Satish Kumar, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

School Scholars Exam Grade Prediction by Enhancing Privacy and Trained Model

Satish Kumar[1]
Research Scholar, Department of CSE, PCST Bhopal[1]

Dr Surendra Singh Vishwakarma[2]
Associate Professor, Department of CSE, PCST Bhopal[2]

Abstract- **The growing availability of educational data has created new opportunities for predicting student academic performance, although privacy protection and appropriate feature selection remain significant challenges. This paper proposes a School Scholar Grade Prediction by Neural Network (SSGPNN) model for privacy-aware student grade prediction. Initially, distributed educational datasets are cleaned by removing sensitive information and transforming categorical attributes into suitable numerical representations. A Genetic Algorithm (GA) is employed to select an optimized set of relevant features and eliminate less informative attributes. The selected features are subsequently used to train a neural network for grade prediction. The proposed model is evaluated at dataset percentages of 12%, 25%, 50%, and 75% using precision, recall, F-measure, and accuracy. Experimental results show that SSGPNN outperforms K-PPD-ERT and achieves maximum precision, recall, F-measure, and accuracy values.**

Keywords— Data mining, Bio Inspired Algorithm, Privacy Preserving Mining, Artificial Neural Network,

I. INTRODUCTION

The rapid advancement of big data technologies and artificial intelligence (AI) has produced both significant opportunities and serious challenges. On the positive side, these technologies have made the collection, storage, processing, and sharing of enormous volumes of heterogeneous data, including images, text, audio, and other digital information, considerably easier and more efficient. Their widespread adoption has contributed to remarkable developments across numerous sectors. In the coming years, big data analytics is expected to become increasingly important in supporting economic growth, technological innovation, and societal development [1]. However, the extensive semantic information embedded in collected, exchanged, and publicly released datasets can also create substantial privacy risks. Malicious entities may analyze, correlate, and combine such information to infer confidential or personally identifiable information about individuals [2].

Consequently, ensuring data security, particularly the protection of individual privacy, has emerged as a major concern in the big data environment. Over the past decade, privacy preservation has therefore received considerable attention from academic researchers, industrial organizations, and government policymakers. In this context, the development of dependable privacy-preserving mechanisms that can safeguard sensitive information without causing an unacceptable reduction in the effectiveness and accuracy of big data analytics continues to be an important and promising research direction [3].

To address the diverse security risks and privacy challenges associated with big data and artificial intelligence, researchers have introduced a variety of protection strategies and technological solutions. From the perspective of the data life cycle, privacy-preserving approaches for big data can generally be classified into four major categories: privacy-preserving data publication, privacy protection

during data storage, privacy-preserving data mining, and privacy protection during data utilization [4]. Among these areas, privacy-preserving data mining (PPDM) has attracted substantial research interest since the concept was first introduced by Agrawal in 2000 [5]. PPDM aims to enable useful knowledge and patterns to be extracted from large datasets while preventing the disclosure of confidential data and sensitive patterns. At the same time, such techniques seek to maintain acceptable computational and analytical efficiency. Building upon this foundation, researchers have developed numerous approaches that integrate privacy-preserving mechanisms with established data-mining and artificial-intelligence techniques. These developments have subsequently extended privacy-aware data analysis to several application domains, including image processing, natural language processing, pattern recognition, and other data-intensive computational applications.

II. LITERATURE SURVEY

In [6], the authors investigated the trade-off between data privacy and utility by adopting a rational framework based on information-theoretic measures. To systematically address this challenge, the privacy–utility relationship was initially formulated as a minimax information-leakage problem. Subsequently, a Privacy-Preserving Attack and Defense (PPAD) game was introduced and modeled as a two-person zero-sum (TPZS) game, representing the conflicting objectives of attackers and defenders. Furthermore, an alternating optimization technique was developed to determine the saddle point of the proposed PPAD framework, thereby providing an effective mechanism for balancing information utility with privacy protection.

In [7], the authors introduced DetectPMFL, a privacy-preserving momentum federated learning framework designed to operate in the presence of unreliable industrial agents. The proposed approach incorporates a detection mechanism for identifying unreliable participants and reducing their negative influence on the federated learning process. Moreover, the privacy risks associated with the learning procedure were mathematically analyzed, with particular consideration given to convolutional neural networks. Based on this analysis, Cheon–Kim–Kim–Song (CKKS) homomorphic encryption was employed to secure the confidential information contributed by participating agents while supporting privacy-aware collaborative learning.

In [8], the authors presented an efficient privacy-preserving subgraph matching approach with integrated authentication for social-network environments. The proposed scheme enables a cloud server to execute subgraph matching queries without gaining access to users' confidential information. In addition to maintaining data privacy, the approach incorporates mechanisms for verifying data integrity and authenticating legitimate users. Consequently, recipients can determine whether received information originates from an authorized sender and verify that the transmitted content has not been altered or manipulated during communication.

In [9], the authors proposed a blockchain-based framework for the secure storage and sharing of students' digital profiles. The approach utilizes important blockchain properties, including decentralization, security, trustworthiness, and resistance to unauthorized modification. Consensus among participating network nodes is established through the Delegated Byzantine Fault Tolerance (DBFT) protocol, whereas controlled sharing of student profile information is supported through an access-control mechanism. Publicly accessible educational information is maintained on the blockchain, while confidential student-profile data is encrypted and stored separately in databases or cloud-based storage systems. This architecture facilitates secure management of individual digital profiles while enabling authorized information exchange among different educational platforms and systems.

In [10], the authors designed and implemented a Privacy-Preserving Automated Stress Monitoring System known as **ARU**. The system supports unobtrusive multimodal information acquisition through multiple devices, including smartphones and affordable wearable sensors. It integrates

different categories of information, such as physiological, behavioral, and social signals, to perform real-time stress assessment and provide useful feedback to individuals experiencing stress. The framework was specifically adapted to Indian academic environments, and an Android-based application was developed to facilitate data collection and generate machine-learning-based insights into users' stress levels. Established assessment instruments, including the Perceived Stress Scale (PSS) and Daily Stress Inventory (DSI), were employed to validate the system's predictions. The study contributes toward conducting real-world, or "in-the-wild," investigations in Indian settings and addresses the limited availability of practical datasets concerning stress-related and mental-health challenges among younger populations.

In [11], the authors addressed distributed learning scenarios in which information is maintained across multiple data sources without requiring the direct exchange of raw records. The study considered horizontally partitioned datasets, where different observations belonging to the same feature space are distributed among separate participating sources. The research particularly focused on classification tasks involving structured healthcare information that can be represented in spreadsheet-based formats. To support secure and scalable collaborative learning, the authors proposed a privacy-preserving distributed machine-learning framework based on the Extremely Randomized Trees (ERT) algorithm. The proposed method introduces linear computational overhead with respect to the number of participating parties and is also capable of processing datasets containing missing values. The framework was termed **k-PPD-ERT (Privacy-Preserving Distributed Extremely Randomized Trees)**, where *k* denotes the number of potentially colluding parties considered within the privacy model.

III. Proposed Methodology

In this section, the proposed SSGPNN (School Scholar Grade Prediction by Neural Network) model is described with the help of the flowchart presented in Fig. 1. Each processing block of the proposed framework is explained in detail, while the different notations used throughout the methodology are summarized in Table 1. In this study, it is assumed that all distributed datasets follow an identical data structure and provide their original data without any modification or manipulation.

Dataset Cleaning: Each dataset available at a different location or server undergoes a series of preprocessing and cleaning operations before the knowledge extraction process. Initially, sensitive and personally identifiable information, such as student name, roll number, contact number, and school name, is removed from the dataset to maintain data privacy [12, 13]. Subsequently, attributes such as city PIN codes are converted into urban or rural categories, while other textual or categorical attributes are transformed into suitable numerical representations. Student marks are also converted into corresponding grade categories. Finally, the relevant features are represented in binary form using 0 and 1, where 1 indicates that a student performs or participates in a particular activity, whereas 0 indicates the absence of that activity.

$$SPD \leftarrow Cleaning\_Distributed\_Data(SRD) \text{ ---- Eq. 1}$$

Generate Genetic Algorithm Population: To train an effective mathematical prediction model, it is important to optimize the input dataset by identifying the most relevant features. In this work, a Genetic Algorithm (GA)-based feature selection technique is employed to optimize the training dataset [14]. Instead of representing candidate feature subsets as wolves, as in wolf-based optimization methods, the Genetic Algorithm represents each possible feature subset as an individual chromosome. Each chromosome consists of a sequence of genes corresponding to the available features in the dataset. A binary representation is adopted, where a gene value of 1 indicates that the corresponding feature is selected, while a value of 0 indicates that the feature is excluded.

An initial population of chromosomes is generated randomly from the complete feature space. Each chromosome therefore represents a different candidate solution for feature selection. The complete collection of these chromosomes forms the initial GA population.

During subsequent optimization, the fitness of each chromosome is evaluated according to its contribution to the prediction performance, and genetic operations such as selection, crossover, and mutation are applied to progressively obtain an optimized feature subset.

$$GP \leftarrow Random\_Population(f, p) \text{ ---- Eq. 2}$$

where GP represents the Genetic Algorithm population, f denotes the total number of available features, and p represents the population size.

Feature Selection

To improve both data privacy and prediction performance, the processed distributed dataset (SPD) is obtained from the respective data owners in a privacy-preserving representation.

The attribute values are represented primarily in binary form as 0 or 1. Such representation minimizes the exposure of sensitive information associated with individual students, schools, and geographical locations.

Different candidate feature subsets are then generated from the available attributes and supplied to the fitness evaluation process using the corresponding distributed datasets.

$$TD \leftarrow Training\_data(GP, SPD) \text{ ---- Eq. 3}$$

where TD represents the training data, GP denotes the Genetic Algorithm population, and SPD represents the cleaned distributed dataset.

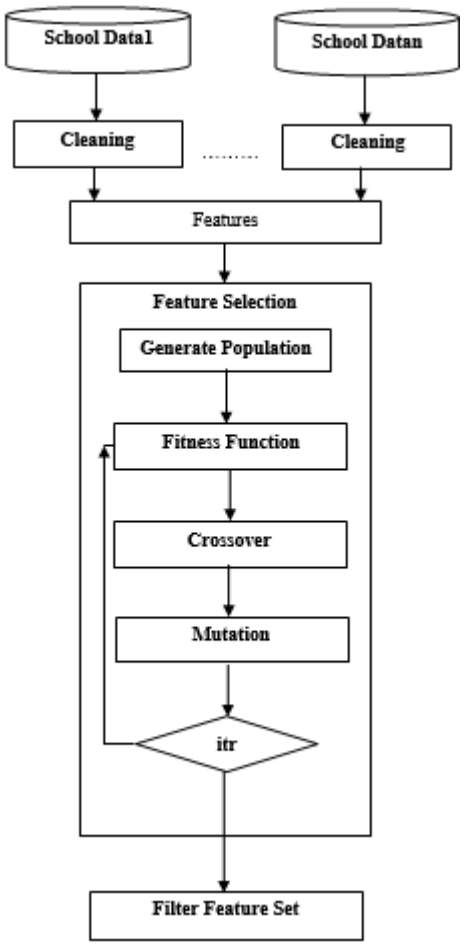

Fig. 1 Block diagram of proposed SSGPNN model.

Fitness Evaluation

The quality of each chromosome in the Genetic Algorithm population is determined through a fitness function. Each chromosome represents a candidate subset of features selected from the available student attributes. The selected features are extracted from the distributed dataset and used to train the neural network. The trained model is subsequently evaluated using the corresponding target grades, and its prediction performance is considered the fitness value of that chromosome.

Chromosomes producing higher prediction accuracy are assigned greater fitness values because their selected features provide more useful information for student-grade prediction. In contrast, chromosomes producing comparatively poor prediction results receive lower fitness values. In this manner, the fitness evaluation enables the Genetic Algorithm to distinguish between useful and less

informative feature combinations without requiring the original sensitive attributes to be directly exposed.

$$H = Fitness(TD, GP) \text{ ---- Eq. 4}$$

where H denotes the fitness values obtained for the chromosomes in the population.

Selection

After fitness evaluation, the chromosomes are ranked according to their respective fitness values. Chromosomes having superior fitness are given a greater probability of being selected for the next stage of optimization. This selection mechanism ensures that useful feature combinations are retained while less effective solutions are gradually removed from the population. An elitism strategy can also be applied to preserve the best-performing chromosome so that the highest-quality solution obtained in the current generation is not lost during subsequent operations [15].

Crossover

Crossover is performed on selected parent chromosomes to generate new candidate feature subsets. In this process, genetic information from two selected chromosomes is exchanged to produce offspring containing characteristics of both parents. Since every gene corresponds to a particular dataset feature, crossover produces new combinations of selected and unselected attributes. This operation enables the Genetic Algorithm to explore different regions of the feature-search space and identify combinations that may provide better prediction performance than their parent chromosomes [15].

Mutation

To maintain diversity within the population and prevent premature convergence toward a local solution, mutation is applied to a small proportion of genes in the offspring chromosomes. During mutation, selected binary values may be changed from 0 to 1 or from 1 to 0. Consequently, previously excluded features can become part of a candidate solution, while selected features can be removed. Mutation therefore provides additional exploration capability and increases the possibility of identifying a more effective feature subset.

Population Update

The newly generated chromosomes are evaluated using the same neural-network-based fitness criterion. Based on their fitness values, suitable offspring are retained in the population, while comparatively weaker chromosomes are replaced. The best-performing solutions are preserved across generations to ensure that the overall quality of the population improves progressively. Through repeated selection, crossover, mutation, and fitness evaluation, the population gradually converges toward an effective combination of student-related attributes.

Optimal Feature Selection

After the optimization process reaches its termination condition, the chromosome with the highest fitness value is considered the optimal solution. The genes having a value of 1 in this chromosome represent the final selected features, whereas genes with a value of 0 correspond to attributes that are excluded from further processing. The resulting optimized feature subset is then supplied as input to the neural network for final training and student-grade prediction. In this way, the Genetic Algorithm reduces unnecessary and redundant attributes while retaining features that contribute most effectively to prediction accuracy.

Training of Neural Network

During training, a neural network will take into consideration both the input training vector and the desired output. Adjust the value of the neuron weight vector for each training set by an amount equal to the number of epochs. The trained neural network was used in a direct manner for the purpose of predicting the session class of student grade.

Proposed SSGPNN Algorithm

Input: SRD

Output: TNN

- SPD‧Cleaning_Distributed_Data(SRD)
- GP ‧Gaussian_population(f, p)
- TD‧Training_data(GP, SPD)
- Loop 1:itr

- H▪Fitness(TD, WP)
- X▪Crossover(H, WP)
- WP▪Mutation(WP, X, H)
- EndLoop
- H▪Fitness(TD, WP)
- NN▪Initialize()
- Loop 1:e // Number of Epoches
- TNN▪Train(NN, T, G)
- EndLoop

IV. EXPERIMENT AND RESULTS

Implementation of this model is done on MATLAB 2016a software. Experiment was done on machine having i3 6th generation processor and 4 GB RAM. Read dataset of student MTS exam [16] from Maharashtra was taken where detail explanation was given in table 2.

Evaluation Parameters In order to compared proposed model with other existing models following set of parameters were evaluated.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F\_Measure = \frac{2 * Prcision * Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Results

Table 1 Precision value based school scholar grade prediction model comparison.

Dataset	K-PPD-ERT [11]	SSGPNN
12%	0.3567	0.7505
25%	0.4969	0.6863
50%	0.4815	0.7082
75%	0.4882	0.6796

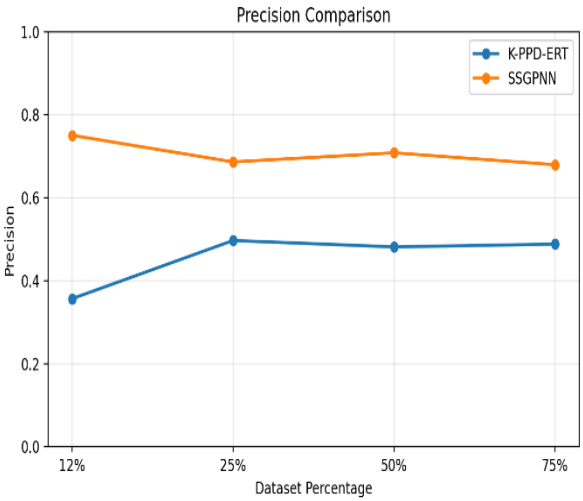

Fig. 1 Precision value based grade prediction comparing models.

Table 1 presents the average precision values obtained by the proposed SSGPNN model and the existing K-PPD-ERT [11] method for datasets of sizes 2, 4, 8, and 12. The results show that SSGPNN consistently achieves higher precision than K-PPD-ERT across all datasets.

Table 2 Recall value based school scholar grade prediction model comparison.

Dataset	K-PPD-ERT	SSGPNN
12%	0.5693	0.7767
25%	0.6824	0.7496
50%	0.6213	0.7491
75%	0.6213	0.6577

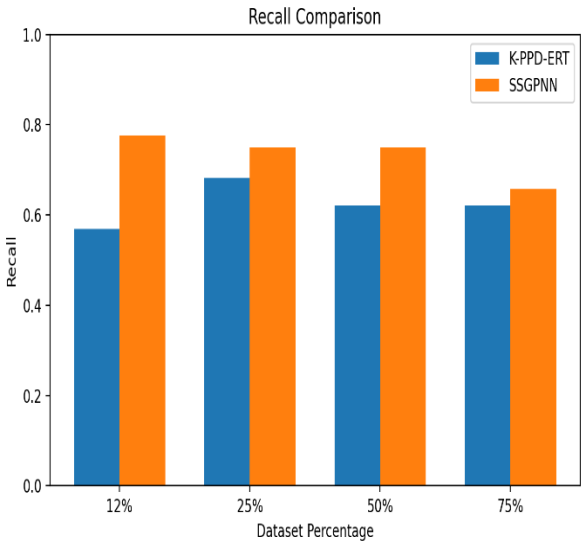

Fig. 2 Average recall value based grade prediction comparing models.

Table 2 recall comparison demonstrates that SSGPNN performs better than K-PPD-ERT [11] for all evaluated datasets. This indicates that SSGPNN has a stronger capability to correctly recognize students belonging to the actual grade categories. Therefore, the proposed approach reduces false-negative predictions and provides more reliable grade classification across different dataset sizes.

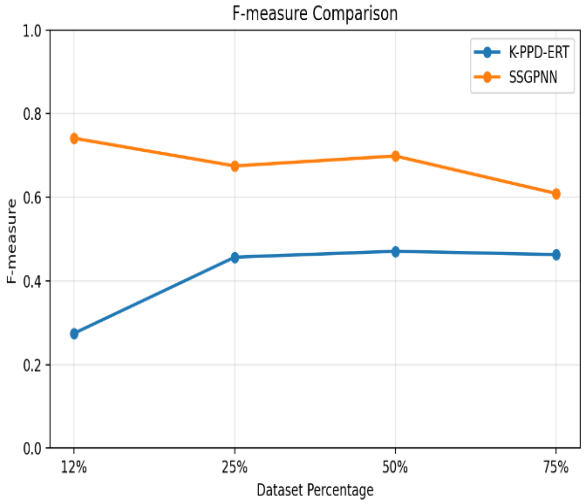

Fig. 3 F-measure value based grade prediction comparing models.

Table 3 F-measure value based school scholar grade prediction model comparison.

Dataset	K-PPD-ERT [11]	SSGPNN
12%	0.2747	0.7414
25%	0.4569	0.6751
50%	0.4711	0.6989
75%	0.4631	0.6093

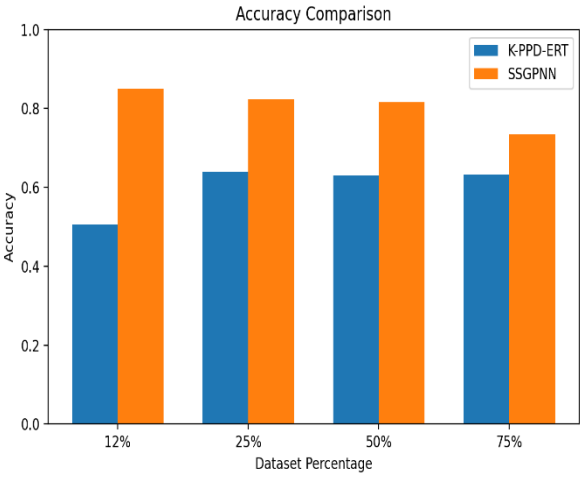

Fig. 4 Average recall value based grade prediction comparing models.

Table 3 and fig. 3 shows that the F-measure results further demonstrate the superiority of the proposed SSGPNN model. Since F-measure provides a combined assessment of precision and recall, a higher value represents a better balance between these two performance measures. For dataset 12%, SSGPNN achieves an F-measure of 0.7414, while K-PPD-ERT obtains only 0.2747. The consistently higher F-measure values indicate that SSGPNN provides a more balanced classification performance and maintains better prediction reliability across varying dataset sizes.

Table 4 Accuracy value based school scholar grade prediction model comparison.

Dataset	K-PPD-ERT [11]	SSGPNN
12%	0.5059	0.8501
25%	0.6404	0.8231
50%	0.6296	0.8160
75%	0.6323	0.7339

Table 4 and fig. 4 shows that the accuracy results also confirm the improved performance of the proposed SSGPNN model. Although the accuracy of SSGPNN decreases slightly as the dataset size increases, it remains higher than the existing method for every evaluated dataset. Overall, the results demonstrate that SSGPNN provides more accurate and dependable student grade prediction than K-PPD-ERT.

V. CONCLUSION

This paper proposed the **SSGPNN** model for privacy-aware and accurate prediction of student academic grades from distributed educational datasets. The framework protects confidential student information through data cleaning and transformation before performing knowledge extraction. A **Genetic Algorithm-based feature selection approach** identifies relevant academic and associated attributes while reducing unnecessary features from the prediction process. The optimized feature subset is then supplied to a neural network for effective grade classification. Experimental evaluation was conducted using 12%, 25%, 50%, and 75% dataset proportions, considering precision, recall, F-measure, and accuracy as performance measures. The comparative results demonstrate that SSGPNN consistently performs better than the existing K-PPD-ERT method across the evaluated dataset proportions. The findings confirm that combining privacy-aware preprocessing, genetic feature optimization, and neural-network classification provides an effective solution for student grade prediction. Future work can investigate larger educational datasets and advanced learning architectures to further improve prediction reliability.

REFERENCES

1. X. Meng, X. Ci. Big data management: Concepts, techniques and challenges [J]. Journal of Computer Research & Development, 2013.
2. Colin Tankard. Big data security [J]. Network Security, 2012(7): 5---8.
3. Elisa Bertino. Big data security and privacy [C]// 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016.
4. Agrawal R, Srikant R. Privacy-preserving data mining [C]// Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM, 2000.
5. Yin Y, Kaku I, Tang J. Privacy-preserving Data Mining [J]. Acm Sigmod Record, 2000, 29(2):439-450.
6. N. Wu, C. Peng and K. Niu, "A Privacy-Preserving Game Model for Local Differential Privacy by Using Information-Theoretic Approach," in IEEE Access, vol. 8, pp. 216741-216751, 2020
7. Z. Zhang, N. He, Q. Li, K. Wang, H. Gao and T. Gao, "DetectPMFL: Privacy-Preserving Momentum Federated Learning Considering Unreliable Industrial Agents," in IEEE Transactions on Industrial Informatics, vol. 18, no. 11, pp. 7696-7706, Nov. 2022.
8. X. Zuo, L. Li, H. Peng, S. Luo and Y. Yang, "Privacy-Preserving Subgraph Matching Scheme With Authentication in Social Networks," in IEEE Transactions on Cloud Computing, vol. 10, no. 3, pp. 2038-2049, 1 July-Sept. 2022
9. M. Jin, C. Cui, G. Yu, X. Li, Y. Zhang and L. Zeng, "A Blockchain-Based Scheme for Secure Storage and Sharing of Student Digital Profiles," 2020 3rd International Conference on Smart BlockChain (SmartBlock), 2020, pp. 209-214.
10. P. K. R. Yannam, V. Venkatesh and M. Gupta, "Research Study and System Design for Evaluating Student Stress in Indian Academic Setting," 2022 14th International Conference on

COMmunication Systems & NETworkS (COMSNETS), 2022, pp. 54-59.
11. AAminifar, M. Shokri, F. Rabbi, V. K. I. Pun and Y. Lamo, "Extremely Randomized Trees With Privacy Preservation for Distributed Structured Health Data," in IEEE Access, vol. 10, pp. 6010-6027, 2022.
12. Nikunj Domadiya, Udai Pratap Rao. "Privacy Preserving Distributed Association Rule Mining Approach on Vertically Partitioned Healthcare Data". Procedia Computer Science Volume 148, 2019.
13. C C Aggarwal, P S Yu, "On static and dynamic methods for condensation-based privacy-preserving data mining," ACM Trans Database Syst, vol. 33, no. 1, 2008
14. S.Mirjalili, S.M.Mirjalili, A.Lewis, "Grey wolf optimizer", Advance in Engineering Software, vol. 69, pp.46–61, 2014.
15. Abdo, Mostafa & Kamel, Salah & Ebeed, Mohamed & Yu, Juan & Jurado, Francisco. "Solving Non-Smooth Optimal Power Flow Problems Using a Developed Grey Wolf Optimizer". Energies 2018.
16. Munde, Vyankat & Kumar, Dr Binod & Shirwaika, Shailaja. (2021). Single Scan Counter Based Association Rule Extraction For Student Learning Analytics. International Journal Of Advanced Research In Engineering & Technology. 11. 2609-2622.