GCU Glasgow Caledonian University

University for the Common Good

# Sand cat swarm optimizer with CatBoost for Sarcoidosis diagnosis

Youssef, Merna; Sharfo, Samer M.; Attar, Hani; Deif, Mohanad A.; Hafez, Mohamed; Solyman, Ahmad

*Published in:*
2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)

*DOI:*
10.1109/EICEEAI60672.2023.10590595

*Publication date:*
2024

*Document Version*
Author accepted manuscript

Link to publication in ResearchOnline

*Citation for published version (Harvard):*
Youssef, M, Sharfo, SM, Attar, H, Deif, MA, Hafez, M & Solyman, A 2024, Sand cat swarm optimizer with CatBoost for Sarcoidosis diagnosis. in *2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*. International Engineering Conference on Electrical, Energy, and Artificial Intelligence, IEEE, 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence, Zarqa, Jordan, 27/12/23. https://doi.org/10.1109/EICEEAI60672.2023.10590595

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

**Take down policy**
If you believe that this document breaches copyright please view our takedown policy at https://edshare.gcu.ac.uk/id/eprint/5179 for details of how to contact us.

Download date: 19. Jan. 2025

# Sand Cat Swarm Optimizer with CatBoost for Sarcoidosis Diagnosis

Merna Youssef
Biomedical Engineering
Cairo University
Giza, Egypt
merna.youssef01@eng.cu.edu.eg

Samer M. Sharfo
Department of Bioelectronics
Modern University of Technology and Information (MTI) University
Cairo, Egypt
samer.86027@eng.mti.edu.eg

Hani Attar
Faculty of Engineering
Zarqa University
Zarqa, Jordan
College of Engineering
University of Business and Technology
Jeddah, Saudi Arabia
Hattar@zu.edu.jo

Mohanad A. Deif
Department of Artificial Intelligence,
College of Information Technology,
Misr University for Science & Technology (MUST)
6th of October City 12566, Egypt
Mohanad.Deif@must.edu.eg

Mohamed Hafez
*Faculty of Engineering FEQS*
*INTI-IU-University*
*Nilai, Malaysia*
mohdahmed.hafez@newinti.edu.my

Ahmad Solyman
School of Computing, Engineering and Built Environment, Glasgow
Caledonian University
Glasgow, UK
ahmed.solyman@gcu.ac.uk

***Abstract*—In the last few years, machine learning has increased in popularity across many disciplines. This paper aims to comprehensively analyze the CatBoost classification algorithm in the context of Sarcoidosis. Analysis was undertaken to evaluate the performance of the CatBoost classification algorithm in comparison to other classifiers. The CatBoost algorithm outperformed other classifiers exploited in this study to identify and differentiate Sarcoidosis. Previous scholarly works ignored missing data observations or filled them with mean values; on the other hand, this study has uncovered that the SIL-2R feature holds significant importance in predicting the occurrence of Sarcoidosis, which improved the selection of treatment and its efficacy. A comprehensive understanding of Sarcoidosis is essential to accurately differentiating symptoms associated with this illness from those associated with other conditions. It is strongly recommended that the CatBoost algorithm be used for sarcoidosis prediction.**

***Keywords*—** Swarm Optimizer, Sarcoidosis, CatBoost

## I. INTRODUCTION

In many fields, using machine learning techniques has become necessary. These tactics give businesses a comprehensive understanding of customer trends, behavior, and internal business operations. Additionally, they play a critical role in promoting the development of new products. Many machine learning techniques, such as AdaBoost, random forest, CatBoost, decision trees, extreme gradient boosting, and K-nearest neighbor, are commonly used in practice[1]. Supervised machine learning has witnessed extensive utilization of classifiers across various domains, including fraud detection, spam email filtering, loan prediction, disease diagnosis, and patient care. This study utilizes diverse machine-learning techniques to identify sarcoidosis disease [2].

Sarcoidosis is an inflammatory disease that can affect multiple organs in the human body and is characterized by persistent inflammation. The precise etiology of this condition remains elusive, and it is characterized by signs affecting both the pulmonary and extrapulmonary systems. This particular medical illness's defining characteristic is the development of an aberrant cell clump called granulomas [3]. Granulomas can develop in many different tissues and organs, but the most commonly affected areas are the eyes, skin, lymph nodes, and lungs. However, it is worth noting that granuloma formation can also affect the liver, heart, brain, and other anatomical regions. Chest X-rays frequently reveal pulmonary manifestations, which may be shown as asymptomatic bilateral hilar lymph node enlargement and/or pulmonary infiltrates. These manifestations may be associated with symptoms such as cough and dyspnea. Peripheral lymphadenopathy, fever, arthritis, cranial nerve palsies, diabetes insipidus, and hypercalcemia are among the extrapulmonary signs and systemic symptoms linked to Sarcoidosis.

The National Heart, Lung, and Blood Institute conducted a thorough review of existing data about the healthcare implications and outcomes of Sarcoidosis in the United States to enhance comprehension of this condition and patient results. In the course of this investigation, differences in outcomes were seen across various demographic categories, including race, ethnicity, gender, and socioeconomic level. Notably, it was shown that African Americans displayed a disproportionate susceptibility to more severe manifestations of the disease. The healthcare burden attributed to Sarcoidosis extends beyond its clinical manifestations, encompassing a substantial influence on emotional, economic, and comorbid disorders connected with the illness. In this particular demographic, there is an observable prevalence of many characteristics, including fatigue, depression, cognitive impairment, pain syndromes, and treatment-related side effects. These factors all contribute to less favorable results [4]. As the 21st century was entered, a substantial accumulation of stored data derived from extensive studies conducted on Sarcoidosis was encountered. To tackle these issues, many statistical techniques have been devised to examine past data and construct predictive models for detecting Sarcoidosis. A study was undertaken by Katsushika et al. [5] to utilize a Deep Learning Algorithm to detect

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Cardiac Sarcoidosis from Echocardiographic films. The results of their investigation suggest that applying a transfer learning strategy combined with a three-dimensional convolutional neural network (3D-CNN) shows promising results as a valuable method for identifying cardiac Sarcoidosis using echocardiographic video analysis[6]. The researchers evaluated the performance of trained and non-pre-trained algorithms using a dataset including 212 echocardiographic films.

In a separate investigation, Chenying Lu et al. [7] performed an empirical analysis to forecast detrimental cardiac occurrences in individuals with Sarcoidosis by utilizing deep learning techniques. The researchers employed the maximum relevance − lowest redundancy strategy to determine which subset of qualities offered the most information while reducing redundancy. Twenty-four features that were gathered manually and 232 features that were retrieved automatically made up the subset. These characteristics were connected to the left ventricle's (LV) structural and functional characteristics. To create classifiers for the goal of classifying endpoints, the researchers used three machine learning (ML) algorithms: multi-layer neural networks (MLP), logistic regression (LogR), and support vector machines (SVM)[8]. Their study showed that automated image processing models—such as LogR, SVM, and MLP—consistently demonstrated higher prediction accuracy than models that relied on human image processing.

The current study substantially adds to the research on machine learning methods for classifying Sarcoidosis. It uses robust data analysis technologies and a vast collection of historical data. Furthermore, a blood protein called the soluble interleukin-2 receptor (sIL-2R) has been suggested as a possible marker for assessing the degree of disease activity in sarcoidosis patients. Assessing blood levels of the soluble interleukin-2 receptor (sIL-2R) holds promise as a significant biomarker for clinical practice. This method can help with the diagnosis and monitoring of Sarcoidosis as well as other specific illnesses [9]. Sarcoidosis is classified into different phases according to the degree of involvement of the lungs and lymph nodes:

- Stage 1 involves the swelling of the pulmonary lymph nodes.
- Stage 2, there is an occurrence of pulmonary lymph node swelling and fibrotic inflammation in the lungs.
- Stage 3 involves the development of fibrosis in the lungs.
- Stage 4 Advanced fibrosis of the lungs.

Several diagnostic techniques, including serum angiotensin-converting enzyme (ACE), fluorodeoxyglucose-positron emission tomography, high-resolution chest computed tomography (CT), and conventional chest radiography, are used to help diagnose Sarcoidosis. Although ACE is frequently employed as a sarcoidosis diagnostic biomarker, its sensitivity is limited. In addition to these disorders, elevated or decreased ACE levels can also be a sign of stomach ulcers, liver inflammation, lymphatic tissue malignancy, and hypothyroidism. On the other hand, reduced ACE levels could indicate a possible sarcoidosis diagnosis.

Immune modulation is greatly aided by the IL-2 receptor, which is made up of three chains: an α (CD25), a β (CD122), and a common γ (CD132). Regulatory T-cells can release soluble IL-2 receptor (sIL-2R) from the cell membrane. This material acts as a stand-in signal for the activation of T cells. Increased T-cell activity is linked to elevated blood levels of sIL-2R, which suggests disease activity.

The molecular mechanism of soluble interleukin-2 receptor (sIL-2R) in the immunopathology of inflammatory illnesses has been explained by several theories. It has been noted that T-cell proliferation is inhibited, which amplifies the immune-stimulatory effects. The information presented clarifies the soluble interleukin-2 receptor's (sIL-2R) role in diagnosing inflammatory diseases like sarcoidosis [10]. Serum sIL-2R measurement is a proper diagnostic technique for Sarcoidosis in clinical practice since it is a biomarker to evaluate the degree of disease activity. In a group of people suspected of having Sarcoidosis, using sIL-2R as a diagnostic biomarker demonstrates sensitivity and specificity in differentiating between patients with and without Sarcoidosis. To look into this, a study evaluating the levels of sIL-2R was conducted on a new group of individuals suspected of having Sarcoidosis. Once a definitive diagnosis of Sarcoidosis or a related illness was established, the sensitivity and specificity of sIL-2R as a sarcoidosis diagnostic marker were assessed by us[11]. A comparison investigation with bloodstream ACE levels was also performed.

To achieve the goal of this study, a set of objectives will be established, which will serve as our scholarly contribution:

- Data science involves exploratory analysis, data cleaning, balancing, and transformation.
- Developing a predictive model using machine learning techniques. Following that, various metrics for model assessment will be applied to evaluate the effectiveness and performance of the implemented models.

The subsequent sections of the paper are organized in the following manner: Section II of the document provides an overview of several machine learning methods, while Section III focuses on presenting designs and nomenclatures. The analytical findings are presented in Section IV, while Section V serves as the concluding section of the work.

## II. MATERIALS AND ALGORITHM

This section will explore machine learning techniques used in the analysis.

### *A. Prediction models*

#### *1) K-Nearest Neighbors -KNN*

A straightforward and adaptable supervised machine learning technique is K-Nearest Neighbors (KNN). It makes predictions or classes based on the average value or majority class of its closest neighbors in a feature space. Large datasets or high dimensions might affect the algorithm's efficiency, and the choice of 'k' and the distance measure is critical. Although KNN is easily interpreted, it may not be scalable to intricate issues, yet it provides a simple overview of classification and regression tasks.

*2) Support Vector Machine -SVM*

Robust supervised learning methods for regression and classification are support vector machines (SVM). SVM looks for a hyperplane that maximizes the margin between classes and divides data into the best possible classes. Using kernel functions, it works well in high-dimensional spaces, is resistant to overfitting, and can be applied to linear and non-linear problems. SVMs are a helpful tool in machine learning because of their adaptability and capacity to handle complex decision boundaries, making them frequently utilized in various areas.

*3) Random Forest*

The Random Forest (RF) technique is a widely recognized ensemble learning approach based on decision trees belonging to the group of bagging-type ensembles [12]. The distinguishing characteristic of Random Forest (RF) compared to conventional decision trees is its distinct methodology for dividing nodes. In the Random Forest (RF) algorithm, the division of each node is performed by utilizing the most optimal predictor from a collection of randomly chosen and specific predictors[13]. Including this supplementary randomness level significantly improves RF's resilience against overfitting [14].

To optimize the performance of the ensemble of trees inside the Random Forest (RF) algorithm, a slight adjustment is implemented to reduce the correlation between the trees. Multiple decision trees are generated using bootstrapped training sets, similar to the bagging technique. However, during the creation of these decision trees, at each node where a split is being contemplated, a random subset of 'm' predictors is selected as potential candidates for the split from the entire collection of 'p' predictors [15]. The Random Forest (RF) methodology can be used for classification and regression problems[16].

Random Forests are known for reducing overfitting and enhancing decision tree generalization performance by introducing randomness in the data used to train each tree and the predictor features considered for splitting nodes.

*4) Adaptive Boosting*

AdaBoost, often called adaptive boosting, is a machine learning ensemble technique that aims to make a reliable and accurate predictive model by enhancing the performance of weak learners. Each training instance is given a weight, which AdaBoost uses to prioritize incorrectly categorized cases over correctly identified ones in later iterations. To produce a more robust and more accurate model, it integrates the predictions of several weak learners, which are usually shallow decision trees. By iteratively modifying the weights of incorrectly identified cases, the model can concentrate on portions of the dataset previously found challenging. The final model is a weighted sum of the individual underperforming learners, which yields a compelling, powerful ensemble, especially for classification problems. AdaBoost is widely used for its simplicity, versatility, and ability to handle complex datasets[17].

*5) Decision Trees*

The decision tree is a popular and adaptable machine-learning approach for regression and classification applications. The algorithm creates a decision tree by reducing the dataset into subgroups according to the most essential traits 18].

When building a decision tree, the algorithm looks for the best feature that splits the data to minimize variance for regression tasks and maximize information gain or Gini impurity for classification tasks. This recursive splitting process continues until a predetermined stopping criterion, such as particular tree depth or the minimal number of samples in a leaf node, is satisfied [19]. Decision trees are valuable because they are easy to interpret, leading to a readily comprehensible tree structure[20]. However, they are prone to overfitting, especially when dealing with dense woods. Decision Trees are frequently used in ensemble methods, such as Random Forests and AdaBoost, to improve their generalization and prediction performance [21].

*6) Gradient Boost*

Gradient Boosting is a robust boosted algorithm for regression and classification tasks. It is derived from a combination of two fundamental concepts: Gradient Descent and Boosting. The key idea behind Gradient Boosting is to build an ensemble model forward stage-wise, iteratively improving the model's performance. The foundation for developing Gradient Boosting came from the initial attempts to generalize an adaptive boosting algorithm to work with various loss functions, as pioneered by references [22] and [23]. The Gradient Boosting algorithm follows a series of steps to enhance the model's predictive ability incrementally. This algorithm is widely used in machine learning and has proven effective in addressing various predictive tasks[24].

*7) Extreme gradient boosting*

Extreme Gradient Boosting, commonly known as XGBoost, is a powerful and efficient machine learning algorithm for classification and regression tasks. It belongs to the family of gradient-boosting algorithms and has gained popularity for its high performance and scalability. XGBoost enhances the traditional gradient boosting method by incorporating regularization techniques and a unique optimization strategy. It combines decision trees as base learners and employs a gradient descent optimization process to minimize a predefined loss function[25].

Critical features of XGBoost include:

- Regularization: XGBoost incorporates L1 (Lasso) and L2 (Ridge) regularization terms in its objective function, helping prevent overfitting and improving model generalization.
- Tree Pruning: The algorithm employs a process known as "tree pruning" to remove unnecessary branches, enhancing model efficiency and reducing the risk of overfitting.
- Parallelization: XGBoost is designed for parallel and distributed computing, making it highly efficient and suitable for large datasets.

- Handling Missing Data: It can effectively handle missing data during training.
- Cross-Validation: XGBoost supports built-in cross-validation, aiding in selecting optimal hyperparameters.

In the context of a dataset with m samples and n features, denoted as $D = \{(X_j, y_j)\}_{j=1}^{m}$, where $X_j(x_{1j}, x_{2j}, \ldots, x_{nj})$ is a vector of $n$ feature sand $y_j(\in) R$ represents the response of $n$ feature, typically binary $(e.g., \text{yes or no})$ or encoded numerically $(0 \text{ or } 1)$.

The samples $(X_j, \text{and } y_j)$ Are assumed to be independently and identically distributed according to some unknown. The objective of the learning task is to train a function that minimizes the expected loss distribution.

Due to its exceptional performance and versatility, XGBoost is widely used in various machine-learning competitions and real-world applications. It has become a popular choice for data scientists and practitioners working on predictive modeling tasks.

*8) CatBoost*

CatBoost is a high-performance machine-learning algorithm for gradient-boosting decision trees. It is particularly well-suited for categorical feature handling, making it advantageous in scenarios where traditional gradient-boosting algorithms might struggle with such features.

Critical features of CatBoost include:

- Categorical Feature Support: CatBoost handles categorical features without preprocessing or one-hot encoding. This feature simplifies the data preparation process, often improving model performance.
- Optimized Learning Process: CatBoost incorporates a specialized algorithm for handling categorical data and employs a variant of gradient boosting that optimizes the learning process, reducing the likelihood of overfitting.
- Robust to Overfitting: CatBoost integrates techniques such as depth-wise growing and ordered boosting, contributing to its robustness against overfitting.
- Built-in Cross-Validation: The algorithm includes built-in cross-validation functionality, making evaluating model performance and tuning hyperparameters easier.
- Efficient GPU Usage: CatBoost is designed to take advantage of GPU acceleration, enhancing training speed and scalability.
- Support for Regression and Classification: CatBoost can be used for regression and classification tasks.

CatBoost has gained popularity in various machine learning applications, mainly when datasets contain categorical features and a robust and efficient gradient boosting algorithm is required. Its ease of use, competitive performance, and ability to handle complex data make it a valuable tool for practitioners and researchers [26]. The general CatBoost idea is to reduce or minimize the expected loss defined:

$$g^t = argmin_{g \in G} \mathcal{L}(H^t + g) \quad (1)$$
$$= \; argmin_{g \in G} EL(yH^{t-1}(X) + g(X))$$

Where (X, y) tests data sampled from training data D, L is a smooth loss function.

Also, the gradient boosting process builds an iterative series of approximations. $H^t$, with function $g^t$, chosen from a group of functions $G$, step size $\alpha$ base predictor.

*B. Optimization with SCSO*

As [27] mentioned, each sand cat in the SCSO (Sand Cat Search Optimization) method represents a solution to a d-dimensional optimization problem represented by a one-dimensional array. The floating-point values in this array correspond to the variable values (x1, x2, xd, …) as in Fig.5. First, the algorithm creates a candidate matrix with the number of sand cat populations required based on the magnitude of the challenge. Then, for every sand cat, a fitness function is calculated. This mathematical function defines the critical variables in problem-solving and directs the setting of these variables for a solution. The fitness function gives each sand cat a value according to its performance. Whether the goal of the problem is minimization or maximization, the fitness function's main objective is to match it. Every iteration, the sand cat with the highest cost value emerges as the top search agent. As a result, search agents try to update their positions by taking the best search agent's position into account in later iterations. This iterative procedure continuously improves the search agents' placements.

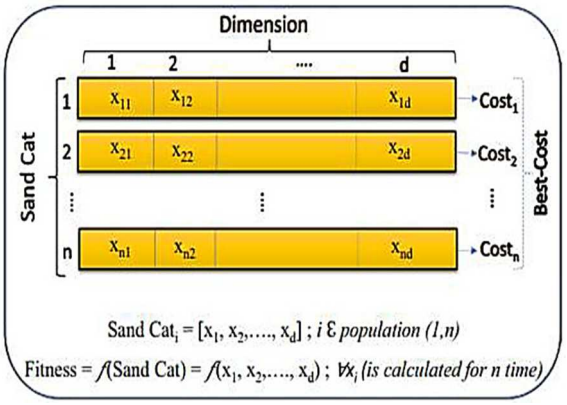


Fig.1. Sand Cat Swarm Optimizer Phases

Interestingly, the optimal answer is represented by the best solution from each iteration, similar to the sand cat closest to its prey. This method encourages effective memory use by avoiding the needless storing of earlier solutions. To ensure efficient problem-solving, the SCSO algorithm essentially uses the idea of sand cat behavior to improve solutions iteratively.

## III. DESIGN AND NOMENCLATURES

In this analysis, a range of evaluation criteria will be explored to evaluate the efficacy of a machine-learning model. The metrics encompassed in this analysis are the confusion

matrix, accuracy, recall, precision, true positive rate, true negative rate, false-positive rate, and false-negative rate.

### A. Confusion Matrix

A confusion matrix provides information about the classifications made by a classifier, including both actual and predicted classifications. It is a fundamental tool for evaluating the performance of a classifier. In Table I, you can see the confusion matrix for the classifier, which is used to assess the classifier's performance[28].

TABLE 1. CONFUSION MATRIX

| | **Predicted** | | |
|---|---|---|---|
| **Actual** | | ***Negative*** | ***Positive*** |
| | ***Negative*** | True Negative (TN) | False Negative (FN) |
| | ***Positive*** | False Positive (FP) | True Positive (TP) |

### B. Model Evaluation Metrics

Model training time, model accuracy, and memory usage are valuable metrics for assessing classifier performance.

Accuracy: This metric assesses the accuracy of the model's predictions by determining the ratio of correctly classified samples to total samples. It is computed using the formula (2):

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \times 100\% \quad (2)$$

True Positive Rate (TPR), or sensitivity or recall, represents the ratio of correctly predicted positive class instances to the total number of positive class instances. A perfect sensitivity score is 1.0, while the lowest achievable score is 0.0.

True Negative Rate (TNR), often called specificity, indicates the proportion of accurate negative predictions from the total number of negative class instances. The highest achievable specificity score is 1.0, while the lowest is 0.0. TPR and TNR are expressed in equations (3) and (4).

$$\text{TruePositiveRate} = \text{TP}/(FN + TP) \times 100 \quad (3)$$

$$\text{TrueNegativeRate} = \text{TN}/(FP + TN) \times 100 \quad (4)$$

Precision signifies the count of correctly predicted positive values concerning the total number of positive class instances, as depicted in equation (5).

False Positive Rate (FPR) represents the count of incorrect positive predictions relative to the total number of negative instances, as illustrated in equation (6).

$$\text{FalsePositiveRate} = \text{TP}/(FNP + TP) \times 100 \quad (5)$$

$$FalseNegativeRate = FP/(FP + TN) \times 100 \quad (6)$$

## IV. RESULT AND DISCUSSION

In this section, two analyses will be conducted to assess the performance of the machine learning algorithms that were previously presented. To commence, an examination of the data shall be initiated by gathering quantitative statistics and detecting any instances of absent values or outliers. Additionally, the equilibrium of the independent variables will be evaluated. After conducting the initial investigation, the presence of missing values and outliers in the dataset was detected. Additionally, a balanced distribution of the independent feature has been observed.

### 1. Exploratory Analysis

Exploratory analysis is crucial in creating a predictive model, providing a comprehensive grasp of the information. In this preliminary stage, our objective is to investigate fundamental inquiries, including (i) identification of features with missing values, (ii) identification of features with outliers, (iii) assessment of the balance in the response feature, and (iv) examination of the general distribution of data points, along with other pertinent observations. The answers to these questions are provided by our findings, which are visually shown in Figure 2.

The distribution of people diagnosed with Sarcoidosis, categorized by neurological involvement, is depicted in Figure 1. Multiple studies have demonstrated a notable rise in the incidence of instances among individuals who engage in smoking, impacting both neurologically involved and noninvolved. Figure 2 depicts the distributions of all features on the whole dataset scale. Figure 3 provides valuable insights regarding the association between features and the target variable. Visualizations facilitate the acquisition of a more comprehensive comprehension of the dataset and its inherent attributes, a pivotal aspect in the preparatory stages of predictive modeling.

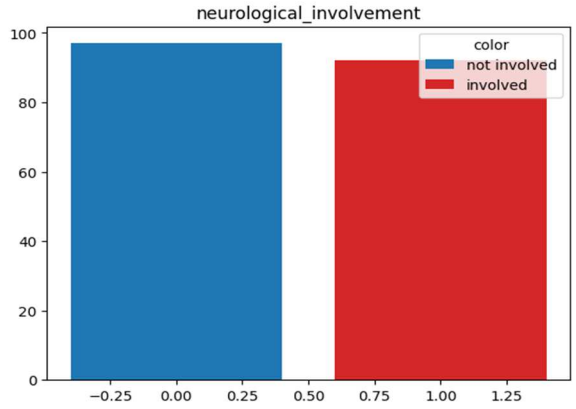


Fig.2. Neurological Involvement in The Dataset

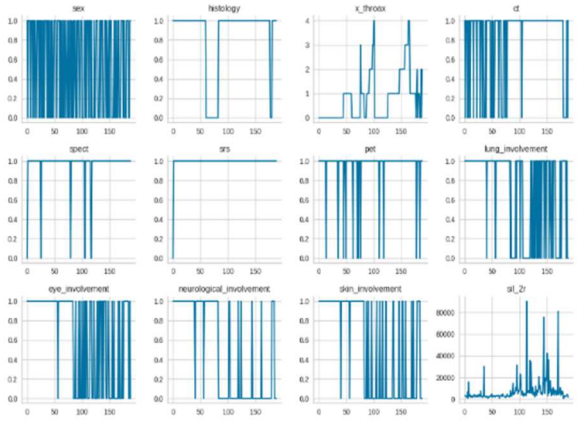

Fig.3. Distributions of Features

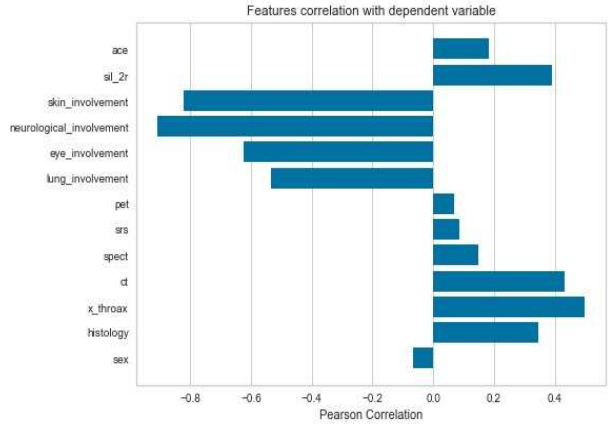


Fig.4. Correlation Between Features and The Target.

### 2. *Data Preprocessing*

No missing data was identified in the dataset. While some outliers were detected during the initial analysis, they were addressed by replacing them with the correct values on a case-by-case basis. This data-cleaning process ensures the reliability and accuracy of the dataset for further analysis and modeling.

### 3. *Cross Validation*

Cross-validation using 20-fold was done. As cross-validation is usually used for model selection, it was also used to enhance the generalization of the model on the available data to ensure data reliability for robust performance estimation as it evaluates the models across several permutations of training and test data subsets and reduces the variance in the performance metrics.

### 4. *Results and Discussion*

The false negative rate is often used in null hypothesis testing, particularly in situations involving multiple comparisons. Precision is a critical metric in the context of the problem addressed in this paper. Predicting that a patient does not have Sarcoidosis when they do can have severe consequences. Additional metrics, such as accuracy and recall, are also presented in Fig.4. Following the training of the models on the training dataset and the prediction of probabilities on the test dataset, the true positive rate and false positive rate were calculated to evaluate the models' performance further.

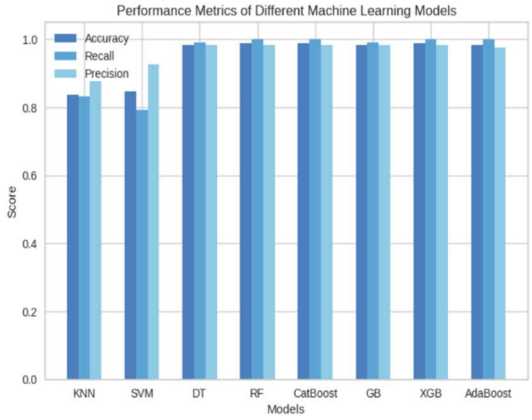


Fig.5. Performance Metrics of Different Machine Learning Models.

### 5. *Optimization*

An optimization technique called SCSO was used to enhance the performance of CatBoost as the CatBoost got the best results among all other models, equaling XGboost. The sand cat swarm optimization, or SCSO, is a recently developed metaheuristic algorithm with an easy-to-implement framework for efficiently finding the optimal solution to optimization issues. It is introduced with balanced behavior in the exploration and exploitation phases. However, the optimization technique did not improve the performance of CatBoost as expected.

## V. CONCLUSION

In summary, the objective of this study was to conduct a comparative analysis of different predictive machine learning algorithms in the context of supervised learning, specifically for the prediction of sarcoidosis disease. The methodology encompassed a comprehensive examination of the gathered data, intending to make predictions about Sarcoidosis using this dataset. The procedure commenced with an initial analysis, providing significant insights into the dataset. The data underwent cleansing, balancing, and transformation processes to be appropriately prepared for predictive modeling. The study proceeded to implement the machine learning algorithms that were previously mentioned, and a range of measures were utilized to evaluate the performance of the models. Among the algorithms examined, the CatBoost classifier had remarkable performance, attaining the maximum accuracy. Additional evaluation measures provided more evidence of the efficacy of this approach. Therefore, the utilization of the CatBoost classifier is proposed as the optimal selection for constructing predictive models within the domain of Sarcoidosis. Considerable emphasis was placed on the importance of the sil-2r characteristic in identifying Sarcoidosis. The role and significance of disease identification and diagnosis were comprehensively elucidated. Moreover, several additional retrieved traits contributed partially but significantly to the classification and differentiation of the condition. Using more extensive datasets and investigating the exact dosage of sil-2r tailored to individual patients instead of depending on approximations presents significant potential and advantages for future research endeavors in this domain.

## REFERENCES

[1] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. 9, no. 1, pp. 381–386, 2020.

[2] P. Ungprasert, J. H. Ryu, and E. L. Matteson, "Clinical manifestations, diagnosis, and treatment of sarcoidosis," *Mayo Clin Proc Innov Qual Outcomes*, vol. 3, no. 3, pp. 358–375, 2019.

[3] M. A. Judson, "The clinical features of sarcoidosis: a comprehensive review," *Clin Rev Allergy Immunol*, vol. 49, pp. 63–78, 2015.

[4] A. K. Gerke, M. A. Judson, Y. C. Cozier, D. A. Culver, and L. L. Koth, "Disease burden and variability in sarcoidosis," *Ann Am Thorac Soc*, vol. 14, no. Supplement 6, pp. S421–S428, 2017.

[5] S. Katsushika *et al.*, "Deep learning algorithm to detect cardiac sarcoidosis from echocardiographic movies," *Circulation Journal*, vol. 86, no. 1, pp. 87–95, 2021.

[6] N. Baghdadi, A. S. Maklad, A. Malki, and M. A. Deif, "Reliable sarcoidosis detection using chest x-rays with efficiencies and stain-normalization techniques," *Sensors*, vol. 22, no. 10, p. 3846, 2022.

[7] C. Lu *et al.*, "Predicting adverse cardiac events in sarcoidosis: deep learning from the automated characterization of regional myocardial remodeling," *Int J Cardiovasc Imaging*, vol. 38, no. 8, pp. 1825–1836, 2022.

[8] Q. I. Ahmed, H. Attar, A. Amer, M. A. Deif, and A. A. A. Solyman, "Development of a Hybrid Support Vector Machine with Grey Wolf Optimization Algorithm for Detection of the Solar Power Plants Anomalies," *Systems*, vol. 11, no. 5, p. 237, 2023.

[9] M. W. Ziegenhagen, U. K. Benner, G. Zissel, P. Zabel, M. A. X. Schlaak, and J. Muller-Quernheim, "Sarcoidosis: TNF-α release from alveolar macrophages and serum level of sIL-2R are prognostic markers," *Am J Respir Crit Care Med*, vol. 156, no. 5, pp. 1586–1592, 1997.

[10] L. E. M. Eurelings *et al.*, "Sensitivity and specificity of serum soluble interleukin-2 receptor for diagnosing sarcoidosis in a population of patients suspected of sarcoidosis," *PLoS One*, vol. 14, no. 10, p. e0223897, 2019.

[11] R. E. Hammam, A. A. A. Solyman, M. H. Alsharif, P. Uthansakul, and M. A. Deif, "Design of Biodegradable Mg Alloy for Abdominal Aortic Aneurysm Repair (AAAR) Using ANFIS Regression Model," *IEEE Access*, vol. 10, pp. 28579–28589, 2022.

[12] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[13] M. R. Parsaei, S. M. Rostami, and R. Javidan, "A hybrid data mining approach for intrusion detection on imbalanced NSL-KDD dataset," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, pp. 20–25, 2016.

[14] J. Singh, S. P. Mohanty, D. K. Pradhan, J. Singh, S. P. Mohanty, and D. K. Pradhan, "Introduction to SRAM," *Robust SRAM Designs and Analysis*, pp. 1–29, 2013.

[15] A. A. Ibrahim, R. L. Ridwan, M. M. Muhammed, R. O. Abdulaziz, and G. A. Saheed, "Comparison of the CatBoost classifier with other machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020.

[16] E. M. O. Mokhtar and M. A. Deif, "Towards a Self-sustained House: Development of an Analytical Hierarchy Process System for Evaluating the Performance of Self-sustained Houses," *ENGINEERING JOURNAL*, vol. 2, no. 2, 2023.

[17] K. Parang, L. I. Wiebe, and E. E. Knaus, "Novel Approaches for Designing 5'-O-Ester Prodrugs of 3'-Azido-2'3'-Dideoxythymidine (AZT)," *Curr Med Chem*, vol. 7, no. 10, pp. 995–1039, 2000.

[18] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Ann Stat*, pp. 1189–1232, 2001.

[19] K. Parang, L. I. Wiebe, and E. E. Knaus, "Novel Approaches for Designing 5'-O-Ester Prodrugs of 3'-Azido-2'3'-Dideoxythymidine (AZT)," *Curr Med Chem*, vol. 7, no. 10, pp. 995–1039, 2000.

[20] M. A. Deif, M. A. A. Eldosoky, H. W. Gomma, A. M. El-Garhy, and A. S. Ell-Azab, "Adaptive neuro-fuzzy inference system controller technique for lower urinary tract system disorders," *J Clin Eng*, vol. 40, no. 3, pp. 135–143, 2015.

[21] D. Witten and G. James, An introduction to statistical learning with applications in R. Springer publication, 2013.

[22] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front Neurorobot*, vol. 7, p. 21, 2013.

[23] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[24] R. E. Hammam *et al.*, "Prediction of wear rates of UHMWPE bearing in hip joint prosthesis with support vector model and grey wolf optimization," *Wirel Commun Mob Comput*, vol. 2022, pp. 1–16, 2022.

[25] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," *Adv Neural Inf Process Syst*, vol. 31, 2018.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337–407, 2000.

[27] A. Seyyedabbasi and F. Kiani, "Sand Cat swarm optimization: A nature-inspired algorithm to solve global optimization problems," *Eng Comput*, vol. 39, no. 4, pp. 2627–2651, 2023.

[28] M. G. Aartsen *et al.*, "Determining neutrino oscillation parameters from atmospheric muon neutrino disappearance with three years of IceCube DeepCore data," *Physical Review D*, vol. 91, no. 7, p. 072004, 2015.