Predicting Dermatological Diseases using Neural Networks

1 Introduction

The aim of this project is to develop a predictive model for six dermatological diseases based on 34 symptoms. The dataset is sourced from the UCI Machine Learning Repository [1]. Accurate prediction of these diseases can aid in early diagnosis and effective treatment planning.

2 Preprocessing Methods

We began by loading the dataset and handling missing values using the `SimpleImputer` with the mean strategy. This ensured all features had valid numerical values. The data was then normalized using `StandardScaler` to ensure uniform contribution from all features. We addressed class imbalance using SMOTE, which oversampled the minority classes. The target variable was one-hot encoded to facilitate multi-class classification.

3 Proposed Solution and Model Architectures

3.1 Model Version 1

The first version of the model used the following architecture:

- Optimizer: Adam
- Loss: SparseCategoricalCrossentropy
- Layers:
  - Dense layer with 64 neurons and ReLU activation
  - Dense layer with 32 neurons and ReLU activation
  - Output layer with 6 neurons and softmax activation

3.2 Model Version 2

The second version of the model used an improved architecture:

- Optimizer: Adam
- Loss: CategoricalCrossentropy
- Layers:
  - Dense layer with 128 neurons and ReLU activation
  - Dropout layer with rate 0.5
  - Dense layer with 64 neurons and ReLU activation
  - Dropout layer with rate 0.5
  - Output layer with 6 neurons and softmax activation

4 Experiments and Results

4.1 Version 1

In the first version, we observed poor performance as seen in the confusion matrix and accuracy plots. The expected issue was the model's inability to learn from the imbalanced data and inadequate architecture.

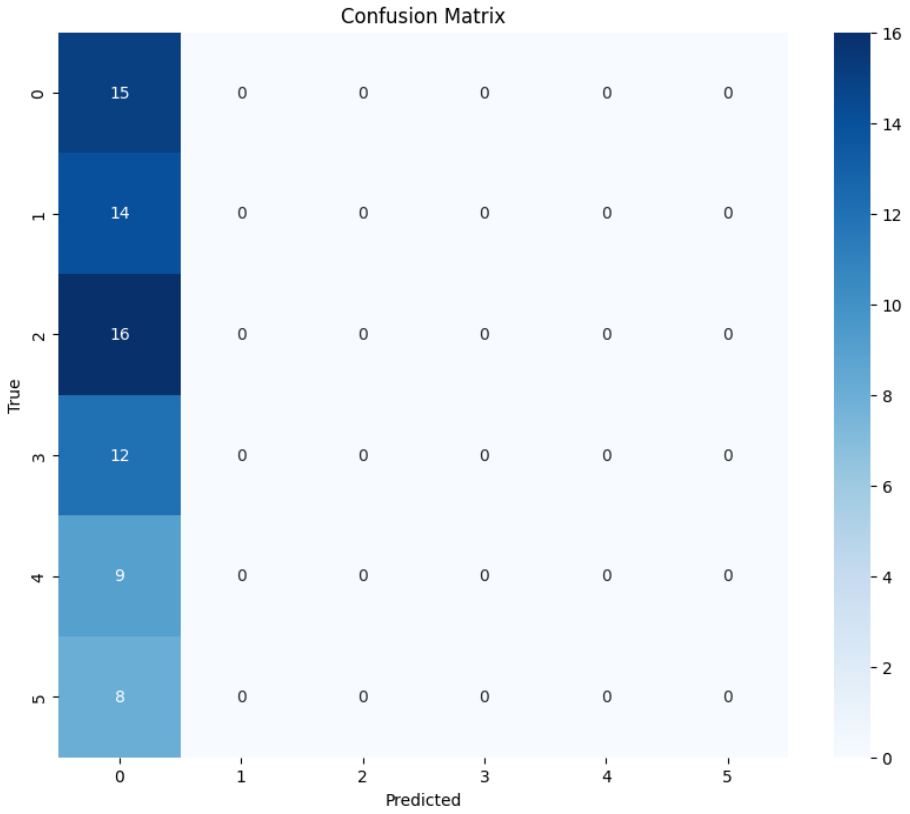

Figure 1: Confusion Matrix for Version 1

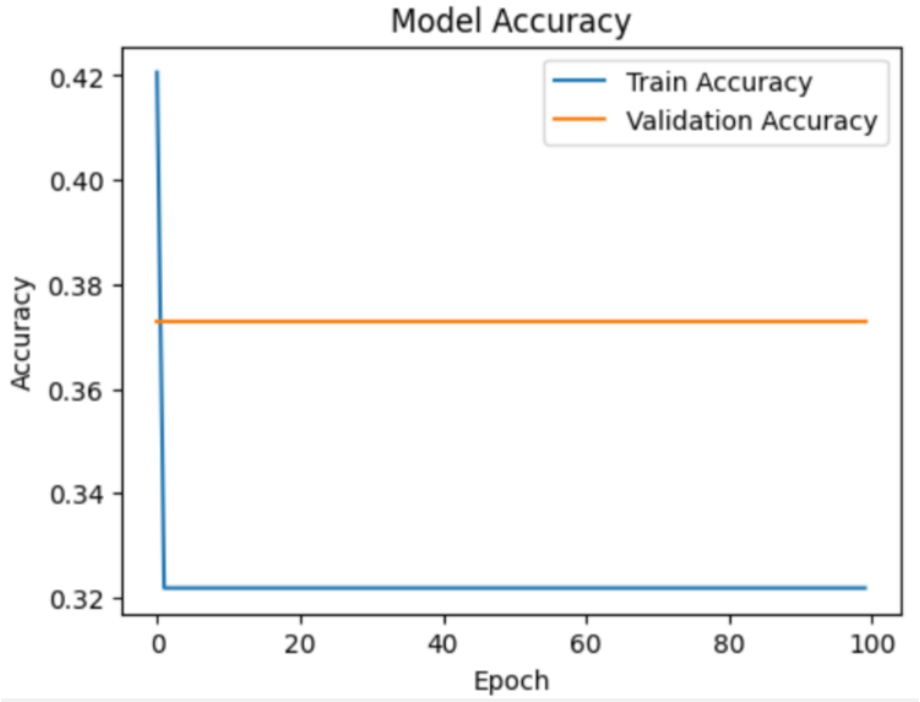

Figure 2: Training Accuracy for Version 1

- **Accuracy:** 0.2027
- **Recall:** 0.1667
- **Precision:** 0.0338
- **F1 Score:** 0.0562

4.1.1 Issues in Version 1

- The model was unable to handle the imbalanced dataset effectively, leading to poor classification performance.
- The architecture was too simple to capture the complexities in the data.
- Lack of dropout layers led to potential overfitting.

4.2 Version 2

In the second version, we improved the model by addressing class imbalance using SMOTE, applying one-hot encoding to the target variable, and using a more complex architecture with dropout layers to prevent overfitting. This resulted in significantly better performance.

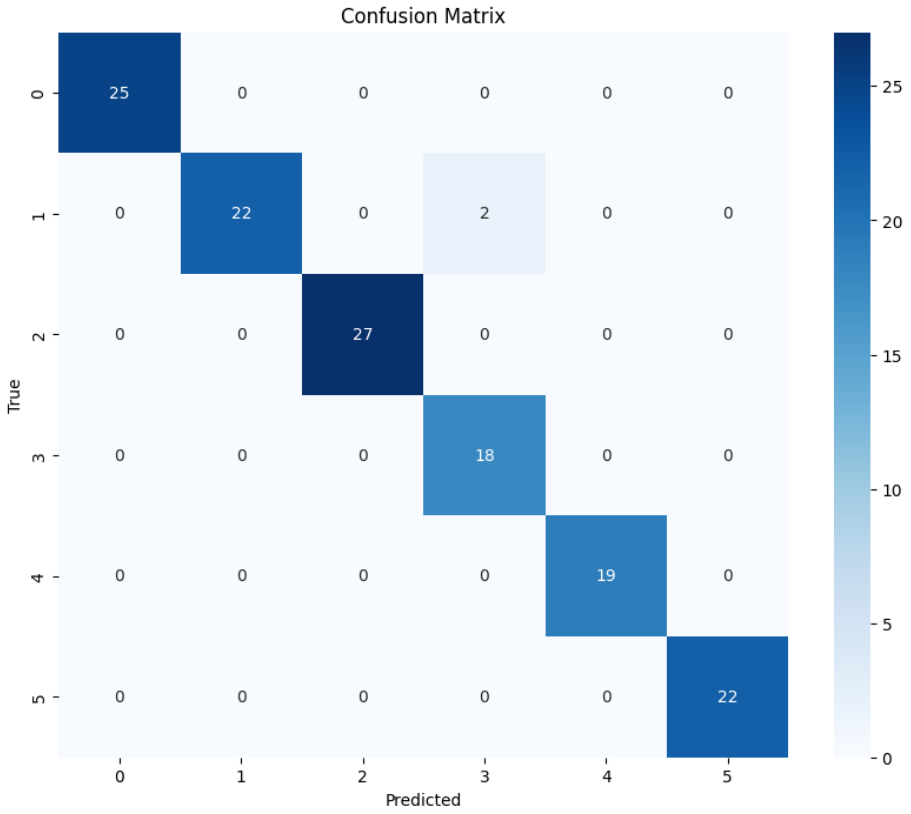

Figure 3: Confusion Matrix for Version 2

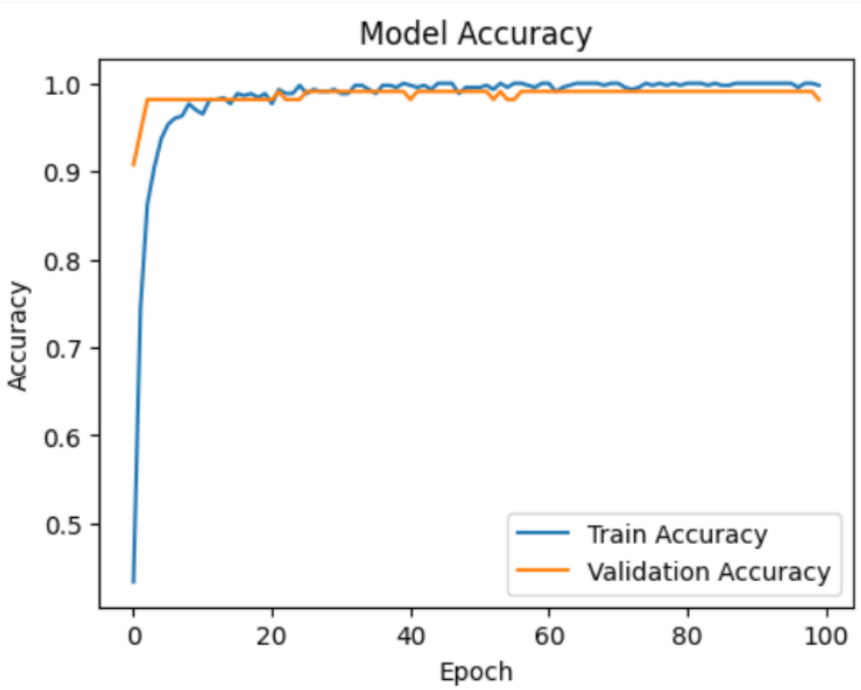

Figure 4: Training Accuracy for Version 2

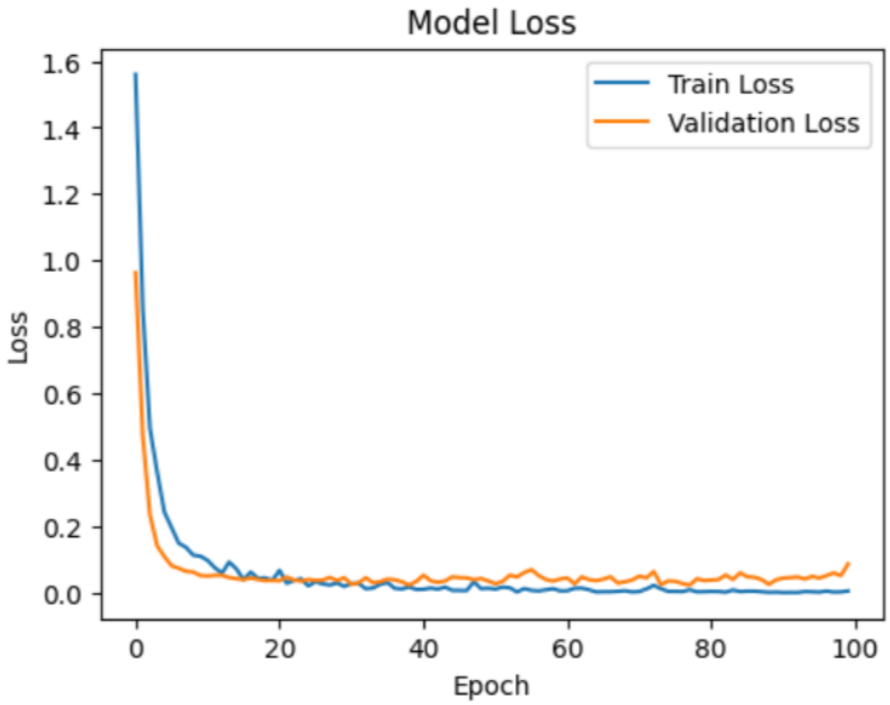

Figure 5: Training Loss for Version 2

- **Accuracy:** 0.9852
- **Recall:** 0.9861
- **Precision:** 0.9833
- **F1 Score:** 0.9840

5 Conclusion

In conclusion, the second version of the model significantly outperformed the first version. The improvements made by addressing class imbalance and using a more robust neural network architecture with dropout layers resulted in higher accuracy, recall, precision, and F1 scores. Future work could explore hyperparameter tuning and incorporating additional features to further enhance the model's performance.

References

[1] UCI Machine Learning Repository. *Dermatology Data Set*. `http://archive.ics.uci.edu/ml/datasets/Dermatology`