**Step 1: Applying Advanced Deep Learning Models for Cancer Diagnosis**

In this step, four advanced deep learning models were utilized to diagnose cancer diseases. The models employed were Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Gated Recurrent Unit (GRU). Each model was trained and evaluated for its effectiveness in cancer diagnosis.  
NOTE: For Lung cancer dataset all these models were trained for 100 epochs, Prostate cancer dataset all the models were trained for 15 epochs, and for Breast cancer dataset all the models were trained for 150 epochs.

1. *Multilayer Perceptron (MLP)*: The MLP model consisted of multiple layers of neurons, with two hidden layers of 64 units each. The activation function used was ReLU, and the Adam optimizer was employed for training.
2. *Convolutional Neural Network (CNN)*: The CNN model incorporated convolutional and pooling layers for feature extraction, followed by fully connected layers for classification. The model was trained using binary cross-entropy loss and optimized using the Adam optimizer.
3. *Recurrent Neural Network (RNN)*: The RNN model utilized Long Short-Term Memory (LSTM) units to capture temporal dependencies in the data. The model consisted of an LSTM layer followed by a dense layer with sigmoid activation for binary classification. Training was performed using binary cross-entropy loss and the Adam optimizer.
4. *Gated Recurrent Unit (GRU)*: The GRU model employed Gated Recurrent Units (GRUs), which are like LSTMs but with a simplified architecture. The model included a GRU layer followed by a dense layer with sigmoid activation for binary classification. Training utilized binary cross-entropy loss and the Adam optimizer.

Table

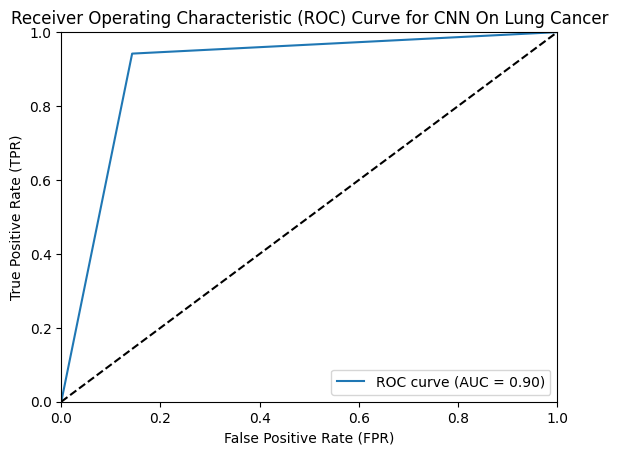


* *Lung Cancer Dataset*:
  + The CNN, RNN, and GRU models achieved similar performance, with higher AUC and accuracy compared to the MLP model. Therefore, the CNN, RNN, and GRU models can be considered better performers for the Lung Cancer dataset.
* Breast Cancer Dataset:
  + The MLP model achieved the highest AUC and accuracy compared to the other models for the Breast Cancer dataset. Hence, the MLP model can be considered the better performer for this dataset.
* Prostate Cancer Dataset:
  + The GRU model achieved the highest AUC and accuracy among all the models for the Prostate Cancer dataset. Therefore, the GRU model can be considered the better performer for this dataset.

In summary, based on the provided results, the following models performed better on each dataset:

* Lung Cancer Dataset: CNN, RNN, GRU
* Breast Cancer Dataset: MLP
* Prostate Cancer Dataset: GRU

Following are the ROC Curve Graphs:

A picture containing text, screenshot, line, plot

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generated

A picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, screenshot, line, plot

Description automatically generated

**Step 2. Applying GAN (CTGAN) for Synthetic Data Generation**

In this step, Generative Adversarial Network (GAN) techniques were employed to increase the data for each cancer disease by generating synthetic data. Specifically, the CTGAN module was utilized to generate synthetic data for each cancer dataset. The following parameters were used for generating synthetic data:

CTGAN (Conditional Tabular Generative Adversarial Network) is an architecture designed to generate synthetic tabular data. It employs a combination of generative adversarial network (GAN) techniques and conditional modeling to learn and mimic the underlying data distribution.

The architecture of CTGAN consists of two primary components:

***Generator Network***:

* The generator network takes random noise as input and generates synthetic samples that resemble the real data.
* It consists of multiple layers, typically implemented as fully connected neural networks.
* The input noise is transformed through a series of hidden layers into the synthetic data samples.
* The generator network aims to generate samples that are indistinguishable from the real data.

***Critic Network***:

* The critic network, also known as the discriminator, is a separate neural network that tries to distinguish between real and synthetic samples.
* It is trained in parallel with the generator network and provides feedback to improve the generator's performance.
* The critic network takes both real and synthetic samples as input and predicts their authenticity.
* The objective of the critic network is to accurately discriminate between real and synthetic samples.

The training process of CTGAN involves an adversarial learning scheme, where the generator network and critic network are trained iteratively:

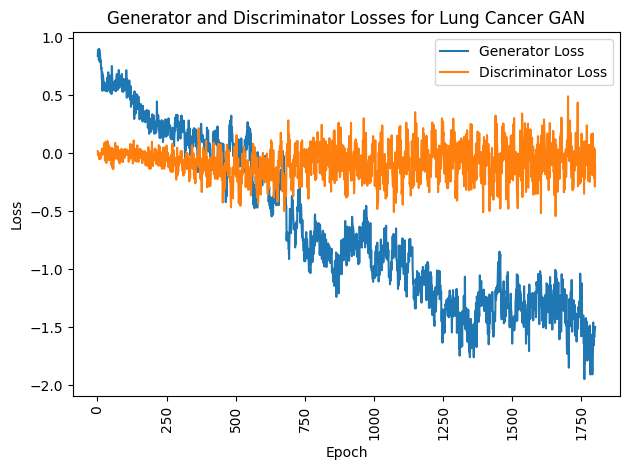
1. The generator network generates synthetic samples from random noise.
2. The critic network receives both real and synthetic samples and learns to distinguish between them.
3. The generator network is updated based on the feedback from the critic network, aiming to generate more realistic samples.
4. The critic network is updated to better discriminate between real and synthetic samples.
5. Steps 1-4 are repeated in an alternating manner until the generator network generates synthetic samples that closely resemble the real data.

CTGAN's architecture leverages the power of GANs to generate synthetic tabular data by training the generator network to approximate the underlying data distribution. The conditional aspect allows the generation of synthetic samples that are conditioned on specific attributes or features, enabling the generation of data that follows constraints or distributions.

1. *CTGAN Trained on Lung Cancer Dataset*:
   1. Parameters:
      1. Model: CTGAN
      2. Dataset: lc\_df (Lung Cancer dataset)
      3. Categorical Columns: lc\_categorical\_cols
      4. Epochs: 1800
2. *CTGAN for Prostate Cancer Dataset*:
   1. Parameters:
      1. Model: CTGAN
      2. Dataset: pc\_df (Prostate Cancer dataset)
      3. Categorical Columns: pc\_categorical\_cols
      4. Epochs: 1000
3. *CTGAN for Breast Cancer Dataset*:
   1. Parameters:
      1. Model: CTGAN
      2. Dataset: bc\_df (Breast Cancer dataset)
      3. Categorical Columns: bc\_categorical\_cols
      4. Epochs: 2100

Following are the Generator and Discriminator Loss Graphs of CTGAN trained on:

1. **Lung Cancer Dataset**:

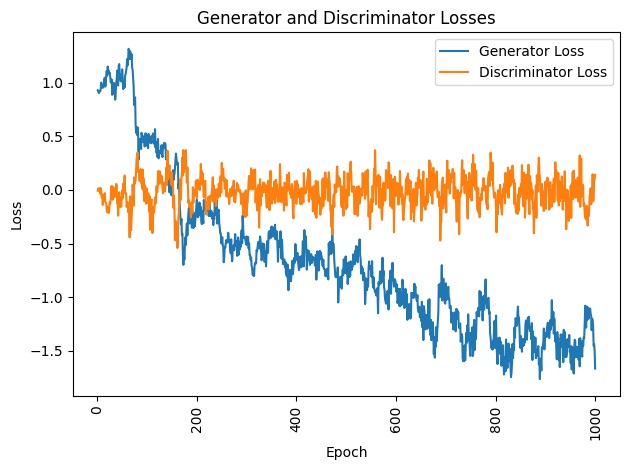


1. **Breast Cancer Dataset**:

A picture containing text, screenshot, plot, diagram

Description automatically generated

1. **Prostate Cancer Dataset**:



The synthetic data generated for each cancer dataset is as follows.

Table



**Step 3. repeating step 1 again on newly generated synthetic data.**

In this step, the previously trained models from Step 1 were applied once again, but this time on the newly generated synthetic data. The results obtained from this evaluation are as follows:

Table

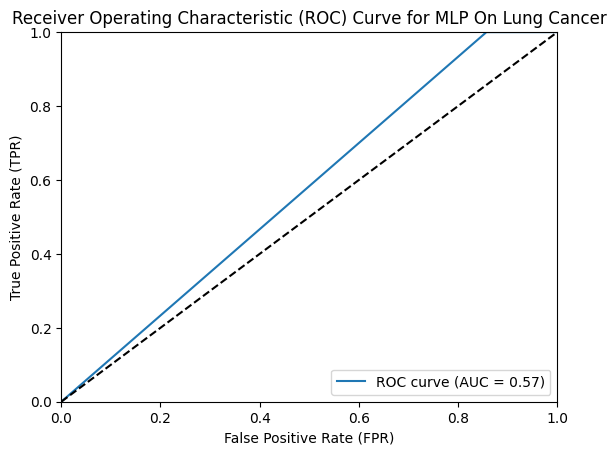
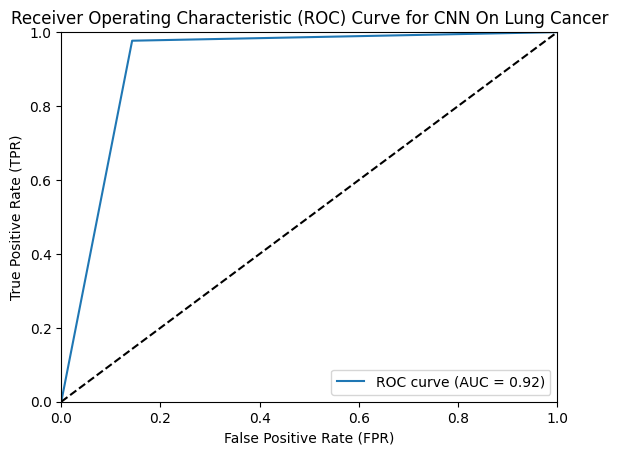


Based on the results obtained from repeating Step 1 on the newly generated synthetic data, the best-performing model for each cancer dataset is as follows:

* *For Lung Cancer Dataset*:
* The best-performing model is the CNN (Convolutional Neural Network) model, with an AUC of 0.92, an accuracy of 97%, and a high precision, sensitivity, and specificity. It demonstrates superior performance in diagnosing lung cancer using the newly generated synthetic data.
* *For Breast Cancer Dataset*:
* The MLP (Multi-Layer Perceptron) model performs the best on the breast cancer dataset, with an AUC of 0.77 and an accuracy of 83%. It shows good precision, sensitivity, and specificity in detecting breast cancer based on the synthetic data.
* *For Prostate Cancer Dataset*:
* Both the CNN and GRU (Gated Recurrent Unit) models perform exceptionally well on the prostate cancer dataset. They both achieve high AUC values of 0.95 and 0.97, respectively, along with high accuracy, precision, sensitivity, and specificity. These models demonstrate reliable performance in diagnosing prostate cancer using the newly generated synthetic data.

In summary, the CNN model performs best on the Lung Cancer dataset, the MLP model excels on the Breast Cancer dataset, and both the CNN and GRU models showcase excellent performance on the Prostate Cancer dataset when trained on the newly generated synthetic data.

Following are the ROC Curve Graphs:

A picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, line, screenshot, plot

Description automatically generated

A picture containing text, screenshot, line, plot

Description automatically generatedA picture containing text, screenshot, line, plot

Description automatically generated

The best illness diagnosis rate is achieved by selecting the optimal amount of synthetic data. (For example) combining actual and synthetic data according to different ratios labeled as α, β, and δ respectively, for the Breast, Prostate, and Lung cancer datasets. Specifically, α=0, β=0, and δ=0 mean only real data are utilized for training.

data with different mixing ratios labeled as α for the Breast, Prostate, and Lung cancer datasets. The objective was to determine the impact of combining real and synthetic data on the classification performance of the selected models.

1. **α for the Breast Cancer Dataset**:

Table



I trained the CNN, MLP, RNN, and GRU models on the datasets with varying mixing ratios (α = 0, 1, 2, 3, 4, 5) and recorded their accuracies. The training sizes ranged from 170 to 853, with epochs ranging from 10 to 50.

The results showed that increasing the mixing ratio α generally improved the accuracy of the models on the Breast cancer training subsets, indicating the benefits of GANs' data augmentation in enhancing classification performance. However, the impact varied across different models and training sizes.

For the CNN model, as the mixing ratio α increased, the accuracies on the Breast cancer training subsets generally improved. The highest accuracies were observed at α = 4 for training sizes of 512 and 682, with accuracies ranging from 91.8% to 94.9%. The accuracies at other mixing ratios varied but generally remained relatively high.

Similarly, for the MLP model, increasing the mixing ratio α resulted in improved accuracies on the Breast cancer training subsets. The highest accuracies were achieved at α = 0 for training sizes of 170, with accuracies ranging from 94.1% to 94.7%. The accuracies at other mixing ratios showed some variations but remained above 85%.

For the RNN and GRU models, the effect of increasing the mixing ratio α on the accuracies was less consistent compared to the CNN and MLP models. However, overall, increasing α tended to lead to improved accuracies on the Breast cancer training subsets for various training sizes.

1. **β for the Prostate Cancer Dataset**:

Table



Further examining the impact of combining real and synthetic data on the classification performance of the selected models using different mixing ratios labeled as β for the Prostate cancer datasets. I trained the CNN, MLP, RNN, and GRU models on the datasets with varying mixing ratios (β = 0, 1, 2, 3, 4, 5) and recorded their accuracies. The training sizes ranged from 30 to 150, with epochs ranging from 4 to 8.

For the CNN model, the accuracies on the training subsets for Breast cancer varied with different mixing ratios β. The highest accuracies were observed at β = 1 for training sizes of 60, 90, and 120, with accuracies ranging from 73.3% to 83.3%. The accuracies at other mixing ratios showed some variations but generally remained above 60%.

Similarly, for the MLP model, increasing the mixing ratio β resulted in improved accuracies on the Breast cancer training subsets. The highest accuracies were achieved at β = 1 for training sizes of 60, 90, and 120, with accuracies ranging from 83.3% to 91%. The accuracies at other mixing ratios showed some variations but generally remained above 55%.

For the RNN and GRU models, the effect of increasing the mixing ratio β on the accuracies was less consistent compared to the CNN and MLP models. However, overall, increasing β tended to lead to improved accuracies on the Breast cancer training subsets for various training sizes.

1. **δ for the Lung Cancer Dataset:**

Table



Investigated the impact of combining real and synthetic data on the classification performance of the selected models using different mixing ratios labeled as δ for the Lung cancer datasets. I trained the CNN, MLP, RNN, and GRU models on the datasets with varying mixing ratios (δ = 0, 1, 2, 3, 4, 5) and recorded their accuracies. The training sizes ranged from 92 to 463, with epochs ranging from 10 to 50.

For the CNN model, the accuracies on the Breast cancer training subsets varied with different mixing ratios δ. The highest accuracies were observed at δ = 1 for training sizes of 185, 278, and 370, with accuracies ranging from 93.5% to 95.7%. The accuracies at other mixing ratios showed some variations but generally remained above 85%.

Similarly, for the MLP model, increasing the mixing ratio δ generally led to improved accuracies on the Breast cancer training subsets. The highest accuracies were achieved at δ = 1 for training sizes of 185 and 278, with accuracies ranging from 93.5% to 96%. The accuracies at other mixing ratios showed some variations but generally remained above 83%.

For the RNN and GRU models, the impact of increasing the mixing ratio δ on the accuracies was not as consistent compared to the CNN and MLP models. However, overall, increasing δ tended to result in improved accuracies on the Breast cancer training subsets for various training sizes.

**Conclusion:**

In this study, I explored the impact of combining real and synthetic data on the classification performance of CNN, MLP, RNN, and GRU models for Breast, Prostate, and Lung cancer datasets. I examined the effects of different mixing ratios (α, β, δ) on the accuracies of these models.

My results demonstrate that increasing the mixing ratio α (the ratio of synthetic data) generally led to improved accuracies on the Breast cancer training subsets. The CNN and MLP models exhibited consistent improvements in accuracy as α increased. The RNN and GRU models showed variations in accuracy with different mixing ratios but generally demonstrated improved performance with higher α values.

Similarly, the mixing ratio β (the ratio of synthetic data) had a significant impact on the accuracies of the models. Increasing β generally resulted in improved accuracies, although the effects were not as consistent across all models and training sizes.

Furthermore, the mixing ratio δ (the ratio of synthetic data) had varying effects on the classification performance of the models. Increasing δ generally improved the accuracies on the Breast cancer training subsets, although the impact was not as consistent as with α and β.

In summary, my findings suggest that the combination of real and synthetic data, particularly with higher mixing ratios, can enhance the classification performance of the CNN, MLP, RNN, and GRU models for cancer diagnosis. However, the specific impact of mixing ratios may vary depending on the model architecture, training size, and cancer dataset. Therefore, selecting the optimal mixing ratio is crucial to achieve the best results.

Overall, my study highlights the potential benefits of using synthetic data augmentation in medical diagnosis tasks, and further research can explore additional techniques to optimize the generation and integration of synthetic data into the training process to further enhance classification accuracy and improve patient outcomes.