



# Stroke Classification with Two-Dimensional Convolutional Neural Networks: Traditional vs. Generative Adversarial Network-based Data Augmentation for Uneven Class Distribution

Estine Kumar,\* Deshant Singh,\* Anurag Sharma\* and Surya Prakash\*

## Abstract

Uneven class distribution in medical data sets such as stroke data, presents a substantial challenge to the classification model, resulting in suboptimal performance and a biased model toward the majority class. This research analyzes the potential of GAN-based data augmentation over traditional data augmentation methods on a 2-dimensional Convolutional Neural Network to improve sampling. GAN-based data augmentation is used to effectively explore the solution space, and these methods generate synthetic samples to balance the class distribution. The integration of these approaches aims to improve the model's robustness and predictability. Experimental analysis shows promising results in classification accuracy, showing the potential of the GAN-based data Augmentation techniques on the proposed CNN model for dealing with uneven class distribution. Deep Convolutional - Generative Adversarial Network (DCGAN), along with Image-Based ISMOTified-GAN (iSMOTified GAN) and Image-Based Markov Chain Monte Carlo (iMCMC-GAN), is compared to classic data augmentation techniques. The positive results demonstrate the potential of GAN-based data augmentation methods as viable approaches to resolving the issues associated with imbalanced datasets. While all augmentation techniques showed improvements in classification accuracy, DCGAN performed the best, achieving an accuracy of 98.01%. The results of this study provide valuable insights into the potential of GAN-generated synthetic data in enhancing classification, addressing both data limitations and privacy concerns.

**Keywords:** DC-GAN; INMF; Synthetic dataset; iSMOTified GAN; iMCMC GAN.

Received: 07 April 2025; Revised: 20 July 2025; Accepted: 02 August 2025

Type: Research article.

## 1. Introduction

Stroke, a leading cause of mortality and disability worldwide, presents significant challenges for early diagnosis and treatment. Early detection plays a critical role in improving patient outcomes, but the scarcity of large, high-quality datasets hampers the development of robust diagnostic models.<sup>[12,14]</sup> In the context of stroke classification, machine learning (ML) and deep learning (DL) techniques offer promising solutions.<sup>[19]</sup> However, the success of these models relies heavily on the availability of sufficient labeled data, which is often limited due to privacy concerns and the high cost of data collection.

Traditional data augmentation techniques, such as rotation, flipping, and scaling, have been widely used to address data scarcity by artificially expanding the available dataset. While these methods have shown effectiveness in improving model generalization, they are limited in their ability to capture the complex variations present in medical images. In recent years, Generative Adversarial Networks (GANs) have emerged as a powerful tool for generating synthetic data that closely resembles real-world datasets. GANs can increase dataset size while preserving the underlying distribution of the original data, offering an innovative solution to the challenges posed by limited medical data.<sup>[23]</sup>

In particular, GAN-based data augmentation have demonstrated considerable success in generating high-quality, realistic images for various domains, including medical imaging. By leveraging the power of GANs, synthetic stroke images can be generated to augment

*The University of the South Pacific, Laucala Campus, Suva, 1168, Fiji*

\*Email: [estinekumar2k@gmail.com](mailto:estinekumar2k@gmail.com) (E. Kumar),

[singhdesant29@gmail.com](mailto:singhdesant29@gmail.com) (D. Singh),

[anuraganandsharma@usp.ac.fj](mailto:anuraganandsharma@usp.ac.fj) (A. Sharma),

[surya.prakash@usp.ac.fj](mailto:surya.prakash@usp.ac.fj) (S. Prakash)

```

id,gender,age,hypertension,heart_disease,ever_married,work_type,Residence_type,avg_glucose_level,bmi,smoking_status,stroke
9046,Male,67,0,1,Yes,Private,Urban,228.69,36.6,formerly smoked,1
51676,Female,61,0,0,Yes,Self-employed,Rural,202.21,N/A,never smoked,1
31112,Male,80,0,1,Yes,Private,Rural,105.92,32.5,never smoked,1
60182,Female,49,0,0,Yes,Private,Urban,171.23,34.4,smokes,1
1665,Female,79,1,0,Yes,Self-employed,Rural,174.12,24,never smoked,1
56669,Male,81,0,0,Yes,Private,Urban,186.21,29,formerly smoked,0
53882,Male,74,1,1,Yes,Private,Rural,70.09,27.4,never smoked,0
10434,Female,69,0,0,No,Private,Urban,94.39,22.8,never smoked,0
27419,Female,59,0,0,Yes,Private,Rural,76.15,N/A,Unknown,0
60491,Female,78,0,0,Yes,Private,Urban,58.57,24.2,Unknown,0
12109,Female,81,1,0,Yes,Private,Rural,80.43,29.7,never smoked,0
12095,Female,61,0,1,Yes,Govt_job,Rural,120.46,36.8,smokes,0
12175,Female,54,0,0,Yes,Private,Urban,104.51,27.3,smokes,0
8213,Male,78,0,1,Yes,Private,Urban,219.84,N/A,Unknown,0

```

**Fig. 1:** Snapshot of Raw, unprocessed stroke data.

limited datasets, potentially improving classification performance. Monte Carlo simulation approaches have also been shown to be effective in generating and analyzing complex transport or medical logistics scenarios under uncertainty.<sup>[26]</sup> However, while GAN-based augmentation holds significant promise, it remains unclear how it compares to traditional data augmentation methods in the context of stroke classification.

This study aims to evaluate the performance of GAN-based data augmentation in comparison to traditional augmentation techniques for stroke classification using a 2D Convolutional Neural Network (CNN). Specifically, we explore the ability of GANs to generate synthetic stroke images that preserve essential characteristics of real data and assess their impact on model accuracy, precision, recall, and F1 score. The results of this study provide valuable insights into the potential of GAN-generated synthetic data in enhancing stroke classification, addressing both data limitations and privacy concerns.

Recently, Generative AI has gained a lot of attention and is being used in various fields like text, audio, and image creation. The technology behind it mainly relies on three

types of deep learning models: Generative Adversarial Networks (GANs),<sup>[7,10]</sup> Long Short-Term Memory (LSTM),<sup>[25]</sup> and Transformer models.<sup>[8]</sup> Among these, GANs are an unsupervised deep learning model<sup>[21]</sup> that consists of two neural networks: the generator and the discriminator. The two networks work against each other in a kind of back-and-forth competition. The generator keeps creating fake data and tweaks its parameters to make the images more realistic, while the discriminator tries to tell the difference between real and fake data. Over time, both networks get better the generator produces more convincing images, and the discriminator becomes sharper at spotting what's real and what's not.

Generative Adversarial Networks are constantly evolving, resulting in several different variations.<sup>[4]</sup> Convolutional Neural Networks (CNNs)<sup>[10]</sup> are considered the leading models for image processing in deep learning, which led to the creation of Deep Convolutional GAN (DCGAN).<sup>[4]</sup> This variant integrates CNNs with GANs, using convolutional networks for both the generator and discriminator to produce more realistic images.

DCGAN has been applied to various data augmentation

**Table 1:** Stroke dataset attributes and descriptions

Attribute	Description	Type
id	Unique identifier for each patient	Numerical
gender	Gender of the patient	Categorical
age	Age of the patient	Numerical
hypertension	0 = No hypertension, 1 = Has hypertension	Binary
heart_disease	0 = No heart disease, 1 = Has heart disease	Binary
ever_married	Marital status (Yes or No)	Categorical
work_type	Type of occupation	Categorical
Residence_type	Urban or rural residence classification	Categorical
avg_glucose_level	Average glucose level in the patient's blood	Numerical
bmi	Body Mass Index (BMI)	Numerical
smoking_status	Smoking status (formerly smoked, never smoked, smokes)	Categorical
stroke	0 = No stroke, 1 = Had a stroke	Binary

**Table 2:** Parameter Settings for Oversampling and Augmentation Techniques.

Oversampling Technique	Types	Parameter
Data Augmentation	Rotation	20
	Width Shift	0.1
	Height Shift	0.1
	Shear	0.2
	Zoom	0.2
	Horizontal Flip	1
SMOTE	K_neighbors	Integer (default = 5)
	Sampling strategy	Float, string, dict, or callable (default = 'auto')
	Random_state	Integer or RandomState instance (default = None)
DCGAN	Total Neurons per Hidden Layer	1024
	Optimizer	Adam
	Loss Function	Binary Cross-Entropy
	Activation	ReLU
	Normalization	1/255
	Learning Rate	0.001
	Total Neurons per Hidden Layer	1024
	Optimizer	Adam
iSMOTified-GAN	Loss Function	Binary Cross-Entropy
	Activation	Leaky ReLU (0.2)
	Normalization	BatchNorm (0.8)
	Learning Rate	0.0002
	Total Neurons per Hidden Layer	1024
	Optimizer	Adam
	Loss Function	Binary Cross-Entropy
	Activation	Leaky ReLU (0.2)
iMCMC-GAN	Normalization	BatchNorm (0.8)
	Learning Rate	0.0002
	Before Oversampling	4621 : 209
	After Oversampling	4622 : 4208

tasks.<sup>[24]</sup> For instance, Qiufeng Wu and colleagues utilized the DCGAN model for enhancing images of tomato leaf diseases, finding that it can generate data that closely mirrors real images, demonstrating its superior performance compared to BEGAN.<sup>[2]</sup> Similarly, Christine Dewi and her team employed the DCGAN model to generate images of traffic signs, aiming to improve traffic sign recognition.<sup>[6]</sup>

In the medical field, classification models typically require large datasets for training. However, issues such as privacy concerns and low incidence rates often result in a lack of sufficient data, which can lead to lower accuracy rates. ChiunLi Chin and colleagues highlighted the need for more stroke data to enhance the overall accuracy of early ischemic stroke detection, emphasizing the challenge posed by the shortage of medical images.<sup>[5]</sup>

Numerous studies have applied the DCGAN model for data augmentation. For example, Sindhura *et al.* used DCGAN to generate CT images of spinal fractures.<sup>[22]</sup> To improve the accuracy of classifying COVID-19 symptoms, Prerak Mann utilized DCGAN to generate chest CT images

for data augmentation.<sup>[16]</sup> In the case of liver cancer, a leading cause of death, Maayan Frid-Adar and colleagues synthesized liver lesion images using DCGAN, boosting accuracy from 78.6 percent to 85.7 percent, a nearly 7 percent improvement.<sup>[9]</sup>

Chen *et al.* developed a GAN-based CNN architecture to generate brain stroke CT images, achieving success with their approach. However, they faced challenges due to the high processing demands of computer hardware, which led to issues with loss values not converging when trying to generate images at higher resolutions.<sup>[4]</sup>

While significant progress has been made in data augmentation using deep learning techniques, most research focuses on image data, such as MRI and CT scans, which are often limited in availability. Moreover, privacy concerns make it difficult for many patients to undergo scans, creating a gap in data accessibility. Our model addresses this by using tabular data, which poses no privacy issues since patient names are not involved. We introduce a novel data transformation method call INMF that converts tabular data

**Table 3:** A comparison of the stroke image.




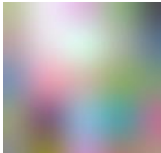

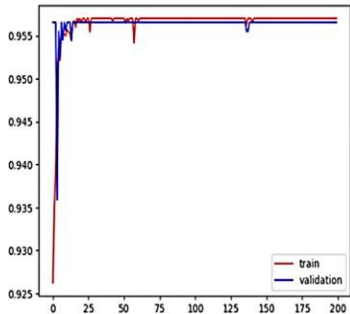
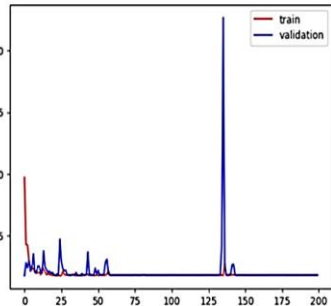
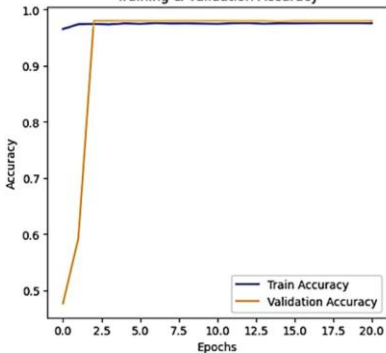
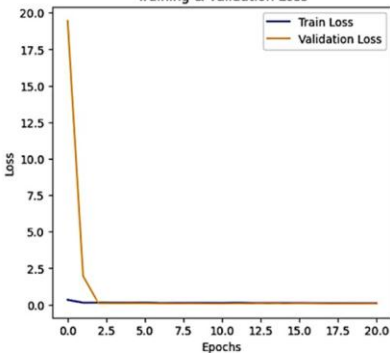
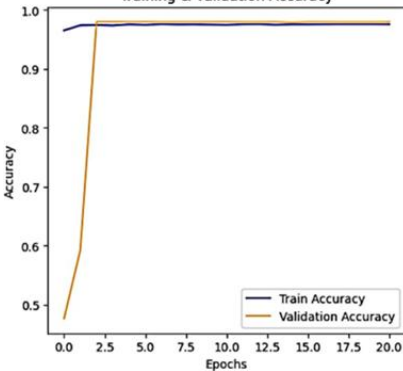
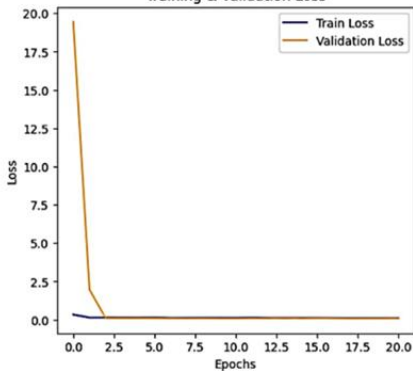
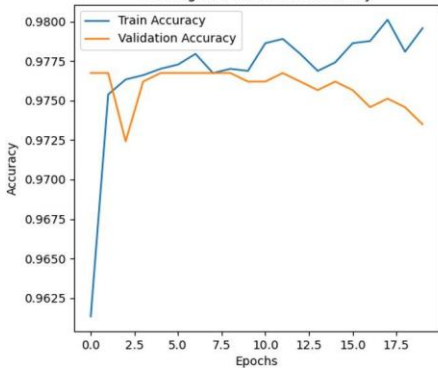
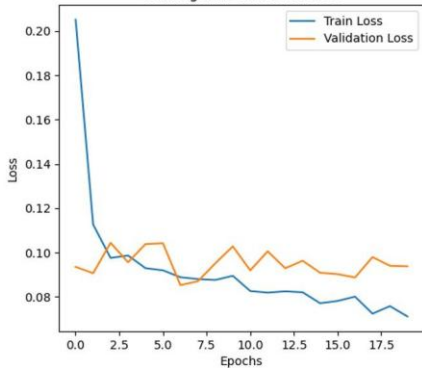
Images	1	2	3	4	5
Image type	Real INMF	Traditional Augmented	DCGAN	iSMOTified GAN	iMCMC GAN
Images					

Table 4: Training and Validation loss and Accuracy.

Traditional Augmented data	<div><div>Training Vs. Validation Accuracy</div><div>Training Vs. Validation Loss</div></div>
DCGAN	<div><div>Training &amp; Validation Accuracy</div><div>Training &amp; Validation Loss</div></div>
iSMOTified GAN	<div><div>Training &amp; Validation Accuracy</div><div>Training &amp; Validation Loss</div></div>
iMCMC GAN	<div><div>Training and Validation Accuracy</div><div>Training and Validation Loss</div></div>

## The Enhanced INMF Algorithm for Image Generation

### Step 1: Input Data

$X$ : Normalized feature matrix with dimensions  $(R \times C)$  (1)

$n$ : Number of components for NMF

### Step 2: Initialization

$W, H$ : Initialize non-negative matrices with random values (2)

$X_{\text{normalized}} = \text{MinMaxScaler}(X)$ : Normalize the input matrix using MinMaxScaler (3)

### Step 3: Non-negative Matrix Factorization (NMF)

$X_{\text{normalized}} \approx W \cdot H$ : Approximate the normalized feature matrix using  $W$  and  $H$  (4)

Iteratively update  $W$  and  $H$  to minimize the reconstruction error: (5)

minimize  $(\|X_{\text{normalized}} - W \cdot H\|_F)$

### Step 4: Image Generation

For each data point  $i = 1$  to  $R$ : (6)

$\text{imgI}_i = \text{Reshape}(W[i, :], \text{dimensions} = (2, 5, 1)) \times 255$  (7)

Convert  $\text{imgI}_i$  to an RGB image:

$\text{imgI}_i^{\text{RGB}} = \text{concat}(\text{imgI}_i, \text{imgI}_i, \text{imgI}_i, \text{axis} = -1)$  (8)

Save each RGB image to the appropriate directory based on class labels  $y_i$ . (9)

into 2D images with RGB values. To our knowledge, no other research has used the INMF method to convert tabular data

to images for DCGAN-based data augmentation for clinical stroke data. We then apply both traditional data augmentation techniques and DCGAN to expand the dataset, addressing data imbalance. Finally, we leverage the high processing power of Convolutional Neural Networks (CNNs) for image processing, enabling more effective recognition and classification.<sup>[3,19]</sup>

## 2. Methodology

### 2.1 Dataset

To carry out this investigation, we obtained the stroke healthcare dataset from Kaggle. The purpose of the dataset is to predict the likelihood of a patient experiencing a stroke based on certain measurable factors.<sup>[1]</sup> An example of the dataset used in our analysis is shown in Fig. 1. The dataset does not contain any identifiable personal information, such as names, addresses, or SSNs, ensuring that there are no confidentiality concerns associated with its use in our research.

This trend underscores the increasing global recognition of oil pollution as a critical environmental issue and the corresponding demand for innovative cleaning technologies. The rising volume of studies emphasizes the relevance of the current research, which contributes to this field by exploring the thermal treatment of oil-contaminated soils and characterizing the extracted oil products. In this work, we thermally treated the oil-contaminated soil in the Karazhanbas field to clean it, and we studied the

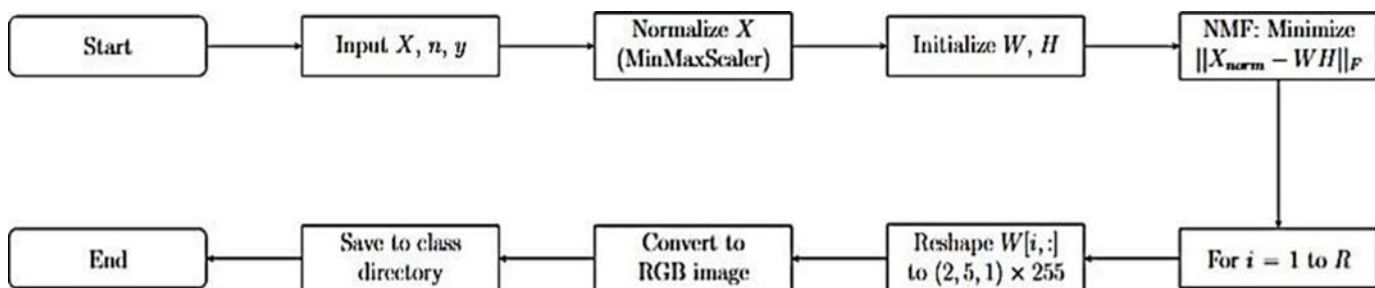


Fig. 2: Flowchart of Enhanced INMF Algorithm for Image Generation

characteristics of the resulting liquid product.

### Dataset description

The dataset used in this study contains records from stroke patients, with a total of 4830 samples prior to oversampling, comprising 4621 non-stroke cases and 209 stroke cases, indicating significant class imbalance. After applying oversampling techniques, the dataset was balanced to 4622 non-stroke and 4208 stroke cases.

The dataset and code are available at: <https://github.com/ECOLS-research-group/Stroke-Classification-Using-2DCNN.git>

The attributes of the dataset are summarized in Table 1. Table 1 presents an outline of the stroke dataset used in this paper with its attributes and description

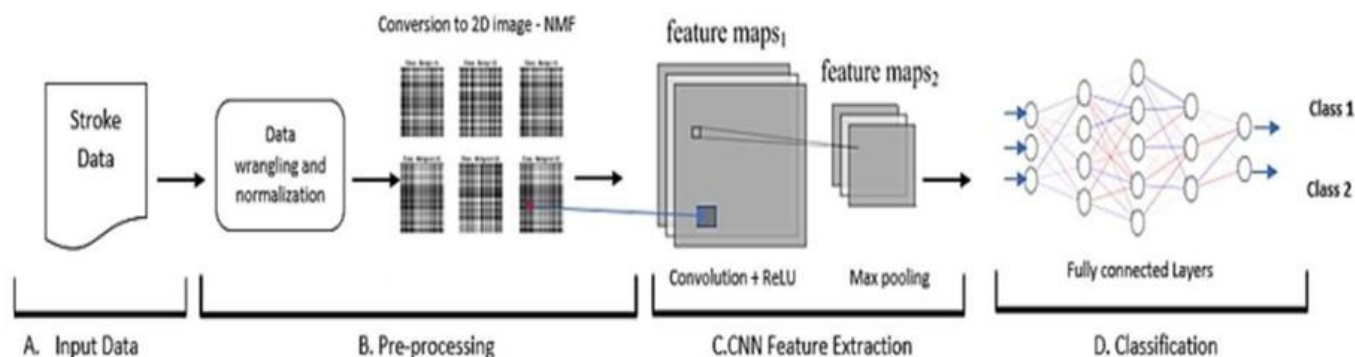
### Pre-processing

During preprocessing, records containing missing values were removed to ensure data quality. Numerical features, such as age, average glucose level, and BMI, were scaled to a range of 0 to 5 to preserve relative differences while standardizing values. Categorical variables were encoded into binary form, with categories represented as 0 or 1 depending on their logical assignment (e.g., 0 for false, 1 for true). This encoding was performed prior to the transformation of the tabular data into image format suitable for input to the convolutional neural network.

### 2.2 Novel data transformation method

Since we are using 1D tabular stroke data, it was crucial for us to transform it into image data for CNN and DCGAN. We





**Fig. 3:** A comprehensive workflow of non-image tabular data classification using CNN.

have used a novel approach called the Image -based Non-Negative matrix factorization to transform the tabular stroke clinical data into images. Further described in section(s) below. The INMF is an enhancement of the Non-negative matrix factorization method.<sup>[18,13]</sup> Refer to Fig. S2 and S3 for INMF architecture and example in [Supplementary Information files](#).



**Fig. 4:** Transformed matrix for 3 data examples from class “Stroke”.

Other data examples can be visualized in the [Supplementary Information files](#). Please see Fig. S4, S5, S6, S7 respectively. Our approach, inspired by the challenges encountered in medical datasets, specifically stroke-related data, seeks to bridge the gap between traditional tabular data and the powerful capabilities of CNN. The INMF method has been introduced in our previous work as well where we simply transformed tabular stroke data into 2-Dimensional image like representations. While NMF has been widely applied in image-based tasks, our work takes a unique approach by using it to convert tabular data into images. This is particularly relevant for medical datasets, such as stroke-related patient data, where each row contains one-dimensional information like age, gender, and health indicators. Our method aims to bridge the gap between traditional tabular formats and the image-based input required by Convolutional Neural Networks (CNNs).<sup>[14]</sup>

The enhanced algorithm and samples are available at: <https://github.com/ECOLS-research-group/Stroke-Classification-Using-2DCNN.git>

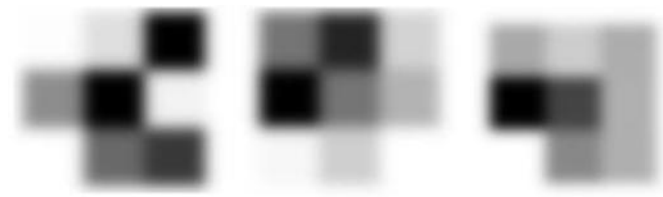


**Fig. 5:** Transformed matrix for 3 data examples from class “No Stroke”.

The data set used in our experiments is a curated collection of patient records, each row representing a 1D data vector characterized by multiple attributes such as age, sex, and health indicators. The flow of the non-image data classification can be visualized in Fig. 2 and 3, followed by illustrations of the transformed matrix in Fig. 4 and 5.

### 3. Experimental analysis

All experiments were conducted using GPU-accelerated TensorFlow with a Keras architecture. A 2D CNN model was employed for feature extraction and classification, following a similar structure to previous deep learning-based approaches. The experiments aimed to compare the effectiveness of traditional data augmentation techniques and GAN-based augmentation in improving classification

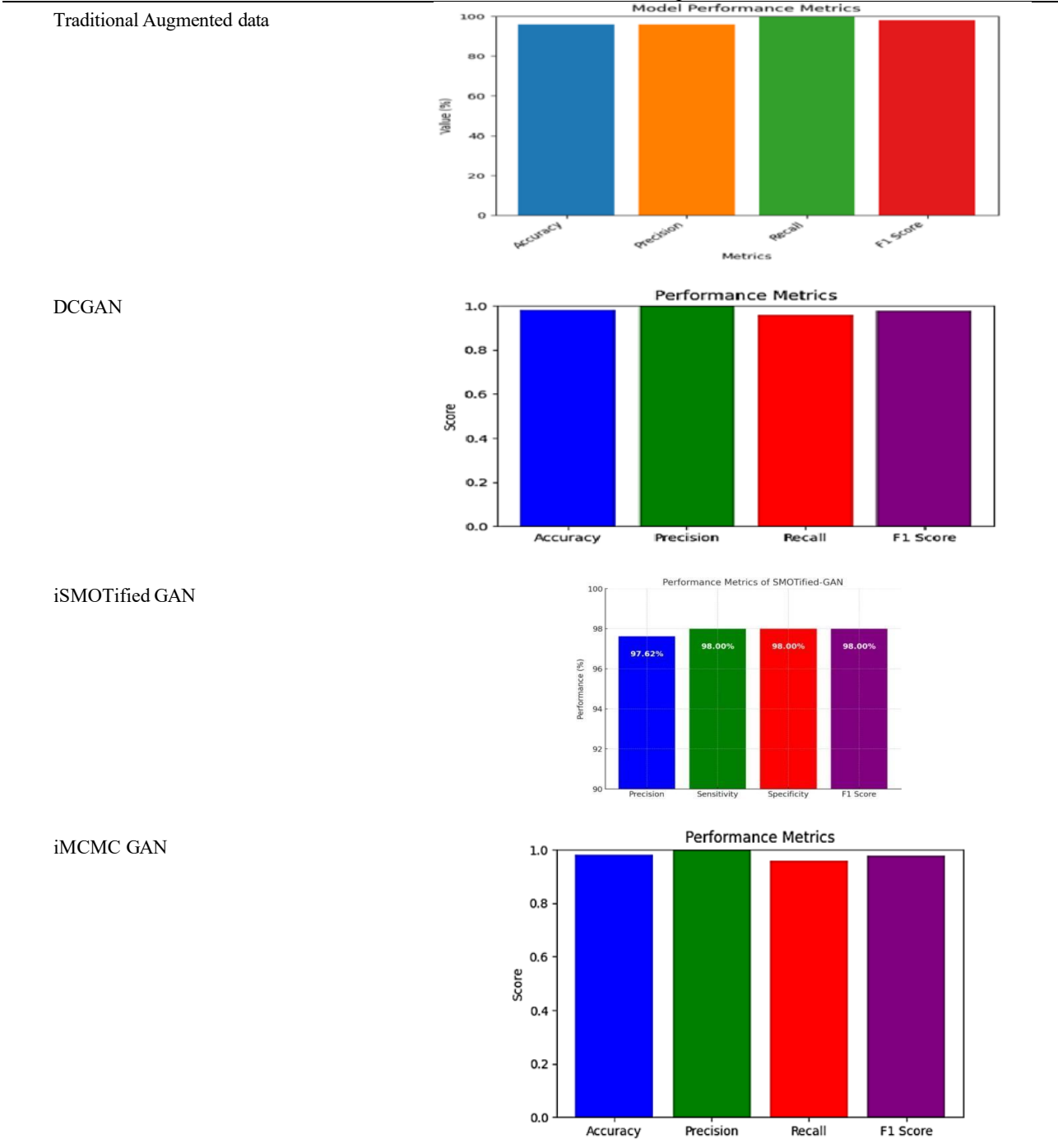


**Fig. 6:** Features learned by the 1st convolutional layer for stroke dataset.

performance. The CNN architectures used for the augmented and GAN-generated datasets have notable differences to optimize performance for each data type. The traditional augmented dataset model uses a larger input size (256x256) with a smaller dropout rate (0.1), as traditional augmentation preserves original data characteristics. In contrast, the GAN-based dataset model adopts a smaller input size (224x224) with a higher dropout rate (0.5) to prevent overfitting on synthetic data. Additionally, the GAN model employs early stopping and learning rate reduction techniques to stabilize training, as GAN-generated images can introduce variability requiring careful optimization. These modifications aim to enhance generalization and performance for both datasets while ensuring that the model adapts effectively to both real and synthetic data distributions. An example of the CNN features learned is depicted in Fig. 6 and 7.

Table 2: presents a summary of the different oversampling

Table 5: Performance Result: Graphs.



techniques and their associated parameters used in the study. The table includes various data augmentation methods, such as rotation, width shift, and horizontal flip, along with their respective parameter values. It also outlines the specific settings for SMOTE, DCGAN, iSMOTified-GAN, and iMCMC-GAN, including details like the number of neighbors, optimizer, loss function, and learning rate. Additionally, the table shows the class ratio before and after oversampling, highlighting the balance between the 'No Stroke' and 'Stroke' classes.

3.1 Oversampling for class imbalance

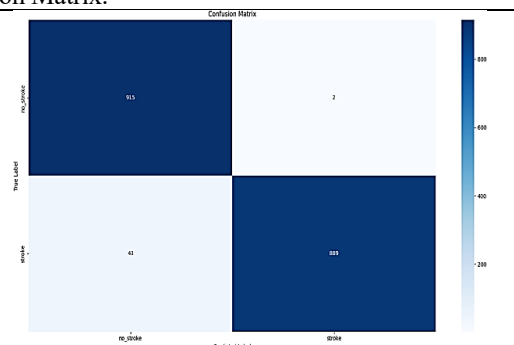
The Stroke healthcare dataset exhibited high imbalance leading to incorrect results as depicted in Fig. 8. To enhance its utility for our CNN model, we deployed SMOTE, data augmentation, Image-based iSMOTified-GAN (iSMOTified-GAN) and Image-based MCMC-GAN (iMCMC-GAN) to increase the size of our minority class.

3.2 Oversampling for class imbalance techniques

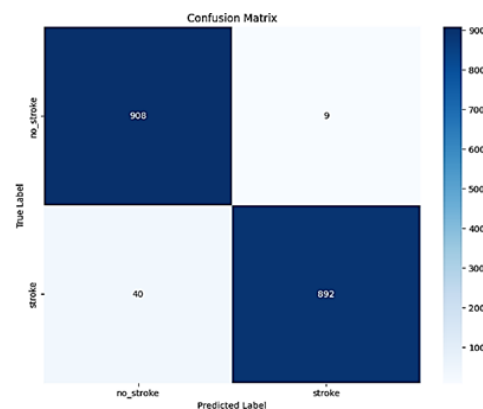
Data Augmentation is any technique used to artificially generate new data from the original dataset. SMOTE is also a type of data augmentation. However apart from SMOTE, we also deployed

**Table 6:** Confusion Matrix.

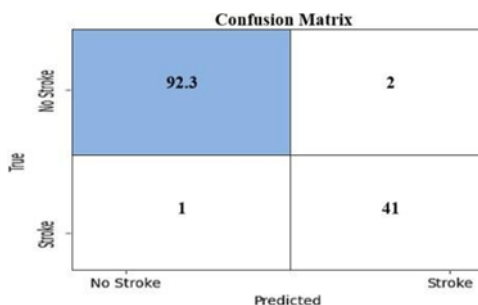
Confusion matrix of the CNN Model when fed with transformed and iSMOTified- GAN dataset



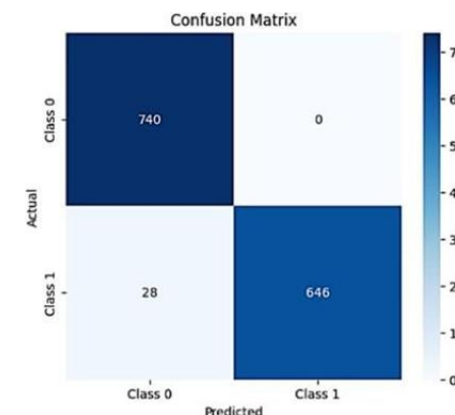
Confusion matrix of the CNN Model when fed with transformed and iMCMC dataset



Confusion matrix of the CNN Model when fed with transformed and traditional augmented dataset



Confusion matrix of the CNN Model when fed with transformed and synthetic dataset from DCGAN



other types of Data augmentation methods such as image rotation, image shift, zoom and others to be able to create more images.

Table 2 depicts the parameter settings for the technique while a comparison of the stroke images using all the methods can be visualized in Table 3.

Synthetic Minority Over-sampling Technique (SMOTE) SMOTE is one of the most commonly used methods to handle imbalanced datasets.<sup>[15]</sup> SMOTE utilizes the k-nearest neighbors'

algorithm to generate synthetic examples by interpolating along the line segments connecting any of the k-minority class neighboring instances.<sup>[15]</sup> We used SMOTE on the original dataset to be able to generate more data. Table 2 depicts the parameter settings for the technique.

### 3.3 iSMOTified-GAN

iSMOTified-GAN<sup>[23]</sup> is a hybrid oversampling technique



**Table 7:** Validation Matrix for the CNN Model

Type of Dataset	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Non-oversampled	92.00	96.00	94.00	96.07
Traditional augmented Data	96.80	100	96.00	98.00
Synthetic Data (DC-GAN)	98.01	95.85	100	98.00
iSMOTified-GAN	97.62	98.00	98.00	98.00
iMCMC-GAN	97.67	98.00	98.00	98.00

designed to address the challenge of class imbalance in image datasets by combining SMOTE with GAN. Hybrid approaches combining domain knowledge and neural architectures have shown promise in guiding model convergence and improving robustness.<sup>[27]</sup> In its original form, iSMOTified-GAN<sup>[20]</sup> was designed to work with tabular data. iSMOTified adapts SMOTE to operate in image feature space, preserving structural and contextual information. The GAN then refines these initial synthetic images, enhancing their realism and diversity through adversarial training.<sup>[23]</sup> This combined approach significantly improves the quality of synthetic data, leading to more balanced training datasets and enhancing the performance of machine learning models on imbalanced medical image classification tasks.<sup>[23,20]</sup> Fig. 9 outlines the flow for sample generation using iSMOTified-GAN.

### 3.4 iMCMC-GAN

iMCMC-GAN<sup>[17]</sup> is a hybrid generative model that combines the sampling power of Markov Chain Monte Carlo (MCMC)<sup>[11]</sup> with the data synthesis capabilities of GAN. iMCMC-GAN leverages MCMC to generate diverse and representative initial samples from the minority class distribution.<sup>[17]</sup> These samples then guide the GANs generator, replacing the traditional random noise input. The GAN further refines these samples, enhancing their realism and quality. By integrating these two approaches, iMCMC-GAN produces high-quality synthetic images that improve the training of classification models, ultimately enhancing their performance on imbalanced medical image datasets.<sup>[17]</sup> Fig. 10 outlines the process flow using the iMCMC-GAN.

### 3.5 DC-GAN based augmentation

A Deep Convolutional Generative Adversarial Network (DCGAN) was trained to generate synthetic stroke images. The generated images were mixed with real samples in

proportion and used to train the CNN classifier. The DCGAN model includes A generator network utilizing transposed convolution layers to generate high-quality synthetic images as illustrated in Fig. 11 below. A discriminator network employing convolutional layers to distinguish between real and synthetic images. A comparison is depicted in Table 3 below between the original image, its augmented and synthetic forms. The DCGAN and CNN Process Flow Diagram for Stroke Classification is illustrated in Fig. 11. Table 3 shows A comparison of the stroke image transformed by INMF, its corresponding augmented form and synthetic image produced by GAN based augmentation techniques: DCGAN, iSMOTified GAN and iMCMC-GAN.

### 3.6 Hyperparameter selection and tuning

Key hyperparameters for GAN-based augmentation methods (DCGAN, iSMOTified-GAN, and iMCMC-GAN) were chosen through a combination of literature-backed defaults and empirical fine-tuning. The main aim was to ensure training stability, high-quality synthetic images, and better classification performance on the CNN model:

- Epochs: Set to 500 for all GAN models, with early stopping used to prevent overfitting once generator loss stabilized.
- Batch Size: Fixed at 64 to balance between convergence speed and computational efficiency.
- Learning Rate: 0.001 for DCGAN; 0.0002 for iSMOTified-GAN and iMCMC-GAN, following common GAN training guidelines.
- Optimizer: Adam optimizer with  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$ , known for stable adversarial training.

Activation Functions: ReLU for DCGAN and Leaky ReLU ( $\alpha = 0.2$ ) for the hybrid models.

**Table 8:** Performance metrics of different algorithms.

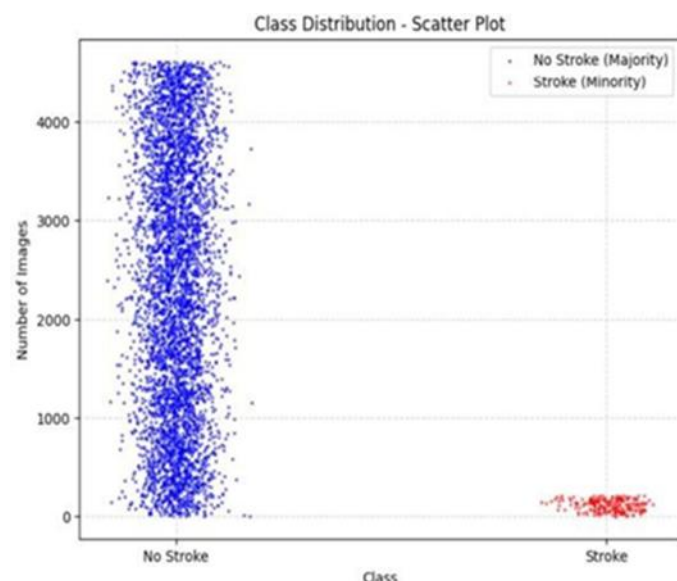
Algorithm	Accuracy	Recall	Precision	F1 Score
Random Forest	95.70%	100%	96.00%	98.00%
Logistic Regression	95.00%	100%	95.00%	98.00%
Decision Tree	93.61%	96.00%	96.00%	96.00%
Naive Bayes	87.50%	88.00%	96.00%	92.00%
Support Vector Machine	95.00%	100%	96.00%	98.00%
Proposed Method: Augmented (of this paper)	96.80%	100%	96.00%	98.00%
iSMOTified-GAN (of this paper)	97.62%	98.00%	98.00%	98.00%
iMCMC-GAN (of this paper)	97.67%	98.00%	98.00%	98.00%
Proposed Method: DC-GAN (of this paper)	98.01%	95.85%	100%	98.00%



**Fig. 7:** Features learned by the 10th convolutional layer for stroke dataset.

- Normalization: Batch normalization with momentum of 0.8 to promote consistent learning across layers.

Although a comprehensive grid search was not feasible, targeted tuning experiments were performed by adjusting each hyperparameter and selecting the setup that produced the best image quality and CNN classification accuracy.



**Fig. 8:** Class distribution for the stroke data before oversampling.

#### 4. Results and discussion

This study evaluates the impact of traditional and GAN-based data augmentation on stroke classification using a 2D CNN. The Stroke Healthcare dataset from Kaggle was used for experimentation. The classification performance of CNN trained with GAN-based augmented images was compared against CNN trained with traditionally augmented images. Additionally, the results were benchmarked against conventional machine learning models, including Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and Nave Bayes, which are often paired with feature selection techniques such as Information Gain, Rough Set, or Weighted Nave Bayes.

As compared to the non-oversampled data, the accuracy of the CNN model improved significantly using various augmentation techniques. Traditional data augmentation boosted the accuracy by 4.8%, with notable improvements in the precision, sensitivity, and F1-score, reflecting a more balanced and accurate model. Specifically, the traditional augmented dataset achieved 96.8% precision, 100%

sensitivity, 96% specificity, and 98% F1-score. Among the synthetic data augmentation methods, DCGAN demonstrated the best performance, increasing the accuracy by 6.01% compared to the non-oversampled dataset. The DC-GAN based synthetic data achieved 98.01% precision, 95.85% sensitivity, 100% specificity, and 98.00% F1-score outperforming both the non-oversampled and traditionally augmented datasets. Other synthetic techniques like iSMOTified-GAN and iMCMC-GAN also showed strong results, with iMCMC-GAN slightly edging out the others, achieving 97.67% precision, 98% sensitivity, 98% specificity, and 98% F1-score. These results underscore the impact of augmentation on improving model performance, with DCGAN achieving the most notable improvement in accuracy this can be compared in Table 5. Table 4 illustrates the accuracy and loss curves for the training and validation sets, providing insight into the learning dynamics of all the augmentation approaches.

**Table 9:** Experimented dataset summary.

Attributes	Instances	Missing Values	Class Ratio (No Stroke: Stroke)
7	9230	None	4622:4208

Table 4 Shows the training and validation loss and accuracy for the 4 augmentation approaches used, this is presented in the form of graphs.

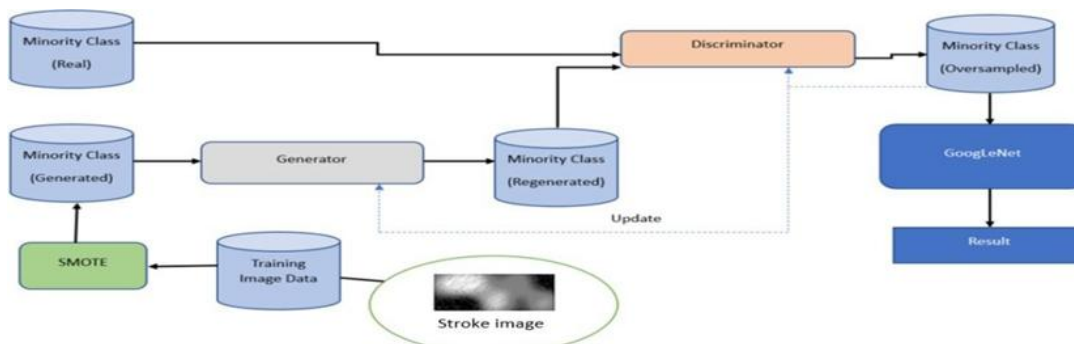
##### 4.1 Confusion matrix of the proposed model

The effectiveness of machine learning methods is evaluated based on several performance metrics. A confusion matrix presents the results of predictions in a classification task, summarizing the accurate and inaccurate forecasts for each class. It provides valuable information about the types of errors made. To assess these parameters, a confusion matrix incorporating actual and predicted data (represented as A, B, C, and D) is constructed. Here, A = True Positive, B = True Negative, C = False Positive, and D = False Negative. Table 6 shows the confusion matrix of the proposed models.

Table 7 summarizes the validation metrics for the CNN model across different types of datasets. The table presents the performance of the model in terms of precision, sensitivity, specificity, and F1-score for each dataset type, including non-oversampled data, traditional augmented data, and synthetic data generated by DC-GAN, iSMOTified-GAN, and iMCMC-GAN. The results highlight the improvements in model performance, with the synthetic data (DC-GAN) achieving the highest precision and specificity, while the iSMOTified-GAN and iMCMC-GAN datasets maintain consistently high scores across all metrics.

##### 4.2 Performance of the proposed model

The performance of the proposed model is depicted in Table 7, where the Accuracy score, Sensitivity, Precision, and the F1 Score of the models are presented. Table 7: summarizes the



**Fig. 9:** Process of sample generation with iSMOTified-GAN.

validation metrics for the CNN model across different types of datasets. The table presents the performance of the model in terms of precision, sensitivity, specificity, and F1-score for each dataset type, including non-oversampled data, traditional augmented data, and synthetic data generated by DC-GAN, iSMOTified-GAN, and iMCMC-GAN. The results highlight the improvements in model performance, with the synthetic data (DC-GAN) achieving the highest precision and specificity, while the iSMOTified-GAN and iMCMC-GAN datasets maintain consistently high scores across all metrics.

Table 5 Shows the performance results for the 4 augmentation approaches used; this is presented in the form of bar graphs.

### 4.3 Result analysis

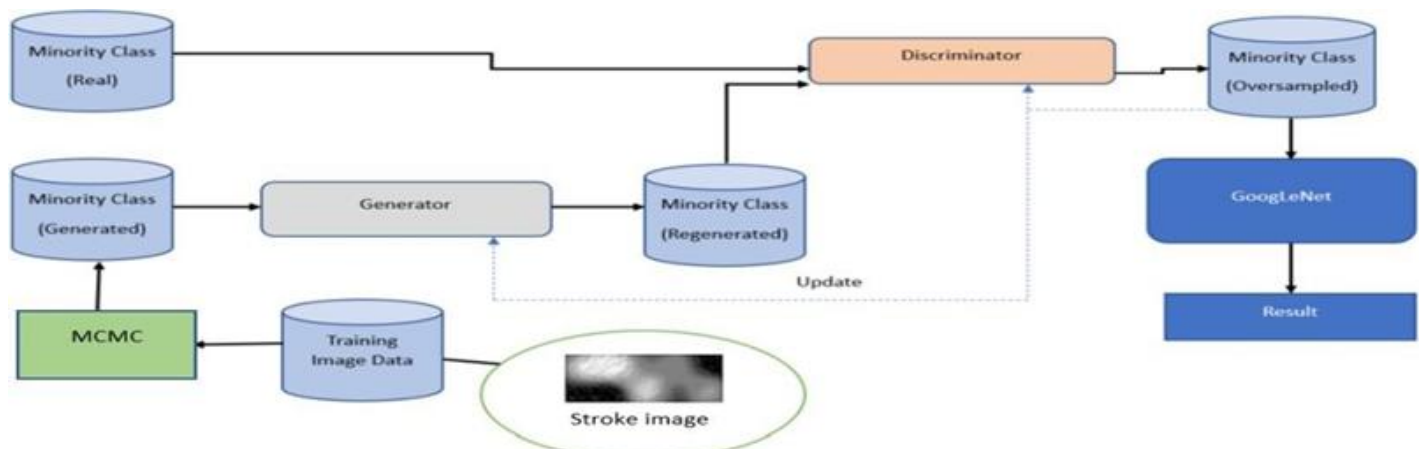
We have conducted a comparison between our model and several traditional machine learning models. The outcomes of this comparative analysis are presented in Table 8 for a comprehensive performance evaluation. Table 8 illustrates that our proposed CNN model surpasses the performance of other classical machine learning models. An experimented dataset summary is also outlined in Table 9.

It is quite evident that the suggested CNN model performs better than other established machine learning models. Upon examination, it can be observed that the

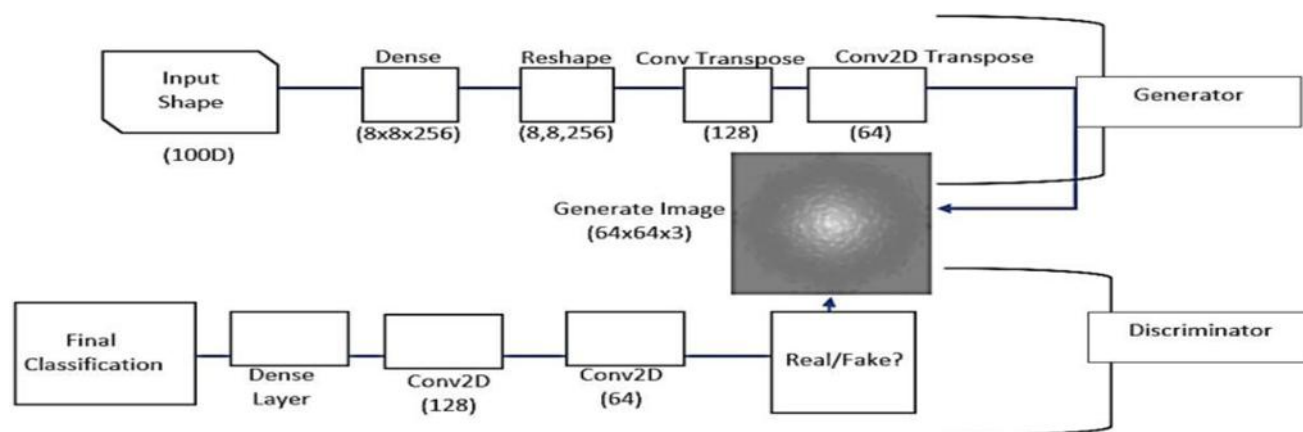
performance of Logistic Regression closely approaches that of the proposed model, while the Naive Bayes algorithm demonstrates the poorest performance among the methods mentioned.

Table 8 presents the performance metrics of various algorithms, including traditional machine learning models and methods proposed in this paper. The table below lists accuracy, recall, precision, and F1 score for each algorithm, showcasing how they compare in terms of classification performance. Among the traditional algorithms, Random Forest and Support Vector Machine show high accuracy, while Naive Bayes has the lowest performance across all metrics. The proposed methods (augmented data, iSMOTified-GAN, iMCMC-GAN, and DC-GAN) demonstrate superior performance, with DC-GAN achieving the highest precision of 100%. These results emphasize the effectiveness of the proposed synthetic data generation techniques for improving classification accuracy and other key metrics.

Table 9 provides a summary of the dataset used in the experiments. The dataset consists of 7 attributes and a total of 9,230 instances. There are no missing values in the dataset, ensuring the completeness of the data. The class ratio between 'No Stroke' and 'Stroke' is 4,622:4,208, indicating a relatively balanced distribution between the two classes. This balanced dataset is crucial for training the models effectively,



**Fig. 10:** Process of sample generation with iMCMC-GAN.



**Fig. 11:** DCGAN Architecture which illustrates the Generator and Discriminator.

ensuring that both classes are represented adequately.

## 5. Conclusion

This study compared traditional data augmentation techniques with various synthetic augmentation methods for stroke classification using Convolutional Neural Networks (CNNs). By transforming raw tabular stroke prediction data into image form, we assessed the impact of different augmentation strategies on model performance. The results show that each augmentation technique led to significant improvements in classification accuracy, with traditional augmentation methods enhancing model performance by increasing diversity in the dataset. The use of synthetic methods, such as DCGAN, iSMOTified-GAN, and iMCMC-GAN, further boosted performance, with DCGAN achieving the highest improvement in accuracy.

Our approach demonstrated that a combination of traditional and advanced synthetic augmentation strategies can effectively enhance the diversity of training samples, leading to improved model generalization. Specifically, the traditional augmentation method improved accuracy by 4.8%, while DCGAN-based augmentation achieved the highest accuracy improvement of 6.01%. The findings suggest that augmentation techniques, including both conventional and generative approaches, can be pivotal in advancing stroke classification models.

In terms of clinical relevance, the ability to generate realistic synthetic stroke data using GANs holds promise for supporting diagnostic models in situations where real patient data is limited due to privacy concerns or data collection challenges. Synthetic data can potentially assist in training models that generalize better to diverse patient populations, helping to reduce bias in stroke prediction tools. However, deploying such models in real-world healthcare settings presents challenges, including ensuring regulatory

compliance, validating model predictions on prospective clinical data, and addressing clinician trust and interpretability concerns. Future work should focus on bridging this gap by collaborating with healthcare practitioners to evaluate the models in clinical workflows and to refine synthetic data generation processes to align with clinical requirements.

While our experiment focused on transforming tabular stroke data, future work could explore how these augmentation techniques perform on other types of structured data, such as time-series datasets, multi-modal healthcare records, and other numerical data representations. As this work progresses, further optimization of GAN architectures, augmentation strategies, and dataset generalization could enhance the effectiveness of this approach. Expanding the use of CNN-based models to diverse datasets would further demonstrate the versatility and potential of these models in predictive analytics beyond traditional image domains.

## Acknowledgments

This work was supported by the University of the South Pacific.

## Conflict of Interest

There is no conflict of interest.

## Supporting Information

Applicable.

## Credit Statement

**Estine Kumar:** Conceptualization, developed the methodology, curated and analyzed the data, conducted the experiments, performed the formal analysis, and prepared the original draft of the manuscript. **Deshant Singh:** Data Analysis, Contributed to some parts of the Methodology and Results. **Anurag Sharma:** provided validation, project



supervision, and contributed to reviewing and editing the manuscript, funding acquisition. **Surya Prakash:** contributed to validation, project supervision, funding acquisition, and also supported the review and editing process. All authors reviewed and approved the final manuscript.

## References

- [1] M. Ashrafuzzaman, S. Saha, K. Nur, Prediction of stroke disease using deep CNN based approach, *Journal of Advances in Information Technology*, 2022, **13**(6), doi: 10.12720/jait.13.6.604-613
- [2] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks, *arXiv*, preprint, 2017, arXiv:1703.10717.
- [3] Ljubomir Buturović and Dejan Miljković. A novel method for classification of tabular data using convolutional neural networks, *BioRxiv*, 2020, 2020-05, doi: 10.1101/2020.05.02.074203.
- [4] Y.-X. Chen, G.-Y. Shih, H.-W. Ting, T.-Y. Chien, Base on GAN Combined with CNN Architecture to Generate Brain Stroke CT Images, *Proceedings of the 2024 8th International Conference on Medical and Health Informatics*, Yokohama Japan, ACM, 2024, 47–51, doi: 10.1145/3673971.3674024.
- [5] C.-L. Chin, B.-J. Lin, G.-R. Wu, T.-C. Weng, C.-S. Yang, R.-C. Su, Y.-J. Pan, An automated early ischemic stroke detection system using CNN deep learning algorithm, *IEEE 8th International Conference on Awareness Science and Technology (iCAST)*, Taichung, Taiwan, China, November 8-10, 2017, 368-372, doi: 10.1109/ICAwST.2017.8256481.
- [6] C. Dewi, R.-C. Chen, X. Jiang, H. Yu, Deep convolutional neural network for enhancing traffic sign recognition developed on Yolo V4, *Multimedia Tools and Applications*, 2022, **81**, 37821-37845, doi: 10.1007/s11042-022-12962-5.
- [7] X. Du, X. Ding, M. Xi, Y. Lv, S. Qiu, Q. Liu, A data augmentation method for motor imagery EEG signals based on DCGAN-GP network, *Brain Sciences*, 2024, **14**, 375, doi: 10.3390/brainsci14040375.
- [8] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs, *arXiv*, 2020, preprint, arXiv:2012.09699.
- [9] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, H. Greenspan, GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification, *Neurocomputing*, 2018, **321**, 321-331, doi: 10.1016/j.neucom.2018.09.013.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, Generative adversarial networks, *Advances in Neural Information Processing Systems*, ACM, **63**(11), 2020, 139-144. ISSN: 0001-0782, doi: 10.1145/3422622.
- [11] G. L. Jones, Q. Qin, Markov chain Monte Carlo in practice, *Annual Review of Statistics and Its Application*, 2022, **9**, 557-578, doi: 10.1146/annurev-statistics-040220-090158.
- [12] S. Kandaya, A. R. Abdullah, N. M. Saad, I. H. Azman, E. F. Shair, N. H. Ali, Analysis of early stroke diagnosis based on brain magnetic resonance imaging using machine learning, *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 2023, **32**, 241-255, doi: 10.37934/araset.32.3.241255.
- [13] H. Kim, H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, *Bioinformatics*, 2007, **23**, 1495-1502, doi: 10.1093/bioinformatics/btm134.
- [14] Estine Kumar, Surya Prakash, and Anurag Sharma. "Stroke Classification Using 2-D Convolutional Neural Networks". Accepted for publication at the *6th International Conference on Communication and Intelligent Systems (ICCIS)*, Springer Nature Singapore, Maulana Azad National Institute of Technology (MANIT), Bhopal, India, Nov 2024, doi: 10.1007/978-981-96-5729-2\_33.
- [15] A. Kummer, T. Ruppert, T. Medvegy, J. Abonyi, Machine learning-based software sensors for machine state monitoring - The role of SMOTE-based data augmentation, *Results in Engineering*, 2022, **16**, 100778, doi: 10.1016/j.rineng.2022.100778.
- [16] P. Mann, S. Jain, S. Mittal, A. Bhat, Generation of COVID-19 Chest CT Scan Images using Generative Adversarial Networks, *International Conference on Intelligent Technologies (CONIT)*, Hubli, India, June 25-27, 2021, doi: 10.1109/conit51480.2021.9498272.
- [17] N. Nishika, A. Sharma, Exploring MCMC guided GAN and comparative analysis for uneven class distribution, *IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAET)*, Kota Kinabalu, Malaysia, August 26-28, 2024, 177-182, doi: 10.1109/IICAET62352.2024.10730311.
- [18] P. Paatero, U. Tapper, Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, *Environmetrics*, 1994, **5**, 111-126, doi: 10.1002/env.3170050203.
- [19] A. Sharma, D. Kumar, Classification with 2-D convolutional neural networks for breast cancer diagnosis, *Scientific Reports*, 2022, **12**, 21857, doi: 10.1038/s41598-022-26378-6.
- [20] A. Sharma, P. K. Singh, R. Chandra, SMOTified-GAN for class imbalanced pattern classification problems, *IEEE Access*, 2022, **10**, 30655-30665.
- [21] David L Silver *et al.* Icml2011 unsupervised and transfer learning workshop, In: Proceedings of ICML Workshop on Unsupervised and Transfer Learning, *JMLR Workshop and Conference Proceedings*, 2012, **27**, 1–15, doi: 10.1109/IJCNN.2011.6033302.
- [22] N. Sindhura D, R. M. Pai, S. N. Bhat, M. Manohara Pai M, Synthetic Vertebral Column Fracture Image Generation by Deep Convolution Generative Adversarial Networks, *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, July 9-11, 2021, 1–4, doi: 10.1109/conecct52877.2021.9622527.
- [23] Deshant Singh and Anurag Sharma. "Improving Imbalanced Tuberculosis Diagnosis Using GoogleNet with



Image-Based SMOTified-GAN”. Accepted for publication at the *3rd International Conference on Advances in Data driven Computing and Intelligent Systems (ADCIS)*, BITS Pilani, K K Birla Goa Campus, India, Sept. 2023.

[24] Q. Wu, Y. Chen, J. Meng, DCGAN-based data augmentation for tomato leaf disease identification, *IEEE Access*, 2020, **8**, 98716-98728.

[25] Y. Yu, X. Si, C. Hu, J. Zhang, A review of recurrent neural networks: LSTM cells and network architectures, *Neural Computation*, 2019, **31**, 1235-1270, doi: 10.1162/neco\_a\_01199.

[26] Surya Prakash, Emergency relief goods transportation strategies—a Monte Carlo simulation approach, In: *Australasian Transport Research Forum*, 2019.

[27] S. Prakash, B. Sharma, An optimized hybrid approach for path planning: a combination of Lyapunov functions and high-level planning algorithms. *Advances in Data-Driven Computing and Intelligent Systems*, Springer Nature Singapore, 2024, 425-436, doi: 10.1007/978-981-99-9524-0\_32.

**Publisher’s Note:** Engineered Science Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits the use, sharing, adaptation, distribution and reproduction in any medium or format, as long as appropriate credit to the original author(s) and the source is given by providing a link to the Creative Commons license and changes need to be indicated if there are any. The images or other third-party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

©The Author(s) 2025