Vol 5, Issue 11, November 2025 www.ijesti.com E-ISSN: 2582-9734
International Journal of Engineering, Science, Technology and Innovation (IJESTI)

Secure and Accurate Privacy-Preserving Record Linkage Using Hardened Bloom Filters

Biplab Biswas
Research Scholar, Department of Computer Science, RKDF University, Ranchi, India.

ABSTRACT

An integral part of many data operations is the linking of personal records from several databases. In order to link individuals without invading their privacy, Privacy-Preserving Record-Linkage (PPRL) methods were created to handle identification problems. It is difficult to design effective PPRL systems that provide a high degree of privacy protection while maintaining excellent linkage quality. Presently, one of the most prominent approaches is PPRL based on Bloom filter encoding (BF), which provides excellent efficiency and linkage quality. The most detrimental of these assaults, however, are pattern mining and graph matching, but both techniques are susceptible to a number of threats. Although there have been several suggestions to fortify BF-based PPRL schemes against these assaults, none of them have conducted a thorough security analysis or managed to maintain the excellent efficiency and linkage quality. In order to increase record linkage accuracy while ensuring privacy, this paper offers an ensemble PPRL architecture that uses phonetic encoding and toughened Bloom filters. This method uses the Dice similarity coefficient in conjunction with the SHA3-384 and SHA3-512 hash algorithms to provide safe approximation matching. The suggested ensemble method achieves a recall of 65.22 percent, an F-measure of 75.0%, and an accuracy of 88.24 percent when tested on the NCVR dataset, surpassing Basic Bloom, Balanced Bloom, and PPRL methods based on Cellular Automata (CA). The outcomes validate the greater privacy protection and more accurate secure record linking provided by the suggested structure.

Keywords: *Data Integration, Matching, Privacy, Linkage, Bloom Filter, Cellular Automata.*

I. INTRODUCTION

With the proliferation of digital technology, several industries, including healthcare, banking, education, government, insurance, telecoms, and e-commerce, have amassed and stored vast amounts of personally identifiable information. Millions of records pertaining to consumers, patients, taxpayers, pupils, or workers may be found in these databases. The need to integrate data from many databases is rising as data-driven decision-making becomes more important to organisations. Combining information from diverse sources can provide a more complete picture of individuals, eliminate duplicate entries, improve service quality, support accurate statistical analysis, and enable sophisticated data mining and machine learning applications. Typos, missing values, acronyms, changes in personal information, inconsistent formatting, and other issues may lead the same individual to have distinct representations in separate databases, making it very difficult to integrate data from several organisations or departments. Despite these differences, record linkage—also called entity resolution or data matching—finds records that belong to the same real-world object. It is essential in many fields, including healthcare analytics, national security, customer relationship management, epidemiological research, census analysis, and fraud detection.

When databases hold sensitive personal information like names, addresses, DOBs, phone numbers, medical records, or financial details—even if record linking delivers substantial benefits—it also brings substantial privacy problems. There are laws and ethical standards in place to preserve people's privacy, and it could be a violation of both to share this information directly between organisations. The transmission, processing, and storage of personally identifiable information is severely limited by data

protection legislation in several nations. Therefore, businesses are hesitant or unable to share private information with other parties, even when doing so might lead to significant societal, health, or financial gains. Organisations may now find matching records without disclosing the original sensitive information stored in their databases thanks to Privacy-Preserving Record Linkage (PPRL), a new field of study that was inspired by these privacy concerns.

Striking a compromise between the two competing goals of protecting the privacy of individuals' personal information and retaining just enough data to reliably locate relevant records is the goal of privacy-preserving record linking. As an alternative to directly comparing the actual attribute values, PPRL approaches encrypt or convert sensitive information into a safe representation before matching. The danger of sensitive information being revealed is minimised since only these encoded values are sent between the participating organisations. High linkage accuracy, robust security against unauthorised disclosure, and computing efficiency fit for large-scale databases are the pillars upon which a PPRL system rests.

A number of methods have been put up to accomplish record linking while protecting individuals' privacy. Methods based on Secure Multi-Party Computation (SMC) and methods based on perturbation are the two main groups into which these approaches fall. Using sophisticated cryptographic protocols, Secure Multi-Party Computation allows several parties to work together to calculate record similarities while keeping their underlying data private. Since sensitive data is encrypted during computation, these approaches theoretically provide excellent privacy guarantees. Many communication rounds, heavy processing overhead, and intensive cryptographic processes are characteristic of SMC methods. These processing needs may become quite costly when database sizes keep increasing into the millions of entries, which limits the practical scalability of systems based on SMC.

Researchers have devised perturbation-based techniques that encode crucial attribute values prior to matching in order to circumvent the computational constraints of safe cryptographic methods. The computational efficiency and scalability of these approaches are often substantially superior, yet they still provide a respectable amount of privacy protection. Some examples of perturbation approaches include Bloom filter encoding, randomisation, phonetic alterations, and hashing. There is always a trade-off between privacy protection and linkage accuracy in these techniques, notwithstanding their computational attractiveness. While minimal disturbance leaves encoded data vulnerable to privacy attacks, excessive perturbation decreases the similarity information needed for accurate matching.

Bloom filter (BF) encoding is the gold standard for privacy-preserving record linking among perturbation approaches. It has numerous practical uses, including healthcare and government data integration initiatives, and has become the most generally accepted methodology. Using a binary bit vector to represent the text, a Bloom filter effectively encodes text while keeping approximate similarity associations. It is a compact probabilistic data structure. First, q-grams, which are overlapping character substrings, are used to encode sensitive attribute information like names or addresses. Next, a collection of independent hash functions maps each q-gram to multiple locations in the Bloom filter. A binary representation is produced by setting the respective bit locations to one. This hides the actual attribute values but keeps enough structural information for approximate similarity comparisons. Next, we use similarity measures such as the Dice coefficient, Jaccard coefficient, or Hamming distance to compare the overlap of the Bloom filter bit patterns of two encoded records to determine their degree of similarity.

A number of significant benefits have contributed to Bloom filter encoding's widespread use. Because it uses approximate matching, it can withstand the common occurrence of misspellings, typos, abbreviations, and missing characters in actual datasets. When used on massive datasets with millions of

records, Bloom filters scale well, use little memory, and provide fast results. For these reasons, Bloom filter encoding has recently gained popularity in the healthcare record linking industry, where the secure integration of sensitive patient data across many institutions is of the utmost importance.

II. REVIEW OF LITERATURE

Ferrante, Anna et al., (2024). The risks and responsibilities linked to the management of personally identifiable information (PII) are increasingly evident to companies. These challenges impact both internal data management techniques and external data connectivity links. Personal data can significantly affect certain connection scenarios. Contemporary methods for privacy-preserving record linkage, such as PPRL-using-Bloom, provide very accurate data connections without the need to reveal or disclose personally identifiable information. In risk-prone scenarios, these methodologies facilitate data integration. We present and analyse several case studies from Australia that demonstrate the use of the PPRL-using-Bloom method to enhance data interconnectivity across various organisations. We outline the fundamental characteristics of the case, any issues and their solutions, as well as any concerns about quality or performance. We also evaluate possible challenges and pathways for advancement. Initiatives that connect state-level datasets with Commonwealth datasets, primary care information with state-level secondary healthcare data, and healthcare data with non-health datasets such as police and criminal justice records exemplify Australian applications employing privacy-preserving record linkage (PPRL-using-Bloom). In highly risk-sensitive environments, instruments such as PPRL-using-Bloom are essential for enabling data linkage. Linkage methodologies and technologies must possess the adaptability to respond to the continuously evolving demands of a world characterised by escalating risks and expectations.

Christen, Peter et al., (2024). Data linking is the act of bringing together records from several databases that pertain to the same thing (often a person). It is not possible to share the highly sensitive datasets needed for applications like health research or government services to other organisations. Users are able to juxtapose encrypted documents while protecting personal information via privacy-preserving record linkage (PPRL), which circumvents this problem. Now that heuristics are at the heart of most PPRL methods, the privacy protections they provide are severely limited and they are open to certain forms of cryptanalysis. Additionally, because to their many user-adjustable parameters, current PPRL approaches are vulnerable to less-than-ideal connection quality and insufficient privacy protections. We provide a new PPRL approach that uses reference q-gram sets generated at random to generate bit-arrays that represent sensitive values. Our approach strikes a compromise between privacy protection, connection quality, and scalability with a single parameter that users may modify. Every other user choice is either data-driven or has very strict limits, with the exception of this one. Our theoretical analysis of the approach and experimental results on several datasets led to these conclusions. Our approach can scale to connect large datasets with strong privacy controls and higher linkage quality, according to the findings. According to the results, our new PPRL approach outperforms previous PPRL methods like Bloom filter encoding in terms of privacy protection, scalability, and linkage quality. One major benefit of our technique is that it only requires one parameter from the user.

Nóbrega, Thiago et al., (2021). Integrating private data from different databases held by different parties is what Privacy-Preserving Record Linkage (PPRL) is all about. New norms and regulations, such as the General Data Protection Regulation, have increased the need for PPRL approaches in many real-world application areas, including healthcare, credit analysis, public policy evaluation, and national security. One example of an unrealistic adversary model utilised by most PPRL systems is the Honest but Curious (HBC) model. This model assumes that all PPRL participants would follow a pre-established procedure for data integration and not try to undermine the secrecy of the data while it is in transit. The HBC model

has limited applicability in real-world contexts due to its reliance on third parties. To overcome the limitations of the majority of adversary models employed by PPRL approaches, we propose a protocol that accounts for covert adversaries, also known as adversaries that might openly cheat by going against the protocol definition. The honest participants, however, should be able to detect this sort of activity using this method. In order to demonstrate the feasibility of this protocol, we employ Blockchain technology and propose an improvement to the Bloom Filter, which is currently the most widely used anonymisation method for PPRL. We demonstrated our contributions' success (in terms of linkage quality) and our ability to detect adversarial harmful conduct during PPRL execution through an investigation that utilised several real-world data sources.

Stammler, Sebastian et al., (2020). Because it enables several data sets to be linked at the record level even when there isn't a unique identifier between them, Record Linkage is extremely helpful when undertaking data analysis in the real world. Consider the following scenario: several healthcare facilities' databases containing patients' medical records that need to be linked using personally identifiable information (such as name, date of birth, or ZIP code). Unfortunately, privacy regulations forbid the transmission of such personally identifiable information (PII) between institutions, therefore it is not possible to engage a trustworthy third party to merge records. It is our recommendation that privacy-preserving record linkage (PPRL) methods be used to link pertinent records without exposing any private information. We provide a foundation for fault-tolerant PPRL by integrating secure multi-party computing with the software used to store medical records.

P. Christen, et al., (2019) The capability to recognise records that pertain to the same item across many databases is crucial in data analytics. Privacy-preserving record linkage (PPRL) refers to methodologies that connect records in a manner that, other from the pairs identified as matches, no confidential information on the entities in the associated databases is disclosed. A widely utilised method in Privacy-Preserving Record Linkage (PPRL) involves encoding confidential information into Bloom filters, facilitating approximate matching using character q-grams. PPRL with Bloom filter encoding has demonstrated both precision and scalability for extensive datasets. Nonetheless, research indicates that Bloom filters employed for Privacy-Preserving Record Linkage (PPRL) are susceptible to assaults that might uncover sensitive information. Prior attack techniques were sluggish and necessitated an understanding of encoding settings; we provide an innovative assault that capitalises on the encoding of data within Bloom filters. Our assault does not need awareness of the encoding mechanism or its parameter configurations employed. It can accurately re-identify q-grams that were not hashable to specific Bloom filter bit locations; utilising these re-identified q-grams, it can subsequently re-identify attribute values with considerable precision. Our approach surpasses previous approaches in speed and can effectively re-identify attribute values from extensive datasets within minutes.

Schnell, Rainer & Borgs, Christian. (2016). A prominent technique for accelerating database queries in chemo-informatics is the XOR-folding of fingerprint bit vectors. To expedite linking for Privacy-preserving Record Linkage (PPRL), folding methods have been employed using Multibit Trees.To safeguard PPRL's Bloom Filters from bit pattern attacks, we propose employing this technique in this study. Utilising authentic data from a German telephone directory, we assess the methodology based on the standards of linkage quality and decoding efficiency. Our research indicates that employing single XOR folding does not diminish connection quality. The folding, meanwhile, completely obstructs bit pattern attacks. An further advantage of folded Bloom Filters is their brevity. Consequently, they utilise fifty percent less memory and facilitate data blocking and connection at a significantly accelerated rate. Consequently, we believe the approach outlined in this study is a promising candidate for forthcoming comprehensive assessments of the cryptographic attributes of PPRL-enabled Bloom Filters.

Niedermeyer, Frank et al., (2014). A common choice for privacy-preserving record linking applications is bloom filter encoded IDs since they allow errors in encrypted identifiers. However, there has been a dearth of research on the topic of Bloom filters' safety. In this work, bloom filters built of bigrams are formalised successfully. In the past, it was widely thought that an attacker could access the whole global dataset utilised to create a sample. Conversely, suppose an adversary is completely oblivious to this dataset's global scope. Instead, we'll act as though our adversary had access to a publicly available list of the most prevalent characteristics. The assault relies on simple statistical analysis and subtle filtering of encrypted bigrams. Previous research only used this method on a small subset of the most popular names; this paper details the process of decrypting a whole database. To prove our point, we breach an encrypted surname database. At the end of our study, we detail Bloom filters that have been strengthened to withstand attacks like these.

R. Schnell, et al., (2009) There has been a recent uptick in the practice of merging disparate databases that include incomplete or extra data about a single individual. Probabilistic record linkage is employed to identify matching record pairings in the absence of unique identifying numbers for these persons. Because of privacy considerations, IDs must be encrypted in many applications. We have created a novel technique that allows for mistakes in identifiers while still protecting privacy through record linking with encrypted identifiers. Bloom filters on identifier q-grams form the basis of the protocol. Results from linkage tests on both real and virtual databases are better than those from phonetic encodings and on par with those from non-encrypted IDs. We put up a method for secure record linking using encrypted identifiers that accounts for identifier mistakes. Many applications could benefit from the protocol's privacy-preserving record linking features due to its easy enhancement and cheap computing burden.

III. PROPOSED FRAMEWORK

This research introduces an ensemble method for safe record matching and linking, which improves toughened Bloom filters' linkage accuracy even more. For approximate record matching, it uses two-factor encoding, which uses phonetic codes and toughened Bloom filter encoding.

Here, the settings for the Bloom filter-based PPRL are decided upon by both sides. The q-grams value, number of hash functions, Bloom filter length, similarity measurements, and other characteristics make up this set. Prior to the PPRL encoding procedure, the databases are queried to determine the common attribute values. The process begins with converting person names into phonetic codes using a phonetic approach (such as Soundex) and then transforming those codes into qgrams. The produced q-grams are subjected to hash algorithms like SHA3-384 and SHA-512 in order to get the hash code. To get the hash value, the modulus of the Bloom filter length is applied to the hash. The hash value is used to set the bit locations throughout the Bloom filters of the relevant attribute values to 1. Afterwards, similarity measurements such as the Dice coefficient are used to compare the resulting Bloom filters. In order to determine which records are matched and which are not, we compare the similarity values to the user-defined threshold.

When matching records, the PPRL approach must be used to prevent the disclosure of personally identifiable information. For PPRL encoded record comparisons, the Bloom filter-based encryption methods are necessary. This study applies cellular automata and basic balanced Bloom filters to PPRL. We compare the ensemble method to current Bloom filter encryption methods for PPRL and propose it as a way to increase accuracy.

The Bloom based PPRL takes the following criteria into account:

Bloom filter length l=30

q-grams=2

Hash functions=2 (SHA 3-512, SHA 3-384)

Padding=Yes

Dice Coefficient threshold value=0.85

The voter registration dataset (NCVR), which includes three identifiers: surname, given name, and middle name, is utilised in secure record matching and linking. The preliminary trials for PPRL encompass 180 entries across the two databases.

The ensemble, cellular automata, fundamental and equitable Bloom filter secure record linking methodologies are analysed concerning f-measure, recall, and precision. The f-measure (F), precision (P), and recall (R) are derived from false positives (FP), false negatives (FN), and true positives (TP) as follows:

$$P = TP/(TP + FP) \quad (1)$$

$$R = TP/(TP + FN) \quad (2)$$

$$F = 2 * (P * R)/(P + R) \quad (3)$$

IV. RESULTS AND DISCUSSION

Table 1: Analysis of Secure Record Linkage Techniques

Techniques	TP	FP	FN	Precision (%)	Recall (%)	F-Measure (%)
Basic Bloom PPRL	10	2	13	83.33	43.48	57.14
Balanced PPRL	12	18	11	40.00	52.17	45.27
CA PPRL	11	3	12	78.57	47.83	59.46
Ensemble PPRL	15	2	8	88.24	65.22	75.00

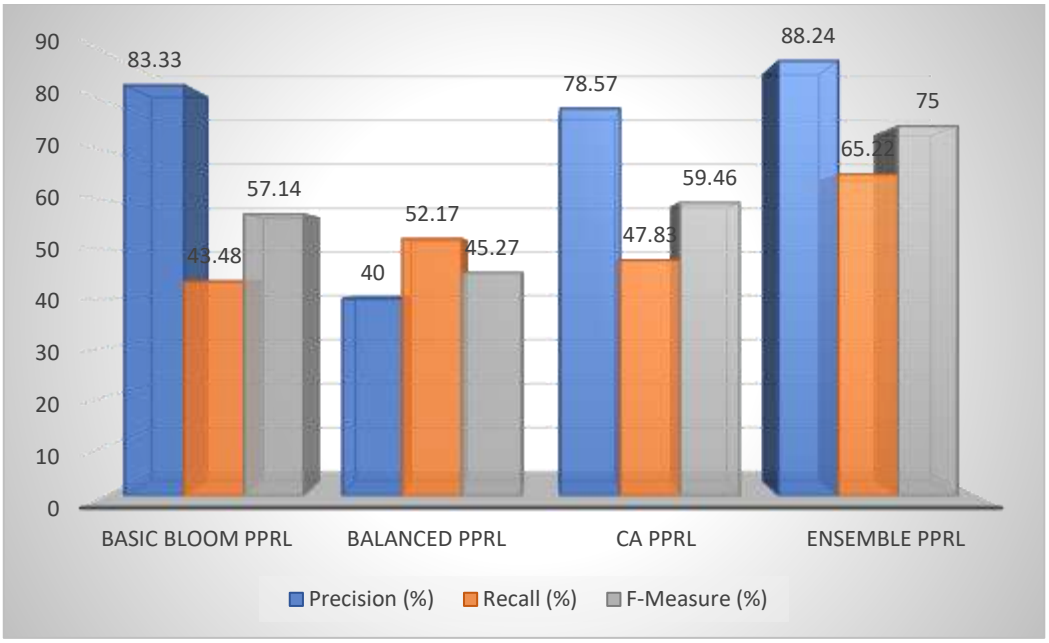

Figure 1: Comparison of Precision, Recall, and F-Measure for Different PPRL Techniques

Table 1 demonstrates that the accuracy of fundamental cellular automata (CA) methodologies surpasses that of balanced and ensemble PPRL at a Dice coefficient (DC) threshold of 0.85. No erroneous positives were detected for basic and CA-based methods, resulting in enhanced accuracy. The occurrence of false positives is more probable in balanced-based PPRL. The ensemble methodology yields aligned records with an increased count of true positives and superior recall compared to current CA, balanced, and fundamental Bloom PPRL methods.

Figure 1 illustrates that the fundamental and CA-based PPRL exhibits superior precision compared to balanced and ensemble methodologies. This accuracy is attained because there are no erroneous positive results for CA and fundamental methods for PPRL. The ensemble PPRL demonstrates the maximum recall, signifying a superior proportion of matched records compared to current balanced, basic, and CA-based PPRL methods.

Table 2: Percentage of Exact Matches for Different PPRL Techniques

PPRL Technique	Percentage of Exact Matches (%)
Basic Bloom PPRL	40
Balanced PPRL	52
CA PPRL	40
Ensemble PPRL	57

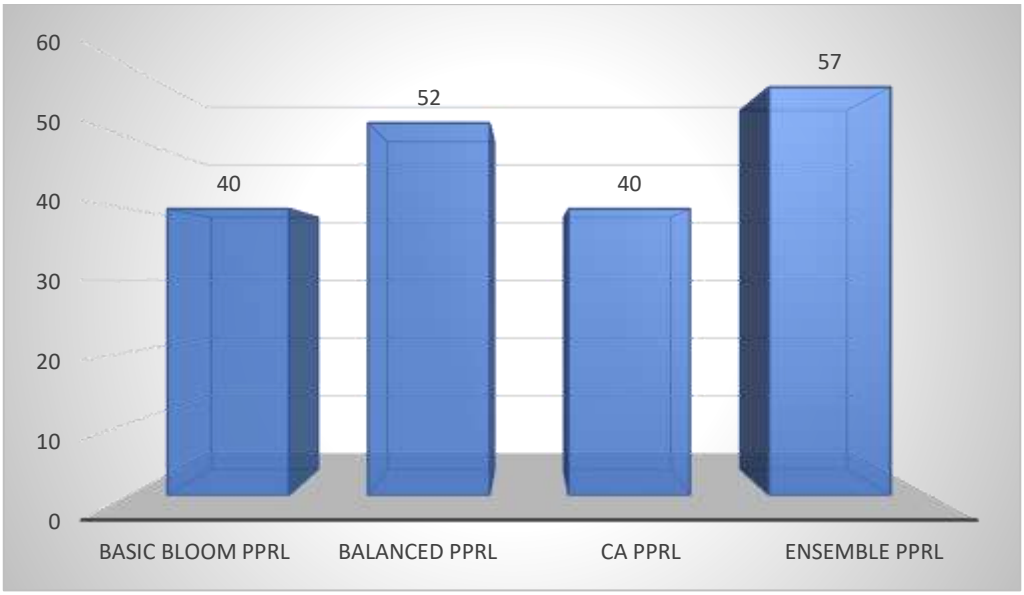

Figure 2: Percentage of Exact Matches for Different PPRL Techniques

The ensemble approach outperforms CA, balanced, and simple Bloom-based PPRL in terms of the number of true positives, as shown in Table 2.

The precision of PPRL methods is affected by the Dice coefficient value, which ranges from 0.85 to 1, as seen in table 3.

Table 3: Variation of F-Measure with Dice Coefficient for Different PPRL Techniques

Dice Coefficient	Basic PPRL	Balanced PPRL	CA PPRL	Ensemble PPRL
0.85	56.00	24.00	54.50	66.50
0.90	56.00	60.00	54.50	68.00
0.95	56.50	57.00	56.00	72.00
0.99	56.00	56.00	56.00	71.50

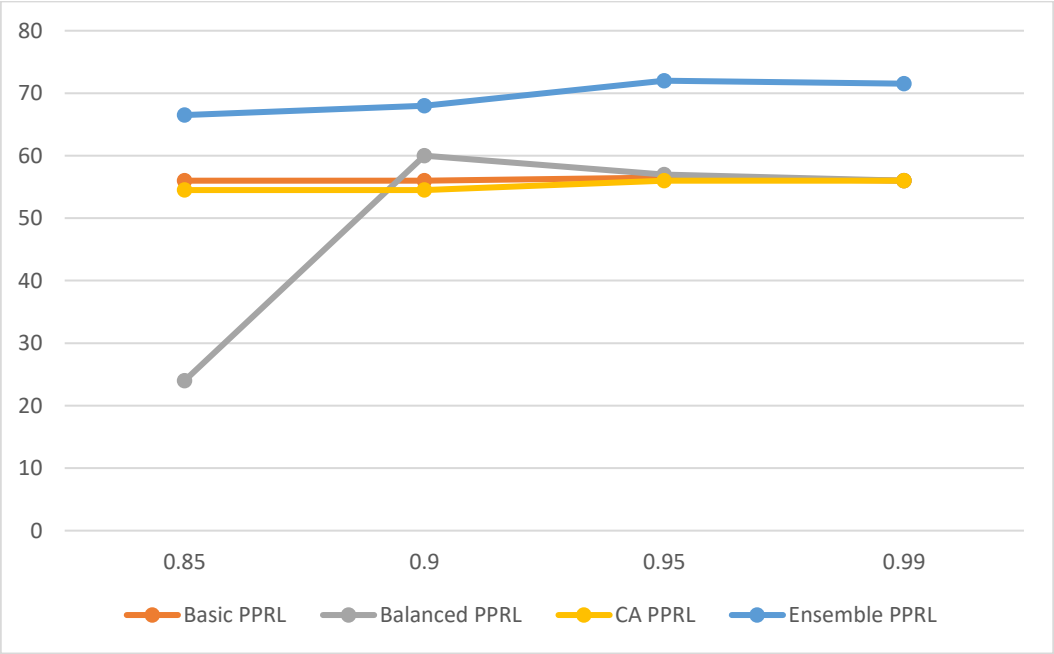

Figure 3: Variation of F-Measure with Dice Coefficient for Different PPRL Techniques

Table 3 shows that at both the 0.95 and 1 thresholds, the ensemble-based PPRL improves the f-measure. Consequently, secure record linking methods based on CA, basic, and balanced all fall short when compared to the ensemble approach.

V. CONCLUSION

This analysis proposed a robust and accurate Privacy-Preserving Record Linkage (PPRL) architecture leveraging an ensemble technique with enhanced Bloom filters. The advocated solution fuses phonetic encoding and Bloom filter-based encryption to provide safe approximate record matching without disclosing critical personal details. Using SHA3-384 and SHA3-512 hash algorithms paired with the Dice similarity coefficient, the framework effectively enhances both privacy security and linkage accuracy. The introduced ensemble PPRL architecture presents a proficient, scalable, and privacy-protecting method for safe record linking. It combines data security with matching fidelity, making it well-suited for applications that deal with sensitive information, such as healthcare, government, census, and financial systems.

REFERENCES

1. K. Tyagi and S. Willis, “Accuracy of privacy preserving record linkage for real world data in the United States: A systematic review,” JAMIA Open, vol. 69, no. 9, pp. 146–152, 2025.
2. A. Ferrante, A. Brown, S. Randall, and J. Boyd, “Linkage for risk-sensitive environments: Australian examples of privacy-preserving record linkage using Bloom filters,” International Journal of Population Data Science, vol. 9, no. 5, pp. 2–15, 2024.
3. P. Christen, S. Ziyad, A. Vidanage, C. Nanayakkara, and R. Schnell, “A single parameter method for secure privacy preserving record linkage,” International Journal of Population Data Science, vol. 9, no. 5, pp. 1–10, 2024.
4. Y. Liang, J. Ma, Y. Miao, D. Kuang, X. Meng, and R. H. Deng, “Privacy-Preserving Bloom Filter-Based Keyword Search Over Large Encrypted Cloud Data,” IEEE Transactions on Computers, vol. 72, no. 5, pp. 3086–3098, 2023.

5. S. Galán, P. Reviriego, S. Walzer, A. Sanchez-Macian, S. Liu, and F. Lombardi, "On the Privacy of Counting Bloom Filters Under a Black-Box Attacker," IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 3, pp. 4434–4440, 2023.
6. S. Stammler, T. Kussel, P. Schoppmann, F. Stampe, G. Tremper, S. Katzenbeisser, K. Hamacher, and M. Lablans, "Mainzelliste SecureEpiLinker (MainSEL): Privacy-Preserving Record Linkage Using Secure Multi-Party Computation," Bioinformatics, vol. 38, no. 3, pp. 2–15, 2022.
7. P. Christen, R. Schnell, T. Ranbaduge, and A. Vidanage, "A critique and attack on 'Blockchain-based privacy-preserving record linkage'," Information Systems, vol. 108, pp. 101930, 2022.
8. T. Nóbrega, C. Santos Pires, and D. Nascimento, "Blockchain-based Privacy-Preserving Record Linkage Enhancing Data Privacy in an Untrusted Environment," Information Systems, vol. 102, pp. 101826, 2021.
9. T. Ranbaduge and R. Schnell, "Securing Bloom Filters for Privacy-Preserving Record Linkage," in Proc. 29th ACM International Conference on Information and Knowledge Management (CIKM), pp. 2185–2188, 2020.
10. P. Christen, T. Ranbaduge, D. Vatsalan, and R. Schnell, "Precise and Fast Cryptanalysis for Bloom Filter Based Privacy-Preserving Record Linkage," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 11, pp. 2164–2177, 2019.
11. K. Irvine, M. Smith, R. Vos, A. Brown, A. Ferrante, J. Boyd, and S. Thackway, "Real world performance of privacy preserving record linkage," International Journal of Population Data Science, vol. 3, no. 4, pp. 64–77, 2018.
12. N. Bhogal and S. Jain, "A Review on Big Data Security and Handling," International Research Based Journal, vol. 6, no. 1, pp. 1–10, 2017.
13. P. Christen, R. Schnell, D. Vatsalan, and T. Ranbaduge, "Efficient Cryptanalysis of Bloom Filters for Privacy-Preserving Record Linkage," in Lecture Notes in Computer Science, vol. 5, no. 8, pp. 628–640, 2017.
14. R. Schnell and C. Borgs, "XOR-Folding for Bloom Filter-based Encryptions for Privacy-Preserving Record Linkage," SSRN Electronic Journal, pp. 3–10, 2016.
15. K. Schmidlin, K. Clough-Gorr, and A. Spoerri, "Privacy Preserving Probabilistic Record Linkage (P3RL): A Novel Method for Linking Existing Health-Related Data and Maintaining Participant Confidentiality," BMC Medical Research Methodology, vol. 15, no. 4, p. 46, 2015.
16. E. Durham, M. Kantarcioglu, Y. Xue, C. Toth, M. Kuzu, and B. Malin, "Composite Bloom Filters for Secure Record Linkage," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 12, pp. 2956–2968, 2014.
17. F. Niedermeyer, S. Steinmetzer, et al., "Cryptanalysis of Basic Bloom Filters Used for Privacy Preserving Record Linkage," Journal of Privacy and Confidentiality, vol. 6, no. 2, pp. 59–79, 2014.
18. P. Christen, D. Vatsalan, and V. S. Verykios, "A Taxonomy of Privacy-Preserving Record Linkage Techniques," Information Systems, vol. 38, no. 6, pp. 946–969, 2013.
19. M. Kuzu, M. Kantarcioglu, E. Durham, and B. Malin, "A Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage," in Proc. 11th Privacy Enhancing Technologies Symposium (PETS), pp. 226–245, 2011.
20. R. Schnell, T. Bachteler, and J. Reiher, "Privacy-preserving record linkage using Bloom filters," BMC Medical Informatics and Decision Making, vol. 9, no. 1, p. 41, 2009.