Choosing the Right Supervised Machine Learning Algorithm for Specific Applications

Dheeraj Vaddepally

dheeraj.vaddepally@gmail.com

Abstract:
Choosing the right supervised learning algorithm is essential for handling certain machine learning problems and applications. Whether the challenge is a classification or regression task determines which method is used. Because they are good at handling categorical results, algorithms like Random Forests, Support Vector Machines (SVM), and Logistic Regression are frequently used for classification problems like spam detection and medical diagnoses. On the other hand, Linear Regression, Gradient Boosting, and Random Forests are frequently used to forecast continuous variables in regression tasks, such as forecasting stock market movements or housing prices.
Computational limitations, comprehensibility needs, and data magnitude and integrity are important factors to take into account while choosing an algorithm. Through the alignment of algorithmic properties with the particular requirements of the application, practitioners can improve model performance and guarantee dependable results. This method highlights the significance of a methodical approach to algorithm selection in supervised learning that is suited to various domains and applications.

Keywords: supervised learning, SVM, linear regression, continuous variables, classification.

I. INTRODUCTION

Machine learning, defined by Samel, A. L. "the field of study that gives computers the ability to learn without being explicitly programmed". Machine Learning a branch of artificial intelligence (AI), refers to the development of computer system capable of performing tasks that typically require human intelligence. Now a days Machine learning gives supporting advancements in a wide range of industries including robotics, healthcare, finance and marketing. Machine learning divided into three primary paradigms like Supervised learning, unsupervised learning and reinforcement learning [1].

Supervised Learning, involves training models using labelled datasets where inputs are paired corresponding with outputs. This model predicts accurate outcomes. Unsupervised learning, involves training models using unlabeled datasets where input datasets are not paired with output data. Reinforcement learning is decision making through interaction with an environment to maximize cumulative rewards. Supervised learning is particularly applicability in solving classification and regression problems [2]. Classification includes tasks such as spam detection, image recognition which assigning discrete labels to data points. Regression includes tasks such as price predictions, weather forecasting which is aim to predict continues outcomes. This paper providing a guide to Supervised learning, comparing key algorithms and highlighting the strength and limitations and discussing their applications. Although , everyone gain a understanding of how to select appropriate supervised learning algorithm.

II. CORE CONCEPTS OF SUPERVISED LEARNING

Supervised learning where a model is trained on labelled data, where each input is paired with corresponding output. This input – output mapping helps models to learn patterns and accurate predictions. It adjusts itself to minimize error and improve accuracy [3].

A main aspect is the division of data into training dataset and testing dataset. Training dataset is used to learns the model and testing dataset evaluate its performance. Also supervised learning include overfitting and underfitting [4]. Overfitting occurs when the model can overfit the trained data failing to generalize the unseen

data. Underfitting arise s when the model is too simple to capture underlaying patterns. Supervised learning has two types of problems one is Classification while other is Regression [3].
Classification algorithms are used to predict a categorical output. Examples include spam detection where the email is spam or not, medical diagnostics where symptoms are used to predict diseases. Regression algorithms are used to predict continues numerical output. Examples include house price prediction based on the location and features, stock market forecasting using historical data [5].

III. ALGORITHM OVERVIEW BY APPLICATION

A. Linear regression

Linear Regression is the most famous algorithm of machine learning that compute the relationship between dependent variable and one or more independent variables. It is used to predict continues values such as salary, age, product, price [6].
First is Simple linear regression which is uses a single independent variable to predict the value of a quantitative dependent variable [7]. The linear regression model is represented by the equation below:
$y = mx + c + e$
where m is slop and, c is intercept value.
In mathematical terms simple linear regression is
$Y = \beta 0 + \beta 1X$
Where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

While Multiple linear regression involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots\ldots + \beta nXn$
where:

- Y is the dependent variable
- X1, X2, …, Xn are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of this algorithm is to find the best Fit Line equation that can help predict the values based in the independent variables [8].
While tasks where the goal is to predict the probability that an instance belongs to a given class or not in such instances Logistic regression is used for classification. It is employed for binary classification where the use of sigmoid function comes into play, that takes input as independent variables and produces a probability value between 0 and 1 [9].

B. Decision Tree

Further a decision tree is a flow chart like structure used to make decisions. It is as well employed for classification and regression problem. Structure of this algorithm is tree like which start with root nodes and end with leaf nodes. The decision node works as root node. It can further be divided into two and more subtree. The sub tree again becomes root node and divides into two or more leaf nodes [5] [10].
Here the root node represents the entire dataset and the initial decision to be made. While the Internal node represents decisions or tests on attributes. Each internal node can have one or more branches. Where branches represent the outcome of a decision or test, leading to another node. At last, the leaf node represents the final decision or prediction. No further splits occur at these nodes.

C. Naive Bayes

Then we have Naïve Bayes classifier which is works based on conditional probability rule. Which is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other [11]. The examples of Naïve Bayes classifier are spam recognition, face detection, character recognition.

The foundational principle i.e. Bayes' theorem of Naïve Bayes classifier is stated mathematically as the following equation:

$P(A|B) = P(B|A)\ P(A) / P(B)$

- P(A) = Priori Probability (Probability of event before event b)
- P(A|B) = Posterior Probability (Probability of event before event b)
- P(B) = Priori Probability of predictor
- P(B|A) = it is likelihood which is the probability of the predictor

D. Support Vector Machine

When it comes to algorithm which works well with both for linear and nonlinear classifications problems the Support Vector machine algorithm comes into play. Using Support vector machine helps make best decision and differentiate "n" spaces into classes. Its adaptable and easy to implement for applications like anomaly detection, spam detection, gene expression analysis, image classification, face detection [3].

The following diagram gives an overview of how the algorithm works and key terms associated with it.

- Support Vector: The nearest positive point and nearest negative points
- Hyper plane: The center line is called Hyper plane.
- Margin: the distance between two parallel lines

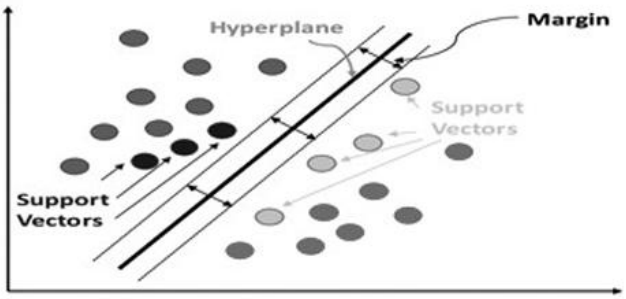

Fig 1. Support Vector Machine [12]

E. Artificial Neural Network

Modelled after neurons in human brain we have Artificial Neural Networks. It contains artificial neurons which are called units. These units are arranged in a series of layers that together constitute the whole Artificial Neural Network in a system [13]. A layer can have only a dozen units or millions of units as this depends on how the complex neural networks will be required to learn the hidden patterns in the dataset. It is a hybrid of supervised, unsupervised, reinforcement approach [2] [13].

It usually consists of the following layers which then gives the output:

- Input layer: It helpful for accept all the inputs.
- Hidden layer: It is a layer between input and output layer.
- Output layer: it is used to show the final result.

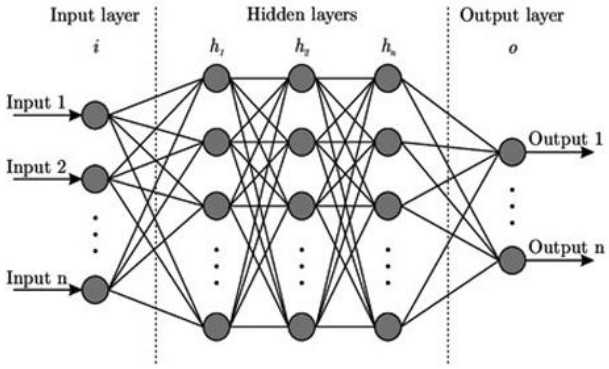

Fig 2. Artificial neural network [13]

F. Gradient Boosting

Then comes the ensemble learning approaches which popularly used for classification and regression-based tasks at hand i.e. Gradient Boosting. It is one kind of ensemble Learning method which trains the model sequentially and each new model tries to correct the previous model. There it has a technique called the Gradient Boosted Trees whose base learner is CART (Classification and Regression Trees) [14]. The below diagram explains how gradient-boosted trees are trained for regression problems. Each tree predicts a label and the final prediction is given by the formula,

y(pred) = y1 + (eta *r1) + (eta * r2) + ... + (eta * rN)

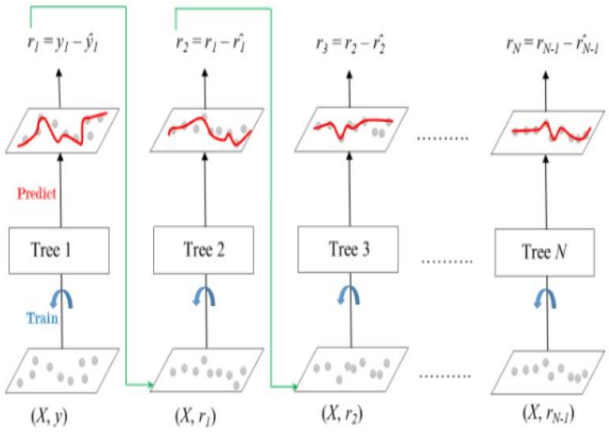

Fig 3. Gradient Boosted Trees for Regression [13]

G. Random forest

At last, we have Random forests or Random Decision Trees which is a collaborative team of decision trees that work together to provide a single output. Originating in 2001 through Leo Breiman, Random Forest has become a cornerstone for machine learning enthusiasts [15].

Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. Random forests are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments [2].

IV. PERFORMANCE EVALUATION METRICS

Now that the understanding of several algorithms is there, we should know how to evaluate how well the algorithm is performing. And evaluating the performance of supervised learning algorithms involves metrics tailored to the problem type:

A. Classification metrics

Classification is the process of categorizing data or objects based on their traits or properties into specified groupings or categories. Classification is a type of supervised learning approach in machine learning in which an algorithm is trained on a labelled dataset to predict the class or category of fresh, unseen data [16]. The primary goal of classification is to create a model capable of properly assigning a label or category to a new observation based on its properties. there is different classification metrics:

First is a confusion matrix, which is a table that summarizes the performance of a classification algorithm and it consists of four metrics:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)

From the confusion metrics further metric are calculated such as: Accuracy, which is a fundamental metric used to evaluate the performance of classification models. It measures the proportion of correctly predicted instances (both true positives and true negatives) among all instances in the dataset [3]. The formula for accuracy is:

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

Whereas Precision quantifies the proportion of true positive predictions among all instances predicted as positive. It is particularly valuable when the cost of false positives is high [12]. The formula for precision is:

Precision = TP/ (TP + FP)

Then, Recall, also known as sensitivity or true positive rate, is a fundamental classification metric that assesses a model's ability to correctly identify all positive instances within a dataset [16]. It quantifies the proportion of true positive predictions (correctly predicted positive instances) among all instances that are actually positive. The formula for recall is:

Recall = TP/ (TP + FN)

F1-Score

The F1-Score combines both precision and recall into a single value, providing a balanced assessment of a model's performance [17]. It is calculated using the harmonic mean of precision and recall:

F1-Score = 2 * (Precision * Recall)/ (Precision + Recall)

ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a model's ability to distinguish between positive and negative classes at various thresholds. The Area Under the ROC Curve (AUC) quantifies the overall performance of a classification model [17]. An AUC of 1.0 indicates perfect discrimination, while an AUC of 0.5 indicates performance equivalent to random guessing [16].

B. Regression metrics

Regression metrics are essential for evaluating the performance of regression models, which predict continuous numerical values. Unlike classification metrics, regression metrics focus on the error between predicted and actual values [1]. Here are some commonly used regression metrics:

Mean Absolute Error (MAE)

In the fields of statistics and machine learning, the Mean Absolute Error (MAE) is a frequently employed metric. It's a measurement of the typical absolute discrepancies between a dataset's actual values and projected values [6].

The formula to calculate MAE for a data with "n" data points is:

$MAE = 1/n \sum i=1n | xi-yi |$

Where:

- xi represents the actual or observed values for the i-th data point.

- yi represents the predicted value for the i-th data point.

Mean Squared Error (MSE)
A popular metric in statistics and machine learning is the Mean Squared Error (MSE). It measures the square root of the average discrepancies between a dataset's actual values and projected values [8]. MSE is frequently utilized in regression issues and is used to assess how well predictive models work. For a dataset containing 'n' data points, the MSE calculation formula is:
$MSE=n1\sum i=1n(xi-yi)2$
Where:

- xi represents the actual or observed value for the i-th data point.
- yi represents the predicted value for the i-th data point.

Root Mean Squared Error (RMSE)
Root Mean Squared Error (RMSE) is the square root of MSE, providing error in the same units as the target variable.

C. Cross validation

Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set [4].
Metric selection depends on application goals. For instance, in medical diagnostics, maximizing precision is often critical to minimize false positives, ensuring patient safety.

V. CHOOSING THE RIGHT ALGORITHM FOR SPECIFIC APPLICATIONS

Selecting the right supervised learning algorithm is essential to successfully solving certain real-world issues. Knowing the application domain and the particulars of the data involved is essential for this approach [18]. Here, we provide a methodical process for selecting the best algorithm depending on different application cases.

A. Medical Diagnostics

The necessity for interpretation and efficiency in healthcare diagnosis influences the algorithm's selection. When forecasting the existence of a disease through the symptoms of a person and medical record, for example, logistic regression is frequently used for binary categorization problems. Because of its comprehensibility and clarity, it may be used in therapeutic settings where it is essential to understand the model's judgements. Support Vector Machines (SVM), which are capable of analyzing high-dimensional medical data and successfully spotting patterns that differentiate between various health issues, may be used for increasingly complicated datasets [4] [19].
Furthermore, Random Forests are useful for medical analytics because of their resilience and capacity to manage missing data, which enable them to produce accurate predictions for a range of patient groups.

B. Financial Modelling

The type of data and the particular indicators of financial performance being examined should be taken into consideration when choosing algorithms for economic analysis. Using prior information for predicting forthcoming patterns, linear regression is frequently utilized when predicting constant outcomes like price movements in stocks or financial indicators [10].
Because Random Forests can handle big datasets and spot irregularities in transactional trends, they are useful for detecting fraud. Furthermore, gradient boosting methods are excellent at managing complicated datasets with non-linear correlations, which makes them appropriate for sophisticated financial models wherein prediction accuracy is crucial.

C. Text Analysis

Algorithms that are capable of handling sparse and highly dimensional information are necessary for analyzing text operations. Because of the convenience of use and efficiency in handling massive amounts of text, naïve Bayes classifiers are often employed for textual classification applications like sentiment assessment and identifying spam. Support Vector Machines are beneficial in advanced text detection jobs as they possess the capacity to establish intricate decision boundaries that enhance classification accuracy across a variety of text corpora [18].

D. Image Recognition

When choosing an algorithm for image recognition tasks, accuracy and computing efficiency must be taken into account. Applications like facial recognition can benefit from the effectiveness of Support Vector Machines with specialized kernels in categorizing pictures in high-dimensional areas [13]. Furthermore, Random Forests may be used to handle the enormous datasets that are common in image recognition applications. They achieve this by combining predictions from several decision trees, which improves overall accuracy while reducing overfitting.

Selecting the best-supervised learning algorithm necessitates a thorough comprehension of the particular application environment and data properties. Individuals may improve model performance and obtain more dependable results in a variety of areas by tailoring algorithm choices to the requirements of every application, whether it be financial models, healthcare diagnoses, analyzing text, or picture identification. This based on applications strategy guarantees that the chosen algorithms not only satisfy technical specifications but also tackle real-world practical difficulties [20].

VI. CHALLENGES IN ALGORITHM SELECTION

Choosing the right algorithms for supervised learning is a complex task that calls for a thorough evaluation of several variables. The features of the database, the particular needs of the application, and the practical limitations that practitioners encounter all effect how efficient an algorithm is [4]. Some of the main difficulties in choosing an algorithm are listed below:

A. Size and Quality of Data

The amount and integrity of the dataset have a big impact on how well algorithms are chosen. Although big datasets might yield insightful information, they can also contribute noise that can cause algorithms to make mistakes. Training trustworthy models requires high-quality, well-annotated data; low-quality data might provide biased results and erroneous predictions. Furthermore, anomalies or missing variables might make the model training process more difficult, requiring careful data pretreatment to guarantee the best possible performance from the algorithms [14].

B. Requirements for Interpretability

When choosing algorithms, interpretability is becoming more and more crucial, especially in delicate industries like medical and financial. To maintain confidence and adhere to legal requirements, stakeholders frequently need clear disclosures of the way algorithms make judgements. In situations where comprehending how choices are made is essential, less complicated, easier-to-understand algorithms work best, as despite their excellent accuracy, complicated algorithms like deep learning usually lack transparency [18].

C. Limitations in Computation

The choice of algorithms is heavily influenced by computational capacity. Certain techniques, like Support Vector Machines (SVM), might be computationally challenging, which restricts their use in contexts having limited facilities or in enormous data sets. Application scenarios require effective algorithms that strike a compromise between performance and resource consumption, especially in situations wherein speed is crucial [16].

D. Hybrid Methods

By using the advantages of many algorithms, hybrid approaches—like ensemble techniques that combine SVM and Random Forests—can improve prediction accuracy. Although these techniques boost accuracy and

resilience, they may also make it more difficult to comprehend the model and raise processing requirements [4]. The difficulty is in successfully balancing these trade-offs to preserve comprehensibility and maximize model performance [14].

E. Accuracy and Speed in Balance

Accuracy and processing speed must be balanced in numerous scenarios. Certain models could be quite accurate, but they need a lot of time and computing power to train and interpret. On the other hand, simple algorithms could produce forecasts more quickly, yet at the expense of less accuracy. Based on the particular needs of the application and its functional limitations, this trade-off needs to be thoroughly evaluated [17]. These difficulties show how difficult it may be to choose the best-supervised learning algorithm for a particular job, highlighting the necessity of a systematic strategy that takes into account several variables affecting model performance and applicability [19].

VII.FUTURE ASPECTS AND CONCLUSION

Supervised learning has emerged a cornerstone of machine learning provide robust solutions for the real-world problems through its classifications and regression capability. Ability to train models with label datasets allows for precise predictions in medical diagnostics, spam detection, stock market forecasting etc. [3]
Several significant developments and advances which will improve the efficacy, effectiveness, and generalizability of machine learning algorithms in a variety of fields are going to impact the prospects of algorithm selection in supervised learning. The emergence of Automated Machine Learning (AutoML) tools is one noteworthy breakthrough. By automating model training, hyperparameter adjustment, and assessment, these tools are anticipated to streamline the method selection process. This will enable professionals to concentrate on strategic choices while allowing the non-experts to implement machine learning approaches more readily. Furthermore, as the need for transparency in AI systems grows, the incorporation of Explainable AI (XAI) methodologies will become crucial [1].

Future algorithms will probably include features that offer transparent information about how they make decisions, increasing stakeholder and user confidence, especially in crucial areas like banking and healthcare. Furthermore, the significance of ethical issues in algorithm selection will only increase, with an emphasis on creating algorithms that give morality, transparency, and bias avoidance top priority. By using the advantages of several algorithms and mitigating their shortcomings, hybrid and ensemble approaches can enhance the accuracy of predictions and resilience even further. Computational technology breakthroughs like edge computing, as well as quantum computing, have the potential to revolutionize algorithm scalability and efficiency, allowing for quicker processing speeds and real-time decision-making [20].

Finally, a trend towards continuous learning frameworks could develop, which would enable models to adjust in real-time in response to fresh data inputs, increasing their long-term usefulness. Taken together, these patterns point to a vibrant future for supervised learning that prioritizes automation, comprehension, and appropriate implementation in a range of applications. In conclusion, a dedication to innovation, openness, and ethical considerations will define the future of supervised learning. In addition to enhancing model performance, these advancements will guarantee that machine learning solutions are applied ethically and successfully across a range of applications, thereby advancing both technology and utility. The above are the advances in supervised machine learning approach and our study emphasized on how depending on the complexity of problem, type of dataset, and other factors play major role in choosing the appropriate algorithm.

REFERENCES:

[1] Y. B. P. K. &. S. O. Singh, "A review of studies on machine learning techniques.," International Journal of Computer Science and Security, vol. 1, no. 1, pp. 70-84, 2007.
[2] M. L. Y. &. J. S. K. Bkassiny, "A survey on machine-learning techniques in cognitive radios.," IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 1136-1159, 2012.
[3] S. B. Z. I. &. P. P. Kotsiantis, "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pp. 3-24, 2007.

[4] J. G. K. S. M. M. L. &. J. D. T. Greener, " A guide to machine learning for biologists," Nature reviews Molecular cell biology, vol. 23, no. 1, pp. 40-55, 2022.
[5] W. &. S. V. C. Maass, "Pairing conceptual modeling with machine learning. Data & Knowledge Engineering," vol. 134, p. 101909, 2021.
[6] G. W. D. H. T. T. R. &. T. J. James, " Linear regression. In An introduction to statistical learning: With applications in python," Cham: Springer International Publishing, pp. 69-134, 2023.
[7] L. &. D. K. Weisberg, "The intersection of equity pedagogy and technology integration in preservice teacher education: A scoping review," Journal of Teacher Education,, vol. 74, no. 4, pp. 327-342, 2023.
[8] P. &. L. J. Roback, "Beyond multiple linear regression: applied generalized linear models and multilevel models in R," Chapman and Hall/CRC, 2021.
[9] P. &. V. T. R. Schober, "Logistic regression in medical research," Anesthesia & Analgesia, vol. 132, no. 2, pp. 365-366, 2021.
[10] C. S. C. P. Y. S. &. M. M. Lee, " Predictive analytics in business analytics: decision tree," Advances in Decision Sciences,, vol. 26, no. 1, pp. 1-29, 2022.
[11] K. G. N. A. H. S. &. R. M. B. Maswadi, "Human activity classification using Decision Tree and Naive Bayes classifiers," Multimedia Tools and Applications, vol. 80, no. 14, pp. 21709-21726, 2021.
[12] R. Szostak, "The basic concepts classification. Advances in Knowledge Organization," vol. 13, no. 10.5771/0943-7444-2020-3-231. , pp. 24-30, 2012.
[13] A. B. K. M. A. V. R. &. S. D. Manoharan, "Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review.," Journal of Energy Storage, vol. 55, p. 105384, 2022.
[14] M. G. H. S. S. U. T. S. S. M. R. H. M. A. M. R. .. &. M. A. Abdolrasol, "Artificial neural networks based optimization techniques: A review. Electronics," vol. 10, no. 21, p. 2689, 2021.
[15] Y. Z. W. L. W. &. H. Y. Chen, " Large group activity security risk assessment and risk early warning based on random forest algorithm," Pattern Recognition Letters , vol. 144, pp. 1-5, 2021.
[16] A. A. O. A. M. R. E. &. H. O. Shehadeh, " Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. An evaluation of modified decision tree, LightGBM, and XGBoost regression.," Automation in Construction, vol. 129, no. 103827, 2021.
[17] M. Z. &. A. A. H. Naser, " Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences," Architecture, Structures and Construction, vol. 3, no. 4, pp. 499-517, 2023.
[18] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," SN computer science, vol. 2, no. 3, p. 160, 2021.
[19] S. K. F. T.-A. M. &. A. N. Nematzadeh, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases.," Computational biology and chemistry, vol. 97, no. 107619, 2022.
[20] S. &. D. R. Shurrab, "Self-supervised learning methods and applications in medical imaging analysis: A survey.," PeerJ Computer Science, vol. 8, no. e1045, 2022.