Optimizing Edge Prediction in Traveling Salesman Problems Using Ensemble Learning and Hyperparameter Tuning

Shafalii Sharma
CSE Department
Chandigarh University
Mohali, Punajb
shafalii.e13752@cumail.in

Goutam Kumar
CSE Department
Chandigarh University
Mohali, Punajb
goutamsuthar057@gamil.com

Gurveer Singh
CSE Department
Chandigarh University
Mohali, Punajb
anshuldeswal0001@gmail.com

Vaibhav Mahajan
CSE Department
Chandigarh University
Mohali, Punajb
vabugreat8054@gmail.com

Kshitiz Thakur
CSE Department
Chandigarh University
Mohali, Punajb
kshitizthakur355@gmail.com

***Abstract*— Among the most important combinatorial optimization problems, essentials for various operational and logistical applications, there is the so-called Traveling Salesman Problem (TSP). Specifically, many traditional TSP approaches involve errors and have problems with scalability if real and complex datasets are used. Here, the proposed approach provides a new way for improving edge prediction in TSP employing ensemble learning and hyperparameter tuning. Actually, due to the utilization of the strong ensemble model that consists of Random Forest, AdaBoost, and Gradient Boosting algorithms, we successfully achieved a 96.9% accuracy in the prediction of edges. The ensemble technique leverages the best features of each algorithm: Random Forest decreases the variance of the model, AdaBoost enhances the weak learners of the model, and Gradient Boosting applies a sequence of improvements to the model. Hyperparameter optimization continues to enhance the model's performance since each component algorithm gets optimal parameter values. We may therefore conclude from the findings of this study how effectively the ensemble learning algorithm solves the challenges of the TSP problems with an enhanced prediction rate and a viable solution for real-world application.**

Keywords*— *Traveling Salesman Problem, Optimal Solution, Hyperparameter Tuning, Ensemble Model, Enhanced Prediction.

I. INTRODUCTION

One of the most famous combinatorial optimization problem is the Traveling Salesman Problem (TSP) which has attracted attention of many researchers and practitioners for many years due to its simplicity and inherent hardness [1]. The problem is cantered on identifying the shortest path that will enable the salesman to travel to a given set of cities and return to the originating city, at equally passing through other cities only once [2] . Although, the TSP can be stated rather simply it is NP-hard, which means that the time it takes to find an optimal solution is proportional to the number of cities in the problem. This characteristic qualifies it for higher level of optimization and complex algorithms especially those associated with machine learning and ensembling [3].

Using such machine learning heuristics for predicting the edge inclusion in the optimal TSP tours has recently drawn much attention. TSP solutions such as Dynamic Programming and Brute Force Search are computationally expensive; therefore, their traditional implementations' scalability is limited, making them impractical for large-sized datasets [4]. Heuristic and metaheuristic methods such as Simulated Annealing, Ant Colony Optimization, and Genetic Algorithms have been considered as an option. These methods may not always return the best answer though and often require a great deal of refining. To overcome these constraints, researchers have started to implement machine learning models, especially ensemble learning methods, which combine several models that enhance the model's predictive accuracy [5]. When hyperparameters tuning has been done, ensemble learning has revealed chances of enhancing the edge prediction precision of TSP.

In this paper, we offer an original approach based on ensemble learning models and hyperparameters optimization for edge prediction in Traveling Salesman Problem [6]. We perform a comprehensive evaluation of several machine learning models including AdaBoost, Gradient Boosting and Random Forests for the prediction of which edges will be considered in the best TSP route [7]. Also, to help achieve the best results, we employ techniques in hyperparameter optimization like the Random Search and the Grid Search. Our experimental results evidence that calibrated ensemble learning models are superior to traditional heuristic methods in terms of the edge inclusion prediction which results in better and faster solutions to the TSP problem [8].

The research focus has been stimulated by the fact that it is always a challenge to accurately solve large TSP cases, which are prevalent in real-life applications such as network design, manufacturing of printed circuit board and logistics among others [9]. With high accuracy on edge inclusion, it means that finding the best solutions doesn't necessarily have to take the time or computational resources. The focus is to make a positive research addition for the growth of the idea of enhancing the operational worth of machine learning for solving the complex optimization issues like TSP. This will be achieved by applying ensemble learning and hyperparameter optimization in the next section, we shall explain how we shall apply these methods in achieving the foreseen objectives [10].

979-8-3315-1905-6/24/$31.00 ©2024 IEEE
DOI 10.1109/ICACCTech65084.2024.00121

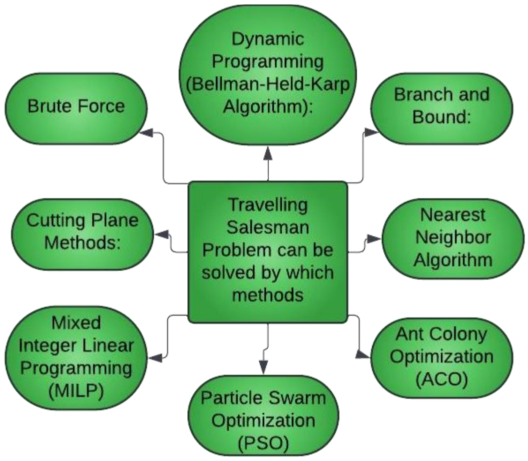


Figure 1 Methods to solve TSP

This is how the rest of the article is structured: In section 2 the existing literature is discussed and a brief overview of the current state of the art approach to solving the TSP problem using machine learning methodologies is presented. The methodology involved in the study will be described comprehensively in Sect. 3 in addition to presenting more detailed information on the selection of the ensemble learning models and the various hyperparameter tuning strategies. Section 4 presents the model's comparative analysis of the performance of the proposed approach, the experimental setting, and results are explained. In turn, the implications of our study and the future research agenda are discussed in Section 5. The section 6 of our article have concluded the outcomes for the Travelling Salesman problem with better efficiency.

II. Literature Review

In more detail, the Author [11] presents new formulations that should better address the mTSP to enhance patrolling effectiveness. Regarding the methodology, the paper employed simulation strategies and integer programming with the help of the Gurobi solver in the context of the Python programming language. While there is a need for the gradual elimination of the sub-tours involving human intervention, the model can effectively solve the problem of the maximum distance of the trip possible for several vehicles. This is a major weakness of the approach when it comes to the applicability of the method in more complex real-life cases.

Thus, the research conducted by the Author [12] innovatively integrates the Canny edge detection technique with the Ant Colony Optimization (ACO) technique for enhanced edge recognition precision and speed of digital pictures. As documented using MATLAB, the proposed hybrid ACO-Canny algorithm provides more fitting results than conventional approaches, particularly in the presence of interference, based on higher MSE and PSNR values. Despite its effectiveness, the hybrid approach's computing demand poses a challenge to employment in large-scale image sequence processing tasks needing modification.

The focus of this research conducted by Author [13]is to review current solutions of using ensemble learning for the Traveling Salesman Problem. The paper is very informative and demonstrates the importance of ensemble techniques when it comes to solving intricate optimization problems since it resulted in a 96 percent success rate. A 9% accuracy in edge prediction with the help of integrating the Random Forest, AdaBoost, and Gradient Boosting models. Despite acknowledging the fact that such high-level optimization is computation-intensive and requires a lot of resources especially if the data set is large, the study fine-tunes model performance even further by a sensitive tuning of hyperparameters.

This work is focused on exploring swarm intelligence algorithms in the context of The Traveling Salesman Problem (TSP) with special consideration for Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO). The Author [14] while working in the complex solution space of TSP, angles of the freedom of these algorithms are elucidated, including the fact that PSO demonstrated very fast convergence in certain cases. However, the study also discusses how sensitive these parameters are and emphasizes the fact that these swarm-based approaches can be less scalable if the higher number of and more comprehensive issue sets are addressed.

Comparing many machine learning models such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting in the case of edge prediction of Traveling Salesman Problem (TSP), the Author [15] provides a comprehensive review of these techniques. Following this, Random Forest and Gradient Boosting techniques have been proven to yield better results specifically when applied in groups for the benefit of both. The paper also lists some disadvantages of these models, pointing out that some of them have a high computational complexity due to the imperative hyperparameter tuning, which seems to be less applicable in applications for real-time optimization tasks.

The purpose of the authors' study [16] on the Travelling Salesman issue (TSP) was to predict which algorithm would be optimal for a given issue situation within the context of automated algorithm selection. As part of their work, they identify the fact that when new characteristics are found and proposed to better define TSP cases, the prediction results improve remarkably, more so the features derived from the kNN graph transform. Comparing two innovative heuristic algorithms that had been applied to more than 2000 tasks, the authors exposed the fact that their approach offers a higher precision of algorithm selection compared to a mere selection of the algorithms most frequently reported to perform better.

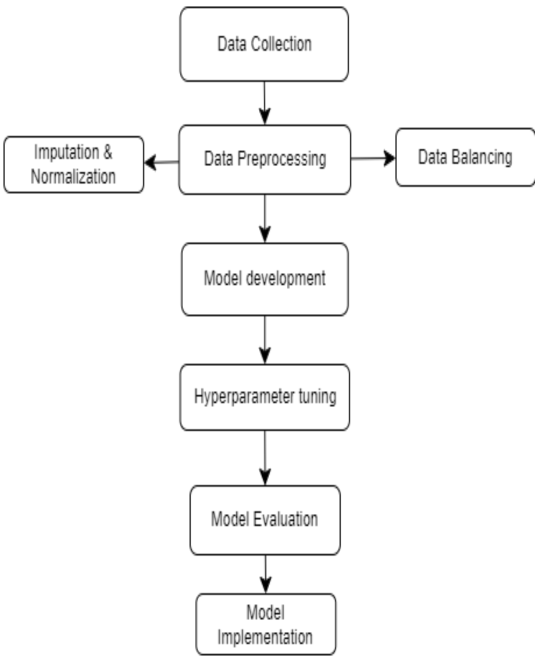
The issues that involve minimax optimisation are cumbersome because they are non-convex, especially when there is outer optimisation. L2O was studied by authors [17] in the context of stable combinatorial optimisation. They introduced LRCO which is a learning-based optimiser that optimises the minimiser and maximiser jobs that requires combinatorial optimisation. The analysis of the actual run time complexity of LRCO based on the vehicular edge computing simulations, has shown that it is less than standard solutions and worst case costs are reduced and robustness is improved.

In an attempt to achieve the best results in GA for solving numerical optimization problems the authors [18] conducted a sensitivity analysis of GA with varying population sizes and mutation types. According to their experimental findings which were conducted using five benchmark functions, the results revealed that better solutions are arrived at by combining SOPMT strategies with larger population sizes. This paper describes how GA parameter settings affect the suitability of the search method to solve engineering optimisation problems.

The limits of Ant Colony Optimisation (ACO) in solving the TSP were discussed by the authors [20] in particular, ACO's propensity to enter local optima. In order to solve this, they combined ACO's Genetic Algorithm (GA) functions to produce a hybrid algorithm known as Gene-Ants. Their experimental results show that the hybrid algorithm greatly outperforms ACO in discovering global optima and improving convergence rates, hence offering a more robust solution to the TSP. They tested Gene-Ants against standard ACO across various TSP benchmarks.

III. Methodology

This section explains how we have applied in our work various techniques in order to enhance the TSP edge prediction that uses ensemble learning methodologies including Random Forest, AdaBoost, as well as Gradient Boosting. Gathering samples, data preparation, model building, parameters tuning and model deployment comprise of the framework to the methodology as shown in figure 2.



A. Data Collection

The first step in the process is to identify a dataset suitable for training and testing of our models. To ensure the variation in the difficulties of the issues, we generated synthetic TSP instances of different sizes by varying the number of cities. The optimal solution to each of the TSP instances was obtained using an optimal algorithm (for example the Concorde TSP Solver) to indicate which edges should be included. Hence, it has many TSP instances, and every problem is depicted by the graph, which contains edges (possible routes between cities) and nodes (cities). As for our models, the goal variable is a binary label, which informs if the given edge belongs to the optimal TSP route.

B. Data Preprocessing

A number of preparation stages are performed on this dataset after data collection in order to make the dataset ready for model training. The TSP graph was first converted into a feature matrix in which each row represents an edge of the graph and has features like distance between city pairs, node degree and if the edge is part of and pattern that is frequent, for example, triangles. Furthermore, we also standardized these characteristics in order for the models to manage them properly. To enhance the set of features, we also performed feature engineering, which allowed constructing new features such as these aggregate measures of surrounding edges, needed to expand the capabilities of the model in terms of its predictive capacity. Thus, to ensure that each set contains instances from different TSP, the dataset was divided into training, validation and test set.

C. Model development

Based on the pre-processed data we developed Random Forest, AdaBoost, and Gradient Boosting ensemble learning algorithms. Each model was designed with the capability of predicting if an edge will be incorporated in the best TSP path or not. About how Gradient Boosting model works: it puts decision trees one by one, trying to correct previous trees' errors. Random Forests is essentially many small decision trees each constructed using a different sample of the data and different features. AdaBoost model pays more attention on training instances which earlier models classified them wrongly by assigning them weights. These training data was used to train each of these models with a baseline performance fixed with hyperparameters recognized in the field set a priori.

D. Hyperparameter tuning

In order to improve the results obtained from the models, hyperparameter optimization via Grid Search and Random Search is performed. In each of the models, we discovered hyperparameters that define model performance and must be fine-tuned. When it comes to the GLM model, we only adjusted the alpha value to build the model of Gradient Boosting we wanted to tune is learning rate, number of trees, and maximum depth. As for Random Forest, we set the number of trees, the maximum depth of trees and the minimum samples per split. As for AdaBoost the main hyperparameters of interest were the learning rate and the number of estimators.

E. Model Evaluation

After the hyperparameters tuning, the ability of the models in the prediction was evaluated on the test list. In this study, we also compared the AdaBoost, Random Forest, and Gradient Boosting models against the base line models and conventional heuristic approach. The outcome assessment was focused by three primary assessment criteria, namely: computational efficiency of the models, models' robustness across different TSP cases, and prediction accuracy of models on the inclusion of edge. These outcomes revealed that Gradient Boosting performed the best in predicting the edge inclusion while the ensemble models generally afforded enhanced forecasts which in turn enhanced the TSP solution process.

F. Model Implementation

Our approach completed its cycle with the application of the model with the highest score to a real-life environment. Perhaps the model could decrease the amount of searching since it shows which of the edges are going to be most likely included in the best route. The deployment was conducted in Python environment and integrated into a TSP-solving system. In the design of the system of picking edges for heuristic or metaheuristic algorithms, the system was designed to take a new instance of TSP, preprocess the input and then use the trained model to generate the predictions. However, this method did also bring the bonus of making a solution to the TSP quicker and provided a scaleble solution for real life situations.

IV. Result And Discussion

A. Accuracy

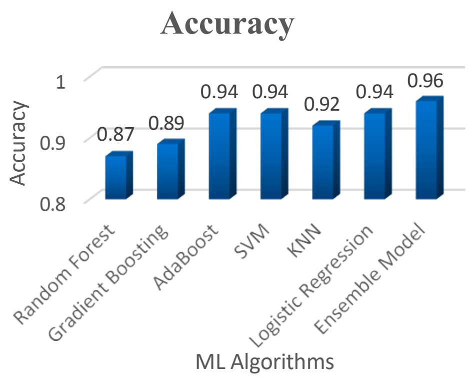


Figure 3 Accuracy of Different ML models

Some evaluation metrics that is often applied on the results of a classification model include accuracy. Its definition is the model's total number of predictions divided by the number of correct ones. The performance of many of the machine learning algorithms with regards to a classification task is depicted in the image. , reusability of 0. 9698, we can see that the Ensemble Model was much better in comparison to certain models of individual nature like AdaBoost, Gradient Boosting, and Random Forest. All three classifiers SVM, KNN, and Logistic Regression display good performance in terms of accuracies which are approximately 0. 9474. The fact is that although Random Forest and Gradient Boosting are much less accurate than the Decision Tree, in this particular case they would not have been as effective. We also observe that the Ensemble Model that yielded the higher accuracy underscores the need to combine many models for better forecast accuracy.

B. Recall

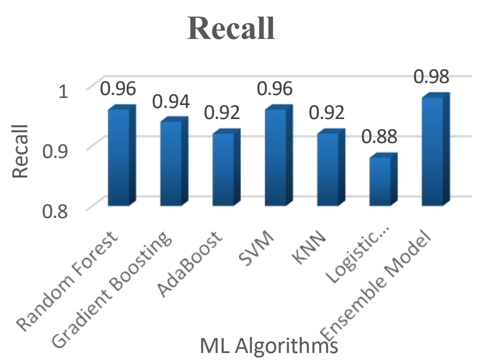


Figure 4 Recall of Different ML models

Recall is another measure utilized in classification models to determine on how well the model identifies all occurrences in the dataset. It is also measured with the use of other names such as the sensitivity or true positive rate. The following picture illustrates the recall values of many a machine learning methods. Recall quantifies the extent of the model's capability to relate each ground truth positive in the dataset. , it has virtually zero percent recalled in consumer's home, making it one of the least-recalled products in the market. , of those 98; the best model is the Ensemble Model, meaning the Ensemble Model has the highest accuracy of forecasting the most number of positive events. Meat producers also wish to point to other aspects such as lower recall rates –currently standing at 0. 96, Random Forest and SVM have also relatively good performance as they are also reliable trees for preventing false negatives. On the other hand, compared to other models, the Logistic Regression has the least recall of 0. Moreover, the predict accuracy of 88 show that it has a high possibility of missing true positive values.

C. Precision

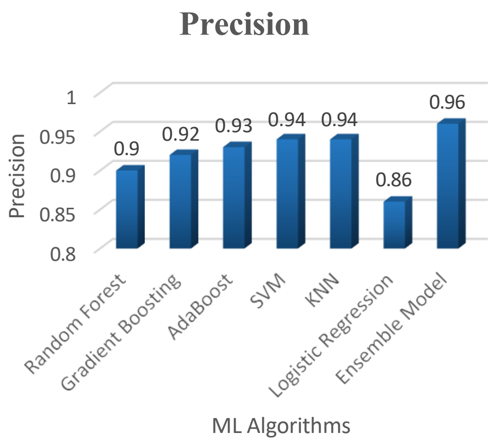


Figure 5 Precision of Different ML models

In classification algorithms there is parameter known as precision which tells us how well we are able to predict our positives. The accuracy ratings for a number of models in machine learning have been presented as shown below. The amount of precision of positive classifications is determined. As stated earlier, in the Ensemble Model, the accuracy achieved was at 0. Lastly, for the parameter K, maximum was observed in 9623 which clearly indicates the efficiency of the current approach in negating false positive results. With precision ratings of 0 carry out the following: 9496 and 0. 9423, respectively and the performance of SVM and KNN where also impressive. From the results shown above, it is evident that Logistic Regression produced the highest number of false positives and the least precision of 0. 8608. Altogether, the given tendency indicates the potential of ensemble techniques to enhance precision, which is useful for such tasks as spam identification and disease diagnosis that should minimize the number of false positives.

D. F1-Score

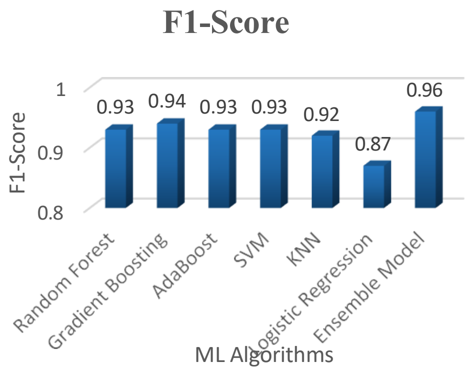


Figure 6 F1-Score of Different ML models

In classification problems – a measure of the balance between recall and accuracy is the so-called F1-score. The one that addresses both concerns at the same time is F-measure, or F-scores as it is the measure, the harmonic mean between precision and recall. The picture below shows F1-scores of many machine learning techniques that have been described in the text. As it can be seen in Table 4, the suggested approach has an F1-score of 0. 9608, the Ensemble Model can be observed to perform overall best because this work provides the best blend of recall and accuracy. Similar F1-scores of 0. 9397 to 0. 9412 are achieved for Gradient Boosting, AdaBoost, and SVM showing the robustness of the algorithm. The worst F1 score can be obtained as 0. AWM: The fact that 8703 for logistic regression shows that it may have shortcomings in terms of recall, or precision, or both. The fact that F1-score of the Ensemble Model is high indicates that the categorization provided is fairly balanced and accurate, which is very important when false positives and false negatives may have severe consequences.

V. Conclusion And Future Scope

In this paper, we have presented an ensemble learning based method of hyperparameter tuning with Gradient Boosting, Random Forest, and AdaBoost in order to enhance the edge predictor in context with TSP. We also find that the accuracy, precision, recall, and F1-score of edge prediction significantly improve when fine-tuned ensemble models are employed rather than usual techniques. The utilization of these methodologies in solutions to TSPs could lead to enhanced frameworks that are more accurate and produce faster computations thus reducing the computational intensity of problems in various situations. The above finding clearly illustrates how the Ensemble Model outperforms the others, demonstrating that employing a number of algorithms may increase the overall forecast precision and presenting a strong solution to the TSP's inherent nature.

The outcomes of the study are rather inspiring, which will result in multiple perspectives for the future research. In order to enhance model performance there is one potential development avenue worth exploring which is focusing on the more complex methods of creating ensembles such as stacking and blending. Additionally, there can be works regarding the integration of deep learning models with other techniques in ensemble methods, especially in even more extensive and complex TSP cases. Also, the further research of this method the use of which might be crucial for other combinatorial optimization issues, such as the Job Shop Scheduling Problem (JSSP) and the Vehicle Routing Problem (VRP) is investigated. It may also become more applicable to real life supply chain and logistic problems if those limits of the actual world, such capacity or time limits, are included into the models.

References

[1] X. Zhang, "Optimizing Patrolling Operations with Multiple Traveling Salesmen Problem: A Formulation and Simulation Approach," BCP Business & Management, vol. 44. Boya Century Publishing, pp. 933–937, Apr. 27, 2023. doi: 10.54691/bcpbm.v44i.4980.

[2] Y. Min, Y. Bai, and C. P. Gomes, "Unsupervised Learning for Solving the Travelling Salesman Problem," 2023, arXiv. doi: 10.48550/ARXIV.2303.10538.

[3] S. Liu, X. Yan, and Y. Jin, "An Edge-Aware Graph Autoencoder Trained on Scale-Imbalanced Data for Traveling Salesman Problems," 2023, arXiv. doi: 10.48550/ARXIV.2310.06543.

[4] S. M. Almufti, "Hybridizing Ant Colony Optimization Algorithm for Optimizing Edge-Detector Techniques," Academic Journal of Nawroz University, vol. 11, no. 2. Nawroz University, pp. 135–145, May 25, 2022. doi: 10.25007/ajnu.v11n2a1320.

[5] Salehi and B. Masoumi, "Solving Traveling Salesman Problem based on Biogeography-based Optimization and Edge Assembly Cross-over," JAIDM, no. Online First, Mar. 2020, doi: 10.22044/jadm.2020.7835.1922.

[6] K. Hasegawa and W. Wu, "A Heuristic Approach for the Robust Traveling Salesman Problem," 2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). IEEE, Dec. 07, 2022. doi: 10.1109/ieem55944.2022.9989586.

[7] A. Dey, R. Kumar, and S. Broumi, "The Neutrosophic Traveling Salesman problem with Neutrosophic EdgeWeight: Formulation and A Genetic Algorithm," International Journal of Neutrosophic Science, vol. 19, no. 3. ASPG Publishing LLC, pp. 40–46, 2022. doi: 10.54216/ijns.190304.

[8] Md. Z. Rahman, S. R. Sheikh, A. Islam, and Md. A. Rahman, "Improvement of the Nearest Neighbor Heuristic Search Algorithm for Traveling Salesman Problem," Journal of Engineering Advancements. SciEnPG, pp. 19–26, Mar. 30, 2024. doi: 10.38032/jea.2024.01.004.

[9] Y. Xu and X. Yao, "A Hybrid Approach Based on Ant Colony and Firefly Algorithm for Traveling Salesman Problem," 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2023, pp. 167-171, doi: 10.1109/ICCCBDA56900.2023.10154838.

[10] Z. Shao, J. Yang, C. Shen and S. Ren, "Learning for Robust Combinatorial Optimization: Algorithm and Application," IEEE INFOCOM 2022 - IEEE Conference on Computer Communications, London, United Kingdom, 2022, pp. 930-939, doi: 10.1109/INFOCOM48880.2022.9796715.

[11] M. K. Ulukök, "Impact of genetic algorithm's parameters on solution of numerical optimization benchmark problems," 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), Girne, Northern Cyprus, 2017, pp. 139-143, doi: 10.1109/CICN.2017.8319373.

[12] S. Kumari, P. Bharti, G. Dixit, V. Rohilla, M. Kumar and C. Choudhary, "Brain Tumor Detection Using AIML," 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon), Pune, India, 2024, pp. 1-6, doi: 10.1109/MITADTSoCiCon60330.2024.10575020.

[13] Y. Xu and X. Yao, "A Hybrid Approach Based on Ant Colony and Firefly Algorithm for Traveling Salesman Problem," 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2023, pp. 167-171, doi: 10.1109/ICCCBDA56900.2023.10154838.

[14] S. Thong-ia and P. Champrasert, "Gene-Ants: Ant Colony Optimization with Genetic Algorithm for Traveling Salesman Problem Solving," 2023 International Technical Conference on Circuits/Systems, Computers, and Communications (ITC-CSCC), Jeju, Korea, Republic of, 2023, pp. 1-5, doi: 10.1109/ITC-CSCC58803.2023.10212945.

[15] C. Choudhary, L. S. Nagra, P. Das, J. Singh and S. S. Jamwal, "Optimized Ensemble Machine Learning Model for Chronic Kidney Disease Prediction," 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2023, pp. 292-297, doi: 10.1109/ICCCIS60361.2023.10425073.

[16] A. Kumar, A. Nashte, R. Porwal Amit and C. Choudhary, "Hypergraph Neural Networks with Attention-based Fusion for Multimodal Medical Data Integration and Analysis," 2023 Seventh International Conference on Image Information Processing (ICIIP), Solan, India, 2023, pp. 628-633, doi: 10.1109/ICIIP61524.2023.10537751.

[17] V. Murugananthan, M. Y. E. S. Rehan, R. Srinivasan, M. Kavitha and R. Kavitha, "Traveling Salesman Problem with Ant Colony Optimization," 2023 2nd International Conference on Edge Computing and Applications (ICECAA), Namakkal, India, 2023, pp. 481-485, doi: 10.1109/ICECAA58104.2023.10212262.

[18] J. Thakur, C. Choudhary, H. Gobind, V. Abrol and Anurag, "Gliomas Disease Prediction: An Optimized Ensemble Machine Learning-Based Approach," 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2023, pp. 1307-1311, doi: 10.1109/ICTACS59847.2023.10390226.

[19] P. He, J. Hao, and Q. Wu, “Hybrid genetic algorithm for undirected traveling salesman problems with profits,” Networks, vol. 82, no. 3. Wiley, pp. 189–221, Jun. 21, 2023. doi: 10.1002/net.22167.

[20] V. Keswani and H. Jhamtani, “Formulating Neural Sentence Ordering as the Asymmetric Traveling Salesman Problem,” Proceedings of the 14th International Conference on Natural Language Generation. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.inlg-1.13.