How to Cite:
Neethidevan, V., & Anand, S. . (2022). Implementing and evaluating the performance of various Machine Learning algorithms with different datasets. *International Journal of Health Sciences*, *6*(S1), 4684–4694. https://doi.org/10.53730/ijhs.v6nS1.5890

Implementing and evaluating the performance of various Machine learning algorithms with different datasets

V. Neethidevan
AP (SLG) MCA Department, Mepco Schlenk Engineering College, Sivakasi,

Dr. S. Anand, M.E., Ph.D.
Professor, Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi

***Abstract*---**Machine learning algorithms are used to train the machine to learn on its own and improve from experience. It involves building the mathematical models to help in understand the data. When these models are applied with tunable parameters to the observed data. Using this program can be considered to be learning from the data. Once the models learned enough from the data given as input, they could be used for predicting and understand different features of new data. The supervised learning involves modelling the relationship between measured features of data and some label associated with data. Once the model is trained with enough data and features, then new data can be given to the model for classification purpose. It is further classified into classification tasks and regression tasks. *Unsupervised learning* involves modelling the features of a dataset without reference to any label, and in this based on some similarity features data are grouped into some form. The similarity features are nothing but distance between the data is very minimum. These models include tasks such as *clustering* and *dimensionality reduction.* Clustering algorithms identify distinct groups of data, while dimensionality reduction algorithms search for more simple representations of the data. In this paper the main focus will be implementing and evaluating performance of different machine learning algorithms like, classification, clustering, Principal component analysis and Support vector machine with different data set implemented under google colab environment. Analyse the performance of various algorithms for their accuracy. In this, SVM support both linear and non linear problems and it outperforms all other algorithms.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
***Corresponding author**: Neethidevan, V.; Email: neethidevan@mepcoeng.ac.in*
Manuscript submitted: 27 Feb 2022, Manuscript revised: 18 March 2022, Accepted for publication: 09 April 2022

***Keywords*---**Machine learning, classification, clustering, dimensality reduction.

Introduction

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed. Two of the most widely adopted machine learning methods are **supervised learning** which trains algorithms based on example input and output data that is labeled by humans, and **unsupervised learning** which provides the algorithm with no labeled data in order to allow it to find structure within its input data.

Machine Learning and Big Data Challenges

Data Collection & Usage

When you collect data from different sources, the user must understand the various aspects of data such as, what are the types of data, purpose of data, whether the data is free from errors etc. To transform incoming data into value-added business insights, need to understand what kind of data you need and how you plan to use it.

Security

With so many people sharing their personal information and millions of bots generating even more online data, it's relatively easy to sway public opinion toward one or another decision.

Data Validation

When the data is collected from different sources, the reliability of data is more important. Each data is validated before applying to any model for various machine learning tasks.

Right Algorithms

Among the several machine learning models available, choosing the algorithm to perform specific tasks is a challenging one. Based on the fast experience and application requirements, right algorithm must be chosen.

Training Dataset

When training machine learning algorithm, need for a good and large training dataset, so that the algorithm can identify the major patterns, information, and insights. If the dataset is used is small, the results might be very biased

Data Noise

With respect to "Pareto's Principle" 80% of a data scientist's valuable time is spent simply finding, cleansing, and organizing data, leaving only 20% to perform analysis."

Literature survey

In [1], the authors applied the various algorithms on survival prediction were compared using the Surveillance, Epidemiology, and End Results (SEER) breast cancer data set. Used machine learning algorithms are: Naive Bayes, J48, Multiobjective Evolutionary Fuzzy Classifier (MEFC), Support Vector Machines (SVM). The most successful algorithm is J48 algorithm. The most successful algorithm is J48 algorithm according to the tests. . In [2], theauthors worked on

the automatic identification of handwritten numbers through computers, and it has a greater application prospect in letter postal identification and financial statements and bank bill processing. They have taken the MNIST handwritten digit database as samples, discusses algorithms KNN, SVM, BP neural network, CNN and their application in handwritten digit recognition. In the training process, the implementation of system is done using KNN with Python, SVM with scikit-learn library, and BP, CNN with Tensorflow, and fine-tunes the algorithm parameters to get the best results for each algorithm. Finally, by comparing the recognition rate and recognition duration of the four algorithms, the advantages and disadvantages of the four algorithms in handwriting recognition are analyzed**.** In [3], the authors, proposed an improved crow search algorithm to optimize the extreme learning machine. Also, a particle swarm algorithm search strategy is proposed to enhance the global search capability. In the latter part of the algorithm iteration, Gaussian function is added, and the penalty coefficient of the function is used for local disturbance, gradually reducing the amplitude of the search trajectory, and then adaptively adjusting the parameters to avoid being attracted by local extremum. Finally, the improved crow search algorithm is used to optimize the hidden layer neurons and connection weights of the extreme learning machine neural network, so as to obtain accurate prediction results. Through function fitting, regression data set fitting and classification data set for classification experiment verification, the proposed algorithm has higher training speed and efficiency. At the same time, this method is not only significantly higher than the traditional ELM method, but also obtains a more compact network structure, which is an effective neural network optimization algorithm.In [4], the authors aimed at the following: first, to reveal the possible features for determining the mode of childbirth, and second, to explore machine learning algorithms by considering the best possible features for predicting the mode of childbirth (vaginal birth, cesarean birth, emergency cesarean, vacuum extraction, or forceps delivery). An empirical study conducted, to explore the relevant features for predicting the mode of childbirth, while five different machine learning algorithms were explored to identify the most significant algorithm for prediction based on 6157 birth records and a minimum set of features. The research revealed 32 features that were suitable for predicting modes of childbirth and categorized the features into different groups based on their importance. Various models were developed, with stacking classification (SC) producing the highest f1 score (97.9%) and random forest (RF) performing almost as well (f1-score = 97.3%), followed by k-nearest neighbors (KNN; f1-score = 95.8%), decision tree (DT; f1-score = 93.2%), and support vector machine (SVM; f1-score = 88.6%) techniques, considering all (n = 32) features. In [5], the authors developed an approach to distinguish, legitimate supplicant from an attacker on the basis of the characteristics of a set of parallel wireless channels, which are affected by time-varying fading. They assessed and compare the performance achieved by different approaches under different channel conditions. Then they used classification methods based on machine learning. They resorted to more conventional binary classifiers, considering the cases in which such messages are either labelled or not. For the latter case, they used clustering algorithms to label the training set. The performance of both nearest neighbor (NN) and support vector machine (SVM) classification techniques is evaluated. In [6], the authors proposed novel meta-algorithm SafePredict, that works with any base prediction algorithm for online data to guarantee an arbitrarily chosen correctness rate, 1 -

ε, by allowing refusals. Allowing refusals means that the meta-algorithm may refuse to emit a prediction produced by the base algorithm so that the error rate on non-refused predictions does not exceed ε. When the base predictor happens not to exceed the target error rate ε, SafePredict refuses only a finite number of times. When the error rate of the base predictor changes through time SafePredict makes use of a weight-shifting heuristic that adapts to these changes without knowing when the changes occur yet still maintains the correctness guarantee. Empirical results show that (i) SafePredict compares favorably with state-of-the-art confidence-based refusal mechanisms which fail to offer robust error guarantees; and (ii) combining SafePredict with such refusal mechanisms can in many cases further reduce the number of refusals. IN [7], Photovoltaic (PV) systems operating in the outdoor environment are vulnerable to various factors, especially dust impact. Authors analysed I-V characteristics of PV strings under various fault states and in soiling condition.The new algorithm can diagnose PV faults using a small amount of simulated labeled data and historical unlabeled data, which greatly reduces labor cost and time-consuming. Moreover, the monitoring of dust accumulation can warn power plant owners to clean PV modules in time and increase the power generation benefits. PV systems of 3.51 and 3.9 kWp are used to verify the proposed diagnosis method. Both numerical simulations and experimental results show the accuracy and reliability of the proposed PV diagnostic technology.In [8], The main focus of this study is to explore the possibilities for prediction of energy prices using various machine learning (ML) algorithms. Exploring the possibilities of predicting the energy price in the open electricity market using four different algorithms namely: Simple Linear Regression, Support Vector Machines (SVM), K nearest neighbor, and Long Short-Term Memory. The main contribution of this work is to develop an ML system that can predict future prices. Realtime data are obtained from the Indian Energy Exchange (IEX) which handles around 30% of energy transactions through online within India under open access. The results are validated from the same which ensures the proper validation of the proposed model. The four models on the Indian Energy Exchange dataset are trained and the results are compared to find the best algorithm with the highest accuracy

Experimental Study of different algorithms
Classification

The data set used for classification as follows.
The dataset contains different flowers, like flower_photo, daisy, dandelion, roses, sunflower, tulips. How to classify images of flowers. It creates an image classifier using a keras.Sequential model, and loads data using preprocessing.image_dataset_from_directory. The following concepts were used in this model.

- Efficiently loading a dataset off disk.
- Identifying overfitting and applying techniques to mitigate it, including data augmentation and Dropout.

Steps involved in a basic machine learning workflow:

1. Examine and understand data
2. Build an input pipeline
3. Build the model
4. Train the model

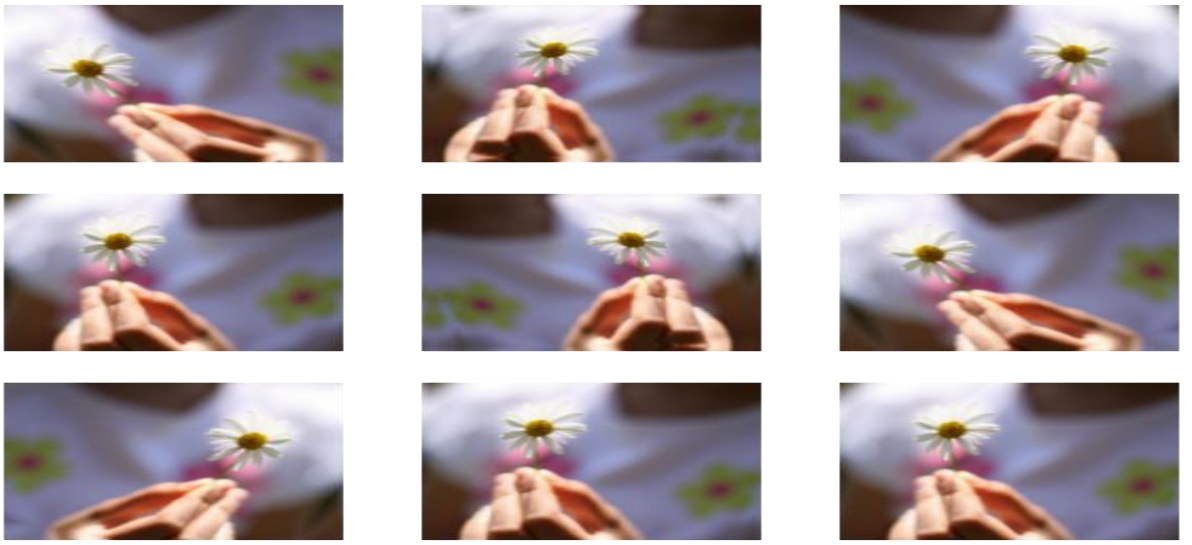
5. Test the model
6. Improve the model and repeat the process

Output generated from the system

Model: "sequential"

Layer (type)	Output Shape	Param #
rescaling_1 (Rescaling)	(None, 180, 180, 3)	0
conv2d (Conv2D)	(None, 180, 180, 16)	448
max_pooling2d (MaxPooling2D)	(None, 90, 90, 16)	0
conv2d_1 (Conv2D)	(None, 90, 90, 32)	4640
max_pooling2d_1 (MaxPooling2	(None, 45, 45, 32)	0
conv2d_2 (Conv2D)	(None, 45, 45, 64)	18496
max_pooling2d_2 (MaxPooling2	(None, 22, 22, 64)	0
flatten (Flatten)	(None, 30976)	0
dense (Dense)	(None, 128)	3965056
dense_1 (Dense)	(None, 5)	645

Total params: 3,989,285
Trainable params: 3,989,285
Non-trainable params: 0



Clustering

The *k*-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional dataset. It accomplishes this using a simple conception of what the optimal clustering looks like:

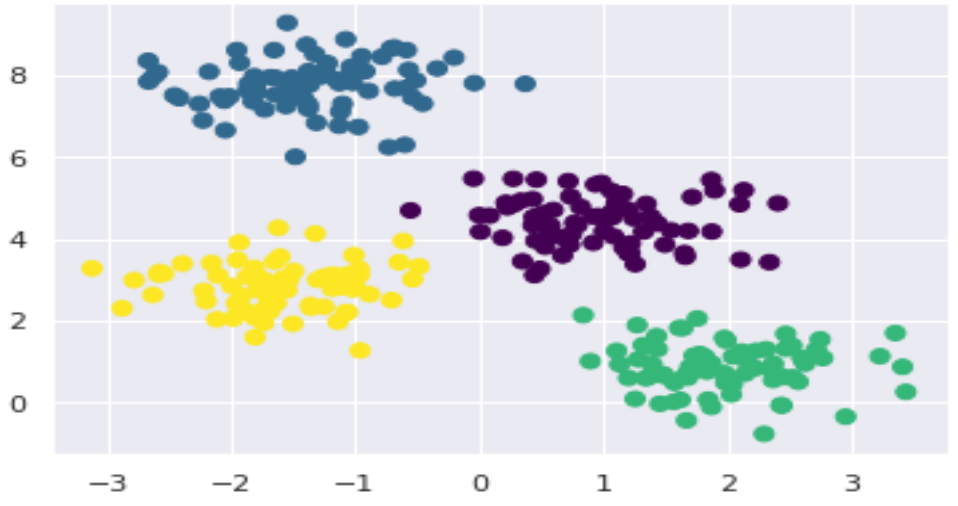
- The "cluster center" is the arithmetic mean of all the points belonging to the cluster.
- Each point is closer to its own cluster center than to other cluster centers.

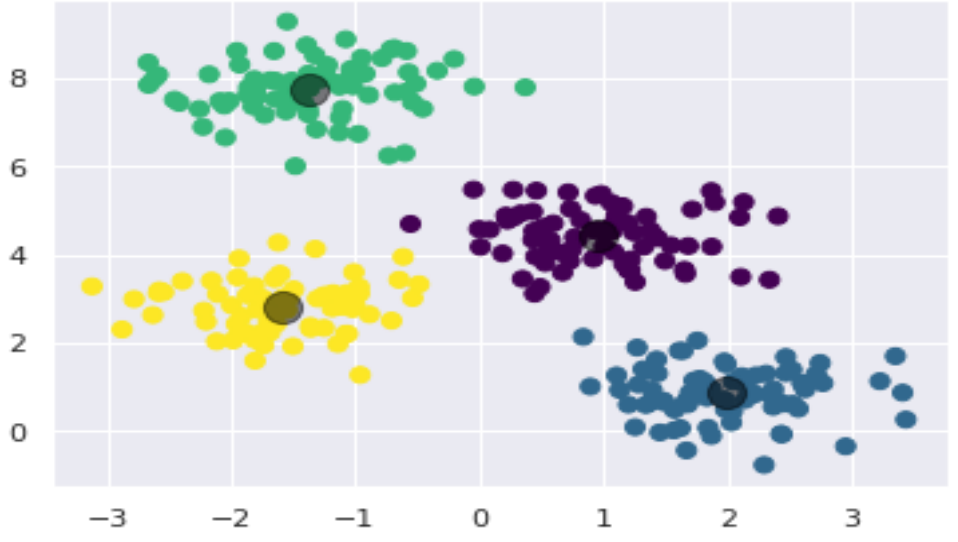
Expectation–maximization (E–M) is a powerful algorithm that comes up in a variety of contexts within data science. *k*-means is a particularly simple and easy-to-understand application of the algorithm, and we will walk through it briefly here. In short, the expectation–maximization approach here consists of the following procedure:

1. Guess some cluster centers
2. Repeat until converged
    a) *E-Step*: assign points to the nearest cluster center
    b) *M-Step*: set the cluster centers to the mean

Here the "E-step" or "Expectation step" is so-named because it involves updating our expectation of which cluster each point belongs to. The "M-step" or "Maximization step" is so-named because it involves maximizing some fitness function that defines the location of the cluster centers—in this case, that maximization is accomplished by taking a simple mean of the data in each cluster.

The literature about this algorithm is vast, but can be summarized as follows: under typical circumstances, each repetition of the E-step and M-step will always result in a better estimate of the cluster characteristics.

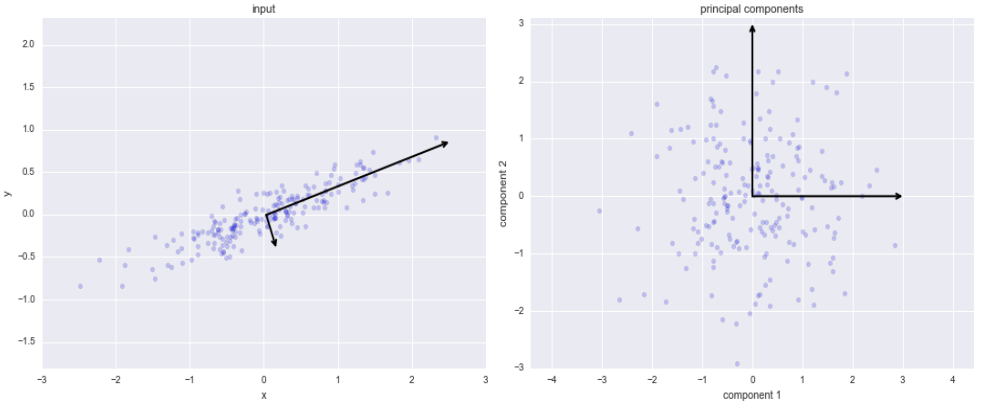




Principal Component analysis

One of the most broadly used of unsupervised algorithms, principal component analysis (PCA). PCA is fundamentally a dimensionality reduction algorithm, but it can also be useful as a tool for visualization, for noise filtering, for feature extraction and engineering, and much more. Principal component analysis is a fast and flexible unsupervised method for dimensionality reduction in data. In principal component analysis, this relationship is quantified by finding a list of the *principal axes* in the data, and using those axes to describe the dataset.

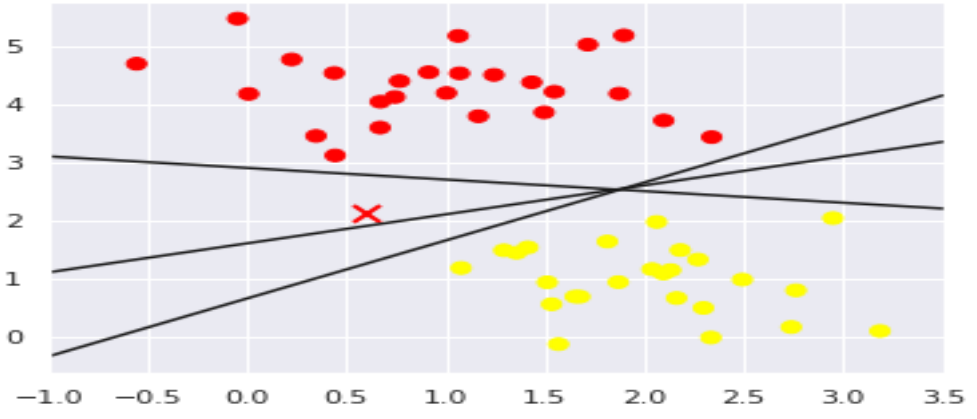
It is clear that there is a nearly linear relationship between the x and y variables.

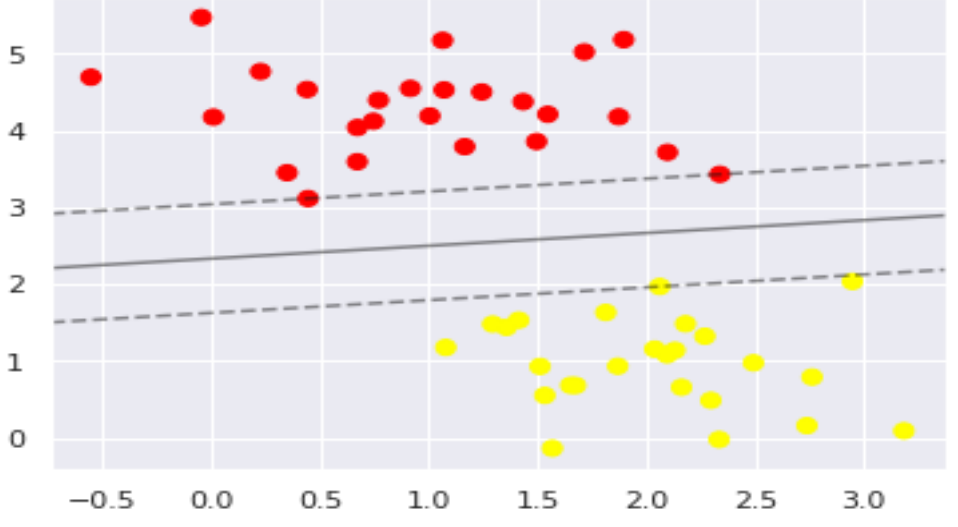


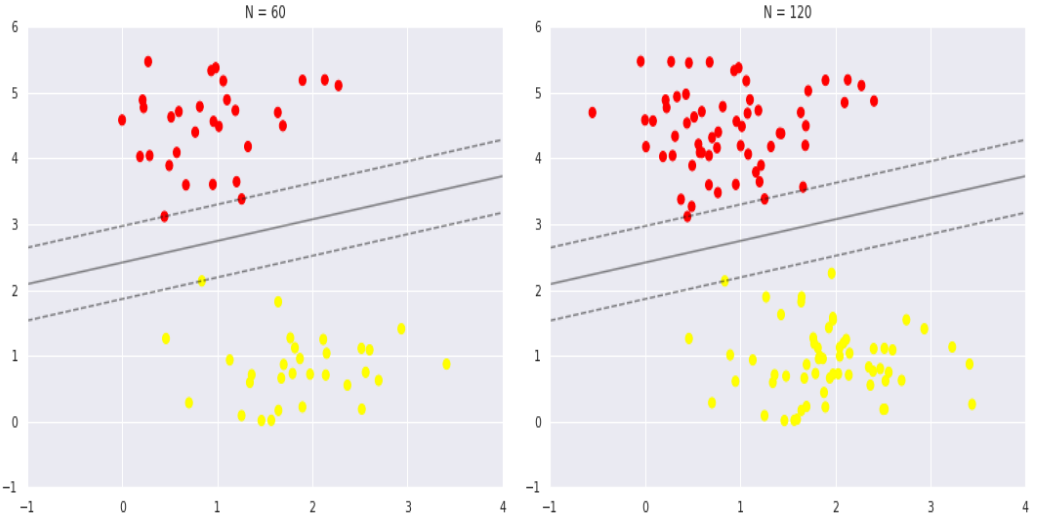
Support vector machine

Support vector machines (SVMs) are a particularly powerful and flexible class of supervised algorithms for both classification and regression. A linear discriminative classifier would attempt to draw a straight line separating the two sets of data, and thereby create a model for classification. For two dimensional data like that shown here, this is a task we could do by hand. But immediately we see a problem: there is more than one possible dividing line that can perfectly discriminate between the two classes!

These are three *very* different separators which, nevertheless, perfectly discriminate between these samples. Depending on which you choose, a new data point (e.g., the one marked by the "X" in this plot) will be assigned a different label! Evidently our simple intuition of "drawing a line between classes" is not enough, and we need to think a bit deeper.


5
4
3
2
1
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


5
4
3
2
1
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0



Conclusion

Thus various Machine learning algorithms were implemented using google colab. The classification algorithm is applied for flowers data set and it classified the flowers in to, like flower_photo, daisy, dandelion, roses, sunflower ,tulips. In clustering concepts, k-means clustering is applied and 4 different clusters are formed. The principal component analysis, is a dimensionality reduction method, tool for visualization, for noise filtering, for feature extraction. The support vector machine is a most important algorithms used for classifying both Linear and Non-Linear problems and it is considered as a best machine learning algorithm and finally concluded that SVM outperforms other machine learning algorithms.

References

1. G. Y. Özkan and S. Y. Gündüz, "Comparision of Classification Algorithims for Survival of Breast Cancer Patients," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), 2020, pp. 1-4, doi: 10.1109/ASYU50717.2020.9259846.
2. W. Liu, J. Wei and Q. Meng, "Comparisions on KNN, SVM, BP and the CNN for Handwritten Digit Recognition," 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications(AEECA), 2020, pp. 587-590, doi: 10.1109/AEECA49918.2020.9213482.
3. L. Cao, Y. Yue, Y. Zhang and Y. Cai, "Improved Crow Search Algorithm Optimized Extreme Learning Machine Based on Classification Algorithm and Application," in IEEE Access, vol. 9, pp. 20051-20066, 2021, doi: 10.1109/ACCESS.2021.3054799.
4. M. N. Islam, T. Mahmud, N. I. Khan, S. N. Mustafina and A. K. M. N. Islam, "Exploring Machine Learning Algorithms to Find the Best Features for Predicting Modes of Childbirth," in IEEE Access, vol. 9, pp. 1680-1692, 2021, doi: 10.1109/ACCESS.2020.3045469.
5. L. Senigagliesi, M. Baldi and E. Gambi, "Comparison of Statistical and Machine Learning Techniques for Physical Layer Authentication," in IEEE

Transactions on Information Forensics and Security, vol. 16, pp. 1506-1521, 2021, doi: 10.1109/TIFS.2020.3033454.
6. M. A. Kocak, D. Ramirez, E. Erkip and D. E. Shasha, "SafePredict: A Meta-Algorithm for Machine Learning That Uses Refusals to Guarantee Correctness," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 2, pp. 663-678, 1 Feb. 2021, doi: 10.1109/TPAMI.2019.2932415.
7. J. Huang, R. Wai and G. Yang, "Design of Hybrid Artificial Bee Colony Algorithm and Semi-Supervised Extreme Learning Machine for PV Fault Diagnoses by Considering Dust Impact," in IEEE Transactions on Power Electronics, vol. 35, no. 7, pp. 7086-7099, July 2020, doi: 10.1109/TPEL.2019.2956812.
8. P. Chaudhury, A. Tyagi and P. K. Shanmugam, "Comparison of Various Machine Learning Algorithms for Predicting Energy Price in Open Electricity Market," 2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE), 2020, pp. 1-7, doi: 10.1109/ICUE49301.2020.9307100.
9. Loganathan. N, Lakshmi. K, Chandrasekaran. N, Cibisakaravarthi.R. S, Priyanga.H .R and Varthini.H.K, "Smart Stick for Blind People," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 65-67, doi: 10.1109/ICACCS48705.2020.9074374.
10. Anantharajan, Shenbagarajan & Gunasekaran, Shenbagalakshmi. (2021). Automated brain tumor detection and classification using weighted fuzzy clustering algorithm, deep auto encoder with barnacle mating algorithm and random forest classifier techniques. International Journal of Imaging Systems and Technology. 31. 10.1002/ima.22582.