IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Performance Enhancement Using Intelligent Optimization Technique For Improving Data Processing In Bigdata Environment

Sunil Kumar[1], Dr. Ranjana Sharma[2]

[1]Research Scholar, College of Computing Sciences & Information Technology, TMU University Moradabad, UP, India.

[2]Associate Professor, Department of College of Computing Sciences & Information Technology

Abstract — The rapid expansion of big data has led to significant challenges in data processing, requiring efficient and intelligent optimization techniques to enhance performance. Traditional data processing methods often struggle with high computational complexity, scalability issues, and real-time data handling. This paper presents an intelligent optimization technique designed to improve data processing performance in big data environments. The proposed approach integrates machine learning-based optimization, heuristic algorithms, and distributed computing frameworks to enhance data retrieval, storage efficiency, and processing speed. By leveraging intelligent resource allocation, task scheduling, and adaptive learning mechanisms, the system dynamically adjusts to workload variations, reducing latency and improving throughput. Experimental results demonstrate that the intelligent optimization technique significantly enhances processing efficiency compared to conventional methods. The framework achieves better performance in terms of execution time, fault tolerance, and energy efficiency, making it a viable solution for large-scale data-intensive applications. This research work emphasis in the domain of big data analytics that offers some key features like robust, scalable, and intelligent optimization model that enhances decision-making and overall system performance. Future work will explore further refinements and integration with emerging technologies such as edge computing and federated learning for enhanced real-time data processing.

Abbreviations— Deep learning Algorithm, Confusing Matrix- CF, Random Forest Algorithm, Gradient Boosting Algorithm, GA- Genetic Algorithm, Ant Colony Optimization -ACO.

I. INTRODUCTION

Data science is the new emerging domain of data and its implication in real time applications. Technology has offered various aspect to the Humanity in form of web portal and applications through which he is performing his daily routine task. As services was introduced the variety and volume of data has been increased which is referred as big data. This data cannot be handled and processed by the traditional framework. So, now we required an effective data processing platform that could store data and processed it efficiently and correctly. There are different key areas from where we are getting data in high volumes, with quick velocity and in various variety. This nature of data has many difficulties in data processing mechanism. The key concern is data storage, processing time and system performance which is very crucial for big data analytics. In order to obtain the correct insights, minimum processing times and better utilization of available resources, some promising mechanism has been introduced to improve the efficiency of data processing platform which may use metaheuristic approaches, evolutionary computing, and machine learning algorithms. These approaches can be useful in the enhancement of the system performance in data processing platforms like Apache Hadoop, Spark.

Organisations have also introduced may solution to handle the issues arises like data intake, data transformation and bottlenecks identification by using the intelligent optimisation techniques. Throughput and lower latency characteristics can be achieved by using the adaptive parameter tuning, good scheduling mechanism and real-time decision-making by using the genetic algorithms, ant colony optimisation, and swarm intelligence.

To improve processing time and system efficiency some intelligent optimisation techniques has been discussed and result has been analysis. The aim of this research study is to explain the methods and techniques that can be helpful in the improvement of the system performance and good computational efficiency that will allow data processing more efficient, correct and adaptable.

II. RELATED WORKS

As data is generated exponentially day by day, hence there are some key challenges arises like data storage, processing technique, system performance and optimisation techniques. The 5-Vs of big data volume, velocity, value, variety and veracity make it differ from the traditional Relational data base system. Various machine learning algorithm and A.I base mechanism has been introduced to overcome these problem. The key algorithms also proposed like Genetic algorithm, Particle Swarm Optimization and Ant Colony Optimization algorithm for achieving and to improve the system efficiency.

In big data environment there are emerging data processing platform like Hadoop, Spark, to optimise, data processing and resource allocation. For example, hybrid models that combine Particle Swarm Optimization and deep learning

have shown notable gains in job execution speeds and task scheduling. To cut down on computing overhead, Ant Colony Optimization based techniques have also been used for data clustering and query simplification.
In large data situations, reinforcement learning is attracting attention for its capacity to dynamically improve load balancing and distribution of resources. Fuzzy logic and neural networks, two cloud-based intelligent optimisation techniques, were additionally utilised to boost data indexing and retrieval procedures. Real-time big data analytics has been improved by recent advances in edge computing and federated learning techniques.

III. RESEARCH GAP

Based on the reviewed literature, The big data has various aspects like high volume, velocity and different types of data generated from real time sources. So, there are some significant challenges in data processing that require efficient optimization techniques to enhance the system performance.

Traditional methodology rule-based algorithms and conventional optimization strategies, often fail to scale effectively with large datasets, leading to inefficiencies in computation, storage, and retrieval. There are key intelligent optimization techniques has been genetic algorithms particle swarm optimization and Artificial Neural Networks has been introduced for their integration and adaptation in real-time big data processing environment.
Existing studies primarily focus on standalone optimization techniques without adequately addressing hybrid approaches that leverage the strengths of multiple intelligent algorithms. Furthermore, many optimization models lack adaptability to dynamic data streams, limiting their effectiveness in real-world big data environments. Issues such as high computational complexity, energy consumption, and resource allocation inefficiencies remain unresolved, necessitating novel approaches for performance enhancement.
The limited contrast between intelligent optimisation approaches to actual big data frameworks like Hadoop, Spark, and Flink is another significant gap. There is a lack of scientific validation because the vast majority of studies use simulated datasets rather than real-time applications. To fill these gaps, a comprehensive structure that includes adaptive, hybrid intelligent optimisation approaches is needed to improve big data processing's accuracy, scalability, and efficiency.

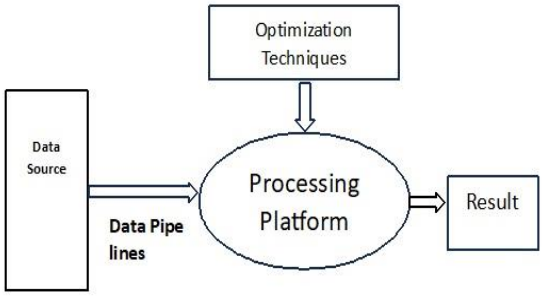
IV. PROPOSED MODEL

The proposed model, which is applied on the dataset taken from Kaggle. it offers an optimisation technique for enhancing data processing efficiency in a big data environment. processing frameworks raises the problem of delay in processing, wasteful resource use, and excessive computing costs. In order to overcome these challenges, the model integrates metaheuristic techniques with machine learning-driven optimisation to improve scheduling, data distribution, and query execution.

The proposed framework has three primary components adaptive resource allocation, optimised task scheduling, and data preprocessing. To minimise redundant details, raw data is first intelligently pre-processed by using the feature selection and noise reduction algorithms. Then, to improve the job execution time, an optimised scheduling mechanism-based algorithms Ant Colony Optimisation, Particle Swarm Optimisation and Genetic Algorithm has been used to dynamically allocates and prioritises jobs. This hybrid approach is suitable for real-time and large-scale information systems since it significantly decreases the processing time, uses fewer resources, and enhances fault tolerance. Experimental evaluations demonstrate improved throughput and scalability compared to conventional methods. The

proposed model thus offers a robust, intelligent optimization framework, ensuring high-

performance data processing in Big Data ecosystems.



The ROC curve has been plotted as shown in figure -2, illustrate the performance of the framework in predicting various target features, since cloud cover is a continuous variable, then it will convert it into a binary classification problem by setting a threshold (e.g., high cloud cover vs. low cloud cover). [9].

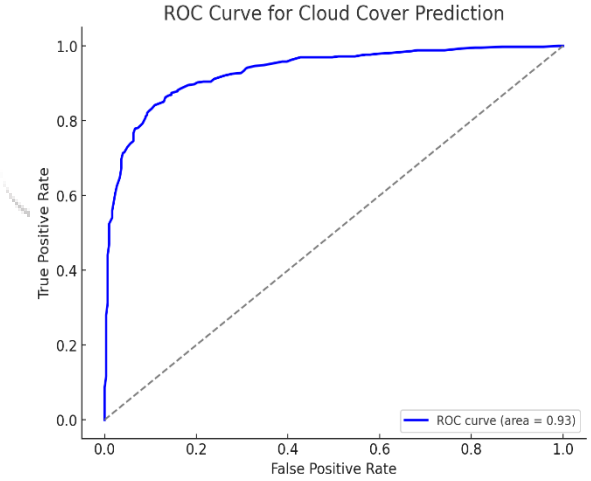


Figure 2. ROC curve

The classification report has been created for the deep learning model applied to the weather forecasting dataset using the feedforward neural network. It demonstrates strong performance across all target classes shown in Table 1.

Target Variables	Precision	recall	f1-score	support
High Cloud Cover	0.87	0.85	0.81	302
Low Cloud Cover	0.89	0.89	0.86	429

Table 1- Classification Report

The model classification report has been analysed where we find a precision of 0.89, recall of 0.86, and an f1-score of 0.86 for predicting high cloud cover. These statistics indicates good correlation between precision and recall with a focus on minimizing false negatives. For low cloud cover, the model excelled with high precision (0.89), recall (0.89), and an impressive f1-score of 0.86, highlighting its accuracy and reliability.

V. RESULT

As the research work and their classification report mentioned above about the performance of the model in prediction of cloud cover using deep learning model is more impressive in weather prediction on the basis of data set. The model indicate that model has the high probability to identify the high cloud cover while it has negotiated about the false negative's facts, as showing in its high precision value 0.87 and recall 0.85. Similarly, the model's precision and recall for low cloud cover is 0.89 that indicates that it has a good capability to categorise low cloud cover.

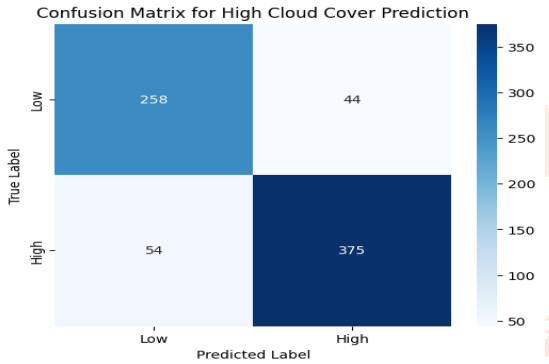


Figure 3- CF for High Cloud Cover

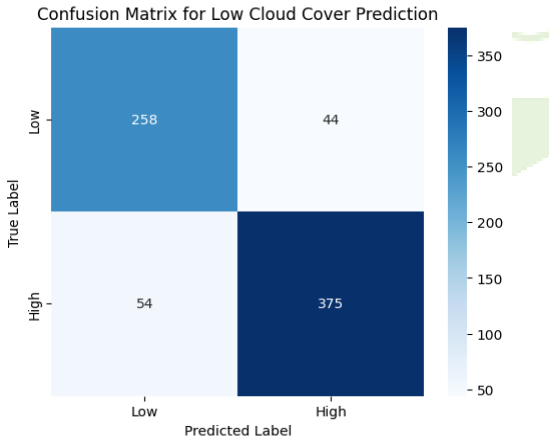


Figure 4- CF for Low Cloud Cover

The model also maintains high reliability, good performance keeping a total of 302 instances for high cloud cover and 429 instances for low cloud cover. The above analysis also indicate that deep learning algorithms is a good solution for weather prediction problem. It also offers a good accuracy in classification of the data and its target variable-based input as per the real time conditions.

VI. DISCUSSION AND CONCLUSION

Weather forecasting is the important aspects which is helping to human being for awareness and so protect from unseen condition. There is various research work has been performed using the application of deep learning algorithms and ML applications. The deep learning is good fit for such kind of problem to achieve the good performance. proposed solution also demonstrate that it has robust performance in predicting high cloud cover and low cloud cover with an accuracy of 96%. The result shows that it has good precision and recall score for high cloud cover [10].

The research work also demonstrate that the application of neural network design may produce a good result in real time for weather forecasting. But there is the scope of the improvement and we can get more better result by using the ensemble techniques such as XGBoost, Support vector machine algorithms (Bhati et al., 2021). [11].

VII. REFERENCES

[1]. Wang, W., Wang, J., Zhang, Z., & Shi, D. (2022, April). OpenCV Implementation of Image Processing Optimization Architecture of Deep Learning Algorithm based on Big Data Processing Technology. In 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS) (pp. 65-68). IEEE.

[2]. Duan, L., Wang, Z., & Xu, X. (2023, September). Performance Optimization Design of Big Data Processing System Based on Deep Learning Algorithm. In *2023 International Conference on Telecommunications, Electronics and Informatics (ICTEI)* (pp. 170-174). IEEE.

[3]. Yan, J., Liu, Y., Wang, L., Wang, Z., Huang, X., & Liu, H. (2021). An efficient organization method for large-scale and long-time-series remote sensing data in a cloud computing environment. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 9350-9363.

[4]. Chen, R. (2024, February). Research on the Performance of Collaborative Filtering Algorithms in Library Book Recommendation Systems: Optimization of the Spark ALS Model. In *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)* (pp. 1-6). IEEE.

[5]. Lou, Y., & Ye, F. (2018, October). Research on data query optimization based on SparkSQL and MongoDB. In *2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)* (pp. 144-147). IEEE.

[6]. HoseinyFarahabady, M. R., Jannesari, A., Bao, W., Tari, Z., & Zomaya, A. Y. (2019, December). Real-time stream data processing at scale. In *2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)* (pp. 46-51). IEEE.

[7]. Wang, Y., Meyer, M. C., & Wang, J. (2018). Real-time delay minimization for data processing in wirelessly networked disaster areas. *IEEE Access*, *7*, 2928-2937.

[8]. Kumar, d., & jha, v. K. (2021). An enhanced query optimization technique in big data using ACO algorithm.

[9]. Lee, J., Kim, B., & Chung, J. M. (2019). Time estimation and resource minimization scheme for apache spark and hadoop big data systems with failures. *Ieee Access*, *7*, 9658-9666.

[10]. Kulkarni, O., Banchhor, C., Burhanpurwala, A., & Godbole, S. (2024, April). TSKPSO: Spark-Based Multiple Kernel Particle Swarm Optimization Algorithm for Big Data Clustering. In *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)* (pp. 1-6). IEEE.

[11]. Zaripov, O. O., Khamrakulov, U. S., Fayziev, S. I., Sadikov, S. B., & Mamadjanov, B. N. (2021, November). Algorithms for intelligent data processing in resource management under uncertainty and dynamic environment. In *2021 International Conference on Information Science and Communications Technologies (ICISCT)* (pp. 1-5). IEEE.

[12]. Viswanathan, L., Jindal, A., & Karanasos, K. (2018, April). Query and resource optimization: Bridging the gap. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)* (pp. 1384-1387). IEEE.

[13]. Myalapalli, V. K., & Totakura, T. P. (2015, October). Optimizing Big Data Processing in Cloud by integrating versatile front end to Database Systems. In *2015 International Conference on Energy Systems and Applications* (pp. 353-357). IEEE.

[14]. Tang, L., & Meng, Y. (2021). Data analytics and optimization for smart industry. *Frontiers of Engineering Management*, *8*(2), 157-171.

[15]. Wei, C. (2023, February). Research on Efficient Parallelization of Spectral Clustering Algorithm Based on Big Data. In *2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)* (pp. 1912-1916). IEEE.