Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Medical Diagnosis using Traditional Machine Learning Methods with Augmentation

Jyotirmay Mishra[1], Dr. Parveen Kumar[2]

Research Scholar (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula, UP, India
Research Guide (Computer Science), School of Engineering and Technology, Shri Venkateshwara University, Gajraula, UP, India
Email: mjyotirmay@gmail.com

Chest X-ray analysis plays a crucial role in diagnosing various pulmonary conditions. With the advent of traditional machine learning algorithms and advancements in data augmentation techniques, this study delves into the efficacy of different augmentation methods in enhancing the accuracy of chest X-ray abnormality diagnosis. This white paper discusses the application of classical ML algorithms and their performance when integrated with diverse data augmentation strategies, aiming to optimize diagnostic accuracy.
Keywords: Chest X-ray, Machine Learning, Data Augmentation, Traditional Algorithms, Medical Imaging, Diagnosis.

1. Introduction

The utilization of machine l-earning in medical imaging, particularly in chest X-ray analysis, has exhibited promising outcomes in diagnosing pulmo--nary abnormalities [1]. Traditional ML algorithms, such as Support Vector Machines (SVM), Random Forest, and k-Nearest Neighbors (k-NN), have been applied to this domain. However, the limited size and variability of medical datasets often hinder model generalization. Data augmentation techniques present a viable solution by expanding the dataset's diversity and size, potentially improving model performance [2].

2. Machine Learning Models

We have used Logistic Regression, Random Forest Classifier, and Support Vector Classifier (SVC). For our diagnosis methodology. Let us delve deeper into each of these machine

learning models, exploring their intricacies and working principles.

Logistic Regression:

Logistic Regression [3] is a foundational algorithm in machine learning, primarily used for binary classification tasks. Contrary to its name, it doesn't perform regression; instead, it models the probability of an instance belonging to a certain class using a logistic (sigmoid) function as shown in figure 1. This model is adept at predicting binary outcomes, such as whether an email is spam or not, based on a set of features. Its core principle lies in fitting a linear decision boundary that separates the classes in the feature space. Despite its simplicity, Logistic Regression is powerful, efficient, and particularly useful when dealing with linearly separable data or when interpretability of results is crucial. It's computationally efficient and less prone to overfitting, especially when the number of features is small.

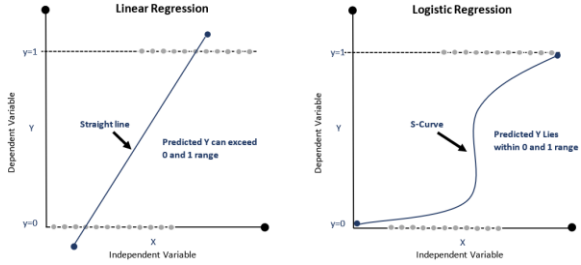

Figure 1: Logistic Regression Vs Linear Regression

Logistic Regression is a fundamental and widely used statistical technique for binary classification problems. Despite its name, it's a classification algorithm rather than a regression one. It's particularly useful when the outcome to be predicted is a categorical variable with two possible outcomes, such as 'yes' or 'no', 'spam' or 'not spam'.

Working Principle:

- Model Representation: Logistic Regression models the probability of a certain class or outcome using a logistic function. It computes the probability that an instance belongs to a particular class.
- Decision Boundary: It separates the classes by a linear decision boundary in the feature space.
- Cost Function: The algorithm minimizes a cost function, often the logistic loss function, to optimize model parameters.

Key Characteristics:

- Linear Classifier: It's a linear classifier, which means it assumes a linear relationship between the features and the log-odds of the target variable.
- Simple and Efficient: Logistic Regression is computationally efficient and less prone to overfitting when the number of features is relatively small.

Random Forest Classifier:

Random Forest is an ensemble learning method [4], renowned for its versatility and robustness in handling both classification and regression tasks. This algorithm operates by constructing multiple decision trees during the training phase and combines their predictions for the final output. Each tree in the forest is trained on a random subset of the dataset, reducing overfitting. The algorithm then aggregates the predictions from these individual trees, typically by a majority vote, to arrive at the final classification. Random Forests are highly flexible, capable of handling large datasets with high dimensionality, and provide insights into feature importance, aiding in understanding the significance of different variables in the dataset. Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the class that is the mode of the classes of the individual trees (Figure 2).

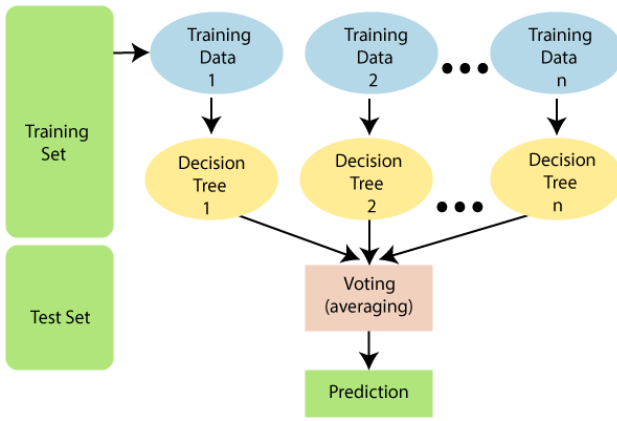

Figure 2: Working principle of random forest algorithm

Working Principle:

- Ensemble Method: It builds multiple decision trees and merges their predictions (bagging).
- Feature Randomness: Each tree is trained on a random subset of features, which helps in reducing overfitting.
- Voting Mechanism: For classification, the final prediction is made by a majority vote of the constituent trees.

Key Characteristics:

- Versatile: Random Forests can handle both classification and regression tasks.
- Robust to Overfitting: They are less prone to overfitting compared to individual decision trees due to the aggregation of multiple trees.
- Feature Importance: They can provide insights into feature importance, aiding in understanding the dataset.

SVC (Support Vector Classifier):

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression analysis [5]. SVC is specifically the classification variant.

Support Vector Machine (SVM) is a powerful supervised learning algorithm known for its effectiveness in both classification and regression tasks. The Support Vector Classifier (SVC) variant specifically addresses classification problems. SVC finds the hyperplane that best separates different classes in the feature space. What sets SVM apart is its ability to handle complex, nonlinear relationships between features through the use of kernel functions. This "kernel trick" enables SVM to map data into higher-dimensional spaces where linear separation becomes possible, even when the original data is not linearly separable. SVM aims to maximize the margin between classes, making it robust against outliers and effective in high-dimensional spaces. Once trained, SVC is memory efficient during prediction, requiring only a subset of data points called support vectors (Figure 3).

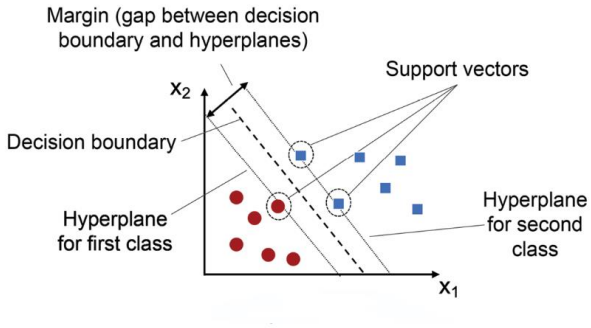

Figure 3: Working Principle of Support vector classifiers

Each of these models has its strengths and limitations. Logistic Regression offers simplicity and interpretability, while Random Forest provides robustness and versatility in handling various data types. SVC, on the other hand, is effective in complex, high-dimensional spaces and offers flexibility through kernel functions. The choice among these models often depends on the specific characteristics of the dataset, the problem's nature, and the trade-offs between model complexity and interpretability. Tailoring the choice to suit the data and problem at hand is crucial for achieving optimal performance in machine learning tasks.

Working Principle:

- Hyperplane: SVC finds the hyperplane that best separates different classes in the feature space.
- Kernel Trick: It can handle nonlinear relationships between features through the use of kernel functions, transforming data into higher dimensions where linear separation is possible.
- Margin Maximization: It aims to maximize the margin between different classes, which enhances its robustness to outliers.

Key Characteristics:

- Effective in High-Dimensional Spaces: It's effective in cases where the number of dimensions is greater than the number of samples.
- Versatile Kernels: Various kernel functions like linear, polynomial, radial basis function (RBF), etc., allow flexibility in modeling complex relationships.

- Memory Efficient for Prediction: Once trained, only a subset of data points (support vectors) is required for making predictions.

Each of these algorithms has its strengths and weaknesses, and their performance can vary depending on the dataset characteristics, feature quality, and problem type. Choosing the most suitable algorithm often involves experimentation and analysis based on the specific problem at hand.

3. Methodology:

This study employs a comprehensive approach to evaluate the impact of diverse data augmentation methods on traditional ML algorithms for chest X-ray abnormality diagnosis. The methodology involves:

Dataset Collection: Gathering a sizable chest X-ray dataset containing both normal and abnormal cases. (Hojjat Salehinejad et al (2019)) collected this dataset from 'Shenzhen No.3 People's Hospital', Guangdong Medical College, and Shenzhen, China. The dataset comprises frontal Chest X-rays (CXRs) classified into two categories: normal and tuberculosis (TB). Among the outpatient clinics, a total of 662 CXRs were recorded, including 336 cases of tuberculosis and 326 normal cases.

Preprocessing: Standardizing image sizes, normalizing pixel values, and segmenting regions of interest to ensure uniformity [7].

Model Development: Implementing SVM, Random Forest, and logistic regression algorithms as baseline models.

Data Augmentation Techniques: Incorporating augmentation methods such as rotation, flipping, zooming, and adding noise to the dataset. The strategic selection of data augmentation techniques significantly impacts the robustness and accuracy of classifiers. By essentially creating an infinite quantity of data from existing samples, augmentation enhances machine learning (ML) classifier models [8]. It serves as a cost-effective approach, compensating for the expenses associated with collecting more data (Figure 4-6).

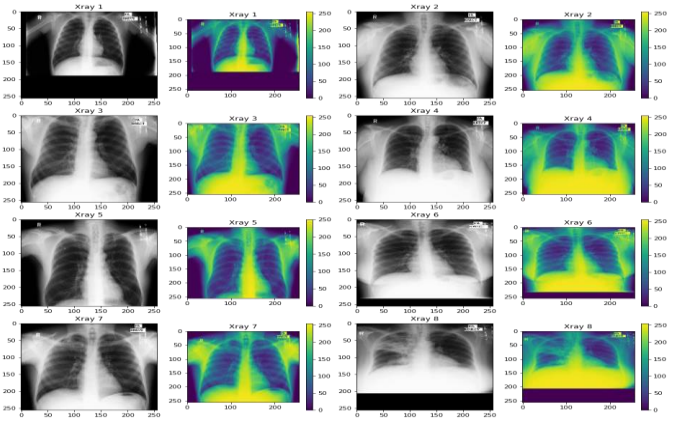

Figure 4: Some of the images from dataset

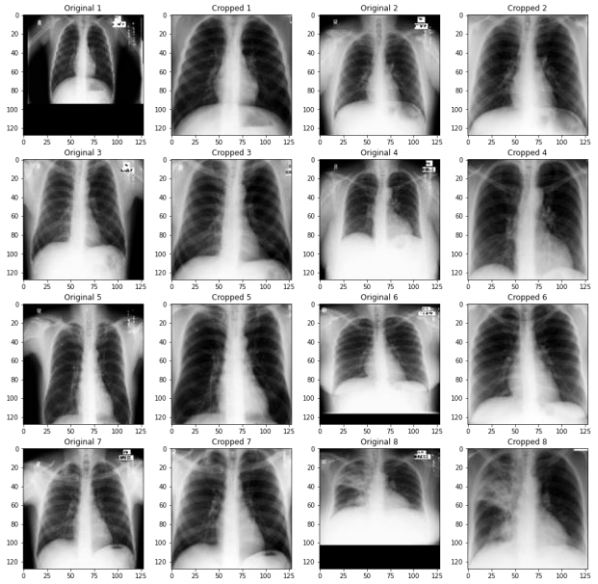

Figure 5: Cropping the Images for Region of Interest

There's a plethora of ways to augment image data, including rotations, adjusting lighting conditions, cropping, masking and more[9]. These operations allow a single image to generate multiple variations, effectively increasing the dataset size. This augmentation process not only enriches the dataset but also aids in combating overfitting issues in classifiers[10, 11]. Ultimately, it enables classifiers to generalize better by exposing them to a wider array of data variations.

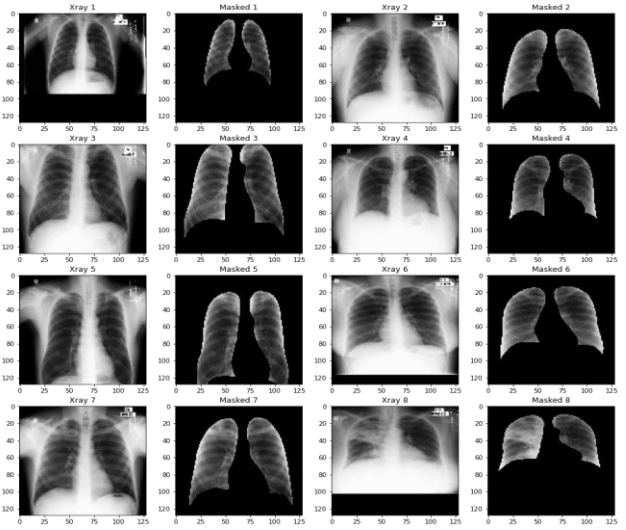

Figure 6: Masked Images

Model Training and Evaluation: Training each model on augmented datasets and evaluating their performance using metrics like accuracy, precision, recall, and F1-score. The

experiments evaluated models' respective performances using various evaluation measures.

1 CV Accuracy: Cross-Validation Accuracy. This represents the average accuracy of the model during cross-validation (often using k-fold cross-validation) on the training data. It indicates how well the model performs on unseen data.

2 Test Accuracy: Accuracy of the model on the test dataset. It showcases how accurately the model predicts on new, unseen data.

3 Area under Curve (AUC): The area under the Receiver Operating Characteristic (ROC) curve. It measures the model's ability to distinguish between classes. A higher AUC generally indicates a better-performing model.

4 Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures the accuracy of positive predictions made by the model.

5 Recall: Recall (also known as sensitivity) is the ratio of correctly predicted positive observations to all actual positives. It indicates the model's ability to identify all positive instances correctly.

6 F1 Score: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall, offering a single metric that summarizes both measures.

4. Results and Discussion

The experimental results demonstrate notable improvements in model performance when trained on augmented datasets compared to unaugmented ones. Specifically, augmentation techniques like rotation and flipping contribute significantly to enhancing model generalization and robustness. SVM and Random Forest models exhibit improved performance with augmented data, showcasing higher accuracy and recall rates in detecting abnormal chest X-ray patterns (Figure 7 and Table 1). We have evaluated our results on three sets of data:

- Unprocessed Images (UI)
- Cropped Images (CI)
- Masked Images (MI)

Results on UI:

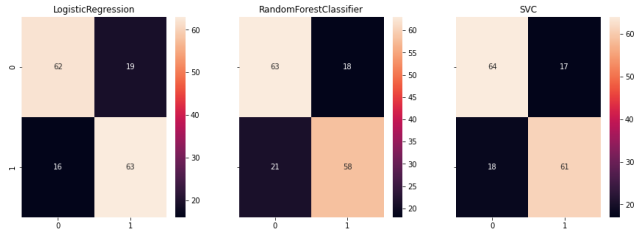

Figure 7: Confusion Matrix of Results obtained by classifiers on Unprocessed Images

Table 1: Results for these unprocessed images (UI)

Sr. No.	Algorithms	CV Accuracy	Test Accuracy	Area under Curve	Precision	Recall	F1 Score
1.	Logistic Regression	0.80	0.78	0.78	0.77	0.80	0.78
2.	Random Forest	0.82	0.76	0.76	0.76	0.73	0.75
3.	Support Vector Machine	0.81	0.78	0.78	0.78	0.77	0.78

This table I presents the performance metrics of different machine learning algorithms on a classification task. Each row corresponds to a different algorithm, and the columns represent various evaluation metrics. Now, interpreting the values in the table:

Logistic Regression: It shows decent cross-validation and test accuracies around 0.80 and 0.78, respectively. The precision, recall, and F1 score are relatively consistent and balanced, indicating a fair overall performance.

Random Forest: It has a slightly higher cross-validation accuracy (0.82) but a lower test accuracy (0.76). Precision and recall are closer, but the F1 score suggests a trade-off between precision and recall.

Support Vector Machine (SVM): SVM demonstrates consistent performance across cross-validation and test accuracies (around 0.81 and 0.78, respectively). Precision, recall, and F1 score are also consistent and balanced, showing a stable performance across different metrics.

These metrics collectively provide insights into the performance and characteristics of each algorithm in solving the classification task, aiding in the selection of the most suitable algorithm based on the evaluation criteria and requirements of the specific problem (Figure 8 and Table 2).

Results on CI:

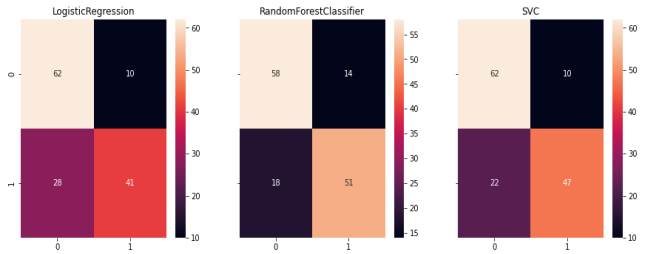

Figure 8: Confusion Matrix of Results obtained by classifiers on Masked Images Images

Table 2: Confusion Matrix of Results obtained by classifiers on cropped Images

Sr. No.	Algorithms	CV Accuracy	Test Accuracy	Area under Curve	Precision	Recall	F1 Score
1.	Logistic Regression	0.74	0.73	0.73	0.80	0.59	0.68
2.	Random Forest	0.79	0.77	0.77	0.86	0.70	0.77
3.	Support Vector Machine	0.78	0.77	0.77	0.67	0.54	0.60

The table presents the performance metrics of three different machine learning algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM). In terms of cross-validation accuracy, Random Forest performs the best with a score of 0.79, closely followed by SVM at 0.78 and Logistic Regression at 0.74. When tested on unseen data, Random Forest maintains its lead in accuracy with a score of 0.77, while both SVM and Logistic Regression exhibit slightly lower but comparable test accuracies of 0.77 and 0.73, respectively. Assessing the models' ability to discriminate between classes, measured by the Area Under Curve (AUC), all three algorithms show similar performance, hovering around 0.77 to 0.73. Looking deeper into precision, recall, and F1 score, Random Forest exhibits the highest values across all three metrics, emphasizing its balanced capability in correctly identifying positive cases, minimizing false positives, and achieving a higher harmonic mean between precision and recall. Logistic Regression follows, with decent precision but relatively lower recall, indicating it identifies positives well but may miss some relevant cases. SVM, while showing competitive accuracy, lags in precision and recall, resulting in a lower F1 score compared to the other models. Overall, Random Forest emerges as the top performer among these algorithms across various evaluation criteria, displaying a robust balance between precision and recall on this dataset (Figure 9 and Table III).

Results on Masked Images:

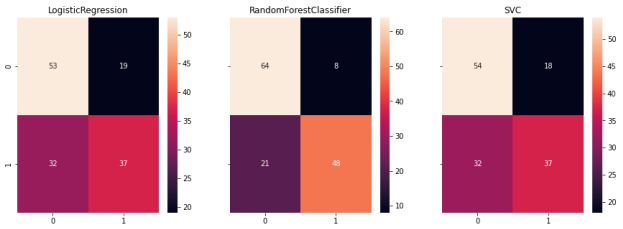

Figure 9: Confusion Matrix of Results obtained by classifiers on Masked Images Images

Table 3: Confusion Matrix of Results obtained by classifiers on masked Images are given in table

Sr. No.	Algorithms	CV Accuracy	Test Accuracy	Area under Curve	Precision	Recall	F1 Score
1.	Logistic Regression	0.67	0.64	0.64	0.64	0.53	0.59
2.	Random Forest	0.77	0.79	0.79	0.70	0.86	0.77
3.	Support Vector Machine	0.68	0.65	0.64	0.67	0.54	0.60

D. Comparison of three methods:

A combined table summarizing the results of the three classification algorithms (Logistic Regression, Random Forest, and Support Vector Machine) on three sets of data: Unprocessed Images (UI), Cropped Images (CI), and Masked Images (MI). The comparison across three distinct image sets—Unprocessed Images (UI), Cropped Images (CI), and Masked Images (MI)—reveals varying performances of three classification algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Notably, Random Forest consistently demonstrates robustness across different image types, maintaining higher accuracy, precision, recall, and F1 score compared to the other algorithms in most cases. Logistic Regression, while exhibiting consistency on Unprocessed Images, experiences a notable performance drop on Cropped and Masked Images, especially in precision and recall. Support Vector Machine shows mixed results, with decreased performance on Cropped and Masked Images compared to Unprocessed Images. These findings suggest that while all algorithms display sensitivity to different image alterations, Random Forest exhibits more resilience and adaptability across varied image modifications, positioning it as a relatively versatile choice for classification tasks involving diverse image datasets (Figure 10).

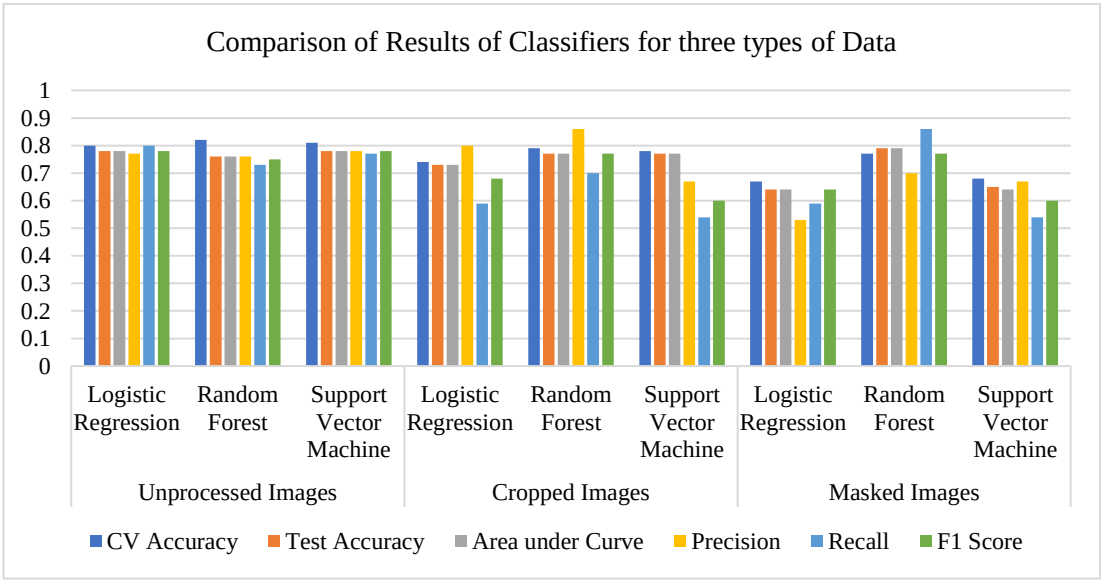

Figure 10: Results of the three classification algorithms on three sets of data.

5. CONCLUSION

This study underscores the potential of traditional ML algorithms in diagnosing chest X-ray abnormalities when coupled with diverse data augmentation techniques. The findings emphasize the significance of augmenting datasets to mitigate overfitting and enhance the generalization capabilities of models in medical image analysis. Future research could explore more sophisticated augmentation methods and leverage deep learning architectures to further advance diagnostic accuracy.

References

1. G. Litjens, T. Kooi, B.E. Bejnordi, A .A .A . Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88 .
2. M. Xu, S. Yoon, A. Fuentes, J. Yang, D. Park, Style-consistent image translation: a novel data augmentation paradigm to improve plant disease recognition, Front. Plant Sci. 12: 773142. doi: 10.3389/fpls (2022) .
3. Boateng, E. Y., & Abaye, D. A. (2019). A review of the logistic regression model with emphasis on medical research. Journal of data analysis and information processing, 7(4), 190-207.
4. Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. In International conference on intelligent data communication technologies and internet of things (ICICI) 2018 (pp. 758-763). Springer International Publishing.
5. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.
6. Yang, F., Lu, P. X., Deng, M., Wáng, Y. X. J., Rajaraman, S., Xue, Z., ... & Jaeger, S. (2022). Annotations of lung abnormalities in the Shenzhen chest X-ray dataset for computer-aided screening of pulmonary diseases. Data, 7(7), 95.
7. S.C. Wong, A. Gatt, V. Stamatescu, M.D. McDonnell, Understanding data augmentation for classification: when to warp? in: 2016 international conference on digital image computing: techniques and applications (DICTA), IEEE, 2016, pp. 1–6 .
8. L. Taylor, G. Nitschke, Improving deep learning with generic data augmentation, in: 2018 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2018, pp. 1542–1547
9. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Inter- national Conference on Machine Learning, PMLR, 2021, pp. 10347–10357 .
10. Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q. Weinberger. On feature normalization and data augmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12383–12392, June 2021.
11. Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. Journal of Big Data, 6(1):1–48, 2019.
12. Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. arXiv preprint arXiv:2001.04086, 2020.
13. Keyu Tian, Chen Lin, Ming Sun, Luping Zhou, Junjie Yan, and Wanli Ouyang. Improving autoaugment via augmentation-wise weight sharing. arXiv preprint arXiv:2009.14737, 2020.
14. Chen Lin, Minghao Guo, Chuming Li, Xin Yuan, Wei Wu, Junjie Yan, Dahua Lin, and Wanli Ouyang. Online hyper-parameter learning for auto-augmentation strategy. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6579–6588, 2019.
15. Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In European Conference on Computer Vision, pages

479–495. Springer, 2020.
16. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database. IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
17. Zheng, Y., Huang, J., Chen, T., Ou, Y., & Zhou, W. (2021). Transfer of Learning in the Convolutional Neural Networks on Classifying Geometric Shapes Based on Local or Global Invariants. Frontiers in computational neuroscience, 15, 637144.
18. Lahoura V, Singh H, Aggarwal A, Sharma B, Mohammed MA, Damaševičius R, Kadry S, Cengiz K. Cloud Computing-Based Framework for Breast Cancer Diagnosis Using Extreme Learning Machine. Diagnostics. 202111(2):241. https://doi.org/10.3390/diagnostics11020241
19. https://ieee-dataport.org/documents/refuge-retinal-fundus-glaucoma-challenge
20. https://www.analyticsvidhya.com/blog/2020/08/top-4-pre-trained-models-for-image-classification-with-python-code/