Cancer cell detection and preparing the model based on the dataset.

The HAM10000 ("Human Against Machine with 10000 training images") dataset is a comprehensive collection of dermatoscopic images of pigmented lesions. These images were obtained from various populations and acquired using different modalities. The dataset consists of a total of 10,015 dermatoscopic images.

The dataset encompasses seven distinct classes of skin cancer, which are as follows:

1. Melanocytic nevi

2. Melanoma

3. Benign keratosis-like lesions

4. Basal cell carcinoma

5. Actinic keratoses

6. Vascular lesions

7. Dermatofibroma

The primary CSV file associated with the dataset is called HAM10000_metadata.csv. It contains essential data related to all the training images, including the following features:

1. Lesion_id

2. Image_id

3. Dx (diagnosis)

4. Dx_type (type of diagnosis)

5. Age

6. Sex

7. Localization

Data analysis of the HAM10000 dataset has yielded several noteworthy observations:

1. The prevalence of skin diseases is highest among individuals around the age of 45, while it is lowest among those aged 10 and below. Moreover, the likelihood of having a skin disease tends to increase with age.

2. Skin diseases are more commonly observed in males compared to females or individuals of other genders.

3. Skin diseases are most visible on the back of the body and least visible on acral surfaces (such as limbs, fingers, or ears).

4. The most frequently diagnosed disease among individuals is Melanocytic nevi, while Dermatofibroma is the least commonly found.

5. The distribution of skin disease types is as follows:

- nv: Melanocytic nevi - 69.9%

- mel: Melanoma - 11.1%

- bkl: Benign keratosis-like lesions - 11.0%

- bcc: Basal cell carcinoma - 5.1%

- akiec: Actinic keratoses - 3.3%

- vasc: Vascular lesions - 1.4%

- df: Dermatofibroma - 1.1%

6. The discovery of skin diseases is attributed to the following types:

- histo: Histopathology - 53.3%

- follow_up: Follow-up examination - 37.0%

- consensus: Expert consensus - 9.0%

- confocal: Confirmation by in-vivo confocal microscopy - 0.7%

Additional observations include:

- The back area is most susceptible to infections and is more prominently affected in males.

- Infections on the lower extremities of the body are more visible in females.

- Infections are also observed in some unspecified regions, affecting individuals of all genders.

- Acral surfaces show the fewest cases of infection, primarily among males, with no such cases observed in other gender groups.

- Benign keratosis-like lesions are most commonly found on the face.

- Apart from the face, other body parts are predominantly affected by Melanocytic nevi.

- The age group between 0-75 years is most affected by Melanocytic nevi, while individuals aged 80-90 are more prone to Benign keratosis-like lesions.

- Melanocytic nevi affect all gender groups the most.

Methodology Followed:

The dataset used in this study exhibited significant class imbalance, particularly focusing on the area of skin disease. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was employed as a balancing technique. SMOTE was chosen over other methods such as undersampling or oversampling due to its ability to generate synthetic data, preserve information, and reduce overfitting. SMOTE works by identifying minority class samples, selecting their nearest neighbors, and creating synthetic samples that interpolate between them based on a distance metric. This approach increases the number of minority class instances and helps alleviate class imbalance, thereby improving the training of machine learning models.

In terms of the model architecture, the input images had a size of 28x28, which is not commonly used by well-known convolutional neural network (CNN) architectures. Hence, a custom architecture was designed. The model consisted of four convolutional layers with decreasing numbers of filters, followed by two dense layers and a softmax layer. Hyperparameters were determined through iterative experimentation. Batch normalization was applied to normalize the outputs of previous layers, enhancing model stability and performance. Dropout layers were incorporated to prevent overfitting by randomly deactivating neurons during training, promoting generalization.

The multiple convolutional layers with increasing filter sizes allowed the model to capture various levels of abstraction from the input images. Max pooling layers were utilized to reduce spatial dimensions, focusing on the most relevant features. The dense layers at the end of the architecture enabled the model to learn complex relationships and make accurate predictions.

Activation functions, particularly ReLU, were employed in the convolutional and dense layers to introduce non-linearity and capture non-linear patterns in the data. The softmax activation function in the final layer produced class probabilities, facilitating multi-class predictions.

The model was compiled with the Adam optimizer, which dynamically adjusted the learning rate based on parameter gradients, enhancing convergence and training efficiency. Sparse categorical cross-entropy loss, suitable for multi-class classification tasks, was selected as the loss function.

Overall, the chosen model architecture and training configuration, along with the application of SMOTE for addressing class imbalance, contributed to the achieved accuracy of 71% on the HAM10000 dataset.