**Skin Lesion distribution:**

The objective of the analysis is to explore the distribution of various types of skin lesions across different demographics including age and gender, and to visualize the localization of these lesions. The insights aim to aid in the development of targeted diagnostic and treatment strategies.

**Methods:**

1. **Data Preprocessing**:
   * The dataset, presumably the HAM10000, consists of dermatoscopic images tagged with metadata including age, gender, lesion type, and localization.
   * Age data was segmented into predefined bins (0-20, 21-40, 41-60, 61-80, 81-100 years) to facilitate analysis by age groups.
2. **Visualization Techniques**:
   * **Seaborn** and **Matplotlib** libraries were used to create visual representations of the data.
   * Styles were standardized using Seaborn’s “whitegrid” to enhance readability and aesthetics.
3. **Analysis by Demographic Features**:
   * **Gender-Based Analysis**: Count plots were generated to compare the frequency of each lesion type by gender. The visualization highlighted differences in lesion occurrence between males and females.
   * **Age-Based Analysis**: Similar to gender, count plots depicted the distribution of lesions across different age groups, helping identify which age groups are most affected by specific types of lesions.
4. **Localization Analysis**:
   * Donut charts with a pseudo-3D effect were crafted to display the localization of lesions on the body, segregated by gender. This approach not only quantifies the data but also provides a percentage-based breakdown, adding depth to the analysis.
   * **Automated Diagnostic System**:

**Develop an automated diagnostic system using various machine learning techniques**

**Overview of Machine Learning Approaches for Skin Lesion Classification**

**Initial Approach with Unsupervised Learning:**

1. **Data Preparation**: Labels were initially dropped from the metadata to apply unsupervised machine learning techniques.
2. **Cluster Analysis**:
   * **Optimal Clusters**: Determined using the elbow curve method.
   * **Evaluation**: The conflict matrix revealed significant misclassification, with an Adjusted Rand Index (ARI) of 0.055, far below the acceptable range of 0.2 to 0.8, indicating ineffective clustering.
3. **Visualization**: Visual inspection of the three clusters showed substantial overlap, confirming poor discrimination between lesion types.

**Findings**:

* Unsupervised learning, specifically K-Means clustering, failed to accurately map skin lesion images to their correct categories, necessitating further investigation or alternative approaches to improve classification accuracy.

**Supervised Learning Implementation:**

1. **Image Metadata Processing**:
   * **Standardization**: Images were resized, normalized, and converted into NumPy arrays to facilitate machine learning processing.
   * **Flattening**: Converted images into a flattened format to create high-dimensional tabular data, where each pixel represents a feature. This step is crucial for traditional models like Logistic Regression, Random Forest, and SVM that require tabular input.
2. **Model Performance**:
   * **Logistic Regression and Random Forest**: Achieved accuracies of 68% and 72%, respectively.
   * **F1 Score Concerns**: The Random Forest model exhibited a 0 F1 score for two groups, highlighting significant mapping inaccuracies.

**Conclusion**: The initial unsupervised approach was supplemented by supervised methods to enhance diagnostic precision. However, the challenges in achieving high classification accuracy with traditional models suggest a need for further refinement of the machine learning strategy or exploration of more sophisticated models tailored to high-dimensional image data.

**Skin Lesion Classification Using CNNs**

**1. Dataset Loading and Preparation:**

* **Image Data**: Loaded using TensorFlow's image\_dataset\_from\_directory for training and validation (80% training, 20% validation).
* **Resizing**: Images resized to 224x224 pixels to match the InceptionV3 input requirements.
* **Normalization**: Pixel values normalized from [0, 255] to [0, 1].

**2. Data Augmentation: was not needed for image metadata processing**

* Implemented to improve model generalization:
  + Random flipping, rotation.
  + Zoom and contrast adjustments.

**3. Model Architecture - InceptionV3 Integration: was not needed for image metadata processing**

* **Base Model**: InceptionV3, pre-trained on ImageNet, without top layers (include\_top=False).
* **Freezing Layers**: Base model layers frozen to prevent weights from updating during initial training phases.
* **Additional Layers**:
  + Global Average Pooling 2D.
  + Dense layer with 128 units followed by a dropout layer for regularization.

**4. Training Strategy:**

* **Learning Rate Scheduler**: Employed to reduce learning rate by a factor of 10 post 10 epochs, facilitating fine-tuned model adjustments.
* **Epochs**: Model trained over 30 epochs, with adjustments in learning rate observed to enhance learning efficiency. (For the model using mage metadata 100 epoch were used)

**5. Performance Evaluation:**

* **Accuracy Improvement**: Notable increase in accuracy observed with additional epochs.
* **Learning Rate Adjustments**: Effectively optimized training progress, plotted for visual inspection.

**6. Experimental Application:**

* **De-identified Data Testing**: Applied model to de-identified images from different databases, achieving approximately 80% prediction accuracy.
* **Data Skewness**: Noted challenges due to skewed dataset with predominant lesion types.

**7. Advanced Model Optimization:**

* **Complex Architectural Enhancements**: Added layers and increased epochs within the CNN framework.
* **Final Model Performance**: Achieved a 0.76 % accuracy with image metadata and 0.74% accuracy with image using an enhanced CNN model with the InceptionV3 base.

**8. Visual Representations and Conclusions:**

* Graphical displays of training/validation accuracy and learning rate adjustments provided insights into model behavior over time.
* Concluded that while initial unsupervised approaches (like K-Means clustering) showed limited success, supervised deep learning techniques significantly improved classification accuracy.