IJSES
Volume 9, Issue 10, pp. 10-15, 2025.
International Journal of Scientific Engineering and Science
ISSN (Online): 2456-7361

Development of a Hybridized Data Balancing Model Using SMOTE and ADASYN for COVID-19 Severity Level Prediction

Ganiyu M.[1], Olabode O.[2], Boyinbode K. O.[3]
[1]Department of Computer Science, Federal Polytechnic, Ile-Oluji, Ondo State, Nigeria- 351110
[2]Department of Data Science, Federal University of Technology, Akure, Ondo State, Nigeria- 340001
[3]Department of Information Technology, Federal University of Technology, Akure, Ondo State, Nigeria- 340001

Abstract— *Development of a Hybridized Data Balancing Model Using SMOTE and ADASYN for COVID-19 Severity Level Prediction. To overcome the limitations of scarce and imbalanced chest X-ray data, Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) sampling were combined to develop a Hybrid data generation and balancing technique (SMOSYN). 809 (538 COVID-19 and 271 Non-COVID-19) chest radiography images were obtained from the COVID-19 patient radiographic images and case note from the isolation centers of Nigeria hospitals. The SMOSYN was implemented to generate and balance the original dataset with realistic, class-conditional synthetic images. The hybridized SMOTE and ADASYN produced high-quality synthetic images and achieved the best data balancing with stability, visual fidelity, and diversity. The results showed that the hybridized Model (SMOSYN) produced lowest FID of 237.38 and highest SSIM of 0.5160 to achieve the best result at the ablation study 16 X 64 compare to other ablations of 16 X 250, 32 X 250 and 32 X 64.*

Keywords—*COVID-19 Severity Prediction, SMOTE, ADASYN, Chest X-ray Imaging, Synthetic Data, undersampling, Data Balancing, overfitting, Machine Leaning.*

I. Introduction

One of the most popular oversampling techniques is the Synthetic Minority Oversampling Technique (SMOTE). SMOTE produces more varied synthetic samples than previous oversampling techniques, which enhances the models' capacity for generalization. Another method for learning from unbalanced datasets is Adaptive Synthetic (ADASYN) sampling. The fundamental concept behind ADASYN is the use of a weighted distribution for various minority class examples based on how difficult they are to learn. This means that more synthetic data is produced for minority class dataset that are more challenging to learn than for those that are simpler. An ensemble of SMOTE and ADASYN was to improve the robustness of oversampling by combining the uniform oversampling of SMOTE with the adaptive focus of ADASYN on difficult regions. This will lead to a better class balance without excessive noise

Class imbalance, which is defined as a notable difference in the quantity of samples across various categories, is a common problem in real-world datasets. This phenomenon is widespread in a variety of fields, such as hardware failure detection [1], software defect prediction [2], financial fraud detection [3] and medical disease diagnosis [4]. One of the main challenges in achieving high accuracy for minority class data points in classification tasks is class imbalance. Datasets are frequently very skewed in the real world; for example, benign tumors greatly outnumber dangerous tumors in medical datasets [5] for cancer diagnosis [6, 7], and legal transactions dominate fraudulent transactions in finance datasets [8]. Statistical models that are built on these highly skewed datasets have a tendency to predict every instance as belonging to the overrepresented class. This makes the model ineffective for identifying instances of the underrepresented class, which is particularly important when the goal is to classify minority instances.

A highly contagious respiratory disease, COVID-19 is brought on by the SARS-CoV-2 virus. When multiple pneumonia cases were connected to a seafood market in Wuhan, China, in December 2019, the illness was first documented there. The precise mechanism of transmission is still unknown, however it is thought to have a zoonotic origin, potentially spreading from bats to humans [9]. Human-to-human transmission has been verified by January 2020, and the virus was swiftly spreading across the globe. The rapid spread and serious public health consequences of COVID-19 led the World Health Organization (WHO) to proclaim it a global pandemic on March 11, 2020 [10]

An effective disease control approach must include early identification of affected persons for timely treatment and isolation to prevent the spread of the disease [11]. Early detection and analysis of infection patterns are essential for designing treatments and managing the spread of sickness [12]. Because of the high infection rate, lack of proven vaccines, and lack of effective therapy, early screening for COVID-19 is crucial to controlling the disease's limited medical resources and halting its spread. The most common symptoms of a 2019-CoV infection are fever, dry cough, dyspnea, chest pain, fatigue, and myalgia; these symptoms are similar to those of SARS-CoV [13, 14]. Less common symptoms include headache, nausea, vomiting, dizziness, stomach pain, diarrhea, loss of taste and smell, and hemoptysis [5].

II. Literature Review

The novel coronavirus disease 2019 (COVID-19) emerged in late 2019 in Wuhan, Hubei Province, China. It was first reported to the World Health Organization (WHO) on December 31, 2019, as a cluster of pneumonia cases with an

unknown cause. The causative agent, later identified as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), is a novel strain of coronavirus not previously detected in humans [15]. Furthermore, COVID-19 is believed to have originated in a wet market in Wuhan, where live animals were sold, facilitating zoonotic transmission from animals to humans. However, the exact source of the virus remains under investigation. By January 2020, human-to-human transmission had been confirmed, leading to a rapid global spread of the virus [16].

On March 11, 2020, the WHO declared COVID-19 a global pandemic. This declaration underscored the rapid international spread and significant public health impact of the virus. Countries worldwide implemented various public health interventions, including lockdowns, travel restrictions, mask mandates, and mass vaccination campaigns to contain the virus and reduce mortality [17]. The detection of COVID-19, caused by the SARS-CoV-2 virus, is crucial for diagnosis, treatment, and controlling the spread of the disease. Various diagnostic techniques have been developed and widely deployed, each with specific advantages and limitations. These methods are broadly categorized into molecular tests [18], antigen tests [19], emerging diagnostic methods [20] and serological tests [21].

Random Oversampling (ROS), which was proposed by Batista et al.[22], balances the dataset by randomly duplicating minority class samples. Despite being easy to use, ROS frequently results in overfitting since the repeated samples may magnify noise in the data and do not provide new information. Notwithstanding its drawbacks, ROS has been demonstrated to perform better than undersampling in specific situations when assessed using measures like AUC.

SMOTE, which creates synthetic minority samples by interpolating between existing minority samples and their k-nearest neighbors, was introduced by Chawla et al. [23] as a solution to the problems caused by ROS. To some extent, SMOTE reduces overfitting by creating a variety of synthetic datasets instead of just copying the ones that already exist. However, there are certain disadvantages to SMOTE which include production of synthetic dataset in areas where classes overlap which harmed the performance of the classifier and produces noisy data. It also frequently produces too many synthetic samples in regions with high concentrations of minority samples. This circumstance makes the overfitting issue worse

Du et al. [24] suggest a safe privacy-preserving SMOTE (SP2-SMOTE) sampling technique in order to protect data privacy. By enabling parties to independently create synthetic samples without disclosing the data, it goes beyond conventional SMOTE and successfully blocks illegitimate label inference using minority-class closest neighbor interpolation. To deal with skewed datasets, Kunakorntum et al.[31] introduced a unique oversampling method called Synthetic Minority based on Probability Distribution (SyMProD). This method eliminates noisy data and uses Z-scores to standardize the data. Then, using the probability distributions of the two classes, the suggested method chooses minority samples. The chosen points and a number of the minority class's closest neighbors are used to create synthetic instances.

Using an Auxiliary-guided Conditional Variational Autoencoder (ACVAE) trained with contrastive learning, Wang et al. [25] provide a novel deep learning (DL) based data balancing method. The authors also looked into an ensemble approach in which a data undersampling method is used after ACVAE creates synthetic positive datasets.

The adaptive synthetic sampling technique (ADASYN), which was proposed by He et al. [26], use a weighted distribution based on how challenging it is to learn various minority class samples. For minority samples that are more difficult to learn, more artificial data is produced. This approach adaptively moves the classification decision boundary towards the challenging samples and lessens the bias brought on by class imbalance. Nevertheless, it has trouble sampling boundary samples and is ineffective at handling noisy data.

Machine learning (ML) is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and statistical models enabling computer systems to perform tasks without explicit instructions. Instead, these systems learn patterns from data and make decisions or predictions based on that data [27]. Supervised Learning is a model that trains using labeled data, allowing the algorithm to learn the mapping between inputs and outputs. Common applications include image classification, speech recognition, and medical diagnosis [28]. Unsupervised Learning involves training models on unlabeled data, where the algorithm identifies hidden patterns or intrinsic structures. It is commonly used in clustering, anomaly detection, and dimensionality reduction tasks such as customer segmentation and topic modeling [29]. Reinforcement Learning is a model where an agent interacts with an environment and learns to make decisions by receiving rewards or penalties. Reinforcement learning has found success in robotics, game playing, and autonomous systems [30].

III. Methodology

Data Source and Description

The original dataset contained 809 (538 COVID-19 and 271 Non-COVID-19) chest radiography images was obtained from the COVID-19 patient radiographic images and case note from the isolation centers in Nigeria hospitals.

Hybridizing SMOTE and ADASYN

Hybridizing SMOTE and ADASYN was used to improve the robustness of oversampling by combining the uniform oversampling of SMOTE with the adaptive focus of ADASYN on difficult regions. This leads to a better class balance without excessive noise.

Procedures of Hybridizing SMOTE and ADASYN

i. Apply SMOTE for Base-Level Oversampling

SMOTE generates synthetic samples uniformly across the minority class, preventing the classifier from overfitting to hard-to-learn regions.

ii. Apply ADASYN to Focus on Hard-to-Learn Regions

ADASYN further generates samples in difficult-to-learn regions where the minority class is underrepresented, improving decision boundaries.

International Journal of Scientific Engineering and Science
Volume 9, Issue 10, pp. 10-15, 2025. ISSN (Online): 2456-7361

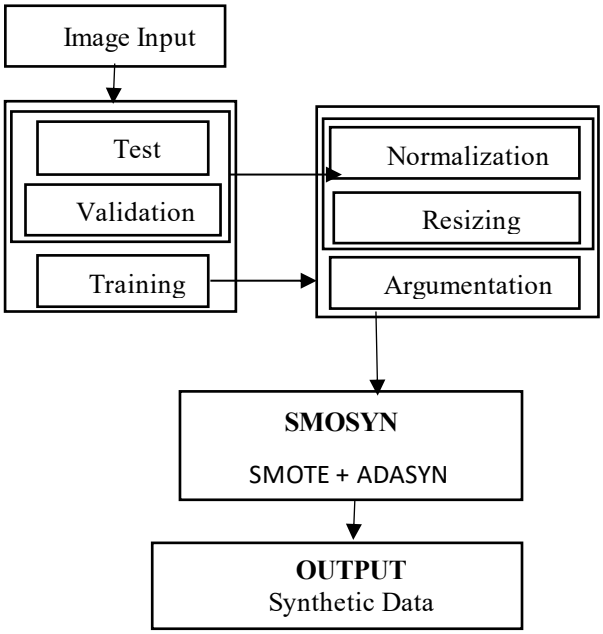

Fig. 1. Conceptualized Framework for the System

iii. Blend the Synthetic Samples

Combine the synthetic samples from both SMOTE and ADASYN with the original dataset. Control the ratio of SMOTE and ADASYN to avoid excessive noise.

Model for Ensemble SMOTE and ADASYN

X_m be the minority class samples.

X_M be the majority class samples.

G_{smote} be the number of synthetic samples generated by SMOTE.

G_{adasyn} be the number of synthetic samples generated by ADASYN.

α be the blending ratio of SMOTE and ADASYN.

i. Compute Oversampling Requirements

Total synthetic samples to be generated:

$$G = \beta \cdot (|X_M| - |X_m|) \quad (1)$$

Where β is the oversampling ratio.

ii. Divide Sampling into SMOTE and ADASYN

Allocate α proportion to SMOTE and $(1-\alpha)$ to ADASYN:

$$G_{smote} = \alpha \cdot G, \quad G_{adasyn} = (1-\alpha) \cdot G \quad (2)$$

where α is a hyperparameter (e.g., α=0.5 for equal distribution)

iii. Synthetic Data Generation

1. SMOTE Generation

Generate synthetic samples using interpolation:

$$x_{new} = x_i + (x_{zi} - x_i), \lambda \sim U(0,1) \quad (3)$$

for randomly chosen minority class neighbors x_{zi}

2. ADASYN Generation

Generate samples in difficult-to-learn regions using the weighted density function:

$$r_i = \frac{k_i}{k}, \hat{r} = \frac{r_i}{\sum r_j}, g_i = \hat{r}.G_{adasyn} \quad (4)$$

iv. Merge the Synthetic Data

Combine the synthetic samples from both SMOTE and ADASYN:

$$X' = X_m \cup X_M \cup X_{SMOTE} \cup X_{ADASYN} \quad (5)$$

where X_{SMOTE} and X_{ADASYN} are the generated synthetic samples.

IV. Results

This section presents the results and performance of the SMOSYN, the developed model and an ablation study on batch size and image size configurations. The results are presented using figures from experimental outputs, followed by comparative analysis.

Ablation Study: Batch Size and Image Size

An ablation study was conducted to examine the effect of varying batch size and image size on model performance. Configurations tested included 16×256, 32×256, 16×64, and 32×64. Larger image resolutions (256) generally produced more detailed outputs but at higher computational cost, while smaller image sizes (64) trained faster but sacrificed realism. Batch size also influenced convergence stability, with 32 performing better than 16 in most cases.

Ablation Study: Batch Size and Image Size 16 X 256

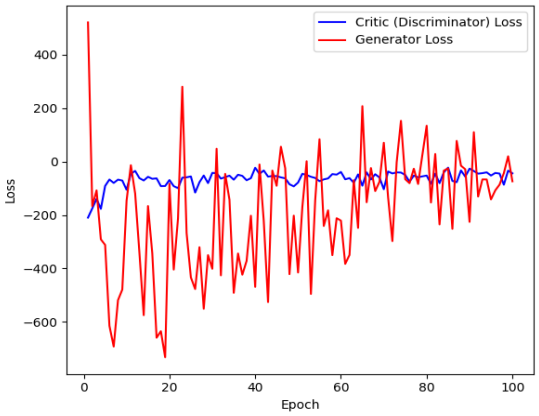

Fig. 2. Ablation Study: Batch Size and Image Size 16 X 256

SMOSYN Resampling with 16-256: Realism Test

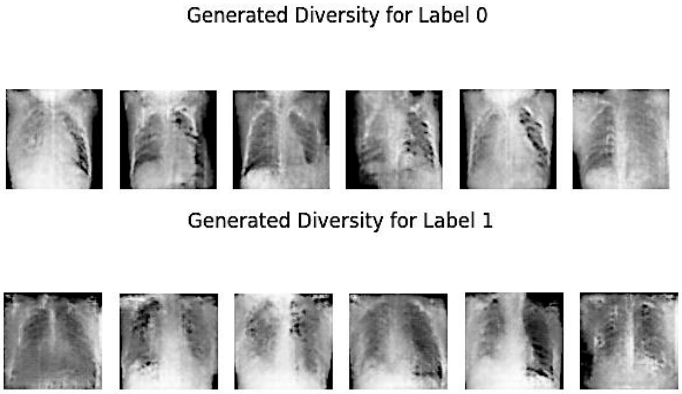

Fig. 3. SMOSYN Resampling with 16-256: Realism Test

Ablation Study: Batch Size and Image Size 32 X 256

International Journal of Scientific Engineering and Science
Volume 9, Issue 10, pp. 10-15, 2025. ISSN (Online): 2456-7361

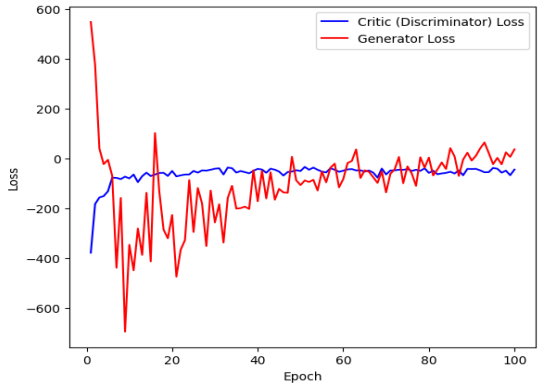

Fig. 4. Ablation Study: Batch Size and Image Size 16 X 256

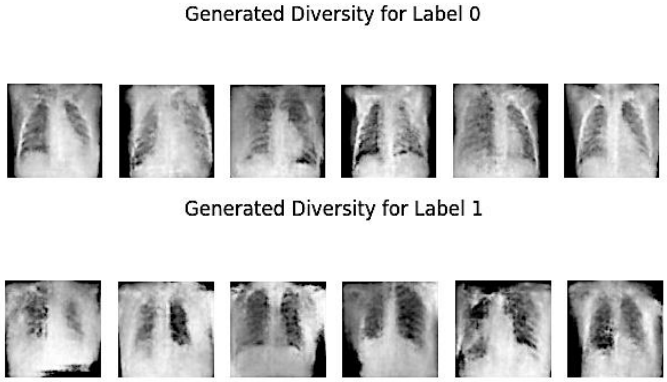

Fig. 5. SMOSYN Resampling with 32-256: Realism Test

Ablation Study: Batch Size and Image Size 16 X 64

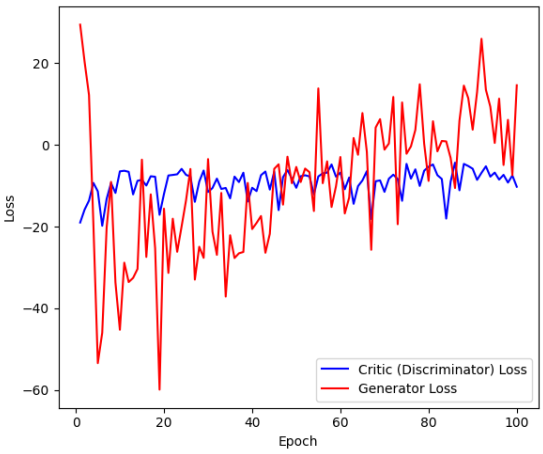

Fig. 6. Ablation Study: Batch Size and Image Size 16 X 64

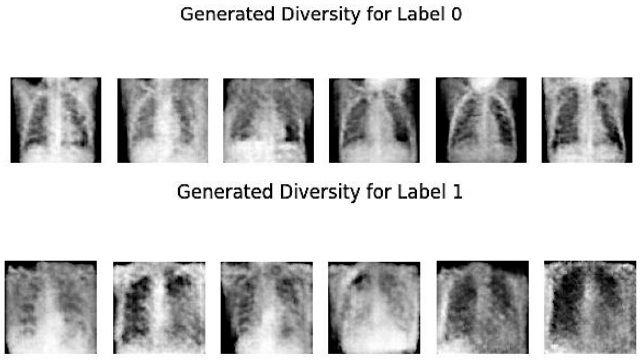

Fig. 7. SMOSYN Resampling with 16-64: Realism Test

Ablation Study: Batch Size and Image Size 32 X 64

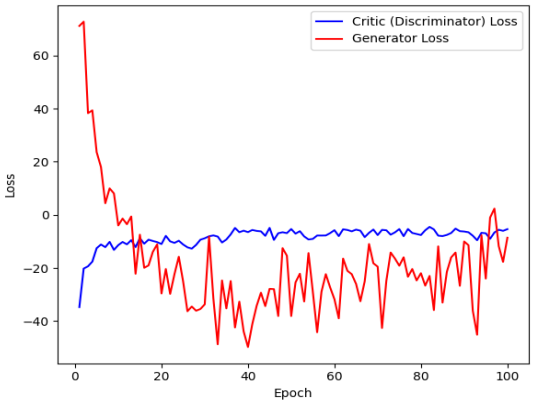

Fig. 8. Ablation Study: Batch Size and Image Size 32 X 64

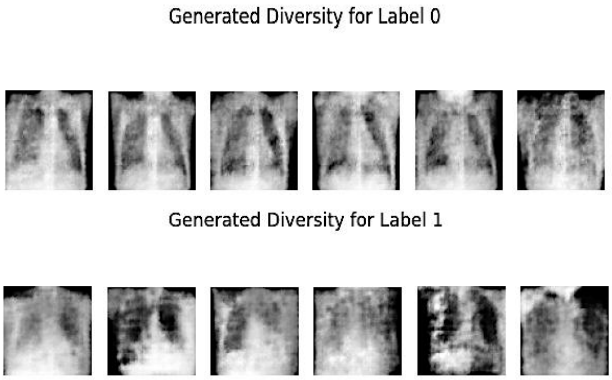

Fig. 9. SMOSYN Resampling with 32-64: Realism Test

TABLE 1. Quantitative evaluation of SMOSYN Ablation

Ablation Study	SSIM	FID
16 X 256	0.3189	333.62
32 X 256	0.3333	332.01
16 X 64	**0.5160**	**237.38**
32 X 64	0.2515	260.79

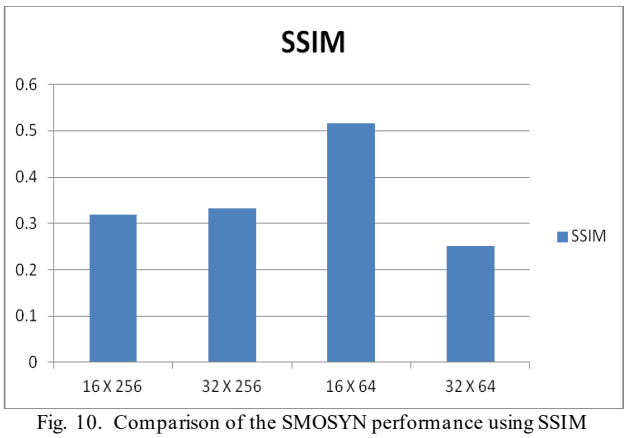

Fig. 10. Comparison of the SMOSYN performance using SSIM

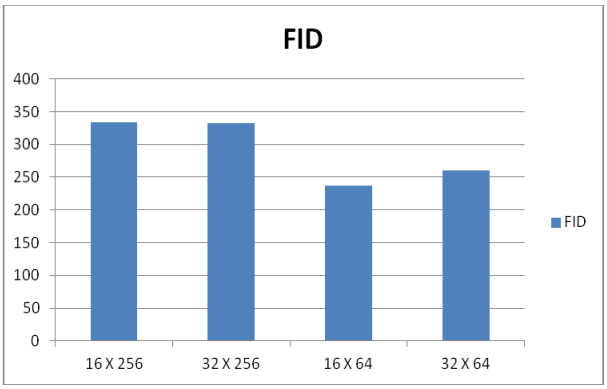

Fig. 11. Comparison of the SMOSYN performance using FID

V. Interpretation of Results

Structural Similarity Index Measure (SSIM)

SSIM values confirmed structural similarity trends. The SMOSYN ablation study 16 X 64 performed best with the highest SSIM of 0.5160 which shows the highest similarity between the real images and the synthetic generated data indicating superior structural fidelity.

Fréchet Inception Distance (FID)

The lower the values of FID reflect closer alignment between real and synthetic data distributions. The hybridized SMOTE and ADASYN (SMOSYN) attained the lowest FID of 237.38 which clearly outperforming the other ablation studies.

VI. Conclusion

This research has introduced a Hybridized Data Balancing Model Using SMOTE and ADASYN to generate more realistic synthetic chest X-ray images for COVID-19 diagnosis and severity level prediction. The hybridized Model (SMOSYN) produced lowest FID of 237.38 and highest SSIM of 0.5160 to achieve the best result at ablation study 16 X 64 compare to other ablations of 16 X 250, 32 X 250 and 32 X 64

An ensemble of SMOTE and ADASYN was used to improve the robustness of oversampling by combining the uniform oversampling of SMOTE with the adaptive focus of ADASYN on difficult regions. This lead to a better class balance without excessive noise.

References

[1] D. Liu, S. Zhong , L. Lin, M. Zhao, X. Fu and X. Liu. "Feature-level SMOTE: augmenting fault samples in learnable feature space for imbalanced fault diagnosis of gasturbines," *Expert System Application* 238 (2), 122023, 2024.

[2] R. P. Ranjan, and N. N. Kumar. "Software bug severity and priority prediction using SMOTE and intuitionistic fuzzy similarity measure," *Appl. Soft Comput*., 150, 111048, 2024.

[3] A. Abd El-Naby, E. E. D. Hemdan and A. El-Sayed. "An efficient fraud detection framework with credit card imbalanced data infinancial services," *Multimed Tools Appl*., 82 (3), 4139–4160, 2023.

[4] S. Fotouhi, S. Asadi and M. W. Kattan. "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *Biomed. Inf*. 90,103089 2019.

[5] U. Naseem, M. Khushi, S. K. Khan, N. Waheed, A. Mir, A. Qazi, B. Alshammari and S. K. Poon. "Diabetic Retinopathy Detection Using Multi-layer Neural Networks and Split Attention with Focal Loss," *In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand*, 18–22 November 2020, Springer: Cham, Switzerland, 2020; pp. 3–14.

[6] A. Panta, M. Khushi, U. Naseem, P. Kennedy and D. Catchpoole. "Classification of Neuroblastoma Histopathological ImagesUsing Machine Learning," *In Proceedings of the International Conference on Neural Information Processing, Bangkok, Thailand*, 18–22 November 2020, Springer: Cham, Switzerland, 2020; pp. 3–14.

[7] M. Khushi, C. E. Napier, C. M. Smyth, R. R. Reddel and J. W. Arthur. "MatCol: A tool to measure fluorescence signal colocalisation in biological systems," *Sci. Rep*. 2017, 7, 1–9.

[8] T. M. Alam, K. Shaukat, M. Mushtaq, Y. Ali, M. Khushi, S. Luo and A. Wahab. "Corporate Bankruptcy Prediction: An Approach Towards Better Corporate World," *Comput. J.* 2020.

[9] P. Zhou et al., "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, pp. 270–273, 2020.

[10] World Health Organization. "WHO Director-General's opening remarks at the media briefing on COVID-19—11March 2020," 2020. [Online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[11] S. Wang et al., "A deep learning algorithm using CT images to screen for COVID-19," *MedRxiv*, 2020, doi: 10.1101/2020.02.14.20023022.

[12] M. A. Khan, M. Alhaisoni, A. Tariq, M. N. Al-Sahaf, A. A. Albesher, and H. Abdalla, "A robust hybrid deep learning features and ensemble learning algorithm for COVID-19 classification using chest X-ray images," *Comput. Biol. Med*., vol. 141, p. 105157, 2022.

[13] R. Girshick, "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, pp. 1440–1448.

[14] X. Xu, C. Yu, J. Qu, L. Zhang, S. Jiang, D. Huang, B. Chen, Z. Zhang, W. Guan, Z. Ling, R. Jiang, T. Hu, Y. Ding, L. Lin, Q. Gan, L. Luo, X. Tang and J. Liu. "Imaging and clinical features of patients with 2019 novel coronavirus," *Eur. J. Nucl. Med. Mol. Imaging,* vol. 47, pp. 1022–1023, 2020.

[15] World Health Organization, "Overview of Testing for SARS-CoV-2 (COVID-19)," 2023. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/hcp/testing-overview.html

[16] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study," Lancet, vol. 395, no. 10225, pp. 689–697, 2020.

[17] Centers for Disease Control and Prevention, "Public Health Interventions for COVID-19," CDC, 2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/php/public-health-interventions.html

[18] World Health Organization, "Recommendations for national SARS-CoV-2 testing strategies and diagnostic capacities," 2021/2022. [Online]. Available: https://www.who.int/publications/i/item/WHO-2019-nCoV-lab-testing-2021.1-eng

[19] U.S. Food and Drug Administration, "At-Home OTC COVID-19 Diagnostic Tests," FDA, 2023. [Online]. Available: https://www.fda.gov/medical-devices/coronavirus-covid-19-and-medical-devices/home-otc-covid-19-diagnostic-tests

[20] F. Krammer and V. Simon, "Serology assays to manage COVID-19," Science, vol. 368, no. 6495, pp. 1060–1061, 2020.

[21] J. S. Chen et al., "CRISPR-Cas12a-assisted rapid and sensitive detection of SARS-CoV-2," Nat. Biomed. Eng., vol. 5, pp. 1228–1235, 2021.

[22] G. E. Batista, R. C. Prati and M. C. Monard. "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations News,* 6 (1), 20–29, 2004.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer. "SMOTE: synthetic minority Over-sampling technique," *J. Artif. Intell. Res*. 16, 321–357, 2002.

[24] W. Du, H. Wang, J. Shen, G. Meng, Y. Guo and W. Zhou. "Secure Privacy-Preserving SMOTE for Vertical Federated Learning," *International Conference on Advanced Data Mining and Applications. Singapore: Springer Nature Singapore*, 301–315. 2024.

[25] A. X. Wang, V. Le, H. N. Trung and B. P. Nguyen. "Addressing imbalance in health data: synthetic minority oversampling using deep learning," *Comput. Biol.Med*. 188, 109830, 2025.

[26] H. He, Y. Bai, E. A. Garcia and S. Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 322–1328. 2008.

[27] T. M. Mitchell. "Machine Learning," *New York, NY, USA: McGraw-Hill*, 1997.

[28] J. D. Kelleher, B. Mac Namee, and A. D'Arcy. "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies," *Cambridge, MA, USA: MIT Press*, 2015.
[29] H. Mohammad-Rahimi, M. Nadimi, M. H. Rohban, E. Shamsoddin, and S. E. Lee, "Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review," *Front. Cardiovasc. Med*., vol. 8, p. 638011, 2021.
[30] R. S. Sutton and A. G. Barto. "Reinforcement Learning: An Introduction," *Cambridge, MA, USA: MIT Press*, 2nd ed., 2018.
[31] I. Kunakorntum, W. Hinthong and P. A. Phunchongharn. "synthetic minority based on probabilistic distribution (SyMProD) oversampling for imbalanced datasets," IEEE Access. 8, 114692–114704, 2020.