Skin Cancer Detection

Jaida Gao

9-25-22

Inspirit AI Research Project

Abstract

The goal of this research project is to predict whether or not a patient has skin cancer through a machine learning model that is developed from an image dataset. Skin cancer is extremely dangerous, as over 9500 people in the US are diagnosed with it daily. If detected early, patients will have a more likely chance of survival. I tested an MLP Classifier, Decision Tree Regressor, a Logistic Regression Model, and a KNN Model to compare various results and ultimately determine the best accuracy. The MLP Classifier had a 74.5% accuracy, the Decision Tree Regressor had a 74.1% accuracy, the Logistic Regression Model had a 68.8% accuracy, and the KNN Model had a 74.6% accuracy (all testing). We can see that the MLP Classifier, Decision Tree Regressor, and the KNN Model had around the same accuracy while outperforming the Logistic Regression Model. However, when comparing training data, there seems to be a large overfitting problem with most of the models.

Introduction

Skin cancer is extremely dangerous, as over 9,500 people in the US are diagnosed with it daily. Skin cancer can spread extremely fast and early detection is more important than ever. The five-year survival rate for early detection is 99% compared to the 68% once the cancer spreads. If patients/doctors are able to use the ML model to detect the presence of cancer, it would benefit everyone, medically and financially.

Background

Right now, biopsies are the most common way for cancer detection, but it comes with risks of hemorrhaging, infections, and damage to nearby organs. There are also imaging tests, but these

expose patients to radioactivity from x-rays and are not that accessible. Researchers at MIT have developed a skin cancer detection DCNN that can detect abnormalities on skin and classify the skin lesions depending on how dangerous they are. However, this program can only use high-resolution images and is not as accessible to everyone.

Dataset

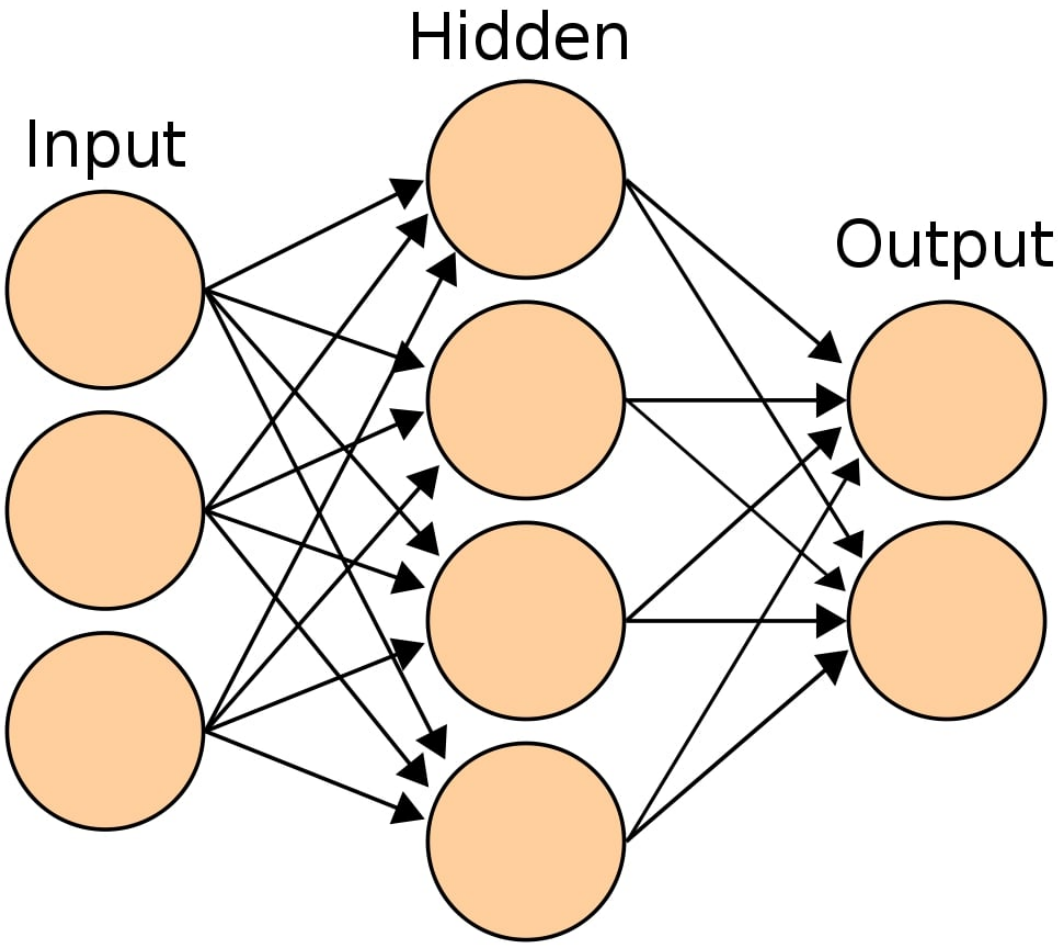
The dataset I used is from Kaggle. It contains processed skin cancer images of malignant and benign skin moles. The data consists of one malignant folder with 1800 224x244 images and one benign folder with 1800 224x244 images; malignant has 300 images for testing data and benign has 360 images for testing data. The rest of the images are all for training. I didn't use the testing folders and just used the training one for both training/testing.

(https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign)

Methodology/Models

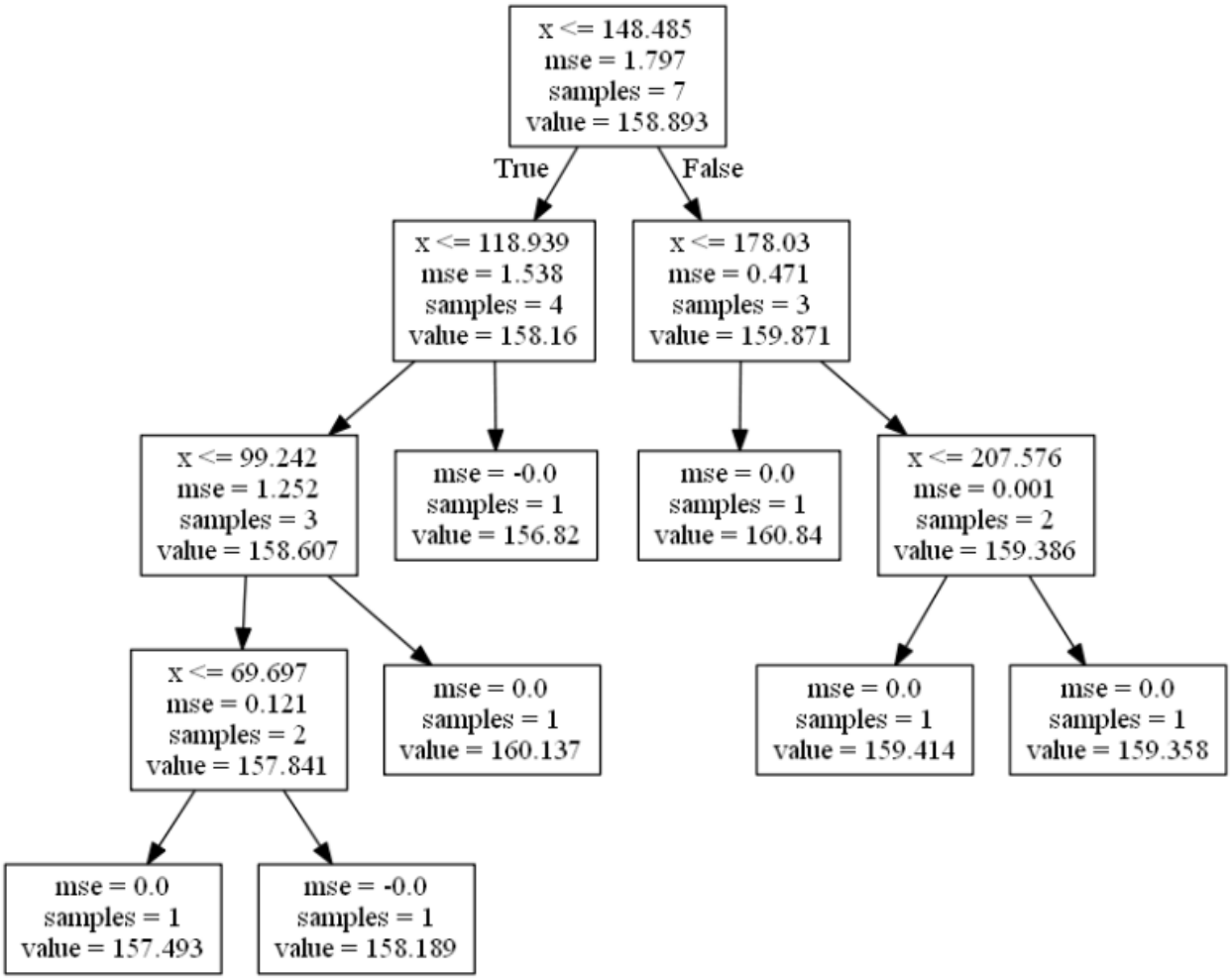
The MLP Classifier takes the pixels of each image and multiplies them by the numbers attached to the edges. It continues to do this through all the hidden layers until you end up with an output

value. Based on this output, it goes back and makes changes to its values. When the output value is too high, it'll lower some of the edge values to achieve a lower output. When If the output is correct, then no changes are made. If the output is incorrect, it'll go back and modify some of the edges to make the output value more accurate. An advantage of this model is that it can capture much more complex patterns. However, it takes more time to train and the predictions are more complex.



(https://coderzcolumn.com/tutorials/machine-learning/scikit-learn-sklearn-neural-network)

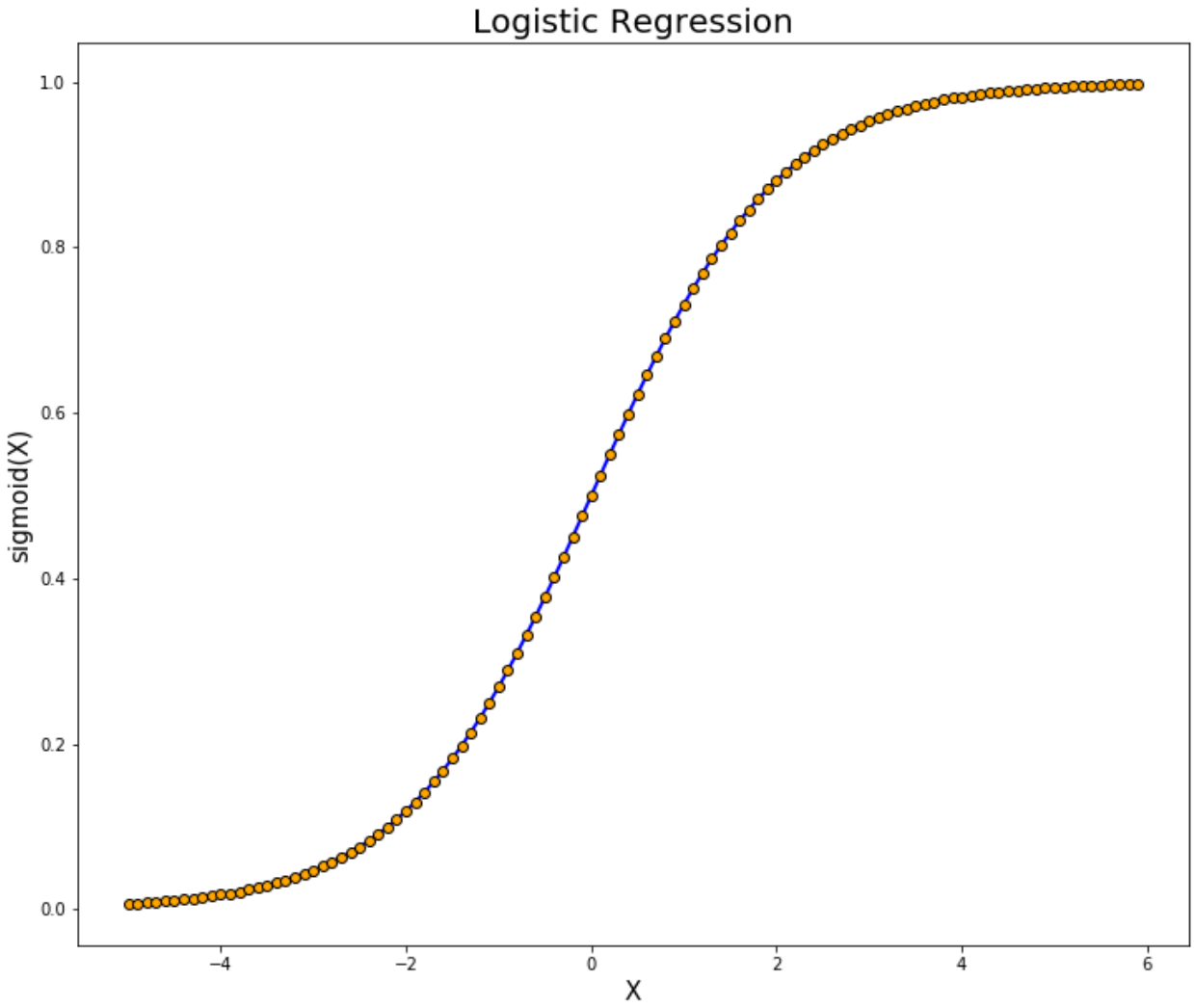
The Decision Tree Regressor takes the images. Each image is represented by its pixels. The pixels are inputs and the model uses the pixels on predefined conditions. The results of these conditions are going to affect the final prediction. An advantage to the Decision Tree Regressor is that it's very simple. However, it's not able to catch complex patterns.



(https://botbark.com/2020/01/01/decision-tree-regression-in-python-in-10-lines/)

The Logistic Regression model takes the pixels of the image and multiples each by a weight. Then, it adds up the products and gets an output from zero to one. The weights are found through the training phase and the result will tell you whether to modify them. If the result is too high,

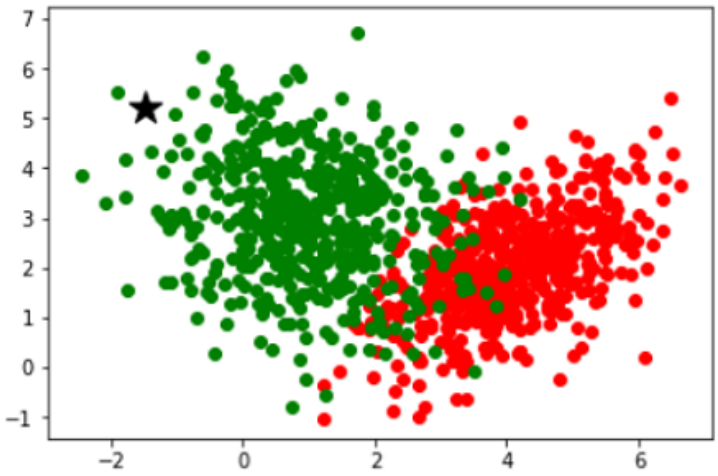
then the weights will be lowered to produce a better accuracy. An advantage to this model is that it's easy to understand the results and it's not too complicated. A disadvantage is that it's not able to capture non-linear dependencies.



(https://towardsdatascience.com/logistic-regression-explained-9ee73cede081)

The KNN Model takes the pictures and finds the images that are the closest to the image that you're trying to predict. The model will choose the majority class of the images that are the closest to the image you have. Then, it will assign the image the label of the class. An advantage

of the KNN Model is that it doesn't require training. A drawback is that it doesn't work well for image specification.



(https://www.analyticssteps.com/blogs/how-does-k-nearest-neighbor-works-machine-learning-classification-problem)

Results and Discussion

This table shows the training and testing results of each model.

Model	Training Accuracy	Testing Accuracy
MLP Classifier	75.3%	74.5%
Decision Tree Regressor	100%	75.3%
Logistic Regression	89.4%	68.8%
KNN Model	82%	74.6%

As we can see, the Decision Tree Regressor produced the highest testing accuracy, but with a major overfitting problem. The KNN Model produced the second highest testing accuracy, but with the same drawback. The Logistic Regression Model had the lowest testing accuracy and a severe overfitting problem, making it not the best model for this particular algorithm. Finally, the MLP Classifier had the third highest testing accuracy and no overfitting problem.

Conclusion

Based on these results, we can determine that the MLP Classifier is the best model as it has the highest testing accuracy with little overfitting. This model was able to pick up on much more complex patterns in the dataset, making it the best model to differentiate benign and malignant skin moles. The 74.5% testing accuracy can be further improved by adjusting the parameters of the hidden layers. The overfitting issue can be solved by adding dropout layers, training in different epochs, or stopping the training prematurely.

Acknowledgments

I would like to thank the InspiritAI mentorship program and Odysseas Drossis for giving me this opportunity and guiding me through this project.

References

L. R. Soenksen *etc.*, Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images, *Science Translational Medicine*, Vol 13, Issue 581.

Mohammad AliKadampur and Sulaiman Al Riyaee, Skin cancer detection: Applying a deep learning based model driven architecture for classifying dermal cell images, *Informatics in Medicine Unlocked*, https://doi.org/10.1016/j.imu.2019.100282

Skin Cancer Facts and Statistics, https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/

Biopsy, https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/biopsy