INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Making a Better Logistic Regression Model Using Ensemble Learning

Sneha Verma[1], Mr. Aman Kumar Sharma[2]
Department of Computer Science, Himachal Pradesh University, India

Abstract: Machine learning is a hot research topic in today's society and an important direction in the field of image classification research. Image classification is a supervised learning method used to classify images. There are many image classification algorithms that perform differently in different conditions. Paper focuses on improving the performance of image classification algorithms by using ensemble learning. Paper analyses the performance of logistic regression model and support vector machine model and then uses ensemble learning to increase the performance of logistic regression model. In the research work, both theoretical and empirical approaches were followed. For the theoretical approach a review of both secondary data as well as data based on results obtained by application on the tools is studied. Secondary data was acquired from the research articles, text books, journals, technical reports, published thesis, websites, e-journals, software tool manuals, conference proceedings and any other research articles published in the related domain. The empirical study was carried out on the set of experiments, using software tools. The results obtained from the experiments were analyzed for the finding of the research. The paper compares the results of three algorithms when tested on same dataset, in same environment and on same system. The final results of the research prove that ensemble techniques give best results and this type of learning is effective.
Keywords: Ensemble Learning, Logistic Regression, Support Vector Machine, Artificial Neural Network, Convolution Neural Network.

I. INTRODUCTION

The recent surge in machine learning research has led to its widespread adoption in various applications, such as text mining, spam detection, video recommendation, image classification, and multimedia concept retrieval.[1] Image classification is an integral part of machine learning.[2] At first instance image classification seems simple i.e. assigning labels to images based on existing already labelled images. But in actual it requires meticulous pixel-level analysis to ensure accurate categorization. This analysis provides valuable insights that drive informed decisions. However, the accuracy of the entire process depends on the quality of training data. Data labelling should be completed accurately in the training phase to avoid discrepancies in the data. In order to facilitate accurate data labelling, publicly available datasets are often used in the model's training phase.[3]There are many image classification algorithms. Different models performs differently in different problems.[4] The performance of these image classification models can be improved using ensemble learning. The core idea behind the need of ensemble learning is that instead of relying on a single model, ensemble methods combine predictions from multiple models to get a more robust and accurate final result. In this paper accuracies of four image classification algorithms will be analysed and main task will be to increase the accuracy of any one algorithm using ensemble learning. Four algorithms used for the purpose are convolution neural network, support vector machine, artificial neural network and logistic regression.

II. TAXONOMY

Let's discuss the four image classification algorithms in brief:

A. Logistic Regression (LR)

Imagine you have a bunch of data and want to predict if something falls into two categories, like "cat" or "not cat" in a picture. Logistic regression is a tool that helps with this. Instead of just saying "cat" or "not cat," it gives you a score between 0 and 1 that shows how likely it is something belongs in that category. The higher the score (closer to 1), the more likely it is a cat. The lower the score (closer to 0), the less likely it is a cat. Therefore logistic regression may be defined as a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.[5] In the 19th century, French mathematician Pierre François Verhulst developed a mathematical formula, the logistic function, to model the growth of human population.

This function has also been applied to understand the behaviour of autocatalytic chemical reactions.[6] Logistic Regression is an extension of linear Regression. Linear regression use a straight line to predict things. The output for Linear Regression must be a continuous value, such as price, age, etc. But in many the predictions are yes or no answers, like having a disease or not. In such cases, a straight line wouldn't work well. In such cases it is not idle to use simple linear regression, because, when the outcome variable is not continuous the assumptions of linear regression i.e., linearity, normality and continuity are violated. To deal with categorical outcome variables, logistic regression was introduced as an alternative to simple linear regression.[7] The biggest difference between logistic regression and linear regression is that linear regression predicts a continuous outcome variable, whereas logistic regression predicts a categorical outcome variable.[8] Unlike linear regression, in logistic regression, we fit a "S" shaped logistic function, instead of a regression line. Linear Regression solves regression problems, while logistic regression solves classification problems. Logistic regression can be of different types, such as binomial (binary), multinomial, or ordinal depending on the nature of outcome variable.

1) When the outcome variable can have only two categories, (e.g., disease present vs. disease absent, dead vs. alive) then binomial or binary logistic regression is used.
2) If the outcome variable have more than two categories (e.g., drug A, drug B, and drug C) which are not ordered then multinomial logistic regression is used.
3) If the outcome variable is ordered (e.g., poor, fair, good, very good, excellent) then ordinal logistic regression is used.[7]

B. Artificial Neural Network (ANN)

For centuries, people have been curious about how the brain works. With the rise of modern electronics, scientists wanted to see if they could create something similar. The first step toward artificial neural networks came in 1943 when Warren McCulloch a neurophysiologist, and a young mathematician Walter Pitts, wrote a paper on how neurons might work. They even built a simple model of neural network using electrical circuits.[9] Hence, Artificial Neural Network (ANN) may be defined as a computational model inspired by the human brain's neural structure. It consists of interconnected nodes (neurons) organized into layers. Information flows through these nodes, and the network adjusts the connection strengths (weights) during training to learn from data, enabling it to recognize patterns, make predictions, and solve various tasks in machine learning and artificial intelligence.[10] Artificial neural networks (ANNs) are really good at finding patterns. This is useful because many things in the world involve patterns, like recognizing faces or figuring out if an email is spam. ANNs can also handle messy information. Imagine a picture that's a little blurry or a sentence with a typo, ANNs can still make sense of it all and come to a good decision.[11] Due to the recent surge in computational capabilities, artificial neural networks have become widely prevalent in modern machine learning, marking a significant shift towards their extensive adoption across various domains.[12] Few areas that uses artificial neural network are: Medical diagnosis and health care, facial recognition, behaviour of social media users is analysed using ANN, stock market forecasting, weather forecasting, robotics and dynamics, etc.

C. Support Vector Machines (SVM)

Support vector machine was developed in 1990s by Vladimir N. Vapnik and his colleagues. In 1995, they published their work in a paper titled "Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing".[13] A support vector machine (SVM) is a machine learning algorithm that uses supervised learning models to solve complex classification, regression, and outlier detection problems. Support vector machines solves these problems by performing optimal data transformations that determine boundaries between data points based on predefined classes, labels, or outputs.[14] Support Vector Machines have been developed in the framework of Statistical Learning Theory.[15] Traditional statistics study the situation when the amount of samples tends to be infinite. However, the amount of samples in our daily lives is usually limited. Different from traditional statistics, Statistical Learning Theory(SLT) is a theory that specializes in studying the laws of machine learning in the case of small samples.

SLT provides a new framework in dealing with the general learning problem.[16] The support vector machine is a novel small-sample learning method, because it is based on the principle of structural risk minimization, rather than the traditional empirical risk minimization principle, it is superior to existing methods on many performances.[17] Support vector machines, being computationally powerful tools for supervised learning, are widely used in classification, clustering and regression problems. SVMs have been successfully applied to a variety of real-world problems like particle identification, face recognition, text categorization, bioinformatics, civil engineering and electrical engineering etc.[18]

D. Convolution Neural Network (CNN)

In the field of deep learning, the CNN is the most famous and commonly employed algorithm.[1] The first ever convolution neural network (CNN) was LeNet. LeNet was proposed by Yann LeCun, in 1998. It was used mainly to recognize digits and handwritten numbers on bank checks.[19] Convolutional Neural Network is an extension of Artificial Neural Network. Both Convolutional Neural Networks (CNNs) and traditional Artificial Neural Networks (ANNs) consist of neurons that autonomously refine their functionality through the process of learning.[20] Convolutional neural networks (CNNs) have significantly contributed to the rise in popularity of deep learning. The key advantage of CNNs over previous models is that it does everything automatically and without human supervision, making it the most popular model. CNNs automatically learn a hierarchy of features that are crucial for classification tasks, eliminating the need for manual feature engineering. This process involves iteratively convolving the input image with learned filters to construct a hierarchy of feature maps. As a result of this hierarchical approach, higher layers in CNNs can learn complex features that are invariant to distortions and translations, enhancing their effectiveness in various tasks.[21] Generally, a CNN consists of three main neural layers, which are convolutional layers, pooling layers, and fully connected layers. These different kinds of layers play different roles.[19] The first two, convolution and pooling layers, perform feature extraction, whereas the third, a fully connected layer, maps the extracted features into final output, such as classification. [22]

1) Feature Extraction
- Input: Images are the most common input format for convolutional neural network.[21]
- Convolution: It applies learnable filters, or kernels, to the input data, which are small grids that slide over the image to extract specific features like edges, textures, or patterns, resulting in feature maps. Multiple filters are used to capture different features. The output of this layer is called feature map.[23]
- Pooling: Pooling is the process where the output of the convolutional layer is down sampled and its key features are extracted. Pooling reduces the dimensionality of the image representation.[24]

2) Classification
- Fully Connected Layer: The fully connected layer is the layer where actual image classification happens. This classification is based on the features extracted in the previous layers.[25] A fully connected layer predict the most suitable label for the given image. This involves flattening the output from the previous layers. The feature maps are typically flattened into a one-dimensional vector. This is done to match the dimensionality between the convolutional/pooling layers and the fully connected layers.[23] Here, fully connected means that all the inputs or nodes from one layer are connected to every activation unit or node of the next layer.[25]
- Output: Finally the classified and extracted images are given as output.

E. Ensemble Learning

Machine learning is a rapidly growing and is a vital field in today's world. Research indicates that traditional machine learning algorithms often exhibit suboptimal performance when trained on imbalanced datasets. Therefore, researchers are exploring innovative learning methodologies like ensemble techniques and deep learning, to address this challenge. Both ensemble learning and deep learning have dominated the machine learning domain.[26] Ensemble learning is a technique that combines multiple machine learning algorithms to produce one optimal predictive model with reduced variance (using bagging), bias (using boosting) and improved predictions (using stacking). In simple words ensemble learning is a process where multiple base models (weak learners) are combined and trained to solve the same problem. This method is based on the concept that weak learner alone performs task poorly but when combined with other weak learners, they form a strong learner and these ensemble models produce more accurate results.[27] Ensemble methods are ideal for regression and classification, where they reduce bias and variance to boost the accuracy of models.[28] The most popular ensemble methods are:

1) Boosting: it is an ensemble technique that learns from previous predictor mistakes ands enhance future predictions. Thus, it improves model's performance.
2) Bagging: it is an ensemble learning technique that enhances the performance and accuracy of machine learning algorithms by reducing variance and preventing overfitting. Bagging involves training multiple models independently on random subsets of the data and then aggregating their predictions through voting or averaging.[28]
3) Stacking: it is an ensemble learning technique that combines the predictions of multiple machine learning models to achieve superior performance compared to individual models. By doing so the strengths of several different learning algorithms are harnessed by the training algorithm.

III. ENVIRONMENT

This experiment is written on google colab using python language. Colab is a hosted Jupyter Notebook service that requires no setup to use and provides free access to computing resources, including GPUs and TPUs. Colab is especially well suited to machine learning, data science, and education.[29] Experiment is conducted using a 64 bit windows 10 system, the processor is Intel(R) Core(TM) i3-7020U CPU @ 2.30GHz and the memory is 8.00 GB RAM. The dataset used for the purpose is taken from the following website: https://www.kaggle.com/ on 1st March 2024 at 05:00 PM. The dataset used contains binary images only. The size of the dataset is 62 MB and it comprises of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centred and occupies about the same amount of space in each image.There are seven categories of emotions in the dataset (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5= Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples.[30]

IV. RESULTS AND ANALYSIS

The research paper[31] "Comparative Analysis of Image Classification Algorithms," was used for analysing the accuracies of four image classification algorithms, mainly convolution neural network, support vector machine, artificial neural network and logistic regression. Given below are the findings of the research paper[31] "Comparative Analysis of Image Classification Algorithms".

The accuracies of four image classification algorithms were as follows:

- Convolution Neural Network: 64.25%
- Support Vector Machine: 52.42%
- Artificial Neural Network: 35.11%
- Logistic Regression: 32.20%

The current research focuses on increasing the accuracy of logistic regression model by adding the features of support vector machine model. This process of improving the results by combining multiple models instead of using a single model is termed as ensemble method. The dataset and environment used in the current experiment are same as used in the research paper "Comparative Analysis of Image Classification Algorithms".

A. Dataset

The dataset consists of seven categories of emotions (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5= Surprise, 6=Neutral). The training set consists of 28,709 examples and the public test set consists of 3,589 examples. The bar graph shown in figure 1 displays the distribution of emotion classes in the training and testing sets.
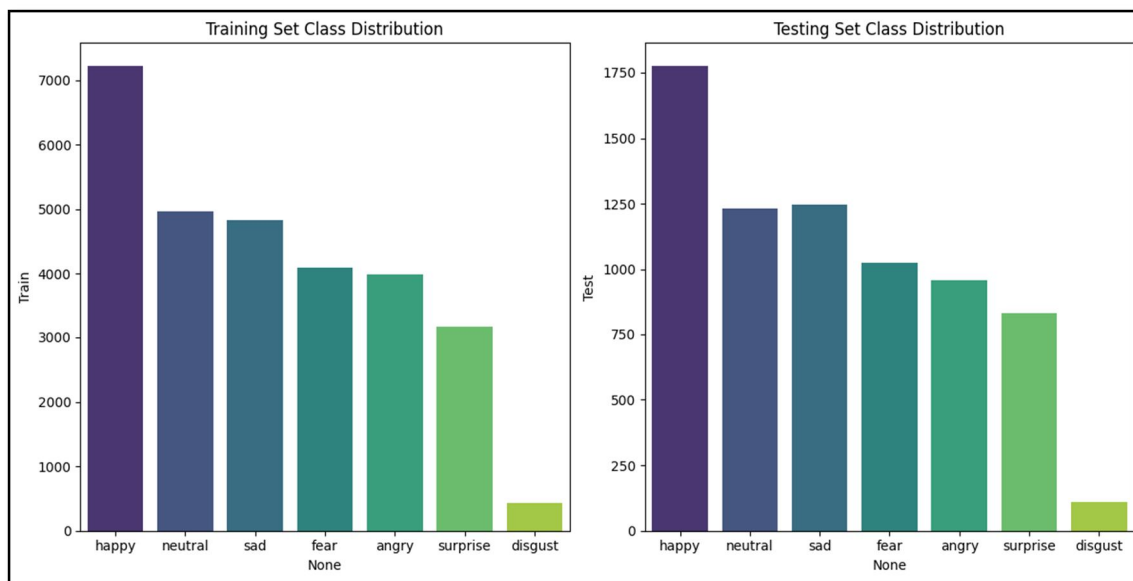

Fig.1 Distribution Of Emotion Classes in Training and Testing Sets.

B. Confusion Matrix

A confusion matrix is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.[32] A confusion matrix shows the actual and predicted counts of each category of emotion.
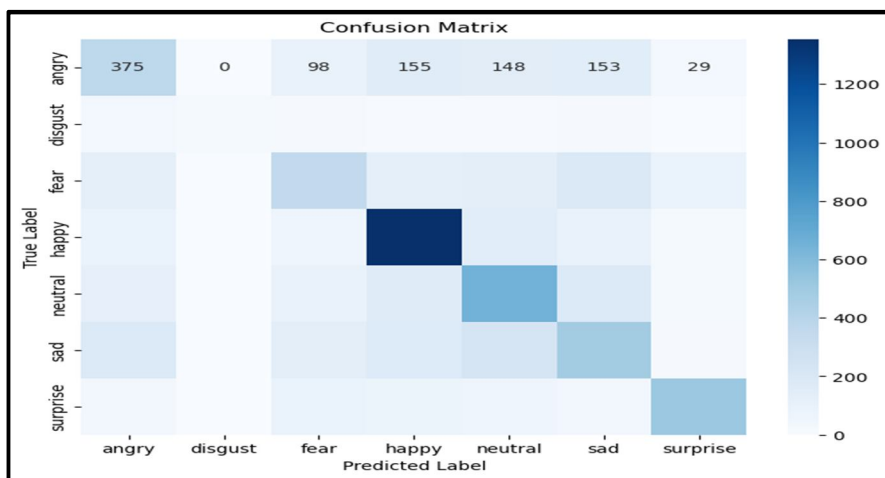

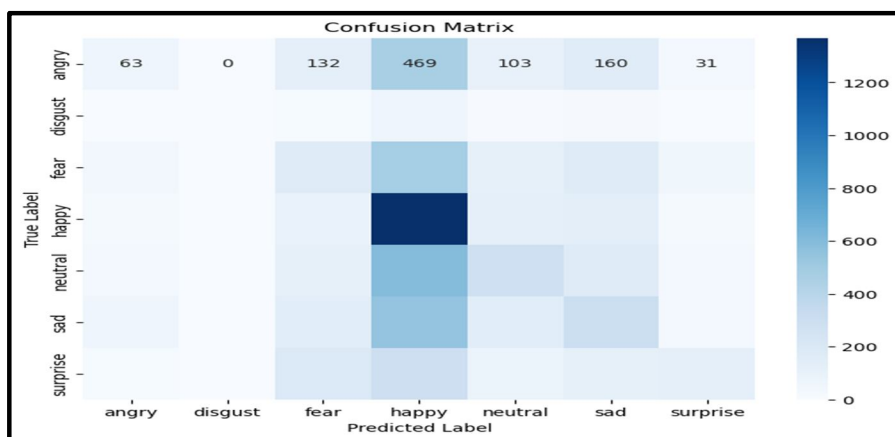Fig.2 Support Vector Machine Confusion Matrix

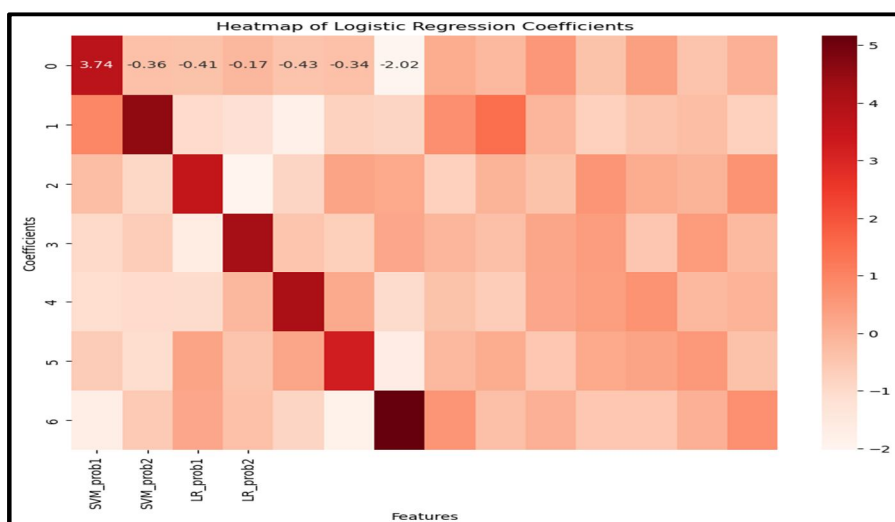Fig.3 Logistic Regression Confusion Matrix

Fig.4 Confusion Matrix of Combine Model (Logistic Regression & Support Vector Machine)

C. Classification Report

The classification report visualizer displays the precision, recall, F1, and support scores for the model. In order to support easier interpretation and problem detection, the report integrates numerical scores with a color-coded heatmap. All heatmaps are in the range (0.0, 1.0) to facilitate easy comparison of classification models across different classification reports.

1) *Precision:* Precision can be seen as a measure of a classifier's exactness.
2) *Recall:* Recall is a measure of the classifier's completeness
3) *f1 score:* The F_1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0
4) *Support:* Support is the number of actual occurrences of the class in the specified dataset. Support doesn't change between models but instead diagnoses the evaluation process.[33]

TABLE 1

Classification Report for Support Vector Machine

	Precision	Recall	F1-Score	Support
0=Angry	0.39	0.39	0.39	958
1=Disgust	1.00	0.21	0.34	111
2=Fear	0.41	0.35	0.38	1024
3=Happy	0.66	0.76	0.71	1774
4=Sad	0.42	0.39	0.40	1247
5= Surprise	0.74	0.62	0.67	831
6=Neutral	0.47	0.53	0.49	1233
accuracy			0.52	7178
macro avg	0.58	0.46	0.48	7178
weighted avg	0.53	0.52	0.52	7178

TABLE 2

Classification Report for Logistic Regression

	Precision	Recall	F1-Score	Support
0=Angry	0.27	0.08	0.13	958
1=Disgust	1.00	0.00	0.00	111
2=Fear	0.19	0.18	0.19	1024
3=Happy	0.36	0.77	0.49	1774
4=Sad	0.33	0.22	0.26	1233
5= Surprise	0.30	0.21	0.25	1247
6=Neutral	0.41	0.17	0.24	831
accuracy			0.32	7178
macro avg	0.41	0.23	0.22	7178
weighted avg	0.32	0.32	0.28	7178

TABLE 3

Classification Report for Combine model of LR and SVM

	Precision	Recall	F1-Score	Support
0=Angry	0.48	0.40	0.44	958
1=Disgust	0.79	0.54	0.64	111
2=Fear	0.50	0.39	0.44	1024
3=Happy	0.62	0.77	0.69	1774
4=Sad	0.49	0.54	0.51	1233
5= Surprise	0.44	0.45	0.45	1247
6=Neutral	0.82	0.67	0.74	831
accuracy			0.56	7178
macro avg	0.59	0.54	0.56	7178
weighted avg	0.56	0.56	0.55	7178

TABLE 4

Final Accuracy of Algorithms

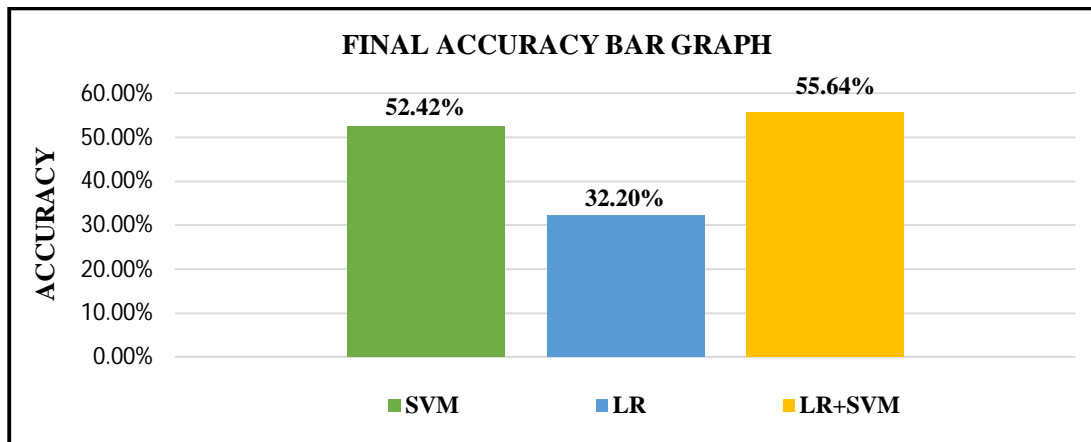| Algorithm	SVM	LR	LR+SVM
Accuracy	52.42 %	32.20 %	55.64 %

Fig.5 Algorithms Accuracy Comparison Bar Graph

V.   CONCLUSION

The research paper briefs about image classification algorithms and how using ensemble learning performance of image classification models can be improved. Paper mainly focuses on improving the accuracy of any one image classification model by using ensemble method. Ensemble techniques are widely used as they deliver better results by just combining two or more features of different models. In the paper logistic regression model and support vector machine model was used for the purpose. The environment and dataset used form implementation purpose was kept same for all the three algorithms. Three main algorithms used in the research paper were logistic regression, support vector machine and a combine model of both the algorithms. In the combine model features of support vector machine were added to logistic regression model. Aim of the research was to improve the performance of logistic regression by adding features of support vector machine model. The result of these image classification algorithms are expressed using confusion matrix, classification reports and graph. On the basis of these results conclusion was drawn out that combine model gave an accuracy of 55.64%. Second best results were given by support vector machines(52.42%) followed by logistic regression(32.20%).

REFERENCES

[1] L. Alzubaidi et al., "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," J Big Data, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[2] R. S. Chugh, V. Bhatia, K. Khanna, and V. Bhatia, "A comparative analysis of classifiers for image classification," in Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering, Institute of Electrical and Electronics Engineers Inc., Jan. 2020, pp. 248–253. doi: 10.1109/Confluence47617.2020.9058042.

[3] "What is image classification? Basics you need to know | SuperAnnotate." Accessed: Dec. 10, 2023. [Online]. Available: https://www.superannotate.com/blog/image-classification-basics

[4] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," Pattern Recognit Lett, vol. 141, pp. 61–67, Jan. 2021, doi: 10.1016/j.patrec.2020.07.042.

[5] "Logistic Regression: Equation, Assumptions, Types, and Best Practices." Accessed: Dec. 06, 2023. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/

[6] E. Y. Boateng, D. A. Abaye, E. Y. Boateng, and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," Journal of Data Analysis and Information Processing, vol. 7, no. 4, pp. 190–207, Sep. 2019, doi: 10.4236/JDAIP.2019.74012.

[7] T. Abedin, M. Ziaul, I. Chowdhury, A. Afzal, F. Yeasmin, and T. C. Turin, "Application of Binary Logistic Regression in Clinical Research Corresponding Author." [Online]. Available: https://www.researchgate.net/publication/320432727

[8] Dalian jiao tong da xue and Institute of Electrical and Electronics Engineers, Proceedings of IEEE 7th International Conference on Computer Science and Network Technology : ICCSNT 2019 : October 19-21, 2019, Dalian, China.

[9] S. B. Maind and P. Wankar, "International Journal on Recent and Innovation Trends in Computing and Communication Research Paper on Basic of Artificial Neural Network", [Online]. Available: http://www.ijritcc.org

[10] "Introduction to Artificial Neural Networks - Analytics Vidhya." Accessed: Dec. 06, 2023. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/

[11] E. Grossi and M. Buscema, "Introduction to artificial neural networks," European Journal of Gastroenterology and Hepatology, vol. 19, no. 12. pp. 1046–1054, Dec. 2007. doi: 10.1097/MEG.0b013e3282f198a0.

[12] A. Goel, A. K. Goel, and A. Kumar, "The role of artificial neural network and machine learning in utilizing spatial information," Spatial Information Research, vol. 31, no. 3. Springer Science and Business Media B.V., pp. 275–285, Jun. 01, 2023. doi: 10.1007/s41324-022-00494-x.

[13] "What Is Support Vector Machine? | IBM." Accessed: May 05, 2024. [Online]. Available: https://www.ibm.com/topics/support-vector-machine

[14] "All You Need to Know About Support Vector Machines." Accessed: May 05, 2024. [Online]. Available: https://www.spiceworks.com/tech/big-data/articles/what-is-support-vector-machine/

[15] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2001, pp. 249–257. doi: 10.1007/3-540-44673-7_12.

[16] Z. Jun, "The Development and Application of Support Vector Machine," in Journal of Physics: Conference Series, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1742-6596/1748/5/052006.

[17] C. Liu, L. Wang, A. Yang, and Y. Zhang, "Support Vector Machine Classification Algorithm and Its Application," 2012.

[18] J. Nayak, B. Naik, and H. S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges," International Journal of Database Theory and Application, vol. 8, no. 1, pp. 169–186, Feb. 2015, doi: 10.14257/ijdta.2015.8.1.18.

[19] A. A. Komlavi, K. Chaibou, and H. Naroua, "Comparative study of machine learning algorithms for face recognition." [Online]. Available: https://inria.hal.science/hal-03620410v

[20] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.08458

[21] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," Computation, vol. 11, no. 3. MDPI, Mar. 01, 2023. doi: 10.3390/computation11030052.

[22] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, no. 4. Springer Verlag, pp. 611–629, Aug. 01, 2018. doi: 10.1007/s13244-018-0639-9.

[23] "What is CNN? Explain the Different Layers of CNN." Accessed: Dec. 13, 2023. [Online]. Available: https://www.theiotacademy.co/blog/layers-of-cnn/

[24] B. Koodalsamy, M. B. Veerayan, and V. Narayanasamy, "Face Recognition using Deep Learning," in E3S Web of Conferences, EDP Sciences, May 2023. doi: 10.1051/e3sconf/202338705001.

[25] "What are Convolutional Neural Networks? | Definition from TechTarget." Accessed: Dec. 13, 2023. [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/convolutional-neural-network

[26] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," IEEE Access, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[27] "Ensemble Learning Techniques. Ensemble can make things simple | by Charu Makhijani | Towards Data Science." Accessed: May 06, 2024. [Online]. Available: https://towardsdatascience.com/ensemble-learning-techniques-6346db0c6ef8

[28] "Ensemble Methods - Overview, Categories, Main Types." Accessed: May 06, 2024. [Online]. Available: https://corporatefinanceinstitute.com/resources/data-science/ensemble-methods/

[29] "colab.google." Accessed: May 05, 2024. [Online]. Available: https://colab.google/

[30] "Find Open Datasets and Machine Learning Projects | Kaggle." Accessed: Dec. 11, 2023. [Online]. Available: https://www.kaggle.com/datasets

[31] S. Verma, "Comparative Analysis of Image Classification Algorithms," Int J Res Appl Sci Eng Technol, vol. 11, no. 12, pp. 1513–1520, Dec. 2023, doi: 10.22214/ijraset.2023.57662.

[32] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," Machine Learning and the Internet of Medical Things in Healthcare, pp. 89–111, Jan. 2021, doi: 10.1016/B978-0-12-821229-5.00003-3.

[33] "Classification Report — Yellowbrick v1.5 documentation." Accessed: Dec. 15, 2023. [Online]. Available: https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 ⊙ (24*7 Support on Whatsapp)