International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Evaluation of Privacy-Preserving Techniques: Bouncy Castle Encryption and Machine Learning Algorithms for Secure Classification of Sensitive Data

Tungar D. V. [1] Patil D. V.[2]

Submitted:22/03/2023 **Revised:**21/05/2023 **Accepted:**11/06/2023

Abstract*:* In the present era of data-driven society, where the protection of sensitive information is of utmost importance, this research paper aims to enhance the effectiveness of data encryption and compare it with machine learning models (ML). This research examines different privacy-preserving methods using the machine learning algorithms and encryption library of Bouncy Castle for securely classifying sensitive data. To encrypt the dataset, the study utilizes the Bouncy Castle providing encryption by RSA-2048 algorithm. The method uses lookup substitution with k-anonymization, which reduces data risks, to improve privacy protection and system performance. This method successfully substitutes sensitive data with anonymized values, protecting privacy while enhancing system efficiency. To implement privacy preservation techniques, a dataset containing employee details is employed, and several classifiers including Adaboost M1, Naïve Bayes, Decision Tree J48, Random Forest and Decision Tree ID3, are utilized. The effectiveness of these algorithms is determined by evaluating their performance using significant metrics. The outcomes show that the Decision Tree J48 method surpasses others in terms of classification performance. Furthermore, the study evaluates the encryption process's efficiency and effectiveness by assessing processing time, encryption time, and decryption time. This investigation gives light on the technique's impact on data security and time overhead. The research's conclusions offer insightful information about the value of data encryption and help decision-makers make sensible choices when selecting appropriate security measures for various use cases. These observations help us comprehend the importance of data encryption and how it helps to protect the security and privacy of that data.

Keywords: *Machine learning, Data encryption, privacy, data privacy, sensitive information security, data security, RSA-2048 algorithm, encryption, predictive accuracy*

1. Introduction

The importance of data security and privacy concerns has grown in the digital age. The hazards connected with data breaches, unauthorised access, and the abuse of personal information. It have received considerable attention as a result of the exponential development in data gathering, storage, and transmission [1].

[1]*Department of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Adagaon, Nashik (Maharashtra), India, Savitribai Phule Pune University, Pune (Maharashtra), India, dipalitungar183@gmail.com*
[2]*Department of Computer Engineering, Gokhale Education Society's R.H. Sapat College of Engineering, Management Studies and Research, Nashik (Maharashtra), India, dipakvpatil17@gmail.com*

The potential effects of compromised data, such as monetary losses, identity theft, and reputational harm, have been highlighted by notable events. Protecting sensitive data including personal identifiers, financial information, and private papers is becoming more and more important to people, organisations, and governments. There is an urgent need to construct strong defences against unauthorised access, interception, and manipulation given the development of technology and the rising value of data. Effectively protecting data security and privacy is a challenging task that calls for all-encompassing solutions that include powerful encryption methods and rigorous adherence to privacy regulations and best practises [2]. It is increasingly important to have reliable security as our reliance on data increases. As data increases in value, it attracts criminal actors looking to exploit weaknesses and acquire unauthorised access as a target. It is becoming more and more important to deploy comprehensive data security solutions that

include encryption, access controls, secure storage, and good data governance practises in order to uphold trust, protect sensitive information, and adhere to privacy legislation.

Sensitive information must be protected from unauthorised access and interception, which is made possible via data encryption. It involves encrypting data using cryptographic methods and keys such that only those with the proper decryption keys may decrypt it. Data encryption is used to reduce the risk of data breaches and unauthorised disclosures for both individuals and organisations. Additionally, privacy-preserving machine learning (ML) algorithm-based solutions provide a way to maintain the confidentiality and privacy of personal data while facilitating learning. These methods strike a balance between data utility and privacy protection by allowing numerous parties to jointly train models without disclosing raw data. It is crucial to evaluate and contrast the efficiency of data encryption and machine learning algorithms in the data-driven world of today. Such analyses aid in comprehending the benefits, drawbacks, and trade-offs of various ways to protecting sensitive data.

2. Literature Survey

Considerations for choosing the right encryption technique include performance indicators, system requirements, desired security level, and complexity. Numerous factors have been taken into account in the extensive study that has been done to evaluate encryption techniques. Abood and Guirguis [3] compared encryption algorithms like AES (Advanced Encryption Standard), RSA (Rivest-Shamir-Adleman), DES (Data Encryption Standard), DSA (Digital Signature Algorithm), TDES (Triple Data Encryption Standard), EEE (Embedded Encryption Engine), CR4 (Confidentiality and Randomness version 4), and ECC (Elliptic Curve Cryptography). They found that BlowFish, AES (Advanced Encryption Standard), RC4 (Rivest Cipher 4), TDES, and Extended Data Encryption Standard (E-DES), offered great encryption speed, with better flexibility, and efficiency, AES being the most reliable.

Similarly, In [4], Riman and Abi-Char analyzed AES, E-DES, DES, and 3DES emphasizing the advantages of E-DES in terms of modest execution, input blocks and larger key sizes, which contribute to boosted security. Additionally, encryption techniques in [5] explored by Dixit et al. include hybrid and traditional methods, highlighting the benefits of AES-ECC in reducing time complexity and space requirements, while the fusion algorithm of DSA-RSA demonstrated improved output and performance. Therefore, through these studies and analyses, researchers have examined and compared various encryption algorithms to assess their strengths and weaknesses, aiding in the selection of the most suitable algorithm based on specific requirements and priorities.

The study [6] based on PPDDM known as privacy-preserving in distributed data mining emphases on conducting data mining on a private database assortment maintained by numerous parties. The tactic [7] follows to the principles of SMC referred to Secure Multiparty Computation as well as data sharing is strictly prohibited excluding the concluding data mining results. Vu et al. embraces a set of SMC protocols in [8], to verify various processes and includes scalar product with secure set union, secure size of set intersection, and secure sum. Alternatively, PPDP does not directly involve data mining referred to privacy-preserving in data publishing. Nonetheless, it focuses on publishing data anonymously to create it valuable for data mining determinations. Also, PPDP mainly emphasizes defence of data privacy, while privacy protection in PPDDM ensures for the data mining method.

The primary goal of statistical disclosure control (SDC) is to protect privacy while releasing statistical tables [9]. Attribute, identity, and inferential disclosures are only a few of the different disclosure kinds that are included in SDC. Identity exposure happens when malicious parties use the made-public information to identify a participant, and whether it constitutes a breach of confidentiality relies on the particulars of how the data was obtained [10]. When considering whether to publish data, statistical authorities primarily consider attribute disclosure. Certain SDC research also investigate non-collaborative query models in which data recipients yield to a particular query. Meanwhile, building an appropriate data mining query in a single try might be difficult, this strategy may not fully meet the information needs of data recipients.

The collaborative paradigm, where data recipients or attackers can submit a series of queries depending on previously received query responses, has thus been the subject of research. This architectural layout enables the database server to efficiently track user requests and determine whether a given query violates privacy rules by taking into account earlier queries [11][12].

3. Methodology

In the present investigation, carefully chosen datasets with varied properties, containing private information like personal indicators, monetary data, or confidential papers, are utilised. These datasets have been created to replicate real-world circumstances with allowing for an in-depth investigation of the efficacy of methods for encrypting data including artificial intelligence. The research emphasises leveraging the Bouncy Castle cryptographic library's privacy-preserving features. It offers a diverse set of cryptographic methods that protect sensitive data during storage, transport, as well as assessment. The highly recognised and secure RSA-2048 (Rivest-Shamir-Adleman) encryption technique is used to secure the data [13]. RSA-2048 is noted for its

robustness and employs a key size of 2048 bits.

Bouncy Castle delivers RSA-2048 implementations that are trustworthy and effective, with strong encryption, digital signatures, decryption, as well as key control features. The incorporation of RSA-2048 into the Bouncy Castle framework allows sophisticated methods of encryption utilising a public-private key pair to perform decryption and encryption, thereby assuring confidentiality as well as fidelity. This study seeks to evaluate the effectiveness of the RSA-2048 encryption algorithm in preserving sensitive details while preserving the privacy and security of data by implementing it in the selected datasets.

3.1 Proposed Method:

The system design begins with a registration and login process, requiring users to enter correct login credentials, as depicted in Figure 1. Before uploading a new dataset, it is important to clear any existing files. After that, the chosen file is uploaded and saved in a database created with MySQL. The submitted file initially appears as the original dataset, however, it must be processed before the RSA-2048 encryption technique offered by Bouncy Castle can be applied.

With RSA-2048 encryption, a public and private key pair is created and used to encrypt data, producing ciphertext. To improve the dataset's security and secrecy, certain data might be encrypted using RSA-2048. At the receiver's end, the system architecture requires the user to log in with the correct credentials and decrypt the dataset using the RSA-2048 public key. For correct key, Bouncy Castle decrypts the file back to its original form.

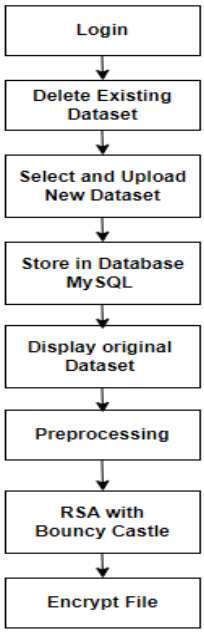

Fig 1: Architecture of Proposed System using BC-RSA-2048

- Dataset: Three unique datasets were used in this experimental study, however, just a single dataset is included in this work for demonstrative considerations. The study's dataset provides specific information about employees, such as their age, job, employer category, zipcode, level of education, salary, gender, and country. Moreover, a scrutiny of the dataset reveals different employee counts based on characteristics such as gender, age, job classification, country, education, as well as employer kind. The dataset provides the file sizes in bytes and timestamps for the uploading, encrypting, and decryption operations. These time stamps and file sizes offer crucial historical data as well as perceptions into the various facets of data management.

3.2 Masking of the data

Masking the data, employs techniques that include data anonymization, as well as lookup substitution, with encrypting is an essential component of the privacy protection method. Data anonymization entails encoding identifiers which connect users to masked data, safeguarding user confidentiality yet preserving the data integrity after the masking process. On the other hand, Lookup substitution covers a database through the use of another lookup table with alternating values corresponding to the initial private information. It facilitates the usage of realistic data despite exposing the source data. The use of encryption, in particular, is highly encouraged for protecting insecure lookup tables, ensuring that data can only be accessed using a password. The data stays illegible when encrypted, yet it can be accessed after decryption. Therefore, To improve complete data security, it is recommended to employ encryption using additional data masking methods [14].

3.2.1 Data anonymization

The process of data anonymization is a critical aimed at protecting sensitive data by altering it in such a way that the original attributes identifying entities are hidden or modified. However, it is important to note that even after anonymization, the data may still contain quasi-identifiers that, when combined with other databases, can potentially lead to re-identification of individuals. To address this concern, the concept of k-anonymity [15] has been presented. It involves grouping the quasi-identifiers in a database in such a way that no less than k-1 records share the identical quasi-identifier values, so preserving privacy [16]. This approach is considered effective for safeguarding individual privacy and is commonly employed in various privacy-preserving data publication methods. Figure 2 illustrates the utilization of k-anonymity techniques as part of the overall privacy preservation process.

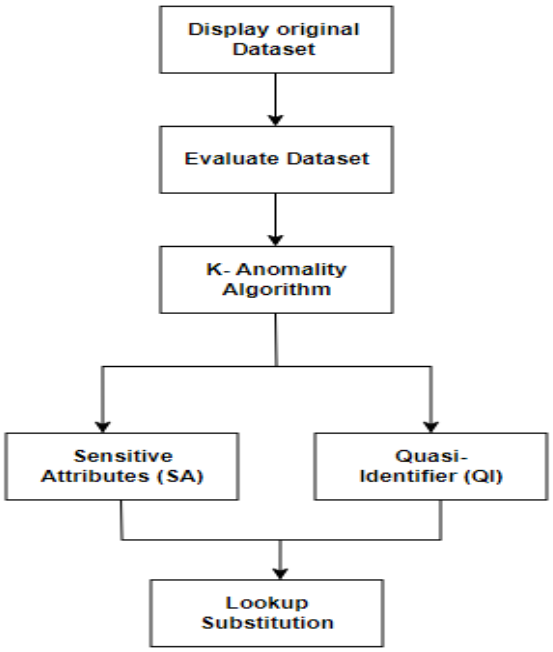

Fig 2: k-Anonymity Dataset Masking

The property of k-anonymity is satisfied by a database when minimum of k-1 records in the table share the identical quasi-identifiers (QIs). Yet, there are cases where databases may not have common quasi-identifiers among the records. To overcome this issue, techniques like generalization and suppression can be utilized to establish common quasi-identifiers among the records. Generalization involves transforming specific values of quasi-identifiers into more generalized or less precise categories, enabling the establishment of common quasi-identifiers. On the other hand, suppression involves selectively removing or obscuring sensitive attributes or quasi-identifiers that could potentially lead to re-identification. By applying these techniques of generalization and suppression, common quasi-identifiers can be established among the records, ensuring the fulfillment of the k-anonymity property.

- SA and QI (Sensitive Attribute and Quasi-identifier): In the study, the management of sensitive information in addition to quasi-identifying data attributes is accomplished through the utilization of two tables: the Sensitive Attribute (SA) table and the Quasi-identifier (QI) table. These tables play a crucial role in the encryption and privacy preservation processes. Let's delve into the explanation of these tables and their respective functions.

A. SA (Sensitive Attribute) Table:

To ensure the protection and confidentiality of sensitive attributes in the dataset, the SA (Sensitive Attribute) table is employed. This table serves as a storage mechanism for the sensitive attributes that require safeguarding. It consists of columns on behalf of various sensitive attributes for instance personal credentials, health histories, monetary statistics, and more. During the estimation, the SA table is populated with the encrypted data of these sensitive attributes, which are obtained by applying encryption algorithms like Bouncy Castle. This ensures that the sensitive information remains secure and inaccessible to unauthorized parties.

TABLE 1: EMPLOYEES' SALARY AS A SENSITIVE ATTRIBUTE

GID	Salary	Count
1	200000	219
2	440000	24
3	150000	91
4	105000	46
5	195000	17
6	335000	69
7	300000	284
8	480000	26
9	550000	22
10	325000	175
11	405000	18
12	500000	61
13	210000	39

B. QI Table (Quasi-identifier):

The QI (Quasi-identifier) table serves as a means of managing quasi-identifiers, which are attributes that, when shared with external information, have the potential to lead to the re-identification of individuals. This table consists of columns demonstrating various quasi-identifying attributes for example occupation, age, zip code, gender, and more. During the assessment process, the QI table is populated with quasi-identifiers' anonymized values, achieved through the application of k-anonymity methods for privacy conservation. Anonymization comprises suppressing or generalizing values within the quasi-identifiers that ensures minimum k-1 records in each group in the QI table using alike quasi-identifier values. This way, the privacy of individuals is maintained while allowing for meaningful analysis of the data.

TABLE 2: A) DIFFERENT ATTRIBUTES IN QI TABLE INCLUDING EDUCATION, EMPLOYER TYPE, AND GENDER

GID	Employer Type	Count
1	State-gov	148
2	Self-emp-not-inc	330
3	Private	3010
4	Self-emp-inc	137
5	Local-gov	232
6	Federal-gov	140
7	Without-pay	2

GID	Education	Count
1	Doctorate	56
2	HS-grad	1311
3	11th	146
4	Bachelors	694
5	Assoc-acdm	134
6	Some-college	839
7	7th-8th	74
8	Prof-school	68
9	Masters	216
10	10th	106
11	Assoc-voc	181
12	9th	61
13	5th-6th	45
14	12th	45
15	1st-4th	22
16	Preschool	2

GID	Gender	Count
1	Male	2674
2	Female	1326

TABLE 2: B) DIFFERENT ATTRIBUTES IN QI TABLE INCLUDING AGE, JOB AND COUNTRY

GID	Country	Count	GID	Job	Count	GID	Age	Count
1	United-States	3567	1	Prof-specialty	513	1	41	90
2	Puerto-Rico	16	2	Craft-repair	501	2	40	106
3	India	90	3	Adm-clerical	500	3	30	106
4	Trinadad&Tobago	2	4	Other-service	391	4	72	11
5	Japan	9	5	Handlers-cleaners	145	5	17	50
6	France	4	6	Protective-serv	71	6	46	74
7	Mexico	75	7	Machine-op-inspct	262	7	38	111
8	Guatemala	14	8	Sales	681	8	36	116
9	Poland	10	9	Exec-managerial	506	9	24	108
10	El-Salvador	11	10	Farming-fishing	112	10	28	123
11	Nicaragua	5	11	Tech-support	101	11	60	41
12	Philippines	30	12	Transport-moving	193	12	42	88
13	Jamaica	11	13	Priv-house-serv	22	13	48	82

To ensure the effective application of techniques for privacy preservation and encryption, the sensitive attributes and quasi-identifiers are segregated into two separate tables: the SA table and the QI table. It allows for independent application of encryption and privacy measures, effectively managing the defence of sensitive data and mitigates privacy dangers. By implementing various security levels with privacy controls on all attributes, the study can analyze the effectiveness of techniques used for privacy preservation and encryption.

3.2.2 Lookup Substitution

In order to protect dataset for sensitive data while still maintaining truthful evidence, a masking procedure is utilized. This technique contains the use of a lookup table, which provides alternative values to replace the original sensitive data. By employing this approach, realistic data can be used without exposing the original information, thereby ensuring the confidentiality of sensitive attributes.

3.2.3 Encryption of Data

By transforming messages into veiled or encrypted texts, the encryption approach serves a critical role in assuring their security [17]. Decryption refers to the process of restoring the encrypted texts to their original form. Operations involving encryption and decryption significantly rely on the use of particular keys. An encryption algorithm's primary objective is to make the decoding process challenging or nearly impossible without knowledge of the related encryption key. Figures 3 and 4 show examples of the essential ideas of encryption and decryption, respectively. Three methods are frequently used in cryptography: hashing, symmetric and asymmetric key encryption.

C. Bouncy Castle:

Various cryptographic libraries are developed and standardized, following suggested algorithmic specifications. The study focuses on BouncyCastle, which is an extensively recognized crypto library that conforms to NIST standards. It also integrates FIPS140-2 Level 1 certified streams designed for cryptographic uses. The study also investigates the susceptibility of the cryptographic library like BouncyCastle against attacks by side channels. It eliminates the need for underlying architectural knowledge or utilisation of constrained sensor data. Although the BouncyCastle library enables cypher executions, they may still be vulnerable to temporal assaults. Since cryptographic techniques frequently work on information in varied quantities, this research focuses on the RSA-2048 encryption technique in order to show the possible extraction of the private key needed for decoding. Timing measurements are taken on the entire RSA-2048 decryption procedure during the BouncyCastle cryptography call while carrying out this investigation. The findings indicate varying time behaviour that is driven by the combination of the data being processed as well as the algorithm being used.

The RSA-2048 algorithm is a secure method for encrypting and decrypting messages using asymmetric key pairs. Following these steps ensures the confidentiality and integrity of the data.

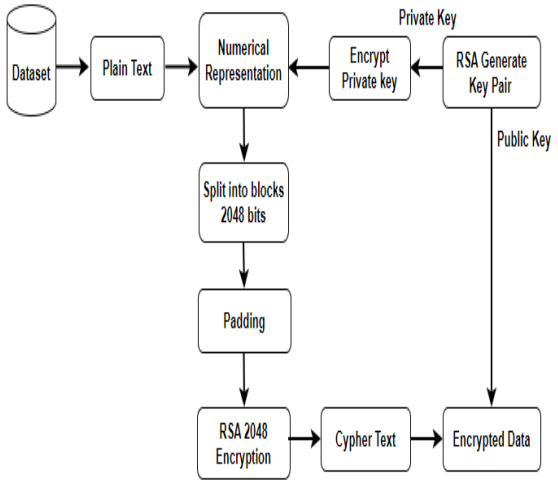

Fig 3: BC_RSA-2048 Data encryption

The encryption algorithm procedure is depicted in Figure 3. Initially, the plaintext is managed through the RSA-2048 encryption algorithm, resulting in a ciphertext and an encryption value corresponding to the input plaintext. The diagram demonstrates the step-by-step process of RSA-2048 encrypting with Bouncy Castle, starting with the creation of a Key Pair, which consists of a public key plus a private key. However, it's important to note that the private key is encrypted. The textual dataset is then transformed into its numerical form divided into blocks according to the block size, which is 2048 bits in the case of RSA-2048. To provide appropriate encryption, each block is padded using techniques such as PKCS#1 v1.5. Modular exponential growth is conducted on each block, leading to a ciphertext block obtained by applying a modulus operation to each block. This operation will continue for every text block, ensuring that the entire dataset is encrypted. Following these steps ensures that messages are securely encrypted.

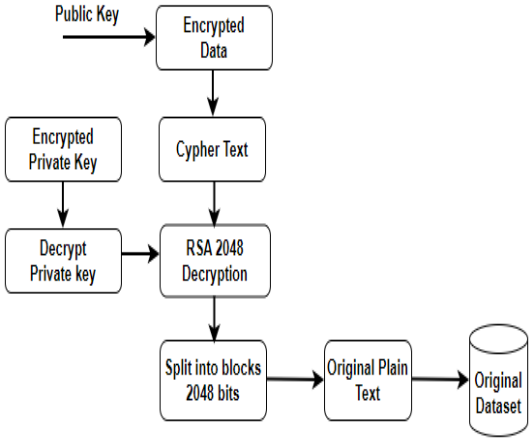

Fig 4: BC_RSA-2048 Data decryption

Figure 4 illustrates how the steps in the decryption process are the opposite of those in encryption. The RSA-2048 decryption algorithm and the matching private key are used to open the encrypted data. The public key is used to initiate the decryption process, taking the ciphertext as input. The ciphertext, like the encryption process, is separated into blocks according to block size. Modular exponentiation of mod n is applied to every block, achieving the plaintext block. The padding from the encryption process is then eliminated from every decrypted block. The decrypted blocks are put together to reconstruct the plaintext's initial numerical format. The numerical representation is transformed again to text in order to obtain the dataset in its original form. This is accomplished by mapping the integers to characters employing a certain encoding technique, like UTF-8. By following these steps, the original dataset can be successfully retrieved through the decryption process, effectively reversing the encryption procedure.

3.3 Machine Learning Model:

In this work, we classified a dataset of employee details using a variety of classification methods, including Adaboost M1, Decision Trees J48, Decision Trees ID3, Naive Bayes, and Random Forest. A dataset with employee data was utilised for testing while another dataset with the same data was used to train the machine learning models. Our goal was to use each algorithm to count the number of employees according to their educational background and employer type. 3000 instances made up the testing dataset. The algorithms' efficiency was assessed based on two criteria: classification accuracy and execution time. Our objective was to evaluate each algorithm's accuracy and efficiency in predicting the employee's educational background and employer type.

AdaBoost.M1 is an ensemble learning technique that combines a number of weak classifiers to produce a powerful classifier. It is also known as adaptive boosting. It is well known for handling difficult classification problems effectively, and it has been routinely used to increase classification accuracy overall [18]. The Iterative Dichotomiser 3 (ID3) algorithm is a well-known decision tree approach that builds the tree based on information gained from category attributes. A modification of the ID3 algorithm [19] known as J48, sometimes known as C4.5, can handle categorical and continuous attributes as well as missing attribute values [20]. A well-known technique, the Naive Bayes classifier [21], use Bayes' theorem to determine the likelihood of a particular class label given the input data. It is a common option for activities requiring classification. Individual regression trees have its limitations, but the Random Forest algorithm circumvents them [22]. It is an ensemble method that is used for both regression and classification tasks. Random feature selection is used during the tree-construction process, and bootstrap trials from the training data are included [23]. The

study uses these classification algorithms in order to assess their precision and efficacy in classifying employee education and employer types, with execution time and error rate acting as assessment criteria.

4. Results and Discussion:

4.1 Lookup Substitution of salary attribute:

Table 3 demonstrates the implementation of lookup substitution for masking, specifically focusing on the sensitive attribute of salary. The table presents alternative values that represent the masked salary, ensuring the dataset's structure and integrity are preserved.

TABLE 3: SALARY LOOKUP SUBSTITUTION TABLE (WITH MASKED VALUES)

GID	Masking_Salary	Count
1	0-100000	233
2	400001-1000000	691
3	300001-400000	1065
4	200001-300000	873
5	100001-200000	1102

For the lookup substitution technique, actual salary numbers are transformed into predetermined ranges. The privacy of people is protected by using salary ranges rather than specific employee salaries since it avoids direct identification of specific wage information. The defined wage ranges and the matching number of employees within each category are shown in Table 4. With this strategy, the dataset can be analysed and shared with outside parties without specific wage information being revealed.

According on the organization's privacy policies, legal requirements, and the required level of granularity in the dataset, certain ranges and the mapping for lookup substitution can be tailored. Lookup substitution ensures the secrecy of sensitive wage information while maintaining the dataset's usability and efficacy.

4.2 Salary attribute encryption:

An efficient strategy is to encrypt the data using the RSA-2048 encryption technique from the Bouncy Castle package to safeguard the security of sensitive information, such as the wage attribute in the employee details dataset. The salary column is encrypted, and the original values are swapped out with their equivalent encrypted ones. A public-private key pair is used in this encryption method, with the public key being used for encryption and the secret key being safely preserved for decryption.

TABLE 4: SALARY ENCRYPTED TABLE USING BC_RSA_2048

Employee ID	Age	Employer Type	Zipcode	Education	Occupation	Gender	Country	Salary
1	39	State-gov	77516	Bachelors	Adm-clerical	Male	United-States	DB75J
2	50	Self-emp-not-inc	83311	Bachelors	Exec-managerial	Male	United-States	vuB52
3	38	Private	215646	HS-grad	Handlers-cleaners	Male	United-States	FxfFt
4	53	Private	234721	11th	Handlers-cleaners	Male	United-States	m4vDB
5	28	Private	338409	Bachelors	Prof-specialty	Female	Cuba	qM/c8
6	37	Private	284582	Masters	Exec-managerial	Female	United-States	Hg/u2
7	49	Private	160187	9th	Other-service	Female	Jamaica	Pinp0
8	52	Self-emp-not-inc	209642	HS-grad	Exec-managerial	Male	United-States	STwMu
9	31	Private	45781	Masters	Prof-specialty	Female	United-States	VeCTQ
10	42	Private	159449	Bachelors	Exec-managerial	Male	United-States	sE4PZ
11	37	Private	280464	Some-college	Exec-managerial	Male	United-States	jpfGR
12	30	State-gov	141297	Bachelors	Prof-specialty	Male	India	1b4jC
13	23	Private	122272	Bachelors	Adm-clerical	Female	United-States	egVuZ

Table 4 displays the encrypted salary figures, guaranteeing the privacy and safeguarding private data from unauthorised access. The encrypted salary values and the original salary data cannot be accessed by anybody without the secret key. Privacy is preserved by encrypting the salary column since the sensitive information is kept safe and unreadable by unauthorised parties. The encrypted wage values are safe even in cases when the dataset is accessed by unauthorised people or exposed to unauthorised disclosure, and they are difficult to decrypt without the associated private key.

Use of the RSA-2048 encryption algorithm from the Bouncy Castle library offers a reliable and well-liked method for protecting sensitive data. This encryption technique improves the dataset's security by guaranteeing the preservation of privacy and preserving the employee information's value for legal uses like analysis and processing. Utilising this encryption technology adds an extra degree of security while balancing data security and data utility.

4.3 Classification of Machine Learning:

The dataset is exposed to classification based on employee education and employer type by using several machine learning classifiers. Algorithms like

Adaboost M1, Naive Bayes, Decision Trees J48, Random Forest, and Decision Trees ID3 are used to complete the classification problem. These are the outcomes of this classification process:

A. Employee Education

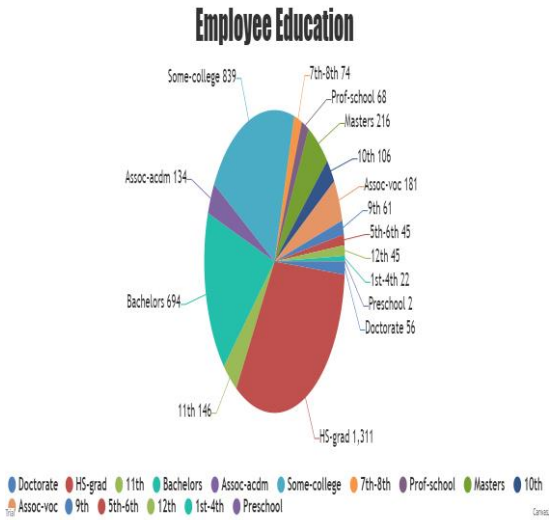

Fig 5: a) Initial Dataset Classification according to education

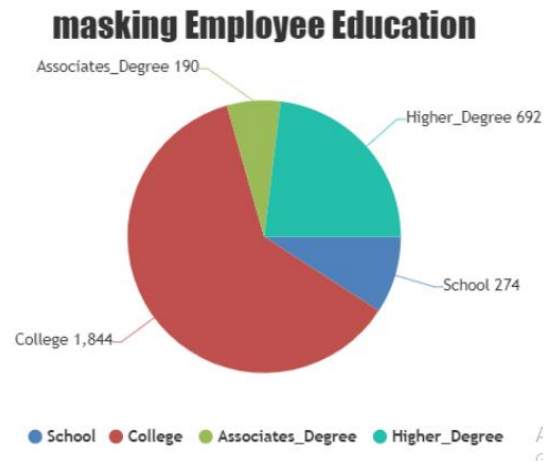

Fig 5: b) Masked Dataset Classification according to education

The classification of the dataset is based on the level of education of the employees and uses preset categories of education, including doctorate, masters, school level, preschool, and others. Figure 5a shows the distribution of employees across each education group and serves as a visual depiction of the employee education classification. According to Figure 5b, those with a high school diploma have the most employees, followed by those with a bachelor's degree and some college education.

Type of Employer

Additionally, the information divides workers into seven different employer categories, including private, local government, state government, and others. The employee count for each type of employment is shown graphically in Figure 6a, with the private employer category having the most workers.

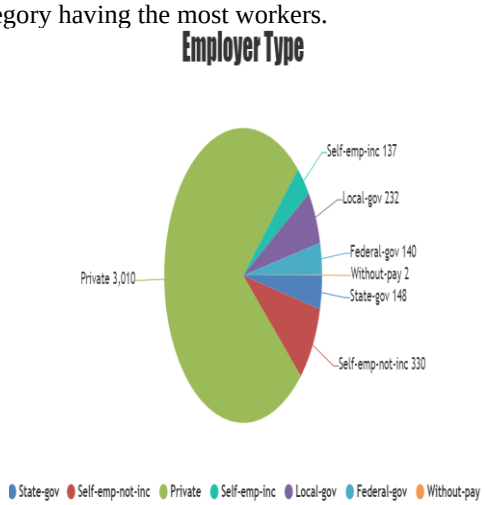

Fig 6: a) Original Dataset Classification of Employer type

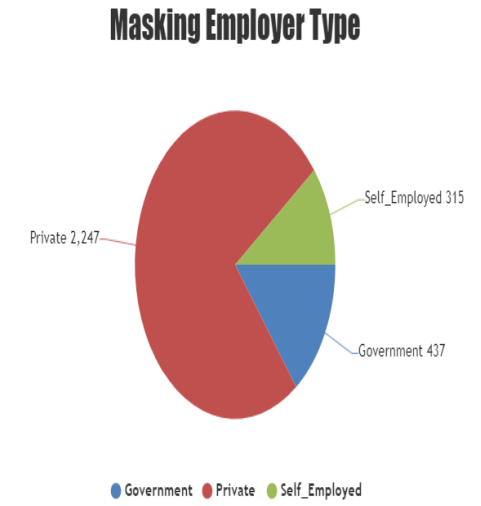

Fig 6: b) Masked Dataset Classification of Employer type

These categorization results enable a thorough knowledge of the properties of the dataset by offering useful insights into the distribution of employees depending on education and company type.

Comparing the classification outcomes of the original dataset with the masked dataset shows how effective the masking technique is at protecting data. Figures 6a and 6b show how sensitive data and identities were transformed in the original dataset and then visually represented, protecting the secrecy of the data. This masking technique successfully hides the original values while maintaining the integrity and structure of the dataset.

4.4 Performance Evaluation

Several assessment measures, such as mean absolute error, root mean squared error, relative absolute error, and root relative squared error are taken into consideration while evaluating each classification algorithm. Table 5 presents these measures and offers

an evaluation and comparison of each algorithm's performance based on them. The findings show that among the algorithms taken into consideration, Naive Bayes (11.0667%) has the largest percentage of cases that are incorrectly classified, followed by Random Forest (8.6%), Decision Trees ID3 (8.3%), Decision Trees J48 (7.2333%), and Adaboost M1 (5.4333%). The task-specific classification needs and the features of the dataset, among other things, should be taken into account while choosing the best classification method.

TABLE 5: MACHINE LEARNING CLASSIFICATION PERFORMANCE

Parameters	Decision Trees J48		Adaboost M1		Naive Bayes		Decision Trees ID3		Random Forest	
Correctly Classified Instances	2783	92.78%	2837	94.57%	2663	88.93%	2841	81.37%	2742	91.4%
Incorrectly Classified Instances	217	7.23%	163	5.43%	332	11.07%	249	8.3%	258	8.6%
Total Number of Instances	3000		3000		3000		3000		3000	

When evaluating the effectiveness of different machine learning algorithms used for classification tasks, evaluation measures are essential. One such indicator, the Kappa statistic, assesses the level of agreement between anticipated and actual classifications while taking chance agreement into consideration. Greater agreement between forecasts and actual values is shown by higher Kappa values, particularly in Adaboost M1. The average absolute difference between expected and actual values is calculated using the mean absolute error (MAE). A closer average match between the predicted and actual values is indicated by lower MAE values, particularly in Random Forest. The average squared deviations between the predicted and actual values are computed using the root mean squared error (RMSE). Better accuracy in capturing the variation between anticipated and actual values is indicated by lower RMSE values, especially in Adaboost M1. The MAE is expressed as a percentage of the mean of the actual values in relative absolute error, which sheds light on the typical relative difference between the anticipated and actual values. Improved accuracy is indicated by lower relative absolute error numbers, as in Naive Bayes. The square root of the variance of the actual values is used to determine the root relative squared error as a percentage. Lower values represent more accuracy. This statistic calculates the ratio of expected and actual values. The performance and accuracy of machine learning algorithms in classification tasks can be evaluated using all of these evaluation indicators.

TABLE 6: ERROR EVALUATION FOR VARIOUS ML ALGORITHMS

Algorithms	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Adaboost M1	0.0116	0.1111	0.2498	99.90%	101.29%
Decision Trees ID3	-0.0078	0.1111	0.2837	111.41%	121.41%
Decision Trees J48	0.0045	0.1123	0.2584	100.98%	104.78%
Random Forest	-0.0104	0.1109	0.2709	99.75%	109.85%
Naive Bayes	0.0056	0.2499	0.4407	99.52%	100.74%

According to the results presented in table 6, various machine learning algorithms exhibit varying performance outcomes through the evaluation. For instance, Naive Bayes demonstrates the highest MAE (mean absolute error) and a high RAE (relative absolute error), suggesting comparatively inferior accuracy. On the other hand, Decision Trees ID3 displays a negative value of kappa statistics, indicating difference in actual and predictions values. Conversely, the specific context as well as classification requirements considered while understanding these metrics. The selection of the utmost fit algorithm should take into account factors beyond the evaluation metrics alone. To gain a visual understanding of the performance measures and the errors associated with each classification algorithm, refer to figure 7. It provides a graphical representation that highlights the variations in performance across the different algorithms.

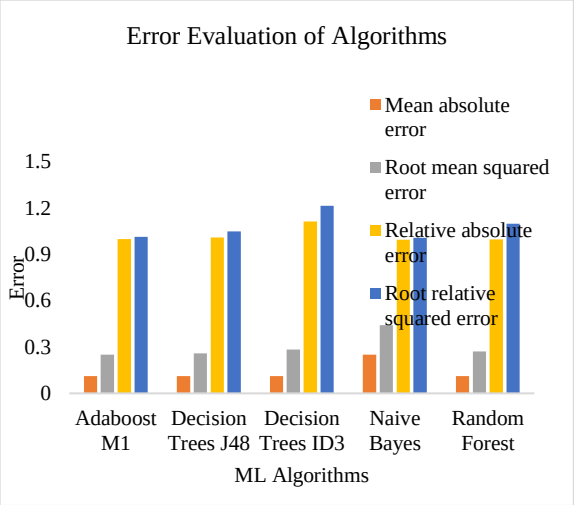

Fig 7: Error Evaluation in various algorithms

4.4.1 Performance Metrics

The performance evaluation of all algorithms in the study encompasses various parameters such as recall,

precision, FPR (False Positive Rate), TPR (True Positive Rate), ROC (Receiver Operating Characteristic) Area and F-measure. These metrics are utilized to measure the effects in the dataset for all education parameters, attained by classification of dataset. Additionally, a it calculates weighted average to provide an overall comparison of all algorithms. Since the weighted average is used for these algorithms, the study summarizes their classification performance, as revealed in the table 7. The table presents the highest ROC and precision of 0.512 and 0.184, respectively, achieved by the Adaboost M1 algorithm. On the other hand, Naïve Bayes exhibits higher True Positive (TP) rate and recall values of 0.111 each. Furthermore, Decision Tree ID3 achieves a high F-measure value of 0.096.

These results provide insights into the performance of each algorithm in classifying the education parameter class in the dataset, highlighting their respective strengths and weaknesses based on the evaluation metrics.

TABLE 7: ALGORITHMS EVALUATED WITHOUT MASKING

Algorithms	Recall	Precision	FPR	TPR	ROC Area	F-Measure
Random Forest	0.086	0.17	0.095	0.086	0.498	0.083
Adaboost M1	0.054	0.184	0.038	0.054	0.512	0.048
Decision Tree ID3	0.093	0.171	0.098	0.093	0.49	0.096
Decision Tree J48	0.072	0.172	0.065	0.072	0.52	0.075
Naïve Byes	0.111	0.162	0.095	0.111	0.401	0.094

The evaluation consequences reveal that the Decision Tree J48 demonstrates superior performance than other algorithms. It achieves better values in metrics such as TP Rate, F-Measure as well as Precision, and Recall indicating improved classification accuracy. Additionally, the algorithm exhibits a higher ROC Area, representing a favorable balance between TPR and FPR. These results suggest that Decision Tree J48 is a strong contender in terms of classification performance. The graphical representation of the performance of the machine learning algorithms can be observed in Figure 8, providing a visual overview of their comparative performance.

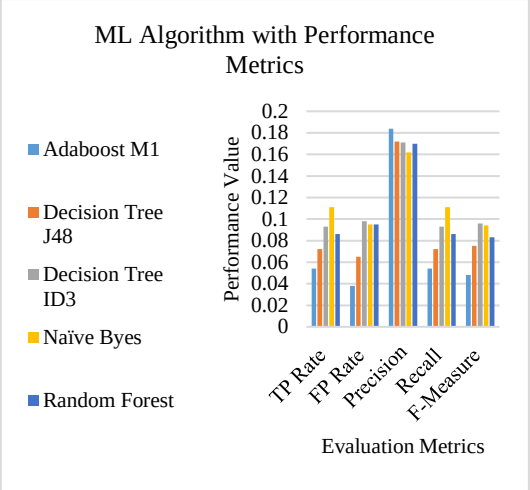

Fig 8: Performance Evaluation of different ML algorithms

4.5 Evaluation of Bouncy Castle RSA-2048 algorithm

The time analysis for BC RSA-2048 includes measuring the time occupied for various operations, such as file upload and deletion, as well as encryption and decryption processes. Table 8 presents three different kinds of datasets (Employee 1, Employee 2, Employee 3) attained by separating a particular dataset into three sections. The table also includes the corresponding file sizes in bytes. It is observed that superior file magnitudes generally need extra processing time for uploading file. The time taken for file upload increases with the increase in file size. Additionally, the different datasets has impact of both the file size as well as system performance on deletion time.

Furthermore, the performance of is evaluated by measuring the time for the encryption and decryption processes of the data. The exact time taken for these operations will depend on various factors, including the size and complexity of the data being processed. By analyzing the time taken for these different operations, it is possible to assess the performance and competence of Bouncy Castle in handling decryption and encryption tasks.

The encryption time refers to the duration for RSA-2048 to encrypt a specific dataset. Similarly, the decryption time is the duration to decrypt the encrypted data as well as restore it to its original form.

Both encryption time and decryption time are important performance indicators for evaluating the efficiency and effectiveness of Bouncy Castle RSA-2048. These metrics can provide insights into the computational resources required for secure data encryption and decryption operations. Detailed information regarding the encryption and decryption time can be found in the provided table.

TABLE 8: DATASET PROCESS TIME USING RSA_2048 IN BC

Dataset	Decryption Time (ms)	Encryption Time (ms)	File Size (KB)	Delete Time (ms)	Upload Time (ms)
Employee 3	959532	1954613	294.43	43	7212
Employee 2	899532	1754613	220.87	40	5349
Employee 1	822761	1559009	220.81	212	5209

To assess the effectiveness of Bouncy Castle using RSA-2048, it is crucial to evaluate the competence of its encryption as well as decryption algorithms. The speed at which data can be encrypted and decrypted plays a significant role in determining the overall performance, particularly in time-sensitive scenarios. Table 8 provides insights into the encryption and decryption times, showcasing an increase in processing time as the file size grows. Specifically, Employee 3 dataset exhibits higher encryption and decryption times compared to the other datasets. For a visual representation of dataset processing time, please refer to Figure 9, which illustrates the time taken to upload the datasets, and Figure 10, which demonstrates the time required for encryption and decryption processes.

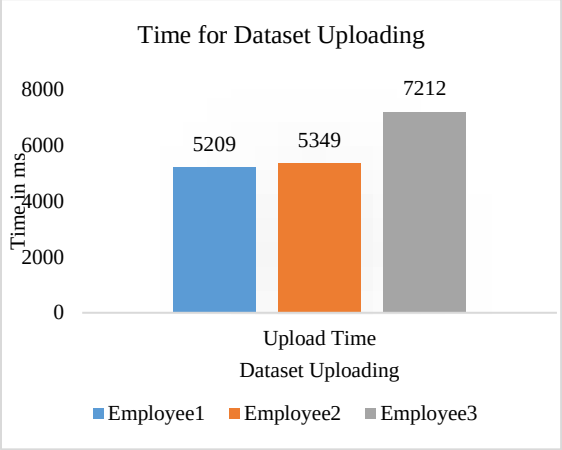

Fig 9: Time taken for uploading the datasets

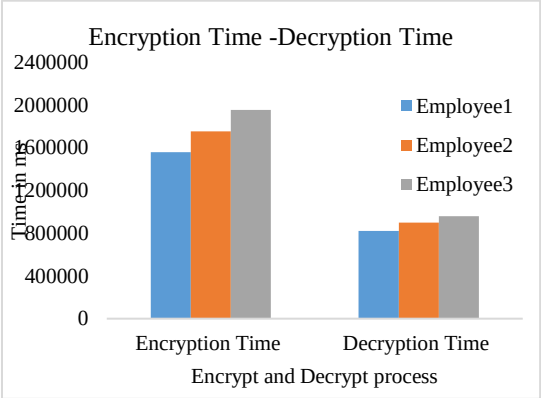

Fig 10: Time taken for Encryption and Decryption of the datasets

The effectiveness of encryption and decryption processes can be influenced by various factors, including the data size, computational resources, and chosen cryptographic parameters. To obtain accurate performance evaluations of Bouncy Castle RSA-2048, it is crucial to conduct thorough benchmarking using realistic scenarios and representative data. Figure 9 visually presents the time evaluation of the system, considering all three datasets.

5. Conclusion

This study uses machine learning algorithms for secure data classification and the Bouncy Castle encryption package to assess privacy-preserving methods. The RSA-2048 encryption algorithm, together with lookup substitution and k-anonymization techniques, is used to improve privacy protection and system speed while reducing data hazards. Adaboost M1, Decision Tree J48, Decision Tree ID3, Naive Bayes, and Random Forest are just a few of the machine learning algorithms used to classify the dataset according to employer type and employee education. Metrics including the Kappa statistic, Mean Absolute Error, Root Mean Squared Error, Relative Absolute Error, Root Relative Squared Error, TPR, FPR, Recall, Precision, F-Measure, and ROC Area are used to evaluate performance.

With higher TPR, Precision, Recall, F-Measure, and a higher ROC Area than the other algorithms, Decision Tree J48 performs relatively better. Analysis of file uploading, encryption, and decryption times is another method used to gauge how effective the encryption process is. The time overhead and data security facets of the encryption approach are discussed in this evaluation. Overall, by highlighting the importance of privacy-preserving strategies in today's data-driven society, this study advances privacy and security in handling sensitive data.

6. References

[1] A. Ali, A. W. Septyanto, I. Chaudhary, H. Al Hamadi, H. M. Alzoubi, and Z. F. Khan, "Applied Artificial Intelligence as Event Horizon Of Cyber Security," in 2022 International Conference on Business Analytics for Technology and Security, ICBATS 2022, 2022. doi: 10.1109/ICBATS54253.2022.9759076.

[2] A. A. Mughal, "Cybersecurity Hygiene in the Era of Internet of Things (IoT): Best Practices and Challenges," Appl. Res. Artif. Intell. Cloud Comput., vol. 2, no. 1, pp. 1–31, 2019.

[3] O. G. Abood and S. K. Guirguis, "A Survey on Cryptography Algorithms," Int. J. Sci. Res. Publ., vol. 8, no. 7, pp. 495–516, 2018, doi: 10.29322/ijsrp.8.7.2018.p7978.

[4] C. Riman and P. E. Abi-Char, "Comparative Analysis of Block Cipher-Based Encryption Algorithms: A Survey," Comput. Fraud, vol. 3, no. 1, pp. 1–7, 2015, doi: 10.12691/iscf-3-1-1.

[5] P. Dixit, A. K. Gupta, M. C. Trivedi, and V. K. Yadav, “Traditional and hybrid encryption techniques: A survey,” Netw. Commun. Data Knowl. Eng. Springer Singapore, vol. 2, pp. 239–248, 2018, doi: 10.1007/978-981-10-4600-1_22.

[6] J. Liu, Y. Tian, Y. Zhou, Y. Xiao, and N. Ansari, “Privacy preserving distributed data mining based on secure multi-party computation,” Comput. Commun., vol. 153, pp. 208–216, 2020, doi: 10.1016/j.comcom.2020.02.014.

[7] M. Blanton, A. Kang, and C. Yuan, “Improved Building Blocks for Secure Multi-party Computation Based on Secret Sharing with Honest Majority,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, pp. 377–397. doi: 10.1007/978-3-030-57808-4_19.

[8] D. H. Vu, T. D. Luong, and T. B. Ho, “An efficient approach for secure multi-party computation without authenticated channel,” Inf. Sci. (Ny)., vol. 527, pp. 356–368, 2020, doi: 10.1016/j.ins.2019.07.031.

[9] Sudhakar, M. ., & Kaliyamurthie, K. P. . (2023). A Novel Machine learning Algorithms used to Detect Credit Card Fraud Transactions. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2), 163–168. https://doi.org/10.17762/ijritcc.v11i2.6141

[10] N. Kaaniche, M. Laurent, and S. Belguith, “Privacy enhancing technologies for solving the privacy-personalization paradox: Taxonomy and survey,” Journal of Network and Computer Applications, vol. 171. Academic Press, p. 102807, Dec. 01, 2020. doi: 10.1016/j.jnca.2020.102807.

[11] S. Murthy, A. Abu Bakar, F. Abdul Rahim, and R. Ramli, “A Comparative Study of Data Anonymization Techniques,” in Proceedings - 5th IEEE International Conference on Big Data Security on Cloud, BigDataSecurity 2019, 5th IEEE International Conference on High Performance and Smart Computing, HPSC 2019 and 4th IEEE International Conference on Intelligent Data and Securit, IEEE, May 2019, pp. 306–309. doi: 10.1109/BigDataSecurity-HPSC-IDS.2019.00063.

[12] B. Ojokoh and E. Adebisi, “A review of question answering systems,” J. Web Eng., vol. 17, no. 8, pp. 717–758, 2019, doi: 10.13052/jwe1540-9589.1785.

[13] G. M. Biancofiore, Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. Narducci, “Interactive Question Answering Systems: Literature Review,” arxiv.org, Sep. 2022, Accessed: May 31, 2023. [Online]. Available: https://arxiv.org/abs/2209.01621

[14] F. J. Aufa, Endroyono, and A. Affandi, “Security System Analysis in Combination Method: RSA Encryption and Digital Signature Algorithm,” in Proceedings - 2018 4th International Conference on Science and Technology, ICST 2018, 2018. doi: 10.1109/ICSTC.2018.8528584.

[15] J. C. Asenjo, “Data Masking, Encryption, and their Effect on Classification Performance: Trade-offs Between Data Security and Utility,” 2017.

[16] K. Arava and S. Lingamgunta, “Adaptive k-Anonymity Approach for Privacy Preserving in Cloud,” Arab. J. Sci. Eng., vol. 45, no. 4, pp. 2425–2432, Apr. 2020, doi: 10.1007/s13369-019-03999-0.

[17] J. Andrew, J. Karthikeyan, and J. Jebastin, “Privacy Preserving Big Data Publication on Cloud Using Mondrian Anonymization Techniques and Deep Neural Networks,” in 2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019, 2019, pp. 722–727. doi: 10.1109/ICACCS.2019.8728384.

[18] M. N. Alenezi, H. Alabdulrazzaq, and N. Q. Mohammad, “Symmetric encryption algorithms: Review and evaluation study,” Int. J. Commun. Networks Inf. Secur., vol. 12, no. 2, pp. 256–272, 2020.

[19] H. Chen, Z. Lin, L. Mo, and C. Tan, “Identification of Colorectal Cancer Using Near-Infrared Spectroscopy and Adaboost with Decision Stump,” Anal. Lett., vol. 50, no. 16, pp. 2608–2618, Nov. 2017, doi: 10.1080/00032719.2017.1310880.

[20] S. CHALO and İ. Berkan AYDİLEK, “A New Preprocessing Method for Diabetes and Biomedical Data Classification,” Qubahan Acad. J., vol. 2, no. 4, pp. 6–18, 2023, doi: 10.48161/qaj.v2n4a135.

[21] N. Khanna, “J48 Classification (C4.5 Algorithm) in a Nutshell,” Medium, 2003. https://medium.com/@nilimakhanna1/j48-classification-c4-5-algorithm-in-a-nutshell-24c50d20658e.

[22] V. K. Vineetha and P. Samuel, “A Multinomial Naïve Bayes Classifier for identifying Actors and Use Cases from Software Requirement Specification documents,” in 2022 2nd International Conference on Intelligent Technologies, CONIT 2022, 2022. doi: 10.1109/CONIT55038.2022.9848290.

[23] R. S. Khairy, A. S. Hussein, and H. T. H. S. ALRikabi, “The Detection of Counterfeit Banknotes Using Ensemble Learning Techniques of AdaBoost and Voting,” Int. J. Intell. Eng. Syst., vol. 14, no. 1, pp. 326–339, 2021, doi: 10.22266/IJIES2021.0228.31.

[24] Brown, R., Brown, J., Rodriguez, C., Garcia, J., & Herrera, J. Predictive Analytics for Effective Resource Allocation in Engineering Education. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/91

[25] Y. L. Pavlov, “Random forests,” Random For., vol. 45, pp. 1–122, 2019, doi: 10.4324/9781003109396-5.