DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018 ***Int. J. Exp. Res. Rev.*, Special Vol. 40: 217-226 (2024)**

Original Article | *Peer Reviewed* | *Open Access*

International Journal of Experimental Research and Review (IJERR)

ISSN: 2455-4855 (Online)
www.iaph.in

ISSN 2455-4855

Securing the Data Using an Efficient Machine Learning Technique

Pinkal Jain[1*] and Vikas Thada[2]

[1]Department of Computer Science & Engineering, Amity University Gwalior -474001, Madhya Pradesh, India;
[2]Department of Computer Science & Engineering, Amity University Gwalior -474001, Madhya Pradesh, India

E-mail/Orcid Id:

PJ, pinku029jain@gmail.com, https://orcid.org/0000-0001-8002-320X; *VT,* vthada@gwa.amity.edu, https://orcid.org/0000-0002-8131-9616

Article History:

Received: 11th Nov., 2023
Accepted: 23rd June, 2024
Published: 30th June, 2024

Keywords:

Privacy, Differential privacy, Decision Forest, Decision Tree, Privacy-Preserving Data Mining, Noisy Data

How to cite this Article:

Pinkal Jain and Vikas Thada (2024). Securing the Data Using an Efficient Machine Learning Technique. *International Journal of Experimental Research and Review, 40(spl.),* 217-226.

DOI:

https://doi.org/10.52756/ijerr.2024.v40spl.018

Abstract: More accessible data and the rise of advanced data analysis contribute to using complex models in decision-making across various fields. Nevertheless, protecting people's privacy is vital. Medical predictions often employ decision trees due to their simplicity; however, they may also be a source of privacy violations. We will apply differential privacy to this end, a mathematical framework that adds random values to the data to provide secure confidentiality while maintaining accuracy. Our novel method Dual Noise Integrated Privacy Preservation (DNIPP) focuses on building decision forests to achieve privacy. DNIPP provides more protection against breaches in deep sections of the tree, thereby reducing noise in final predictions. We combine multiple trees into one forest using a method that considers each tree's accuracy. Furthermore, we expedite this procedure by employing an iterative approach. Experiments demonstrate that DNIPP outperforms other approaches on real datasets. This means that DNIPP offers a promising approach to reconciling accuracy and privacy during sensitive tasks. In DNIPP, the strategic allocation of privacy budgets provides a beneficial compromise between privacy and utility. DNIPP protects privacy by prioritizing privacy concerns at lower, more vulnerable nodes, resulting in accurate and private decision forests. Furthermore, the selective aggregation technique guarantees the privacy of a forest by combining multiple data points. DNIPP provides a robust structure for decision-making in delicate situations, ensuring the model's effectiveness while safeguarding personal privacy.

Introduction

Personal information has been increasingly acknowledged for quite some time. The societies based on data are constantly spewing out the intimate details of ourselves. New technology such as data mining takes advantage of personal data and can offer personalized services or products in different sectors like web search engines or healthcare (Abadi et al., 2016). Advanced data mining techniques have the potential to improve medical services for patients. However, external knowledge from healthcare databases during mining could unintentionally compromise patient confidentiality (Jain et al., 2023). Although mining electronic medical records holds promise for exploring disease relationships and medical treatments, it also raises concerns regarding the exposure of confidential patient information (Abouelmehdi et al., 2018).

Privacy-preserving data mining addresses the problem by employing techniques that maintain the data's secrecy while providing valuable business insights (Jain et al., 2024). Classification is one of the basic techniques for data mining and is crucial in predictive analytics(Bu et al., 2021). The popular model of classification decision trees has excellent accuracy, but it may also have privacy risks because it requires counting (Bettini et al., 2015; Bonawitz et al., 2020). The robust framework known as differential privacy is useful in checking individual privacy leakages; differential privacy checks against breaches of privacy by ensuring that any changes made to individual records do not bias calculations based on data

***Corresponding Author:** pinku029jain@gmail.com

(Jain et al., 2015). Initially introduced for statistical database security purposes, this idea has now become common in PPDM, involving clustering, classification, and deep learning (Feng et al., 2005; Yavanamandha et al., 2023; Mondal et al., 2023; Kumar et al., 2023). In recent years, building differentially private tree-based models has been a successful endeavor (Claerhout et al., 2005). However, most existing approaches typically overlook the issue of allocating an adequate privacy budget, which can sometimes negatively impact the overall performance of the model (Gupta et al., 2020; Cui et al., 2019).

This article suggests an alternative construction for private trees that represents a more refined approach to budget allocation.

1. Our contributions also encompass creating an algorithmic model for budget allocation that allocates different budgets to nodes depending on their position within the tree based on their position within the tree, thereby reducing performance degradation due to improper budget allocation.
2. We suggest a method for selective aggregation to enhance the generality and prediction accuracy of ensemble models, as well as an iterative approach to facilitate speedup in the process.
3. To verify the efficiency of our classification model that ensures privacy preservation and individual protection, we perform simulation experiments on real datasets.

The paper's structure is as follows: Section 2 contains reviews of related works; the remainder of 3 introduces the preliminaries; and the remainder of 4 mainly describes our proposed DNIPP scheme and the system's threat model. Sections 5 discuss the construction of private decision trees, the selective aggregation process, and the evaluation of the accuracy and efficiency of DNIPP. Finally, Section 6 presents the paper's conclusion.

Literature Review

At present, various techniques are employed to put data under data security, such as anonymization techniques (Cui et al., 2019). Procedures that make generalizations over data to safeguard privacy characterize these data anonymization methods. Nonetheless, they can't effectively defend themselves against attacks because modeling the attacker's background knowledge poses a challenge (Miller et al., 2009). Differential privacy provides a strong and practical definition of privacy protection by preventing attackers from extracting precise individual information from computation results (Jain et al., 2015; Yadav and Singh, 2023). This concept has thus gained considerable attention in the realm of privacy-preserving data mining (PPDM) (Malin et al., 2004).

While decision trees are renowned for their transparency in data mining, this attribute can pose a threat to privacy when attackers exploit it to extract information (Yang et al., 2018). To solve this problem, some decision tree algorithms with differential privacy have been proposed. For instance, the SuLQ-based ID3 algorithm was proposed by (Jain et al., 2015), which used differential privacy when evaluating attribute information gain by including Laplacian noise in computing query results (Sharma et al., 2018). However, its effectiveness continuously decreases because it lowers significantly the classification accuracy (Li et al., 2015).

To tackle these problems, DiffP-ID3 and DiffP-C4.5 were developed with an exponential mechanism for selecting splitting attributes to maximize classification accuracy while protecting individuals' privacy at the same time (Tayefi et al., 2017). Besides that, some approaches use ensemble methods like random forests that can help reduce the negative impact of noises on the model's behavior. Freidman and Schuster came up with an efficient way to construct a differentially private ID3 classifier that reveals its efficacy across datasets of different sizes. Alternatively, a few authors proposed a differentially private random forest algorithm that randomly selects split attributes among internal nodes. (Feng et al., 2005) devised a differentially private ensemble method to enhance model accuracy by reducing privacy requirements. Certain methodologies concentrate on reducing the randomness inherent in the exponential mechanism. Fletcher and Islam proposed an alternative by advocating for the use of local sensitivity, as opposed to global sensitivity, in calculating the score function's sensitivity (Yin et al., 2018). Furthermore, they recommended the creation of a random forest with soft sensitivity.

Despite these advancements, the majority of current algorithms fail to account for noise tolerance at varying depths within trees. They introduced an adaptive budget allocation method that continuously allocates privacy budgets for queries and provides consistent accuracy results. However, this approach causes additional spending on privacy parameter calculations, and attempting to optimize the allocation for each query is still an unsolved problem (Zhu et al., 2020). The main goal of this paper is to bridge this gap by developing a well-tuned strategy for allocating privacy budgets so that they are more effective.

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018

This section discusses two techniques employed in differential privacy that focus on its foundational concept (Zhang et al., 2020). Then, we will discuss the Gini Index, which is one of the important metrics used in selecting optimal split attributes during tree construction (Zheng et al., 2017).

Differential Privacy

The differential privacy technique ensures that adding or removing any record from a dataset has a negligible effect on computation outcomes. Consequently, it prevents the extraction of precise individual information from the results.

Definition 1: Differential Privacy

Differential privacy concerns a randomized computation F, where Range(F) encompasses all possible outcomes. Given adjacent datasets D_1 and D_2 differing by one record ($|D_1 \Delta D_2| = 1$), if algorithm F satisfies:

$Pr(F(D1) \in S) \leq e\varepsilon \cdot Pr(F(D2) \in S)$

For any subset S of Range (F), F is said to uphold ε-differential privacy. Here, ε denotes the privacy budget, inversely proportional to the level of privacy protection.

Definition 2: Sensitivity

The sensitivity of a function f: D→Rd, operating on an arbitrary domain (D) and producing a (d)-dimensional real number vector, is defined as:

$\Delta f = \max D1, D2$ where $|D1 \Delta D2| = 1$ $||f(D1) - f(D2)|| 1$

Usually, to achieve (ε)-differential privacy for numerical queries, noise drawn from a calibrated Laplace distribution is added to the query results.

Definition 3: Laplace Mechanism

For a function f: D→Rd, where D is an arbitrary domain, the Laplace mechanism ensures ε-differential privacy and is defined as:

$F(D) = f(d) + Laplace(\varepsilon \Delta f)$

However, for non-numerical queries, the exponential mechanism is employed to maintain ε-differential privacy.

Definition 4: Gaussian Mechanism

For every domain D, given an arbitrary function f: D→ V^n, the function F offers ∈-differential privacy, the Gaussian noise is as follows-

$$P(Y) = \frac{1}{\sqrt{2\pi\gamma}} e^{-(Y-\mu)^2/2\gamma^2}$$

Definition 5: The Exponential Mechanism

Suppose we have a random mechanism (M) with dataset (D) as input and entity object r∈ Range as output, and a score function q (D, r) assigning scores to each output, with δq representing its sensitivity, The mechanism M maintains ε-differential privacy if:

M (r, q) ={return r with probability ∝ exp($2\Delta q$ εq(D,r))

We have reformulated this expression using different mathematical symbols to provide a fresh and effective perspective.

Gini Index

When making a decision tree, most decisive thing to consider is determining the best split attribute selection criteria. CART used the Gini Index as a criterion. This index, which measures the "purity" of samples, takes the following form:

$Gini(D) = 1 - \sum i = 1npj2$

The symbol pj represents the proportion of the jth sample in the sample set. For attribute A, the Gini Index can be defined as follows:

$GD, C(a) = \sum v = 1V |D||Da = v| \cdot Gini (Da = v, c)$

Where, |Da=v| means a subset of samples with an equaling attribute v, and |D| is for all instances. We calculate this subset's Gini index using the formula Gini (Da = v, c).

Thus, when building decision trees, one should select candidate attributes that minimize the Gini index before and after division. This approach provides optimal attribute selection throughout the tree construction process.

Information Entropy

In the data analysis domain, entropy becomes an important measure to know how uncertain our data is. It measures, essentially, how much surprise or randomness exists in a dataset. The more we know about a dataset, the lower its entropy. Greater dataset uncertainty or unpredictability increases entropy. Mathematically, entropy can be expressed using the following formula:

$$E_P = \sum_{j=1}^{n} -P_j \log_2 P_j$$

Information Gain (IG)

The term "Information Gain (IG)" is a pivotal factor in the development of decision trees. It stands in an inverse relationship with entropy, a measure of uncertainty. The process of computing information gain is recursive, continuing until the leaf nodes of the decision tree reach an entropy value of 0, indicating no further splitting is necessary. The calculation of information gain is crucial for each decision tree node, computed as:

$$IG = Ep - (mi/n) * (Ec)i$$

Where:

Ep is the entropy of the original dataset

mi is the total number of instances in each of the i-th children datasets

n is the total number of instances in the parent dataset.

(Ec)i denotes the entropy of i-th child dataset.

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018

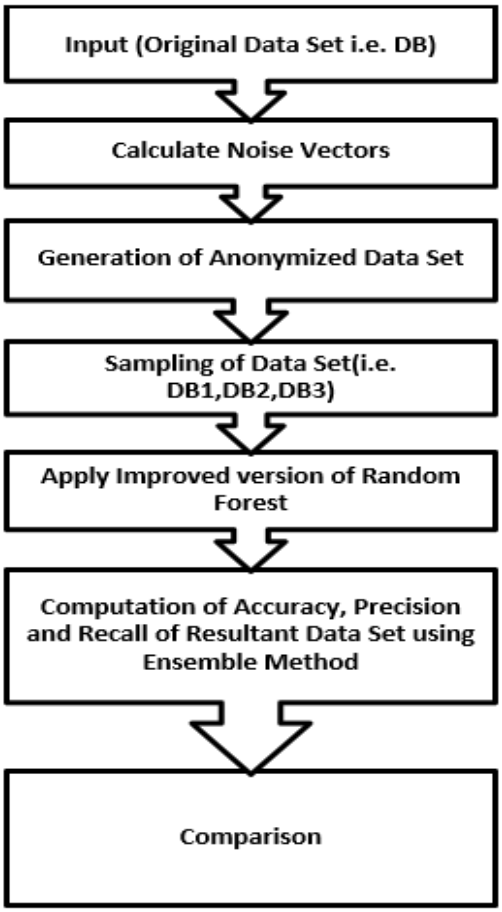

Figure 1. Flowchart of proposed work.

One can compute information gain using either Gini impurity or entropy, but usually the former produces more accurate results. This is what is done in this work by introducing a new scheme called DNIPP.

Proposed Work

It ensures that malicious investigators cannot extract individual privacy information from data sets. Therefore, it helps to build decision trees with strong utility preservation and privacy preservation as proposed in this work. This prevents malicious analysts from extracting individual privacy information from the datasets. The core idea of the DNIPP scheme is to selectively aggregate disjoint subsets into a forest. This strategy mitigates the potential performance degradation that a single private tree might encounter due to the additional randomness introduced for privacy protection. During tree construction, data miners continuously submit queries along with privacy budgets. Nevertheless, once the privacy budget is exhausted, additional queries become impractical. Moreover, leaf nodes and internal nodes have differing levels of tolerance to noise. Therefore, we propose a new budget allocation strategy that assigns a larger privacy budget to nodes at deeper levels, partially mitigating the problem of excessive noise introduced by leaf nodes. The flowchart of the proposed model is shown in Figure 1.

Methodology

For instance, take a record set named Heart Disease Dataset including information on 1024 patients about heart disease characteristics. It contains 14 features ranging from both numerical and categorical values. Both numerical and categorical attributes are present in this dataset. The DNIPP scheme proposed in this work enables the creation of decision trees with strong utility preservation and privacy preservation. It prevents malicious analysts from extracting individual privacy information from the datasets. The main idea behind DNIPP lies in selectively aggregating disjoint subsets into a forest which mitigates potential performance degradation that could be due to extra randomness brought into a single private tree by other means. introduced for privacy protection. This work presents a novel technique for building highly accurate decision forests while ensuring data privacy. Our approach prioritizes privacy in leaf nodes, which are particularly vulnerable to noise introduced for privacy protection.

Dataset

The dataset is considered here as the Heart Disease Dataset. The dataset contains information about 1024

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018

Table 1. Original Heart Disease Dataset i.e., DB.

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

Table 2. Noisy Dataset i.e., DB.

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target	noise
1020	53.19	1	1	140	221	0	1	164	1	0.0	2	0	2	1	-5.80
1021	60.65	1	0	125	258	0	0	141	1	2.8	1	1	3	0	0.65
1022	46.95	1	0	110	275	0	0	118	1	1.0	1	1	2	0	-.043
1023	51.71	0	0	110	254	0	0	159	0	0.0	2	0	2	1	1.711
1024	55.35	1	0	120	188	0	1	113	0	1.4	1	1	3	0	1.35

Table 3. After Row Sampling First Sample of Dataset i.e., DB1.

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target
276	56.41	1	0	132	207	0	1	168	1	0.0	2	0	3	1
784	54.32	1	2	150	232	0	0	165	0	1.6	2	0	3	1
856	63.51	0	2	120	211	0	0	115	0	1.5	1	0	2	1
795	63.40	1	1	128	208	1	0	140	0	0.0	2	0	2	1
477	58.50	1	2	128	229	0	0	150	0	0.4	1	1	3	0
796	38.28	1	1	135	203	0	1	132	0	0.0	1	0	1	1
893	54.33	1	0	128	204	1	1	156	1	1.0	1	0	0	0
828	43.57	1	2	130	233	0	1	179	1	0.4	2	0	2	1
179	57.88	0	0	134	409	0	0	150	1	1.9	1	2	3	0

Table 4. After Row Sampling Second Sample of Dataset i.e., DB2.

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target
231	56.33	1	1	120	236	0	1	178	0	0.8	2	0	2	1
241	66.69	1	2	152	212	0	0	150	0	0.8	1	0	3	0
742	63.99	1	0	130	330	1	0	132	1	1.8	2	3	3	0
179	57.88	0	0	134	409	0	0	150	1	1.9	1	2	3	0
170	47.78	1	0	150	247	0	1	171	0	1.5	2	0	2	1
476	57.10	1	0	165	289	1	0	124	0	1.0	1	3	3	0
839	45.73	1	0	140	261	0	0	186	1	0.0	2	0	2	1
819	59.72	0	0	170	225	1	0	146	1	2.8	1	2	1	0
366	62.27	1	2	112	230	0	0	165	0	2.5	1	1	3	0

Table 5. After Row Sampling Third Sample of Dataset i.e., DB3.

	Age	Sex	Cp	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	target
319	53.37	0	2	128	216	0	0	115	0	0.0	2	0	0	1
702	70.46	0	1	160	302	0	1	162	0	0.4	2	2	2	1
296	64.72	1	0	120	237	0	1	71	0	1.0	1	0	2	0
262	48.68	1	0	122	222	0	0	186	0	0.0	2	0	2	1
867	49.97	1	1	110	235	0	1	153	0	0.0	2	0	2	1
687	55.24	1	0	125	300	0	0	171	0	0.0	2	2	3	0
419	61.82	0	2	160	360	0	0	151	0	0.8	2	0	2	1
545	47.57	1	1	110	229	0	1	168	0	1.0	0	0	3	0
367	46.17	1	1	110	229	0	1	168	0	1.0	0	0	3	0

patients and their attributes are related to heart disease. It has 14 features both in numerical form and categorical. This dataset consists of both numerical and categorical variables. Numerical variables include Age, Trestbps

(resting blood pressure), Chol (serum cholesterol), Thalach (maximum heart rate), and Oldpeak (ST depression induced by exercise). Categorical attributes include Sex, Cp (chest pain type), Fbs (fasting blood sugar), Restecg (resting electrocardiographic results), Exang (exercise-induced angina), Slope (slope of peak exercise ST segment), Ca (number of major vessels colored by fluoroscopy), Thal (thallium stress test result), and target (presence or absence of heart disease). The heart disease dataset is shown in Table 1.

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018

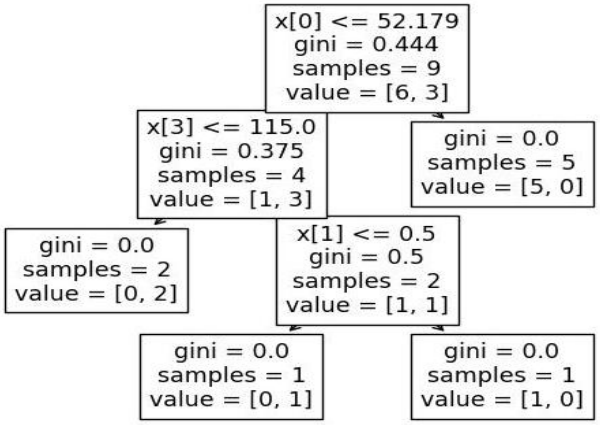

Figure 2. Decision Tree corresponding to row sampled data DB1.

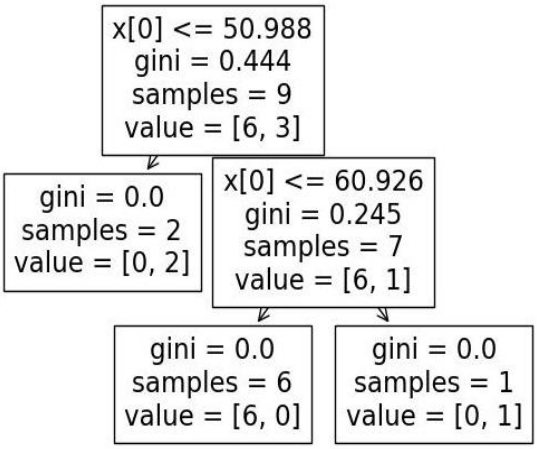

Figure 3. Decision Tree corresponding to row sampled data DB2.

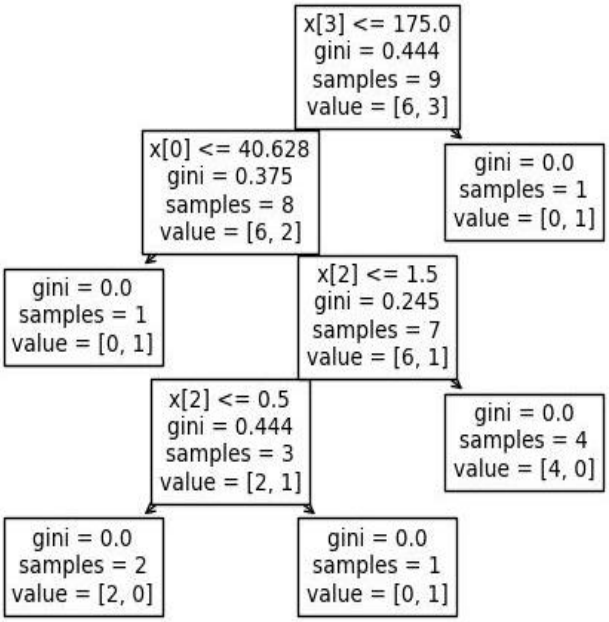

Figure 4. Tree corresponding to row sampled data DB3.

Modified Dataset

A sensitive feature like “age” is identified that has to be anonymized by dual noise integration which is shown in Table 2.

Three types of sampling have been considered for the noisy data in the experimental analysis that is row-wise sampling, column-wise sampling, and combined sampling. Three samples have been generated for each type of sampling. The purpose of generating samples of noisy data sets is to feed each sample data set into a decision tree classifier. A decision tree is generated for each sample of the data set. In this technique, the final prediction is done based on aggregation of all the predictions generated by all the decision trees.

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018

Results & Discussions

Experimental Evaluation and Analysis

This section provides a detailed evaluation of the performance and effectiveness of the proposed algorithm (DNIPP). We assess the algorithm using several metrics such as precision, recall, accuracy, and F1-score. This overall analysis gives an insight into what the algorithm can do best and some limitations it may have. Moreover, experimental studies are performed to show fast computation of important parameters, including Gaussian noise, information entropy, Gini impurity, information gain, and hyperparameter tuning. These experiments demonstrate the scalability of the algorithm's computations as well as its computational efficiency.

At random forest classification various features will be assessed by their importance therefore feature importance vector will be found by –

Table 7. Feature Importance Value.

Name of Feature	Value representing feature importance
age	0. 08777226
Sex	0. 03875317
CP	0. 1577776
Trestbps	0. 06804852
Chol	0. 05536884
Fbs	0. 00838654
Restecg	0. 01418357
Thalach	0. 08071837
Exang	0. 06114553
Oldpeak	0. 12830501
Slope	0. 05535863
Ca	0.05535863
Thal	0.12655215
target	0.12562981

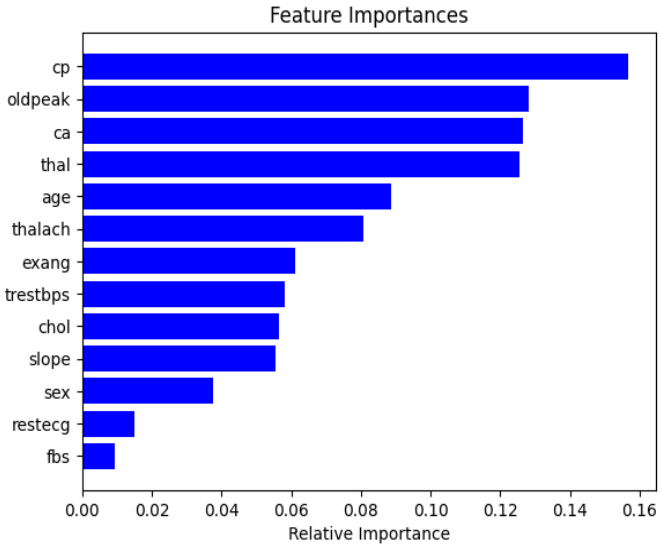

Figure 5. Feature Importance Graph.

Evaluation Metrics and Criteria

Accuracy is a vital metric for assessing the performance of a classification model, including our proposed approach. It indicates the proportion of correct predictions made by the model. For our DNIPP algorithm, the accuracy score is 0.961089, indicating that it correctly classifies 96.11% of the instances.

This surpasses the accuracy achieved by the baseline BDPT method, which stands at 0.78. This significant improvement demonstrates the effectiveness of our privacy-preserving mechanisms in maintaining high accuracy while protecting sensitive data.

Figure 6 visually compares the accuracy scores achieved by different privacy-preserving models. Figure 6 visually compares the accuracy scores achieved by different privacy-preserving models.

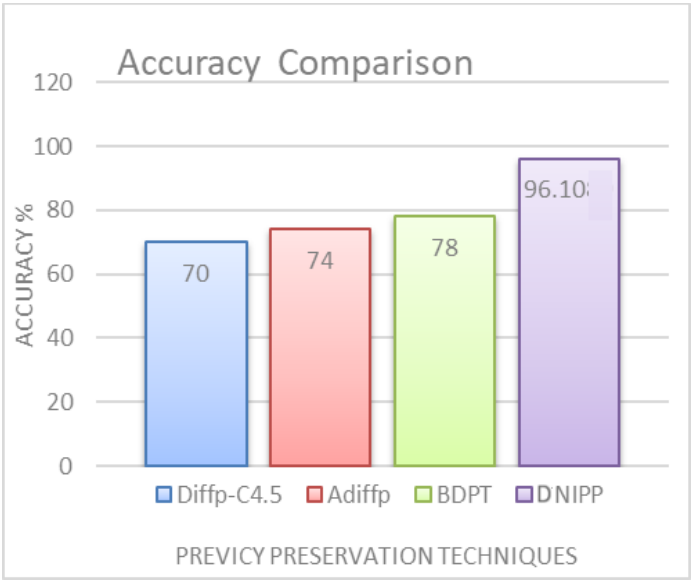

Figure 6. Comparison of accuracy between proposed and existing systems.

Various performance metrics, such as F-1 score, recall, support, and precision, have been computed to evaluate the proposed technique DNIPP. Figure 7 shows the comparison among various performance metrics.

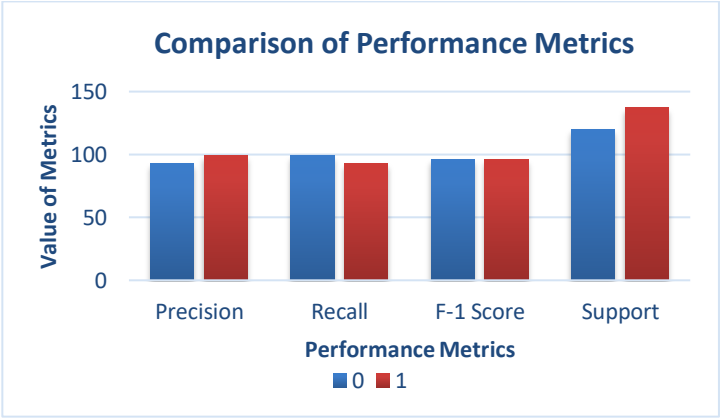

Figure 7. Comparison of performance matrices.

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018

Conclusion & Future Work

This work introduces a new method of constructing decision trees with data privacy. To ensure confidentiality, our work focuses on privacy in leaf nodes, which are most affected by noise. As trees grow deeper and fewer samples are available per node, noise introduces itself into leaf nodes, increasing their susceptibility to noise distortions. Hence, we propose a selective noise integration strategy that adds little noise to the leaves while balancing the trade-off between personal data protection and accuracy. In addition, our selective aggregation technique allows us to choose trees that contribute the most positively to the overall performance of the forest. This ensures that, despite preserving privacy, the aggregated forest remains highly accurate. Experimental results indicate that, compared with previous methods, this approach achieves an excellent balance between privacy and utility.

Acknowledgment

The authors would like to express their gratitude to their family members for their invaluable support and assistance, as well as to the reviewers for their feedback.

Conflict of Interest

The authors declare no conflict of interest.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. *ACM, In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308-318. https://doi.org/10.1145/2976749.2978318.

Abouelmehdi, K., Beni-Hessane, A., & Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *Journal of Big Data, 5*(1). https://doi.org/10.1186/s40537-017-0110-7.

Bettini, C., & Riboni, D. (2015). Privacy protection in pervasive systems: State of the art and technical challenges. *Pervasive and Mobile Computing, 17*, 159-174. https://doi.org/10.1016/j.pmcj.2014.08.004.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., & Seth, K. (2019). Towards federated learning at scale: System design. In Proceedings of Machine Learning and Systems (MLSys) 2020.

Bu, Z., Wang, H., & Long, Q. (2021). On the convergence of deep learning with differential privacy. *arXiv preprint arXiv*:2106.07830.

Claerhout, B., & De Moor, G. J. E. (2005). Privacy protection for clinical and genomic data. *International Journal of Medical Informatics, 74*(2-4), 257-265. https://doi.org/10.1016/j.ijmedinf.2004.06.010.

Cui, L., Qu, Y., Nosouhi, M.R., & Yu, S. J.W.G. (2019). Improving data utility through game theory in personalized differential privacy. *Journal of Computer Science and Technology, 34*(2), 272-286. https://doi.org/10.1007/s11390-019-1918-1.

Feng, Q., He, D., Zeadally, S., & Khan, M.K.N. (2019). A survey on privacy protection in blockchain system. *Journal of Network and Computer Applications, 126*, 45-58. https://doi.org/10.1016/j.jnca.2018.10.020.

Gupta, R., Tanwar, S., Al-Turjman, F., & Italiya, P. A. S. W. (2020). Smart contract privacy protection using AI in cyber-physical systems: Tools, techniques and challenges. *IEEE Access, 8*, 24746-24772.

Jain, P., & Nandanwar, S. (2015). Securing the clustered database using data modification technique. In *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 1163-1166. https://doi.org/10.1109/CICN.2015.331.

Jain, P., & Shakya, H. K. (2023). Achieving privacy preservation in data mining using hybrid transformation and machine learning technique. *MSEA, 71*(4), 7883.

Jain, P., Shakya, H. K., & Lala, A. (2023). Advanced privacy preserving model for smart healthcare using deep learning. *In Proceedings of the IEEE International Conference IC3I 2023*. https://doi.org/10.1109/IC3I59117.2023.10397954.

Jain, P., Shakya, H. K., Nigam, A., Chandanan, A. K., & Murthy, C. R. (2022). Machine learning based privacy preservation in data mining. *CIMS, 28*(12), 350-360.

Jain, P., Thada, V., & Lala, A. (2023). Design of advanced privacy preserving model for protecting privacy within a fog computing scenario. *Proceedings of the IEEE International Conference UPCON 2023*. https://doi.org/10.1109/UPCON59197.2023.10434728.

Jain, P., Thada, V., & Motwani, D. (2024). Providing Highest Privacy Preservation Scenario for Achieving Privacy in Confidential Data. *International Journal of Experimental Research and Review, 39*(spl.) 190-199. https://doi.org/10.52756/ijerr.2024.v39spl.015.

Jain, Pinkal, and Shakya, Harish Kumar (2022). A Review of Different Privacy Preserving Techniques in Data Mining. Paper presented at the International Conference on Innovative Computing & Communication (ICICC) 2022. Retrieved from SSRN: https://ssrn.com/abstract=4021149.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Yang, H. (2019). Advances and open problems in federated learning. *arXiv preprint* arXiv:1912.04977.

Kumar, A., Dutta, S., & Pranav, P. (2023). Supervised learning for Attack Detection in Cloud. *Int. J. Exp. Res. Rev.*, *31*(Spl Volume), 74-84. https://doi.org/10.52756/10.52756/ijerr.2023.v31spl.008

Li, H., Dai, Y., & Lin, X. (2015). Efficient e-health data release with consistency guarantee under differential privacy. IEEE, *In 2015 17th International Conference on E-health Networking, Application & Services* (HealthCom), pp. 602-608. https://doi.org/10.1109/HealthCom.2015.7454576.

Malin, B. A. (2004). An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. Journal of the American Medical Informatics Association, 12(1), 28-34. https://doi.org/10.1197/jamia.M1603.

Sharma, S., Chen, K., & Malin, B. A. (2004). An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future. *Journal of the American Medical Informatics Association*, *12*(1), 28-34. https://doi.org/10.1197/jamia.M1603.

Miller, A. R., & Tucker, C. (2009). Privacy protection and technology diffusion: The case of electronic medical records. *Management Science*, *55*(7), 1077-1093. https://doi.org/10.1287/mnsc.1090.1014.

Miller, A. R., & Tucker, C. (2009). Privacy protection and technology diffusion: The case of electronic medical records. *Management Science*, *55*(7), 1077-1093. https://doi.org/10.1287/mnsc.1090.1014.

Mondal, S., Nag, A., Barman, A., & Karmakar, M. (2023). Machine Learning-based maternal health risk prediction model for IoMT framework. *Int. J. Exp. Res. Rev.*, *32*, 145-159. https://doi.org/10.52756/ijerr.2023.v32.012

Sheth, A. (2018). Practical approaches to privacy-preserving analytics for IoT and cloud-based healthcare systems. *IEEE Internet Computing*, *22*(2), 42–51. https://doi.org/10.1109/MIC.2018.112102519.

Tayefi, M., Tajfard, M., Saffar, S., Hanachi, P., & Ali, R. (2017). Association of hs-CRP with coronary heart disease: A data mining approach using decision tree algorithm. *Computer Methods and Programs in Biomedicine*, *141*, 105–109. https://doi.org/10.1016/j.cmpb.2017.02.001.

Yadav, R., & Singh, R. (2023). Enhancing Software Maintainability Prediction Using Multiple Linear Regression and Predictor Importance. *Int. J. Exp. Res. Rev.*, *36*, 135-146. https://doi.org/10.52756/ijerr.2023.v36.013

Yang, Y., Zheng, X., Guo, W., Liu, X., & Chang, V. (2018). Privacy-preserving fusion of IoT and big data for e-health. *Future Generation Computer Systems*, *86*(SEP), 1437–1455. https://doi.org/10.1016/j.ins.2018.02.005.

Yavanamandha, P., Keerthana, B., Jahnavi, P., Rao, K. V., & Kumar, C. R. (2023). Machine Learning-Based Gesture Recognition for Communication with the Deaf and Dumb. *Int. J. Exp. Res. Rev.*, 34(Special Vol.), 26-35. https://doi.org/10.52756/ijerr.2023.v34spl.004

Yin, C., Xi, J., Sun, R., & Wang, J. (2018). Location privacy protection based on differential privacy strategy for big data in industrial internet of things. *IEEE Transactions on Industrial Informatics*, *14*(8), 3628–3636. https://doi.org/10.1109/TII.2018.2794700

Yuan, J., Yu, S. (2013). Privacy Preserving Back-Propagation Learning Made Practical with Cloud Computing. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 106*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-36883-7_18.

Yuksel, B., Kupcu, A., & Ozkasap, O. (2017). Research issues for privacy and security of electronic health services. *Future Generation Computer Systems*, *68*, 1-13. https://doi.org/10.1016/j.future.2016.08.011.

Zhang, C., Li, S., Xia, J., Wang, W., Yan, F., & Liu, Y. Batchcrypt (2020). Efficient homomorphic encryption for cross-silo federated learning. In Proceedings of the USENIX Annual Technical Conference (USENIX ATC 20), pp. 493–506. https://doi.org 10.5555/3485970.3486018.

Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. IEEE, *In 2017 IEEE International Congress on Big Data*

(BigData Congress), pp. 557-564. https://doi.org/10.1109/BigDataCongress.2017.85

Zhu, T., Ye, D., Wang, W., Zhou, W., & Yu, P.S. (2020). More than privacy: applying differential privacy in key areas of artificial intelligence. https://arxiv.org/abs/2008.01916.

How to cite this Article:

Pinkal Jain and Vikas Thada (2024). Securing the Data Using an Efficient Machine Learning Technique *International Journal of Experimental Research and Review*, *40*(spl.), 217-226.

DOI: https://doi.org/10.52756/ijerr.2024.v40spl.018