A Highly Improved Distortion Minimizing Technique for Watermarking Relational Databases

Kavya.S.P [1], Prabu.K [2], Madesh Kumar.A [3], Muthu Logesh.M [4], Raghupathi Natha Krishnan [5]
[1] Assistant Professor (Sr.G), [2, 3, 4, 5] Students,
[1, 2, 3, 4, 5] Department of Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, Tamilnadu, India
[a)] kavyaspcbe@gmail.com , [b)] prabuksp06@gmail.com, [c)] madesh067@gmail.com, [d)] lokeshsangeeth2001@gmail.com, [b) e)] raghumurugan02@gmail.com

Abstract: These days, ownership protection is a significant problem. Intelligent mining techniques must be utilized on the data removed from the relational databases in order to uncover fascinating patterns (often concealed in the data) which considerably help those making decisions to create sensible, precise and appropriate judgments. Data sharing between involved parties and information retrieval specialists (also known as "pots") is growing as a result. This study focuses primarily on decoding accuracy; only after that is watermarking efficiency estimated. Here, the MD5(Message-Digest 5) algorithm is employed for both encoding and decoding, enabling decoding to achieve the highest level of accuracy. The outcome of the experiments will demonstrate its effectiveness in achieving the highest level of decoding accuracy.
Keywords: Right Protection, Database Watermarking, Data Usability Constraints, Data Quality, Ownership Protection, Reed– Solomon, MD5-Message Digest 5 algorithm.

I. INTRODUCTION

Traditional business data processing has seen significant success with the usage of database technologies. The possibility of utilising this technology in fresh operational disciplines is growing. Database mining techniques are one comparable application area that is going to be very important in the near future. A growing number of businesses are building extremely large databases of company data, including consumer information, transaction histories, sales records, medical records, etc.; these databases are assessed in GBs and even TBs; these databases work as an implicit gold mine of important business information. Watermark embedding and watermark verification are the two stages of the database steganographic process. During the watermarking process, the steganography W is incorporated onto the source database with a key K (often a private that is only known to the owner). The copyrighted database is then made available to the general audience. A verification procedure is used to determine who is in charge of a questionable database. The suspect database is used as input in this case. The recovered itemset is then equated with the original watermark data and the embedded watermark (if present) is retrieved using the secret key K (the same one used during the implantation stage).Numerous applications, such as ownership assertion, fingerprinting, fraud detection, and tamper detection, may benefit from the usage of digital watermarks for relational databases [11].

Bob2 requests that the data owner, Alice3, specify strong usability criteria so that Alice will receive reliable data. Alice desired to have a greater bandwidth on the deceptions made while the embedding of a preprocess for maximal watermark robustness, which is only feasible if she places soft usability limits [1], [2]. Any modeling method for a watermark that aims to reduced bandwidth enables an invader (Mallory4) to minimally tamper with or remove the watermark going above the allowed frequency band. Last but not least, Bob & Alice have different requirements: Bob needs "minimal almost no errors in the pre - processed data, while Alice aims to generate copyrighted object hold great control. As soon as Alice establishes the usability criterion so that the embedded watermark is both reliable and causes little distortion to the underlying data, the concession bandwidth is attained. For a data owner, it is not only difficult but also constrained to analyze the semantics of each operation and utilize that information to set usability limits. Recall that a watermark's robustness is determined by how accurately it can be decoded, which in turn depends on the amount of manipulation bandwidth available [12].

It is suggested that a statistical method be implemented in the existing system to partition a database into as many nonintersecting, distinct groups of item sets as is feasible. The itemset splitting method relies on using unique marker itemset, leaves them harmful to steganography synchronizing problems, especially in the case of row addition and removal attempts because the placement of the flag itemset is disturbed by such kinds of attacks. Similar types of errors may be decreased if flag itemset are retained during the steganography phase and they can be used to generate the information segments once more during the fingerprint decryption phase [6]. But making use of the preserved marker item set to put the segments back together goes against the need for "blind decoding" of the steganography. A similar problem in the decoding process is caused by the threshold fashion for bit decoding, which uses thresholds that were determined at random and without regard to any optimality criteria. This method of controlling deformations introduced into the data during watermark embedding uses the idea of usability bounds on data. However, by executing massive attacks on a huge number of rows, An attacker may interfere with or damage the watermark. However, the data owner's set usability boundaries serve as the foundation for the decoding accuracy. [13]. If an attacker breaches these bounds, the decoding delicacy is also compromised. The main drawback of this approach is that each type of action that will use the data must have its own set of

accessibility criteria, which must be specified by the data owner [3]. This study has some limitations, including considerable distortion, a lack of decoding power, and working power disputes over watermarked information in the event of cumulative attacks.

The proposed work provides a new watermark decoding algorithm that is not constrained by usability issues (or available bandwidth)[12]. Because of this, Alice may more easily utilize our way to declare usability criteria for a particular database only one time for each imaginable type of anticipated activity. Additionally, it guarantees that the signature might only marginally change the actual data rather than jeopardizing the resilience of the newly applied imprint. The proposed method constructs each bit of a multi-bit steganography in every named row (with a numbered character) in order to achieve the highest level of robustness possible in the event that a bushwhacker is ever able to effectively deceive the steganography in a specific named region of the itemset[8]. Proposed work also shows the resilience of steganographic approach by testing the decyphering delicacy of our proposed system under various violent attacks using a actual-world itemset. Additionally, while the event of a cumulative attack when Mallory inserts his unique imprint into Anna's database that includes a watermark [7], it gives methods for settling power conflicts.

II. LITERATURE REVIEW

A decision tree, more properly known as a classification tree, is often used to develop a classification model that anticipates the value of a dependent feature (variant) based on the values of the independent (input) attributes. Decision tree classification (DT) is the name of this method (variables). This fixes a widespread issue with supervised categorization. Because the dependent attribute's possible classes (values) and number are mentioned. [4]. A decision tree structure has leaves that represent classifications and branches that represent groups of qualities that lead to those classifications. Using Support Vector Machines (SVM) to Classify: SVMs are used to develop a model that predicts the class label of examples throughout the test group [4]. One of the effective classifiers for 2-class classification is this classification technique. SVM is capable of handling both linear and nonlinear classification issues. Using K-Nearest Neighbor (KNN) as a classification method is one type of supervised classification method. Commonly, Devijver and Kittler's KNN (Devijver & Kittler, 1982) [6] uses the Euclidean distance notion. The k closest in the Euclidean distance learning set vectors are chosen for each row of the test itemset, and the classification is decided by a majority vote, with ties being broken at random. All of the candidates are subjected to a vote if there are any disagreements for the nth closest vector.

The process of obfuscating a small piece of digital information in a bit stream to prevent a playback device or a bystander from spotting it. Digital watermarking embeds a persistent, undetectable signal into the picture and sound tracks of the video as it passes through the server. Digital watermarking, in the opinion of R. Agarwal, is an effective strategy for preventing the piracy of digital assets including pictures, movies, music and texts as well as protecting digital data from unauthorised copying and alteration. [5]. When the watermark is implanted, the data and the watermark are one and the same [7]. There has been a lot of interest in relational database security as a result of the growth in the usage of these databases via the Internet. Data may be produced in an endless quantity of duplicates from a "source" without suffering any quality degradaxtion, making it readily distributable and feignable [7]. In order to guarantee intellectual safety, damage discovery and traitorous identification and ensure relational data integrity, digital steganography for database systems were established [8].

III. PROPOSED WORK

Pre-processing, Watermark Bits Generation, Data Partitioning, Choosing a Itemset for Watermarking, Watermark Embedding, Watermark Decoding and ultimately Majority Voting make up the proposed work. The non-numerical itemset that contains both numerical and picture data is provided for pre-processing in pre-processing. Pre-processing is a reliable step that eliminates extraneous data from the itemset and image artifacts, as illustrated in Fig. 1.

A. Data Partitioning

Since there are only two columns in this data partitioning example—column-1 has all the QIs[11] and column-2 only has the SA—bucketization can be ignored. It is possible to extrapolate the benefits of slicing over bucketization as follows: Slicing can be used to stop membership leak by splitting attributes into more than two columns. Our empirical analysis of an actual data set demonstrates that membership disclosure is not prevented by bucketization. Second, slicing can be utilized without a clear division between the delicate attribute and the QI characteristics, in contrast to bucketization [14], which needs it. It is frequently difficult to discern among QIs & SAs because there isn't a single extrinsic public repository that can be utilized to determine who traits the attacker already knows for data sets similar to the census data. Slicing may be advantageous for such data. Finally, by enabling a field to have two very different specific QI characteristics and the sensitive attribute, property linkages between the sensitive values and the QI attributes are kept.

The item set is divided into three parts based on three different criteria in this phase: (1) the user-specified bucket count M; (2) for M buckets, the harmony p of buckets assigned to the head bucket; (3) the proportion of query mass c represented by the head bucket. The distribution is primarily divided into a head and tail, with the head containing data values corresponding to a percentage of the overall query mass that ranges from 70 to 90% according to c. The tail is made up of mass in the ratio 1-c.

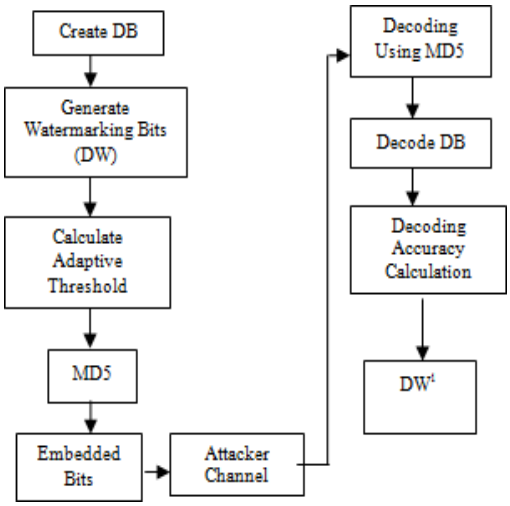

Fig.1. Watermark Encoding and Decoding Stages.

To be precise, the user must specify the value of c before the algorithm can start. In the case of the remaining M-1 buckets, QSB will divide the head bucket into a maximum of M*p buckets and the tail into a minimum of M*(1-p) buckets.

B. Selection of Itemset for Watermarking

The itemset is selected for watermarking using the two methods listed below.

- Hash Value Calculation
- Threshold Calculation.

Threshold Computation: This stage makes use of the adaptive threshold approach. Create adaptive threshold values using centroid to locate tuples in a database.

$$\text{Adaptive Threshold} = C^* \text{Mean} + \text{Standard Deviation} \quad \textbf{(1)}$$

Where, Mean is the average of the tuples, Standard Deviation is the average of the tuples, and C is the constant tuple.

Hash Value Calculation: In this measure, the selected itemset is submitted to the cryptographic hash technique MD5 and only those data items with an even hash value are manually selected. By hiding the identicalness of the watermarked tuples from a hacker, this achieves 2 objectives: (i) it increases the steganography protection; and (ii) it decreases the number of tuples that need to be stegnographed in order to minimize itemset deformation. Itemset is utilized to choose and store tuples with even hash values [11]. The tuple-based itemset I, which is a subset of itemset D, is not physically separated from the other itemset components. If D. Mallory attempts to distort the watermark information by changing the input data in a way that flips the data extraction criterion, Alice might not be able to get the stegnographed tuples in the watermark recognition facet. This is because the number of tuples that are selected also depends on the value of the data selection threshold. Mallory, on the other hand, is unaware of the confidence factor (c), hence he possibly be distorted the watermark with a probability of P by randomly attacking a particular segment of the watermarked data. This chance is decreased [10] by using the attribute selection limit T and even hashes values.

C. Generation of Watermark Bits

Non-binary cyclic error-correcting codes called Reed-Solomon (RS) codes are used for creating watermark bits. Source symbols are analysed as coefficients of a polynomial p(x) over a finite field in Reed-Solomon coding. The actual plan was to oversample p(x) at n > k unique places to produce n code symbols from k source symbols, broadcast the sampled points, and apply interpolation methods at the receiver to reassemble the original message. For a positive number m > 2 and the variables n, k, t, Reed- Solomon (R-S) codes can be written as

$$(n, k) = (2^m - 1, 2^m - 1 - 2t) \quad \textbf{(2)}$$

n - k = 2t is Number of equality indicators,
t is Error correcting capacity of the code.

An R-S code's for generating linear function is provided as:

$$g(X) = g_0 + g_0 X + g_0 X^2 + \ldots\ldots + g_{2t-1} X^{2t-1} + g_0 X^{2t} \quad \textbf{(3)}$$

In R-S codes, a subclass of Bose, Chaudhuri, and Hocquenghem(BCH) codes, the degree of the turbo code and the total number of integrity characters are connected. Since the generating polynomial has degree, the roots of the polynomial must strictly be consecutive powers of. Here is a list of the g(xroots:)'s. No particular power of two, including the root, must be used

as the starting point [13].

D. Encoding

Using the MD5 and Adaptive threshold technique to encode tuples and annex watermarking bits.

MD5: The 128-bit (16-byte) hash value produced by the widely used MD5 message-digest algorithm is mentioned in text form as a 32-digit hexadecimal integer.

Padding Bits: To make the input message 448 modulo 512 bits lengthy, padding bits are added (trotted out). Padding is still used even if the message's length is 448 mod 512 before. In order to prevent a message's bit length from conflicting with 448 mod 512, "0" bits are inserted at the conclusion of the message after a simple "1" bit to complete the padding process. least one bit. A minimum of one bit and a maximum of 512 bits make up the final addition.

Append Length: The result of step 1 is supplemented with a 64-bit indication of the message's span. The limited 64 bits will be used, if the message length is greater than 264. The length of the final message, which was paddingd with bits and b, is a precise integer of 512 bits. The length of the input message will be an exact multiple of 16 (32-bit) words.

Initialize the MD Buffer: A 4-word buffer is used to calculate the text authentication (A, B, C, and D). There are four 32-bit registers: A, B, C, and D. The default settings for these registers are as follows:

Word A: 01 23 45 67
Word B: 89 ab cd ef
Word C: fe dc ba 98
Word D: 76 54 32 10

Process Message In 16-Word Blocks: 4 functions shall be is explained. Each one requires 3- 32-bit words as input and outputs a single 32-bit word.

F (X, Y, Z) = XY or not (X) ZG (X, Y, Z) = XZ or Y not (Z)

H (X, Y, Z) = X xor Y xor Z

I (X, Y, Z) = Y xor (X or not (Z))

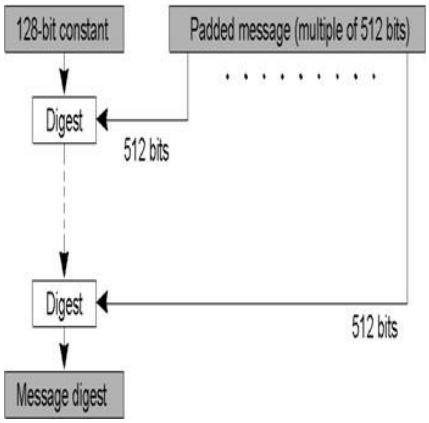

Fig.2. The structure of MD5 algorithm (Rivest,1992).

E. Watermark Decoding

The watermarked bit is not made available to the algorithm for deciphering. As seen in Fig.2, the final bit is subjected to MD5 decoding. There are some values that may have a progressive degree of assurance because they may solidify recognizable elements[9], analogous to existing words, applying a dictionary approach to look-up the hash and equate with grasped-existing values. However, deciphering an MD5 hash without grasping / understanding the pioneer value of the ciphered string is not minutely precise. Few MD5 reverse look-up databases include millions of hashes with the decoded values that correspond to each one. This approach is typically regarded as the simplest because it may be completed in few milliseconds. An alternative method examines the ciphered MD5 elements using a more brute-force method known as "rainbow tables." Although neither method is guaranteed to result in a successful decoding, the chance that it might has led the National Security Agency (NSA) to classify MD5 as technically unsafe [10].

IV. EXPERIMENTAL RESULT

The experimental results for the proposed work are computed here. 50000 tuples from a real-world itemset were hand-

selected for this subset. This work was done on a server running the Java platform with a Pentium(R) Dual-Core CPU running at 2.10GHz and 4GB of RAM. The task's deciphering accuracy is calculated and displayed in figure 3 below. The majority voting mechanism is used in the decoding resultants of the watermarked itemset to eliminate deciphering faults (if any) brought on the malicious assault (or attacks).

The results of the decoding are best shown in Fig. 3. As a result, it blocks the greatest number of assaults and flaws and decodes the largest amount of the supplied itemset.

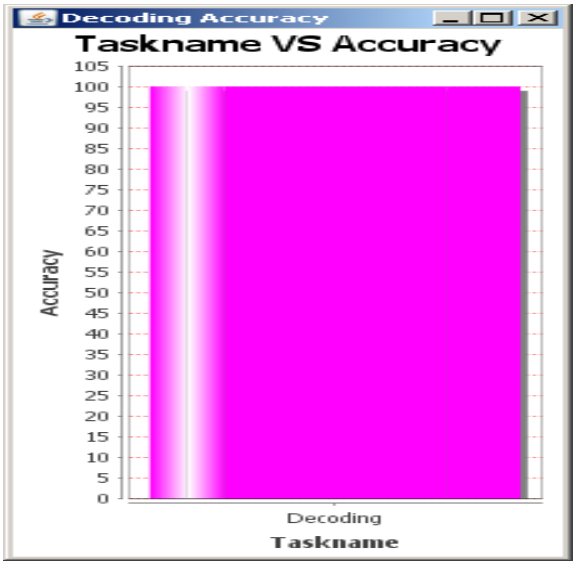

Fig.3. Decoding Result.

V. CONCLUSION AND FUTURE WORK

The Stehnographed data is subject to major attacks in many different forms. When there are several attacks, the security method in the proposed work aids in resolving ownership disputes over watermarked itemsets. Additionally, this method offers optimum delicacy for deciphering with minimal distortion. By evaluating the decoding sensitivity of the watermarking strategy under various sorts of severe attacks utilizing data from the actual world, it establishes the soundness of the technique. Additionally, this offers options for resolving ownership disputes in the event of an additive attack when Mallory adds his own signature to Alice's stegnographed database.

VI. REFERENCES

1. Kesavaraj.G, Sukumaran.S, "A study on classification techniques in data mining", Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 30 January 2014.
2. Adelaja Oluwaseun Adebayo, Mani Shanker Chaubey, "Data Mining Classification Techniques On Theanalysis Of Student's Performance", Volume 7, Issue 4, April 2019, Global Scientific Journal.
3. Bhaskaru.O, M.Sree Devi.M, "Researchon Classification Techniques in Data Mining", Volume-8, Issue-6S4, April 2019, ISSN: 2278-3075, International Journal of Innovative Technology and Exploring Engineering (IJITEE).
4. Perveen.S, Shahbaz.M, Guergachi.A, Keshavjee.K, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", International Journal of Computer Science and Information Technologies.
5. Gajendra Sharma, Umesh Hengaju, "Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes", Volume: 12 Issue: 01 Pages: 4509-4518 (2020) ISSN: 0975-0290, International Journal of Advanced Networking and Applications.
6. Quan Zou1, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, Hua Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques", Volume 9, November 2018. https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full.
7. Sindhu, "Impact of Data Mining in Business Intelligence", International Journal of Innovative Research in Technology Volume 8 Issue 5, ISSN: 2349-6002, October 2021.
8. David Bradley, "Privacy and data mining", International Journal of Business Intelligence and Data Mining, August 12, 2022.
9. Yurong Zhong, "The analysis of cases based on decision tree", IEEE explore, E-ISSN: 2327-0594, 23 March 2017.
10. Brijain R Patel, 2Mr. Kushik K Rana, "A Survey on Decision Tree Algorithm For Classification", International Journal of Engineering Development and Research, Volume 2, Issue 1, ISSN: 2321-9939, 2014.
11. Bhu Lakshmi.D, Arundathi.S, Dr.Jagadeesh, "Data Mining: A prediction for Student's Performance Using Decision Tree ID3 Method", International Journal of Scientific & Engineering Research, Volume 5, Issue 7, July 2014.
12. Rudy Setiono and Huan Liu. Fragmentation problem and Automated Feature Constructions.
13. Syed Zubair Ahmad Shah, Mohammad Amjad, Ahmad Ali Habeeb, Mohd Huzaifa Faruqui, Mudasir Shafi, "Algorithms

for Frequent Pattern Mining of Big Data", International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 12, Pages 7355-7359, 2017.
14. Kajal Singh, Anukriti Mukherjee, "Reliable Algorithms for Machine Learning Models: Implementation Research in Data Science", International Journal of Recent Technology and Engineering (IJRTE), Volume 10, Issue-6, March 2022.