplicates, creating a structured framework for subsequent analysis. This organization facilitates the examination of duplicate counts and ensures clarity in data relationships.

3.2. Privacy preserving implementation

As a privacy-preserving measure, the original dataset containing names and dates of birth was transformed using a 32-bit Bloom filter. This process converted sensitive attributes into hashed representations, ensuring that no personal data was directly accessible during analysis. Specifically, concatenated attributes such as first name, last name, and date of birth components (day, month, and year) were hashed into a single 32-bit representation.

3.3. Threshold-based classification

To facilitate record linkage, similarity scores were computed using the Dice coefficient, which quantifies the similarity between binary representations of hashed attributes. A series of thresholds (0.1, 0.2, 0.3, and 0.4) were applied to classify record pairs as potential matches or non-matches. Record pairs with Dice coefficients below the threshold were considered matches, while those above were labeled non-matches. Precision, recall, and accuracy metrics were computed for each threshold, allowing the selection of optimal linkage parameters.

3.4. Machine learning classification

The machine learning pipeline implemented in this study leverages the XGBoost classifier, a state-of-the-art gradient boosting algorithm optimized for classification tasks on imbalanced datasets. To address the class imbalance, the pipeline integrated ADASYN (Adaptive Synthetic Sampling), which generated synthetic samples for the minority class, focusing on instances in regions of high classification difficulty. The feature engineering process included encoding input data using binary Bloom filter representations and computing derived metrics such as bit density to capture relevant data patterns effectively. Performance evaluation utilized metrics such as precision-recall curves and confusion matrix, providing a comprehensive assessment of the model's predictive capabilities.

3.5. Evaluation metrics

The evaluation metrics included accuracy, precision, recall, and confusion matrix, providing a comprehensive assessment of model performance. Threshold-specific metrics, such as true positives, true negatives, false positives, and false negatives, were calculated to analyze the impact of different Dice coefficient thresholds.

4. RESULTS AND DISCUSSION

3.1. Dataset

The dataset was analyzed to distinguish between unique and duplicated entries. The input dataset, containing 7 columns (Id, IdLinked, FirstName, LastName, DobDay, DobMonth, DobYear), was filtered and cleaned to remove rows with invalid or missing birth year data, ensuring all values in DobYear were numeric. The dataset was divided into unique records (where Id = IdLinked, totaling 4,750 records) and duplicated records (where Id ≠ IdLinked, totaling 11,241 records). After applying filtering and cleaning steps, the

dataset was transformed into a concise format with two columns: ConcatenatedDetails and IsDuplicated. The ConcatenatedDetails column combines the FirstName, LastName, DobDay, DobMonth, and DobYear fields into a single concatenated string, providing a unique representation of each record's personal information. The IsDuplicated column served as a binary indicator, marking records as either unique (0) or duplicated (1) as shown in Table 1.

ConcatenatedDetails	IsDuplicated
Aabar3031973	1
Aaliyahottens14121918	0
Aaliyahottens14121918	1
Aarobbstovm1931973	1

Table 1. Dataset concatenated with a unique or duplicate record.

3.2. Privacy preserving implementation

The dataset was further processed to include a hashed representation using a 32-bit Bloom filter. Each record's concatenated details, derived from the fields FirstName, LastName, DobDay, DobMonth, and DobYear, were hashed into a 32-bit numeric representation (BloomFilter32Bit). This transformation produced a dataset with three columns: PrincipalDetails (containing the concatenated details), IsDuplicated (indicating whether a record is unique or duplicated), and BloomFilter32Bit (the 32-bit numeric hash) (Table 2). The processed datasets were divided into principal, unique, and duplicated subsets, saved as CSV files for further analysis. This approach enables a compact and privacy-preserving representation of personal information, facilitating efficient comparisons in large-scale data processing tasks.

ConcatenatedDetails	IsDuplicated	BloomFilte32Bit
Aabar3031973	1	1643689933
Aaliyahottens14121918	0	3567187259
Aaliyahottens14121918	1	3567187259
Aarobbstovm1931973	1	48879422

Table 2. Dataset concatenated with single or duplicate record and bloom filter the 32 bits.

3.3. Threshold-based classification

The Dice coefficient was employed to measure the similarity between records based on their 32-bit Bloom filter representations, which encode concatenated personal information fields. The coefficient, ranging from 0 to 1, quantifies the overlap in bit-level representation, with higher values indicating greater similarity. The distribution of Dice coefficients across all pairwise comparisons was analyzed, revealing a skewed pattern, with most coefficients concentrated in lower ranges (Figure 1). Thresholds of 0.4, 0.3, 0.2, and 0.1 were visually assessed in a histogram, demonstrating their impact on the classification of pairs into matching or non-matching categories.

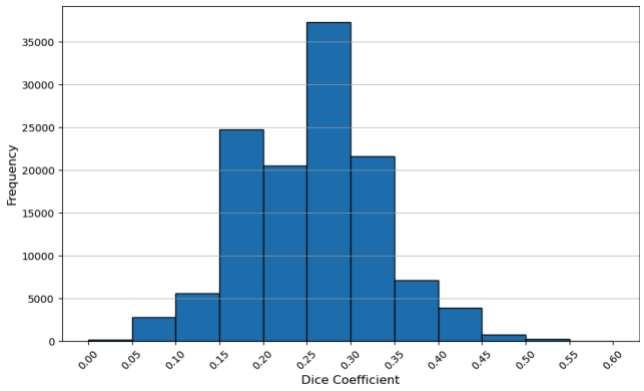

Fig. 2. Histogram of the dice coefficient of all records

Processing the entire dataset was computationally extensive due to the scale of pairwise comparisons. The complete dataset would have required approximately 72 days of computation time using TALC's resources (which allows only 24 hours), making it infeasible for this analysis (Figure 2). As a result, a sample of 3,000 records was selected to reduce computational demands, resulting in 4,498,500 comparisons. This approach balanced the need for robust analysis with practical time and resource constraints, ensuring meaningful insights could still be obtained.

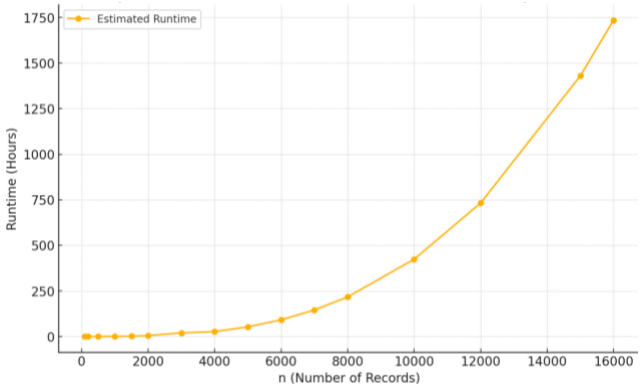

Fig. 2. Runtime for threshold in hours for pairwise comparisons

The results demonstrated that using the Dice coefficient as metric to determine the thresholds, classified duplicate records effectively. Precision ranged from 0.84 to 0.85, indicating a strong ability to correctly identify true duplicates, while recall varied between 0.02 and 0.96, reflecting sensitivity to duplicate detection. Accuracy reached 0.82 across thresholds, showcasing consistent overall performance (Table 3). These findings underscore the utility of Bloom filters combined with the Dice coefficient for privacy-preserving record linkage, even under significant computational limitations.

Threshold	Precision	Recall	Accuracy
0.1	0.857878	0.025854	0.169891
0.2	0.856889	0.289499	0.356183
0.3	0.854286	0.753513	0.681836
0.4	0.849778	0.965653	0.826030

Table 3. Results for threshold, precision, recall and accuracy.

3.4. Machine learning classification

The classification model demonstrated robust performance in handling unbalanced data and accurately distinguishing between classes. The input variables (X) were BloomFilter32Bit, BF_countsOf1s, BF_Density, Hamming_Similarity and bloomFilter32Bit_Binary. The target variable (Y) represented the binary classification labels (unique or duplicated), with class imbalance evident in the initial dataset. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied, generating synthetic samples for the minority class and balancing the dataset. This preprocessing step ensured equitable representation of both classes during model training, reducing bias and enhancing model reliability.

The XGBoost classification model, trained on the oversampled dataset, effectively utilized the engineered features to achieve high classification accuracy. The confusion matrix heatmap highlighted the model's ability to distinguish between true positives and true negatives with minimal misclassification (Figure 3). Precision reached 0.80, indicating a low rate of false positives, while recall reached 0.83, reflecting strong detection of the minority class. The overall accuracy exceeded 0.72, validating the importance of feature selection, balanced inputs, and advanced modeling techniques in achieving reliable classification outcomes (Table 3).

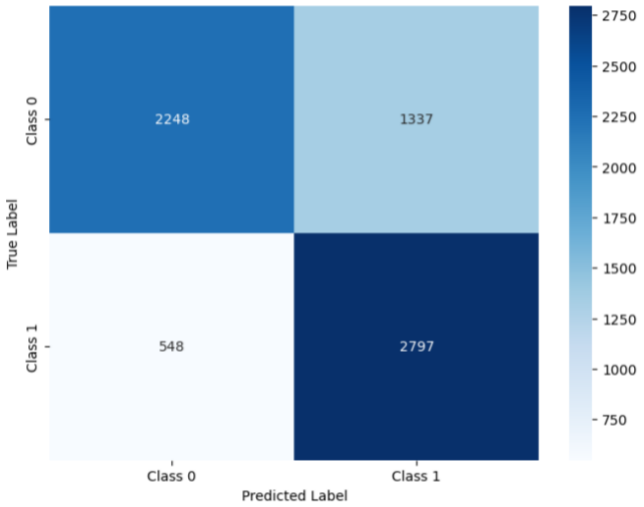

Fig. 3. Confusion matrix for the XG Boost Classifier.

Classification	Precision	Recall	Accuracy
0	0.804006	0.627057	0.727994
1	0.676584	0.836173	

Table 4. Results for XG Boost, precision, recall and accuracy.

5. CONCLUSIONS

This study compared the performance of threshold-based classification and the XGBoost machine learning model for PPRL of homelessness data, highlighting their strengths and computational demands. The threshold approach, combined with Bloom filters and the Dice coefficient, provided reliable results, achieving precision up to 85% and accuracy of 82.6% at a threshold of 0.4. However, its computational cost was significant, requiring approximately 72 days to process the full dataset on available resources, necessitating the use of a smaller sample for analysis. This limitation emphasizes the need for scalable methods in large-scale data linkage tasks.

In contrast, the XGBoost model leveraged advanced feature engineering and class-balancing techniques, such as ADASYN, to achieve precision and recall above 80%, with an overall accuracy of 72.8%. While computationally less intensive for large datasets, its reliance on feature extraction and training highlighted the trade-off between efficiency and complexity. These findings suggest that threshold-based methods are well-suited for resource-constrained environments, whereas machine learning models like XGBoost offer a powerful alternative for settings with sufficient computational capacity. Both approaches underscore the potential for integrating fragmented homelessness data to improve policy decisions while safeguarding privacy.

6. REFERENCES

[1] P. Omerov, Å. G. Craftman, E. Mattsson, and A. Klarare, "Homeless persons' experiences of health- and social care: A systematic integrative review," *Health Soc. Care Community*, vol. 28, no. 1, pp. 1–11, 2020, doi: 10.1111/hsc.12857.

[2] V. Stergiopoulos *et al.*, "Effect of Scattered-Site Housing Using Rent Supplements and Intensive Case Management on Housing Stability Among Homeless Adults With Mental Illness: A Randomized Trial," *JAMA*, vol. 313, no. 9, p. 905, Mar. 2015, doi: 10.1001/jama.2015.1163.

[3] S. Strobel, "Characterizing people experiencing homelessness and trends in homelessness using population-level emergency department visit data in Ontario, Canada," *Health Rep.*, vol. 32, no. 82, 2021.

[4] T. Aubry *et al.*, "A Multiple-City RCT of Housing First With Assertive Community Treatment for Homeless Canadians With Serious Mental Illness," *Psychiatr. Serv.*, vol. 67, no. 3, pp. 275–281, Mar. 2016, doi: 10.1176/appi.ps.201400587.

[5] S. Fazel, J. R. Geddes, and M. Kushel, "The health of homeless people in high-income countries: descriptive epidemiology, health consequences, and clinical and policy recommendations," *The Lancet*, vol. 384, no. 9953, pp. 1529–1540, Oct. 2014, doi: 10.1016/S0140-6736(14)61132-6.

[6] M. A. Mabhala, A. Yohannes, and M. Griffith, "Social conditions of becoming homelessness: qualitative analysis of life stories of homeless peoples," *Int. J. Equity Health*, vol. 16, no. 1, p. 150, Dec. 2017, doi: 10.1186/s12939-017-0646-3.

[7] A. Nishio *et al.*, "Causes of homelessness prevalence: R elationship between homelessness and disability," *Psychiatry Clin. Neurosci.*, vol. 71, no. 3, pp. 180–188, Mar. 2017, doi: 10.1111/pcn.12469.

[8] M.-A. Dionne, C. Laporte, J. Loeppky, and A. Miller, "A review of Canadian homelessness data, 2023," no. 75, 2023.

[9] S. K. Agrawal and C. Zoe, "Housing and Homelessness in Indigenous Communities of Canada's North," *Hous. Policy Debate*, vol. 34, no. 1, pp. 39–69, Jan. 2024, doi: 10.1080/10511482.2021.1881986.

[10] H. Canada, "Substance-related poisonings and homelessness in Canada: a descriptive analysis of hospitalization data." Accessed: Aug. 27, 2024. [Online]. Available: https://www.canada.ca/en/health-canada/services/opioids/hospitalizations-substance-related-poisonings-homelessness.html

[11] S. Gaetz, E. Dej, and T. Richter, *Homelessness Canada in the State of 2016*. Toronto, ON, CA: Canadian Observatory on Homelessness Press, 2016.

[12] S. C. Government of Canada, "Homeless Shelter Capacity in Canada from 2016 to 2022, Infrastructure Canada." Accessed: Aug. 27, 2024. [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410035301

[13] *Homelessness data snapshot: the National Shelter Study 2022 update*, [Cat. No.: T94-60/2024E-PDF]. Ottawa, Ontario: Infrastructure Canada, 2024.

[14] "Point in Time Count," Calgary Homeless Foundation. Accessed: Nov. 05, 2024. [Online]. Available: https://www.calgaryhomeless.com/discover-learn/research-data/data/point-in-time-count/

[15] "2022-2023 Report To The Community | Calgary Drop-In Centre." Accessed: Aug. 28, 2024. [Online]. Available: https://calgarydropin.ca/2022-2023-report-to-the-community/

[16] C. J. Frankish, S. W. Hwang, and D. Quantz, "Homelessness and Health in Canada: Research Lessons and Priorities," *Can. J. Public Health.*, vol. 96, no. S2, pp. S23–S29, Mar. 2005, doi: 10.1007/BF03403700.

[17] J.-S. Lee and S.-P. Jun, "Privacy-preserving data mining for open government data from heterogeneous sources," *Gov. Inf. Q.*, vol. 38, no. 1, p. 101544, Jan. 2021, doi: 10.1016/j.giq.2020.101544.

[18] L. Nitz and A. Mandal, "Bloom Encodings in DGA Detection: Improving Machine Learning Privacy by Building on Privacy-Preserving Record Linkage," *JUCS - J. Univers. Comput. Sci.*, vol. 30, no. 9, pp. 1224–1243, Sep. 2024, doi: 10.3897/jucs.134762.

[19] R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," *BMC Med. Inform. Decis. Mak.*, vol. 9, no. 1, p. 41, Dec. 2009, doi: 10.1186/1472-6947-9-41.

[20] S. M. Randall, A. M. Ferrante, J. H. Boyd, J. K. Bauer, and J. B. Semmens, "Privacy-preserving record linkage on large real world datasets," *J. Biomed. Inform.*, vol. 50, pp. 205–212, Aug. 2014, doi: 10.1016/j.jbi.2013.12.003.

[21] E. A. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom Filters for Secure Record Linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 12, pp. 2956–2968, Dec. 2014, doi: 10.1109/TKDE.2013.91.

[22] R. Schnell and C. Borgs, "Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, Barcelona, Spain: IEEE, Dec. 2016, pp. 218–224. doi: 10.1109/ICDMW.2016.0038.

[23] D. Karapiperis and V. S. Verykios, "An LSH-Based Blocking Approach with a Homomorphic Matching Technique for Privacy-Preserving Record Linkage," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 909–921, Apr. 2015, doi: 10.1109/TKDE.2014.2349916.

[24] D. Vatsalan and P. Christen, “Scalable Privacy-Preserving Record Linkage for Multiple Databases,” in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai China: ACM, Nov. 2014, pp. 1795–1798. doi: 10.1145/2661829.2661875.

[25] M. Kuzu, M. Kantarcioglu, E. A. Durham, C. Toth, and B. Malin, “A practical approach to achieve private medical record linkage in light of public resources,” *J. Am. Med. Inform. Assoc.*, vol. 20, no. 2, pp. 285–292, Mar. 2013, doi: 10.1136/amiajnl-2012-000917.

[26] D. Vatsalan, P. Christen, and V. S. Verykios, “A taxonomy of privacy-preserving record linkage techniques,” *Inf. Syst.*, vol. 38, no. 6, pp. 946–969, Sep. 2013, doi: 10.1016/j.is.2012.11.005.

[27] K. Koneru and C. Varol, “Privacy Preserving Record Linkage Using MetaSoundex Algorithm,” in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, Mexico: IEEE, Dec. 2017, pp. 443–447. doi: 10.1109/ICMLA.2017.0-121.