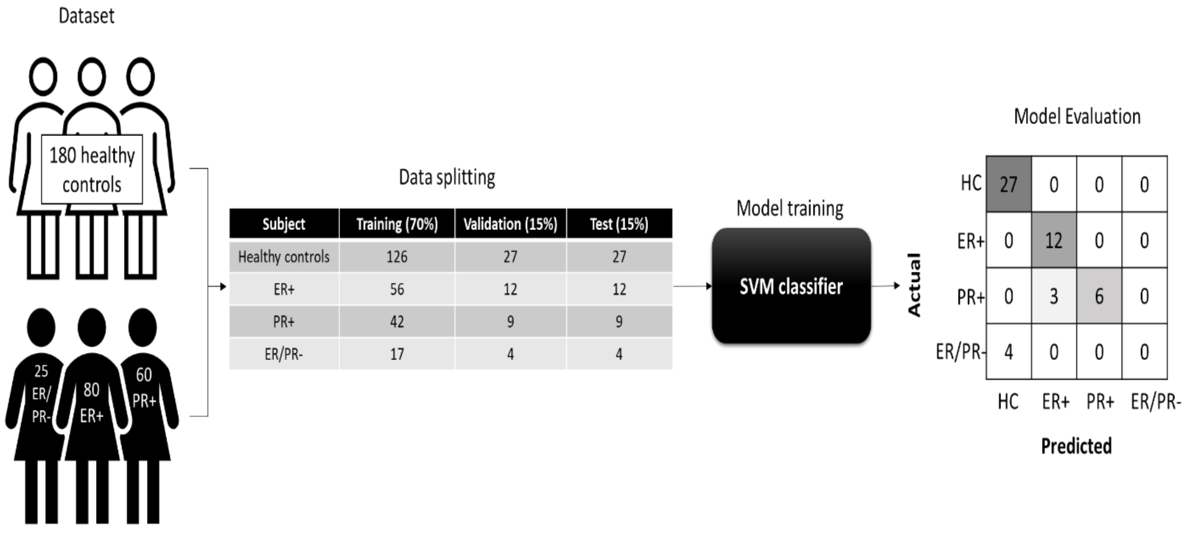
1. Introduction

With the advancement of artificial intelligence (AI), machine learning (ML) and deep learning (DL), the healthcare sector is increasingly recognizing their potential to take on multiple laborious roles such as clinical decision support [1], treatment recommendations [2], hospital in- and out-patient management [3], and so on, which can significantly reduce the work burden faced by healthcare professionals. As evidenced by the recent COVID-19 pandemic, 80% of healthcare professionals worldwide are reported to experience burnout [4]. Usually, a large dataset is required to train these ML and DL models, and while some studies have successfully shown superior model performance in performing disease classification, the dataset used to train the model may be susceptible to the doppelganger effect [5].

Data doppelgangers are samples that are fundamentally similar but are not duplicates [5]. They can be seen in biomedical data, where a group of samples from the same disease class have similar characteristics and traits, resulting in data bias towards this group of samples. As a result, when using this dataset for research, scientists will be unable to accommodate other groups of samples with the same disease but different characteristics and traits [5]. Similarly, when these data doppelgangers are used to train ML and DL models, they undoubtedly achieve high model performance results, but their reliability may be questionable due to these doppelgangers. In usual practice, the dataset used to train ML or DL models is divided into three parts: training, validation, and test [6]. The training set is used to train the model, the validation set is used to tune the model during model training, and the test set is used to evaluate the model's performance. To ensure that the model evaluation is unbiased, the test set must be isolated and completely unseen by the model until it completes its training. This is analogous to a student being tested on a completely new set of questions during an exam rather than seeing the exam questions ahead of time, which would be cheating. However, if large number of data doppelgangers are present in both the training and test sets, the model will produce high classification performance even if the dataset is re-segmented multiple times to produce a new training and test set because the doppelganger effect will persist if data doppelgangers are not removed [5]. Hence, even if the model gave very promising results, its prediction may not be generalizable due to the data bias introduced by these data doppelgangers. When this phenomenon occurs in ML and DL models, the data doppelgangers are referred to as functional doppelgangers as it confounded the prediction outcome [5].



Subject	Training (70%)	Validation (15%)	Test (15%)
Healthy controls	126	27	27
ER+	56	12	12
PR+	42	9	9
ER/PR-	17	4	4

	HC	ER+	PR+	ER/PR-
HC	27	0	0	0
ER+	0	12	0	0
PR+	0	3	6	0
ER/PR-	4	0	0	0

Figure 1: schematic drawing of ML-based breast cancer identification process.

Figure 1 depicts the doppelganger effect using breast cancer as an example. The dataset includes both healthy controls and three types of breast cancer patients: those with estrogen receptors (ER+), those with progesterone receptors (PR+), and those without estrogen or progesterone receptors (ER/PR-). Here, we assume that the characteristics of ER+ patients are consistent across training, validation, and test data, as they are for healthy controls, PR+, and ER/PR- patients. The dataset is then divided into 70:10:10 training, validation, and test data, and Support Vector Machine (SVM) is chosen as the classifier for this task. After model evaluation, the model achieves approximately 86% accuracy, which is considered to achieve a high classification performance. However, looking at the resulting confusion matrix, we can see that the SVM classifier had 0% sensitivity for the ER/PR- class because it predicted these samples as healthy controls. This demonstrates that the SVM classifier is affected by the doppelganger effect of ER+ patients and only recognizes the ER+ and some PR+ patients as having breast cancer, whereas ER/PR- patients with different characteristics are not recognized as having breast cancer. Hence, if the ML system in Figure 1 is used as a clinical decision support tool, we will misdiagnose all ER/PR- patients and delay timely intervention. Some PR+ patients who are misclassified as ER+ patients will also receive incorrect treatment, resulting in money being wasted on treatment that is ineffective for the patient.

2. Data doppelganger identification and proposed solution

Fortunately, methods for detecting functional doppelgangers and removing them from the dataset used to train the model are available. The most notable method is using pairwise Pearson's correlation coefficient (PPCC) to determine how similar the samples are across the training, validation, and test sets [5]. Suspected doppelganger samples are those with a high PPCC score. Wang et al. had demonstrated in multiple work in using PPCC to remove the doppelganger effect in renal cell carcinoma proteomics data [5], biomedical gene expression RNA-seq data [7], and microarray data of leukemia and Duchenne muscular dystrophy gene expression [8]. In all his work, he employed a R package called DoppelgangerIdentifier (DI) throughout all of his study, but it is regrettably limited to proteomics, RNA-Seq gene expression, and microarray dataset [7]. Another study by Waldron et al. [9] used the R package doppelgangR to find duplicates in breast, ovarian, colorectal, and bladder cancer profiles, however doppelgangR was similarly restricted to gene expression data. Hence, we propose a series of solutions, as depicted in Figure 2, to lessen the impact of functional doppelganger on all types of biomedical datasets.

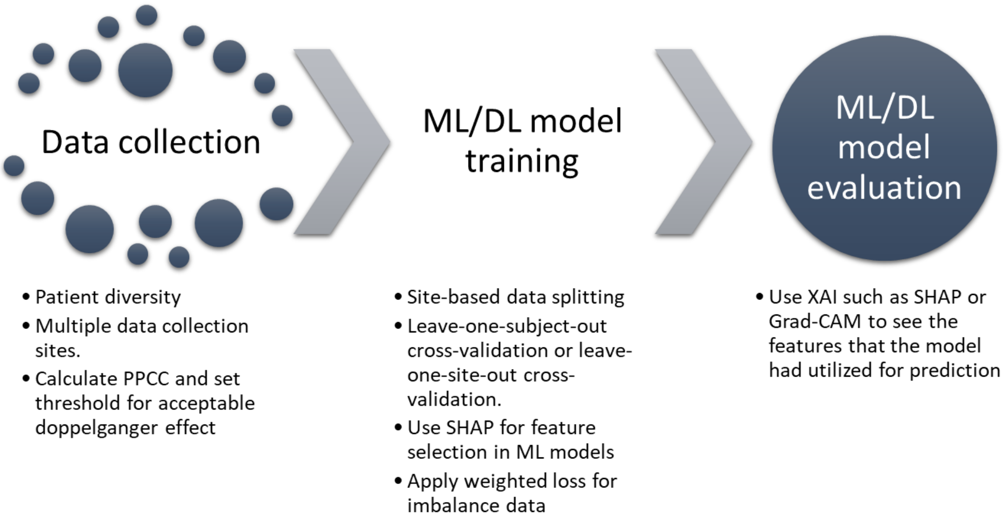

Data collection
ML/DL model training
ML/DL model evaluation
• Patient diversity
• Multiple data collection sites.
• Calculate PPCC and set threshold for acceptable doppelganger effect
• Site-based data splitting
• Leave-one-subject-out cross-validation or leave-one-site-out cross-validation.
• Use SHAP for feature selection in ML models
• Apply weighted loss for imbalance data
• Use XAI such as SHAP or Grad-CAM to see the features that the model had utilized for prediction

Figure 2: Proposed solutions for avoiding the "doppelganger effect" when developing ML or DL models for medical purposes.

2.1 Data collection

Care should be taken throughout the initial data gathering procedure to ensure the high quality of the data required to build an ML or DL model. It is ideal to have a diverse patient population in terms of race, ethnicity, and location. However, while patients are being recruited, a dataset imbalance frequently arise as a result. For instance, data collecting in Singapore will be skewed toward Chinese, who make up 76.8% of the population [10], which could result in functional doppelgangers. Collaboration between various collection locations may therefore aid in reducing data bias. As such, we recommend that data be collected from both government and private hospitals in order to cover a wider range of patient diversity in the local population, as data collected from only one hospital will only reach out to communities in its immediate vicinity. The partnership between Tan Tock Seng Hospital (TTSH) and Wisma Geylang Serai is one of the examples where TTSH specially caters to the Malay communities in Singapore [11]

After the data is collected, we can proceed to calculate the PPCC between the dataset across difference collection site. Using the study by Wang et al. [5] as reference, we can create new datasets based on 0%, 20%, 40%, 60%, 80%, and 100% data doppelgangers from the original dataset. This is done in order to establish the acceptable data doppelganger threshold that will not artificially inflate the performance of the ML or DL model in the subsequent steps.

2.2 ML/DL model training

For model training, we recommend that the data to be split according to site, instead of compiling all the samples together and randomly group them into training and validation set. This is to prevent doppelganger effect within the same site. Assuming we have data from five collection sites, we can evaluate the model using leave-one-site-out cross-validation, with four sites serving as the training set and the remaining one serving as the test set. This process is repeated five times with different training and test sets to ensure that all data from each collection site has been trained and tested. If the data is small, we can also consider leave-out-subject-out cross-validation, in which the number of times the process must be repeated equals to the number of subjects in the data.

Another distinction to be made is the difference between the datasets used in the development of ML and DL models. Unlike ML, the DL model does not require feature extraction and selection procedures [12]. DL models can process raw high-dimensional data such as images and signals without overfitting, whereas ML models, which suffer from the curse of dimensionality, require high-dimensional data to be decomposed into low-dimensional features for model training [13]. Hence, if we are using dataset that is compatible to ML models, such as clinical data or questionnaire data [14], we can apply an Explainable artificial intelligence (XAI) method known as SHapley Additive exPlanations (SHAP), during the feature selection process to identify the top few significant features that are critical for predicting a specific disease outcome. According to Loh et al. [14], SHAP is the most implemented XAI technique for ML models; SHAP being the overall most implemented XAI technique includes 35 out of 45 studies that proposed using ML models. Figure 3 shows an example of XAI output where Zeng et al. [15] use SHAP to determine which features are critical in the prediction of various complications following pediatric congenital heart surgery. According to the graph, despite the fact that surgical time is the most important predictor, it is ineffective in predicting whether a patient will develop an infectious complication after surgery. Hence, SHAP provides important feature analysis to consider for feature selection, which can also check for doppelganger effect because if data doppelganger exists, the model performance will be high regardless of which feature is selected [5], so the SHAP value for each feature will be ranked almost equally and will not be considered significant for prediction.

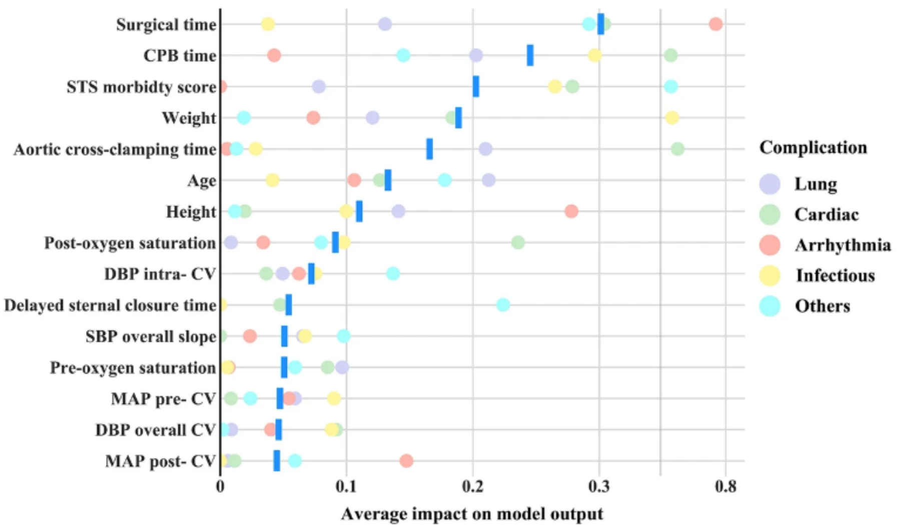


Figure 3: SHAP analysis of top 15 features adopted from Wang et al. [15].

In the case of DL model development, we recommend training the model with weighted loss to account for the class imbalance in the dataset, which means assigning more weight to the minority class and less weight to the majority class so that the model is less likely to be affected by functional doppelganger and skip the minority class.

2.3 ML/DL model evaluation

In the previous section, we discussed how to train and tune models using leave-one-site-out and leave-one-subject-out cross-validation. We can use the same validation approach to determine our model's final performance. However, SHAP is also an additional step in this section to explain the prediction made. While we did our best to reduce the doppelganger effect, it is more important that the model is picking up the necessary information for prediction. ML and DL models have been dubbed the "black box" model in the past due to their poor interpretability, which has hampered their adoption in healthcare due to a lack of evidence-based diagnosis [16]. SHAP is thus a useful tool for providing diagnostic evidence as well as a measure to ensure that the model is selecting the correct features to make a reliable prediction [17]. By doing so, we can also ensure that the prediction is not influenced by the doppelganger effect; otherwise, the features identified by SHAP as important for outcome prediction would make no clinical sense.

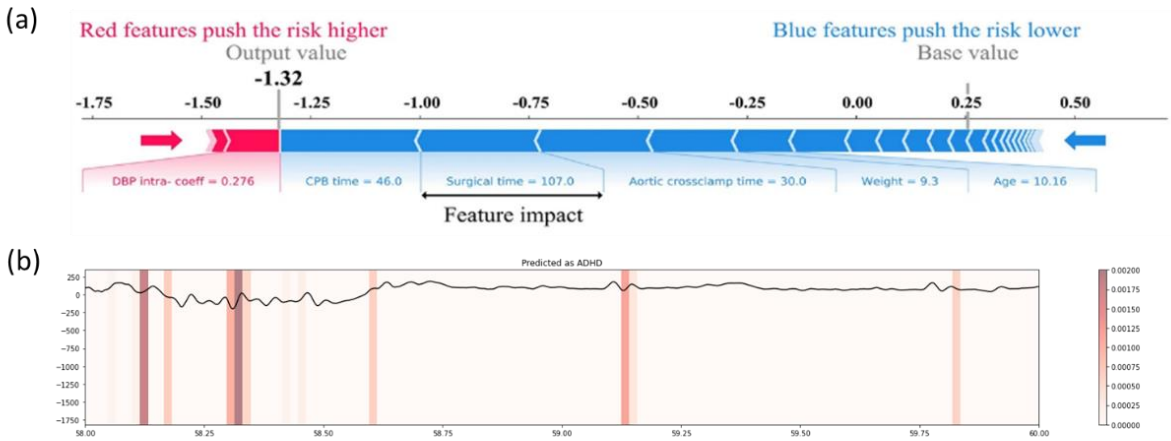


Figure 4: (a) Individual SHAP analysis adopted from Wang et al. [15]. (b) Grad-CAM analysis of ECG Signal.

For example, when we look at the individual SHAP analysis by Zeng et al. [15] in Figure 4a, we can clearly see that the individual is classified as having a low risk of surgical complication by their proposed model due to having a cardiopulmonary bypass (CPB) time of 46 minutes and a surgical time of 107 minutes. SHAP provides clinically relevant explanations because longer CPB time of more than 5 hours is also identified as a risk factor for postoperative acute kidney injury, and longer surgery time indicated higher surgical complexity. Figure 4b depicts another XAI technique known as gradient-weighted class activation mapping (Grad-CAM), which is commonly used in conjunction with DL models, specifically convolutional neural networks (CNN) [18]. Our team provides this Grad-CAM analysis, which shows the significant region of the electrocardiogram (ECG) signals in time that is responsible for this ECG segment to be classified as an ADHD segment. Implementing these XAI techniques can thus overcome the doppelganger effect and increase confidence and trust in AI adoption in healthcare.

3. Other examples of data doppelganger

Despite the negative impact data doppelgangers have on gene expression data, biomedical imaging studies intentionally introduce doppelgangers into their dataset, particularly through data augmentation, to address data scarcity [19]. As such, data augmentation techniques have been applied for brain tumor magnetic resonance imaging [20], and COVID-19 chest X-ray images [21]. In this case, data doppelgangers are not necessarily considered a liability in the medical field.

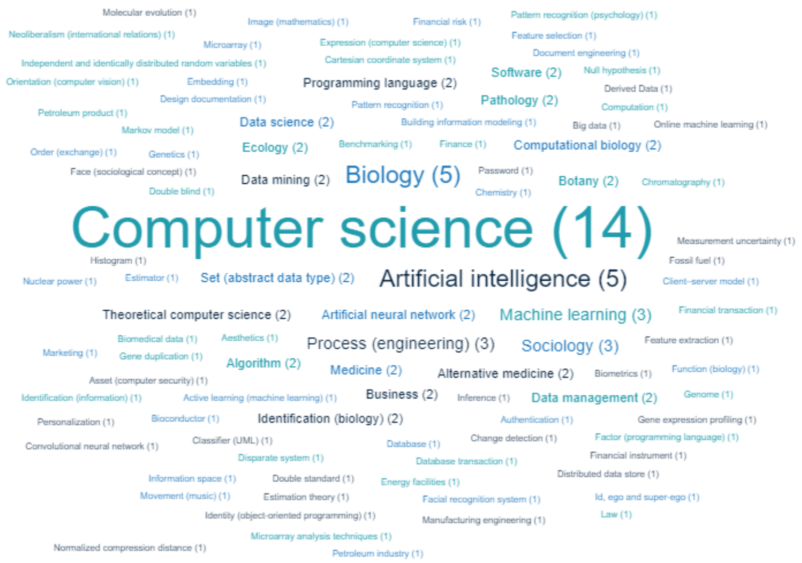


Figure 5: Top field of study for data doppelganger by lens.org

Additionally, data doppelgangers are not just found in biomedical data. We conducted a rough literature search using Lens.org, narrowing our search to journal papers and looking for the term "Data Doppelganger" in the title or abstract. As a result, some notable examples include benchmarking facial recognition tools in differentiating between twins or non-twins who look alike [22], behavioral authentication system in user authentication that distinguishes the user profile from a set of doppelgangers [23], and automatic generation of "digital doppelganger" for financial contracts from the transaction counterparties as a part of financial risk monitoring [24]. Therefore, in light of this, data doppelganger may or may not be useful depending on the application.

References

[1] M. A. Musen, B. Middleton, and R. A. Greenes, “Clinical Decision-Support Systems,” in *Biomedical Informatics*, London: Springer London, 2014, pp. 643–674. doi: 10.1007/978-1-4471-4474-8_22.

[2] F. T. de Dombal, D. J. Leaper, J. R. Staniland, A. P. McCann, and J. C. Horrocks, “Computer-aided Diagnosis of Acute Abdominal Pain,” *BMJ*, vol. 2, no. 5804, pp. 9–13, Apr. 1972, doi: 10.1136/bmj.2.5804.9.

[3] O. S. Pianykh *et al.*, “Improving healthcare operations management with machine learning,” *Nat. Mach. Intell.*, vol. 2, no. 5, pp. 266–273, May 2020, doi: 10.1038/s42256-020-0176-3.

[4] R. F. Mollica and G. L. Fricchione, “Mental and physical exhaustion of health-care practitioners,” *Lancet*, vol. 398, no. 10318, pp. 2243–2244, Dec. 2021, doi: 10.1016/S0140-6736(21)02663-5.

[5] L. R. Wang, L. Wong, and W. W. Bin Goh, “How doppelgänger effects in biomedical data confound machine learning,” *Drug Discov. Today*, vol. 27, no. 3, pp. 678–685, Mar. 2022, doi: 10.1016/j.drudis.2021.10.017.

[6] H. W. Loh, C. P. Ooi, S. G. Dhok, M. Sharma, A. A. Bhurane, and U. R. Acharya, “Automated detection of cyclic alternating pattern and classification of sleep stages using deep neural network,” *Appl. Intell.*, Jun. 2021, doi: 10.1007/s10489-021-02597-8.

[7] L. R. Wang, X. Fan, and W. W. Bin Goh, “Protocol to identify functional doppelgängers and verify biomedical gene expression data using doppelgangerIdentifier,” *STAR Protoc.*, vol. 3, no. 4, p. 101783, Dec. 2022, doi: 10.1016/j.xpro.2022.101783.

[8] L. R. Wang, X. Y. Choy, and W. W. Bin Goh, “Doppelgänger spotting in biomedical gene expression data,” *iScience*, vol. 25, no. 8, p. 104788, Aug. 2022, doi: 10.1016/j.isci.2022.104788.

[9] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. Birrer, “The Doppelgänger Effect: Hidden Duplicates in Databases of Transcriptome Profiles,” *J. Natl. Cancer Inst.*, vol. 108, no. 11, p. djw146, Nov. 2016, doi: 10.1093/jnci/djw146.

[10] W.-Y. Lim, S. Ma, D. Heng, V. Bhalla, and S. K. Chew, “Gender, ethnicity, health behaviour &amp; self-rated health in Singapore,” *BMC Public Health*, vol. 7, no. 1, p. 184, Dec. 2007, doi: 10.1186/1471-2458-7-184.

[11] Tan Tock Seng Hospital, “Enhancing the Health of the Malay Community,” *National Healthcare Group*. https://www.ttsh.com.sg/Community-Health/Central-Health-Stories/Pages/Enhancing-the-Health-of-the-Malay-Community.aspx (accessed Dec. 20, 2022).

[12] O. Faust, H. Razaghi, R. Barika, E. J. Ciaccio, and U. R. Acharya, “A review of automated sleep stage scoring based on physiological signals for the new millennia.,” *Comput. Methods Programs Biomed.*, vol. 176, pp. 81–91, Jul. 2019, doi: 10.1016/j.cmpb.2019.04.032.

[13] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, “Machine Learning and Integrative Analysis of Biomedical Big Data,” *Genes (Basel).*, vol. 10, no. 2, p. 87, Jan. 2019, doi: 10.3390/genes10020087.

[14] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, “Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022),” *Comput. Methods Programs Biomed.*, vol. 226, p. 107161, Nov. 2022, doi: 10.1016/j.cmpb.2022.107161.

[15] X. Zeng *et al.*, “Explainable machine-learning predictions for complications after pediatric congenital heart surgery,” *Sci. Rep.*, vol. 11, no. 1, p. 17244, Dec. 2021, doi: 10.1038/s41598-021-96721-w.

[16] J. Varghese, “Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption,” *Visc. Med.*, vol. 36, no. 6, pp. 443–449, 2020, doi: 10.1159/000511930.

[17] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” May 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” Oct. 2016, doi: 10.1007/s11263-019-01228-7.

[19] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, “Differential Data Augmentation Techniques for Medical Imaging Classification Tasks.,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, vol. 2017, pp. 979–984, 2017, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/29854165

[20] P. Huang, X. Liu, and Y. Huang, “Data Augmentation For Medical MR Image Using Generative Adversarial Networks,” Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.14297

[21] M. Elgendi *et al.*, “The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective,” *Front. Med.*, vol. 8, Mar. 2021, doi: 10.3389/fmed.2021.629134.

[22] S. M. Sami, J. McCauley, S. Soleymani, N. Nasrabadi, and J. Dawson, “Benchmarking human face similarity using identical twins,” *IET Biometrics*, vol. 11, no. 5, pp. 459–484, Sep. 2022, doi: 10.1049/bme2.12090.

[23] M. M. Islam, R. Safavi-Naini, and M. Kneppers, “Scalable Behavioral Authentication,” *IEEE Access*, vol. 9, pp. 43458–43473, 2021, doi: 10.1109/ACCESS.2021.3065921.

[24] P. Kavassalis, H. Stieber, W. Breymann, K. Saxton, and F. J. Gross, “An innovative RegTech approach to financial risk monitoring and supervisory reporting,” *J. Risk Financ.*, vol. 19, no. 1, pp. 39–55, Jan. 2018, doi: 10.1108/JRF-07-2017-0111.