

Performance Optimization of Machine Learning Algorithms by using Crayfish Optimizer

Pravin Game¹, Shrutika Jadhav², Hritik Halli³, Vedanti Bhosale⁴ and Pranav Kaple⁵

¹⁻⁵Pimpri Chinchwad College of Engineering, Pune, India

Email: pravin.game@pccoepune.org, shrutika.jadhav21@pccoepune.org, hritik.halli21@pccoepune.org, vedanti.bhosale22@pccoepune.org, pranav.kaple21@pccoepune.org

Abstract—The need for optimization in machine learning algorithms is essential to enhance prediction accuracy and efficiency across various applications. The Crayfish Optimization Algorithm addresses parameter tuning and feature selection challenges in Linear Regression, Support Vector Machine, K-nearest neighbors, and Isolation Forest in this work. The study uses four datasets representing different domains ranging from medical domain to banking sector. The Crayfish Optimization Algorithm is incorporated into the training process to optimize hyperparameters and reduce feature redundancy. For the Wisconsin Breast Cancer Dataset, the optimized SVM model achieves an accuracy of 97.1%, while the optimized LR model attains 98.07% accuracy. In the case of the Credit Card Fraud Detection Dataset, the optimized KNN model reaches 99.92% accuracy, and the optimized DT model achieves 97% accuracy. For the Glaucoma Image Dataset, the optimized SVM model exhibits a significant increase in precision by 5%, attaining the highest precision value of 95%. The optimized SVM and LR models demonstrate lower mean squared error scores in the Stock Market Dataset. These findings highlight the potential of the COA in enhancing the reliability and adaptability of machine learning models across diverse domains.

Index Terms—Machine Learning, Classification, Crayfish Optimizer, K-Nearest Neighbors, Support Vector Machine, Isolation Forest, Linear Regression.

I. INTRODUCTION

Anomaly detection, with machine learning as an important tool, has contributed significantly to detecting irregularities in diverse fields. The performances of four machine learning algorithms, namely Isolation Forest, Support Vector Machines, Logistic Regression, and k-Nearest Neighbors, will be reviewed using datasets such as credit card fraud detection, breast cancer diagnosis, stock market prediction, and glaucoma detection. It consists of algorithms that can precisely detect anomalies from data by outlining outliers and abnormal patterns. In addition, performance is measured and needs to be accurate and reliable across different applications. As a result, they have gained importance in improving decision-making procedures across a range of domains. In order to reduce risks or make better decisions, anomaly detection is also essential for jobs requiring great precision, dependability, and speed. It has many uses in the financial industry, as identifying fraudulent activity may prevent losses well in time and healthcare, where such detection may save a life. Anomaly detection generally enhances operational efficiency and security by identifying outliers and irregularities. It supports industry predictive maintenance, enhances data integrity, and minimizes errors.

Effective algorithms enable real-time monitoring and timely responses to issues. These systems will allow an organization to minimize risks and maximize benefits.

These datasets on credit card fraud transactions, tumor classification, financial trend prediction, and glaucoma progression detection indicate that the algorithm is helpful in early diagnosis, mitigation of risks, and actionable insight. A critical analysis of machine learning algorithms used in various domains demonstrates how well they work to solve practical issues. The Crayfish Optimization Algorithm greatly increased the accuracy and reliability of these algorithms. Benchmarks were established for anomaly identification in important sectors like healthcare and finance to offer practical standards for future applications. Additionally, it provided insight into how optimization incorporates machine learning models' robustness and practicality for real-time or industrial applications. These investigations have highly emphasized detecting anomalies and showing optimization that some important problems from several perspectives can only be solved together. This research develops an anomaly detection approach to address big real-world problems with transformative effects across healthcare and finance.

II. LITERATURE REVIEW

To improve breast cancer diagnosis, a new classifier called HHO-SVM is introduced in [1], by blending Harris hawk's optimization and SVM for better data analysis. The HHO algorithm determines the SVM parameters in the first phase. The optimized SVM algorithm diagnoses a breast tumor as benign or malignant in the second phase. Testing on real breast cancer data showed promising accuracy rates: 98.24% with normalization scaling and 99.47% with different datasets, but clinical settings is needed to confirm its reliability and applicability.

Various machine learning models for predicting breast cancer using the Wisconsin Breast Cancer Dataset are evaluated in [2]. Among the classifiers evaluated, the Voting classifier which combines both Logistic Regression and Support Vector Machine is at the top with 96.50% accuracy and low error level. The accuracy has also been improved to 98.77% of the Voting classifier by further optimization, thus showing the crucial role of machine learning in the prediction of breast cancer. However, these require testing on various datasets to be reliable.

A new approach using the Bat Algorithm (BA) to optimize Support Vector Regression (SVR) model parameters for the stock market prediction is introduced in [3]. The BA-SVR hybrid model yields better results than the traditional SVR models in the forecast of closing prices of 18 stock indexes of the Chinese market. It gives superior performance across short-term, mid-term, and long-term forecasts when tested on data ranging from January 2016 to December 2020. It is a challenge to accurately predict rock fragmentation in blasting operations to optimize mean fragment size (MFS). In [4], by leveraging support vector regression (SVR) techniques as the foundational prediction tools, coupled with five optimization algorithms—grid search (GS), grey wolf optimization (GWO), particle swarm optimization (PSO), genetic algorithm (GA), and slap swarm algorithm (SSA)—to enhance prediction performance and optimize hyper-parameters. Results demonstrate that the GWO-SVR-based model outperforms others in predicting MFS, achieving high accuracy levels. A model that detects the abnormal behaviors in educational data, particularly among college and university students is introduced in [5]. Firstly, an optimized Isolated Forest algorithm is proposed, combined with a neural network to enhance detection capabilities. Secondly, weight settings for each tree are optimized using discrete particle swarm optimization, leading to improved model accuracy and efficiency. Lastly, the proposed WOS-I Forest demonstrates outstanding performance, achieving an accuracy of 98.6% with a training time of 1 second.

Various ML models for identifying abnormal behavior in student education data using the isolation forest model are evaluated in [6]. The optimized Weighted Optimization Isolated Forest (WOS-I Forest) algorithm is combined with a neural network for anomaly detection. Comparisons with LOF, K-means, and KNN on various datasets show that the WOS-I Forest algorithm achieves 97% accuracy, efficiently identifying abnormal behavior among college students. To improve Breast Cancer prediction results, Gases Brownian Motion Optimization Algorithm approach was implemented in [7]. It was attempted to utilize the Gases Brownian Motion Algorithm to solve the KNN problem to optimize the parameters. Various classification algorithms were investigated and developed for detection. The above approach obtained better results than GA (Genetic Algorithm), ICA (Imperialist Competitive Algorithm), PSO (Particle Swarm Optimization) of 97.79%.

In reference to wastewater treatment addressed in [8], the main goal is to investigate the absorption capacity of Powdered Activated Carbon for removing phenol as a toxic molecule using artificial intelligence that will allow the prediction and optimization of parameters that have a significant effect on the phenol removal mechanism. 11 different distance metrics were optimized which included Euclidean, Manhattan, Jaccard and many more. R, R-squared RMSE, MSE and R-squared adjacent value were calculated. After optimization, model optimized by the distance "Jaccard" with distance weight "squared inverse" was considered as best one.

III. PROPOSED SYSTEM

The proposed model integrates the Crayfish Optimization Algorithm (COA) with SVM, LR, KNN and DT to find the best hyperparameter for each algorithm. Further these ML models are trained on these parameters. The optimized models were evaluated on four different data splits: 90-10%, 80-20%, 70-30%, and 60-40%.

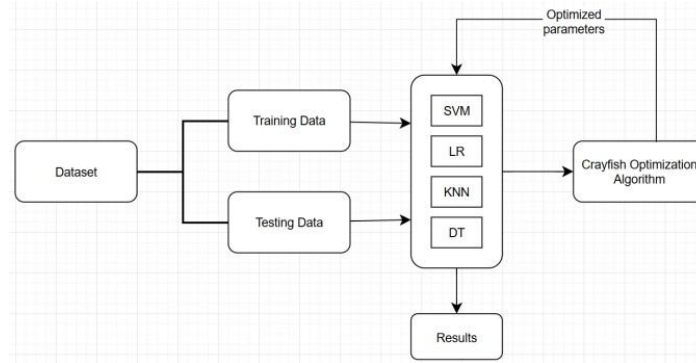


Fig. 1. Flowchart

A. Crayfish Optimization Algorithm

The Crayfish Optimization Algorithm (COA) is a nature-inspired optimization algorithm developed based on the behavior of crayfish. The COA mimics the social behavior of crayfish, particularly their tendency to seek out the best environment for survival.

Crayfish Optimization Algorithm working is as follows:

- Formation of problem: This step includes objective function to be optimized, dataset, ML model.
- Initialization: Create a population of crayfish individuals, each of which stands for a possible fix for the optimization issue. These answers may correspond to various configurations of hyperparameters.
- Evaluation: Evaluate the fitness of each crayfish individual based on its performance on the ML model.
- Exploration and Exploitation: Examine for potential solutions by combining exploration and exploitation tactics. During exploration, crayfish individuals explore the solution space to discover new potential solutions. Exploitation involves intensifying the search around promising solutions to converge towards the optimal solution. This involves exploring different hyperparameter configurations while also exploiting promising solutions discovered so far.
- Movement: Update the positions of crayfish individuals based on their exploration and exploitation. This involves adjusting hyperparameters.
- Reproduction: Crayfish individuals with higher fitness values are more likely to reproduce. This helps maintain diversity in the population while also focusing on promising areas of the search space.
- Termination: Establish a threshold for the optimization process to end, such as the number of iterations reached or the achievement of an acceptable outcome. When the condition is satisfied, stop the algorithm and return the best parameters.

B. Optimization of algorithms by COA

COA is particularly useful for optimizing machine learning algorithms through improved hyperparameter tuning. Among the factors that are essential for defining the machine learning performance model such as the learning rates, regularization coefficients, kernel parameters, and feature weights. Grid search, random search, and other conventional hyperparameter tuning techniques are computationally costly and can produce mediocre outcomes. If applied to ML algorithms, COA systematically adjusts the parameters based on objective function optimizing for measured accuracy, precision, or error metrics against a validation set. Then it identifies the optimal combination of parameters that maximizes the model's performance. COA also comprises mechanisms to escape local optima that maintain diversity within the solution population, thus ensuring a global search for the best parameters. With the incorporation of COA, the advantages ML models attain are improved generalization ability, faster convergence rates, and higher predictive accuracy as realized from the above applications both in medical diagnostics, fraud detection, and stock market forecasting tasks. Due to its flexible and adaptable nature, COA is one of the best optimizers for a large number of ML algorithms. Most notably, it has the potential to be applied to quite powerful SVM, LR, DT, and KNN algorithms which will highly increase their capabilities and effectiveness while solving real- world problems.

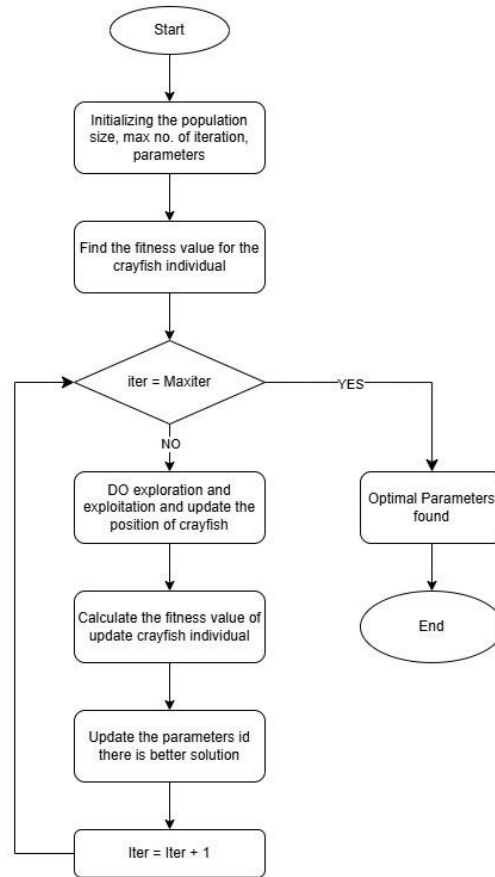


Fig. 2. Flowchart for Crayfish Optimization Algorithm

IV. DATASETS

The proposed model was tested on four different datasets to understand the how the machine learning algorithms which are optimized using COA work on them:

A. Wisconsin Breast Cancer Dataset

Wisconsin diagnostic Breast Cancer (WDBC) dataset, which is available at the University of California, Irvine Machine Learning Repository consists of 569 observations out of which 357 are benign and 212 are malignant and 32 features.

- ID number
- Diagnosis (M = malignant, B = benign)

These 10 features are computed for each cell nucleus: radius, fractal dimension, perimeter, concave points, smoothness, compactness, concavity, area, symmetry, texture.

The standard error, worst (mean of 3 largest values) and mean of these features were computed for each image, resulting in 30 features.

B. Glaucoma Dataset

A preprocessed dataset which consists of the fundus retinal image of the eye is used to test the performance of the model. It consists of a total 455 training images both segregated into separate folders as class_0 and class_1. There are a total of 64 testing images. These images are converted into a numerical array using the NumPy library and passed on to the optimizer.

C. Credit Card Fraud Detection Dataset

In the dataset, European cardholders including credit card transactions done in September 2013. In this set of data, 492 scams from 284807 transactions have occurred over the preceding two days. The transaction dataset is skewed and just 0.172% of the transactions are fraud. It accepts just numerical input variables transformed by

PCA. The key elements obtained by PCA are V1, V2, V3, V4, V5, V28; 'Time' and 'Amount' are only features that remain unaltered by PCA.

D. Stock Market Prediction Dataset

The dataset AAPL.csv contains historical financial data for Apple Inc. It consists of 3732 observations and 64 features. For the purpose of analysis, only a subset of features is considered relevant. These features include, open: Opening price of the stock on a particular day, close: the closing price of the stock on the same day, high: the Highest price the stock reached during the trading day, low: the Lowest price the stock reached during the trading day, volume: The number of shares traded during the day, close_forecast: A forecasted value of the closing value for a future date.

V. RESULTS

A. Experimental Setup:

- Programming Language: Python 3.7 or higher.
 - Libraries and Frameworks:
 - Scikit-learn for implementing ML models (SVM, LR, KNN, DT).
 - Crayfish Optimization Algorithm (COA) implemented in Python.
 - Pandas and NumPy for data manipulation and processing.
 - Data Handling: CSV or Excel files for datasets (e.g., WBCD, credit card fraud, glaucoma, stock market).
- For classification type problems like breast cancer detection, credit card fraud detection and glaucoma detection, five performance metrics were used to assess efficiency of the different models. These metrics are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \times 100 \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$Matthews's\ correlation\ coefficient\ (MCC) = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (5)$$

Here,

TP (True Positive): Observation is positive, and is predicted to be positive. TN (True Negative): Observation is negative, and is predicted to be negative. FP (False Positive): Observation is negative, but is predicted to be positive. FN (False Negative): Observation is positive, and is predicted to be negative.

For regression type of problems like stock market prediction, three performance metrics are used to assess the efficiency of the different models. These metrics are defined as follows:

$$Mean\ Squared\ Error(MSE) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

$$Mean\ Squared\ Error(MSE) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (7)$$

$$R - square = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y - \bar{y})^2} \quad (8)$$

Here,

n = total no. of data points

y_i = observed value

$\hat{y}_i$ = predicted value

$\bar{y}$ = mean of observed values

The optimized models were evaluated on four different data splits: 90-10%, 80-20%, 70-30%, and 60-40%. The performance of each model was assessed on these splits, and the split that produced the best results for each dataset are presented here. The tables below present the results corresponding to the best performing data split for each dataset.

TABLE I: PERFORMANCE OF ML ALGORITHMS FOR WISCONSIN BREAST CANCER DATASET (ON 70-30% SPLIT)

Metrics	SVM	COA- SVM	LR	COA- LR	DT	COA-DT	KNN	COA- KNN
Accuracy	96.4%	97.1%	96.01%	98.07%	56.7%	78.36%	95%	98%
Precision	96.7%	97.2%	95.98%	98.05%	67%	88.2%	98%	94%
Recall	93.5%	98.05%	96.54%	97.01%	61%	100%	90%	100%
F1-Score	95.1%	97.6%	97.02%	97.06%	64%	61%	94%	97%
MCC	92.4%	93.7%	94.25%	95.32%	53%	53%	91%	96%

From the table I, it can be noted that for the 70-30% split results, optimized models are far better than their respective non-optimized counterparts in all of the metrics. COA-SVM has surpassed other models on most of the metrics with a 70-30% split. Highest values can be found on accuracy 97.1%, precision 97.2%, recall 98.05%, F1-score 97.6% and MCC: 93.7%. The COA-LR model is closely followed with good accuracy 98.07% and precision 98.05%, though having recall 97.01% and F1-score 97.06% relatively lower than COA-SVM. Models like COA-DT and COA-KNN have noticed improvement over the non-optimized versions but are still not catching up to COA-SVM on most metrics.

TABLE II: PERFORMANCE OF ML ALGORITHMS FOR CREDIT CARD FRAUD DETECTION DATASET. (ON 90-10% SPLIT)

Metrics	SVM	COA- SVM	LR	COA-LR	DT	COA- DT	KNN	COA- KNN
Accuracy	90.2%	93.7%	94.87%	95.42%	95.7%	97%	99%	99.92%
Precision	1.3%	2%	91.98%	92.01%	100%	100%	94%	89.92%
Recall	90.4%	90.4%	90.41%	91.2%	96%	97%	65%	65.44%
F1-Score	2.6%	4%	85.65%	89.54%	98%	99%	77%	75.75%
MCC	10.3%	13.2%	88.56%	90.74%	92%	94%	78%	76.68%

From the table II, it is evident that the COA-KNN model performed superior to all others on the 90-10 split. It showed the highest accuracy at 99.92%, meaning it has high predictability strength. However, although with high accuracy, its precision is only at 89.92% and its F1-score is 75.75%, which may indicate that it also lacks some strength in correctly classifying positive instances. However, COA-DT was also excellent with high accuracy (97%) with good values of precision, recall, and F1-score metrics, which therefore gives it reliability over multiple measures. The standard models like LR and SVM demonstrate relatively reasonable performance but are drastically improved while optimized by COA, as presented in the COA-LR and COA-SVM models.

TABLE III: PERFORMANCE OF ML ALGORITHMS FOR GLAUCOMA DATASET. (ON 90-10% SPLIT)

Metrics	SVM	COA- SVM	LR	COA- LR	DT	COA-DT	KNN	COA- KNN
Accuracy	89.1%	91.3%	90.1%	92.3%	53.12%	60.93%	84%	86.5%
Precision	90.4%	95%	92.4%	94.5%	62.5%	62.5%	94%	94.3%
Recall	86.3%	86.3%	91.3%	92.1%	15.6%	15.62%	72%	75%
F1-Score	88.3%	90.4%	89.4%	91.5%	25%	25%	82%	83.8%
MCC	78.2%	82.8%	88.7%	90.1%	29%	30%	70%	71%

From Table III, it can be noted that for this evaluation, optimized models perform better than their non-optimized counterparts across most metrics. The COA-LR model records the highest accuracy (92.3%) and precision (94.5%) with a strong F1-score of 91.5% and MCC of 90.1%, closely matching COA-SVM in performance which reaches 91.3% accuracy. COA-DT and COA-KNN, though improved over their non-optimized versions, do not match COA-SVM or COA-LR; COA-DT only reaches 60.93% accuracy, while COA-KNN achieves 86.5%. The basic DT model shows the lowest performance, with accuracy at 53.12% and MCC at 29%. These results indicate that optimized models such as COA-SVM and COA-LR lead in most metrics, emphasizing the significant role optimization plays in enhancing model performance.

TABLE IV: PERFORMANCE OF ML ALGORITHMS FOR STOCK MARKET PREDICTION DATASET. (ON 90-10% SPLIT)

Metrics	SVM	COA- SVM	LR	COA- LR	DT	COA-DT	KNN	COA- KNN
Mean Squared Error	505.55	252.05	79.21	4.45	15736.5	15808.8	1989.39	2172.38
Mean Absolute Error	9.614	8.395	6.5	1.06	92.7	93.08	26.37	27.31
R-squared	0.914	0.957	0.96	1	-1.23	-1.24	0.61	0.69

From Table IV, it is evident that the COA-LR model performed well on the 90-10% split for forecasting in the stock market. It was reported to have the lowest Mean Squared Error at 4.45 and Mean Absolute Error at 1.06 with its highest R-squared value of 1. While the standard LR model also presents a very good R-squared of 0.96, it can be seen that the COA-LR has shown a major improvement in error metrics. The other models showed moderate performances-their error rates and R-squared values are not as favorable for the COA-LR. Of course, the decision tree-based models DT and COA-DT have performed terribly; it shows negative R-squared values that mean the fitted models are highly poor.

VI. CONCLUSION

This paper shows how optimization algorithm, in particular the Crayfish Optimization Algorithm (COA), can greatly improve machine learning models performance on a wide range of datasets. COA-LR outperformed other models in detecting breast cancer using the Wisconsin Breast Cancer Dataset (WBCD), achieving the highest accuracy (98.07%) and F1-score (97.06%) on a 70-30% split. For credit card fraud detection, COA-KNN achieved the highest accuracy of 99.92% on a 90-10% split. COA-LR outperformed all the other models for glaucoma detection with an accuracy of 92.3% and an MCC of 90.1%. With an R-squared value of 1 and the lowest Mean Squared Error (4.45) and Mean Absolute Error (1.06) for the stock market prediction regression task, COA-LR finally showed that it can withstand the challenges of regression tasks. The results consistently show that models with COA optimization outperform those without, highlighting the crucial role that optimization algorithms play in helping machine learning models become more accurate, reliable, and effective in real-world scenarios.

REFERENCES

- [1] S. Almotairi, E. Badr, M. Abdul Salam, H. Ahmed, "Breast Cancer Diagnosis Using a Novel Parallel Support Vector Machine with Harris Hawks Optimization", Mathematics, July 2023.
- [2] K.M.M. Uddin, N. Biswas, S.T. Rikta, S.K. Dey, "Machine learning-based diagnosis of breast cancer utilizing feature optimization technique", Computer Methods and Programs in Biomedicine Update, 2023.
- [3] J. Zheng, Y. Wang, S. Li, H. Chen, "The Stock Index Prediction Based on SVR Model with Bat Optimization Algorithm", Algorithms 2021.
- [4] E. Li, F. Yang, M. Ren, X. Zhang, J. Zhou, M. Khandelwal, "Prediction of blasting mean fragment size using support vector regression combined with five optimization algorithms", Journal of Rock Mechanics and Geotechnical Engineering (JRMGE), 2021.
- [5] Y. Huang, Y. Xue, P. Li, "Subspace Analysis Forest for Hyperspectral Anomaly Detection.", Acta Geodaetica et Cartographica Sinica, pp. 416-425, 2021.
- [6] D. Solyali, A. Mollaei, "A simulation model based on experimental data to determine the optimal tilt angle for a fixed photovoltaic panel," Arch. Adv. Eng. Sci, pp. 1-12, 2023.
- [7] Abdolrazzagah-Nezhad, Majid, Shokooh Pour Mahyabadi, and Ali Ebrahimpoor. "Improving KNN by Gases Brownian Motion Optimization Algorithm to Breast Cancer Detection." Data Science: Journal of Computing and Applied Informatics 4.1 (2020): 1-15.
- [8] Zamouche M., Chermat M., Kermiche Z., Tahraoui H., Kebir M., Bollinger J.C., Amrane A. Mouni L., "Predictive Model Based on K-Nearest Neighbor Coupled with the Gray Wolf Optimizer Algorithm (KNN_GWO) for Estimating the Amount of Phenol Adsorption on Powdered Activated Carbon" Water 2023 15(3):493.