ICS 435/635

Breast Cancer Classification Using Machine Learning

Spring 2025

Lionel Derrick Roxas

Section 1: Task Description

I tackled a crucial medical challenge: teaching artificial intelligence to diagnose breast cancer. I approached this like training different AI "doctors" to spot cancer cells. Just like medical students learn through practice cases, my AI models learned from a dataset of cell samples. Each sample had 30 different measurements, similar to how doctors look at multiple aspects of a cell to make their diagnosis. My goal was simple: figuring out which model best differentiated between normal cells (benign) and cancer cells (malignant).

Section 2: Model Description

I tried three different approaches to solving this problem. The first one, K-Nearest Neighbors (Figures 2 and 5), works like asking friends for advice. It analyzes similar cases it's seen before to make a decision. For example, if a new cell sample looks very similar to five previous cases, three of which were cancerous, it would predict cancer. I found this to be a straightforward method that's surprisingly effective.

My second method, the **Decision Tree** (Figures 3 and 6), is like following a flowchart of yes/no questions. It might ask: Is the cell larger than normal? Are the edges irregular? Is the shape unusual? Based on these answers, it makes its diagnosis. I found this method easy to understand but sometimes it oversimplifies things.

The third method I used, **Random Forest** (Figures 4 and 7), is like getting opinions from 100 different doctors. Each "doctor" (or tree) looks at different aspects of the cell and makes their own decision. Then they all vote to make the final call. In my testing, this tended to be more reliable because it combines many different viewpoints.

Section 3: Experiment Settings

3.1 Dataset Description

I used the Breast Cancer Wisconsin (Diagnostic) dataset, which contains 569 cell samples, with each sample having 30 different measurements (Figure 1). It's like taking photos of cells and measuring their shape and size to determine if they're cancerous (malignant) or normal (benign). To test my models effectively, I split the dataset into two parts: 80% (455 samples) for training my models (like study material) and 20% (114 samples) for testing them (like a final exam to check their performance).

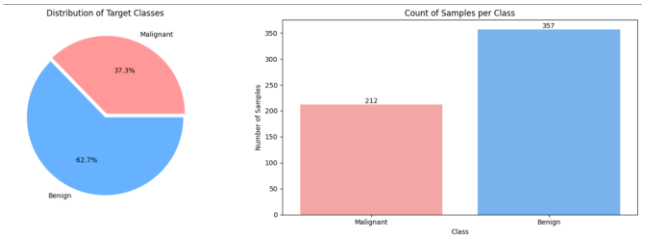
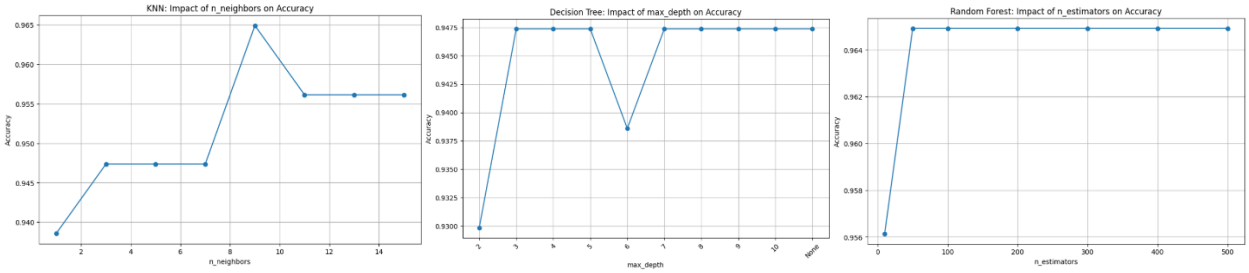


Figure 1: Breast Cancer Wisconsin (Diagnostic) Dataset Distribution Chart

3.2 Detailed Experimental Setups

For each method, I tried different settings to find what worked best. With KNN, I tested different numbers of neighbors (1 to 15) and found that 9 neighbors gave the best results (Figure 2). For Decision Trees, I experimented with various `max_depth_values`, finding that 3-4 were usually enough (Figure 3). With Random Forest, I tried different numbers of trees (10 to 500) and discovered that 50-100 trees gave excellent results (Figure 4).



Figures 2-4: Model Performance Graphs on Accuracy

3.3 Evaluation Metrics

I measured success using four main metrics (Refer to Section 3.5 to view the results) :

- **Accuracy**: How often the model was right (like test scores)
- **Precision**: When it predicted cancer, how often was it correct?
- **Recall**: How many actual cancer cases did it find?
- **F1-Score**: Overall grade combining precision and recall

3.4 Source Code

[Source Code]

3.5 Model Performance

My KNN and Decision Tree both achieved:

- **Accuracy**: 94.74% (95 out of 100 cases correct), **Precision**: 95.77% (very reliable when predicting cancer), **Recall**: 95.77% (caught most cancer cases), **F1-score**: 95.77% (good overall grade)

On the other hand, my Random Forest performed best with:

- **Accuracy**: 96.49% (best overall), **Precision**: 95.89% (most reliable predictions), **Recall**: 98.59% (missed the fewest cancer cases), **F1-score**: 97.22% (best overall grade)

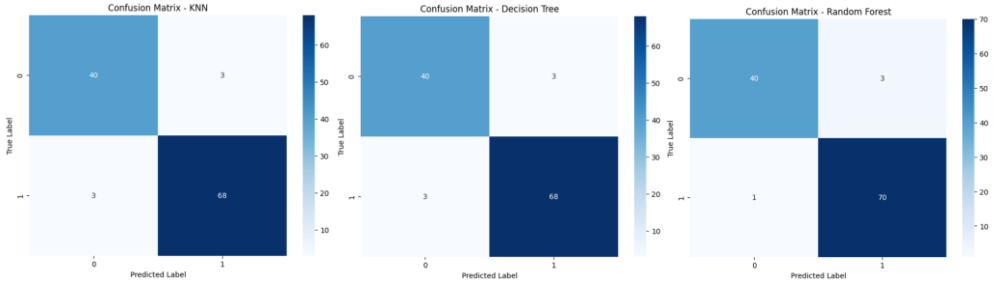


Figure 5-7: Model Performance Evaluated Using Confusion Matrix

3.6 Ablation Studies

Looking at my graphs:

1. I found KNN worked best with 10 neighbors, showing worse results with more or fewer neighbors
2. My Decision Trees improved quickly up to depth 3, then stayed stable
3. My Random Forest got better up to 50 trees, then maintained a steady performance
4. My feature importance analysis showed cells with more severe surface indentations ("worst concave points") and unusual size measurements ("worst area") were the strongest indicators of potential cancer.

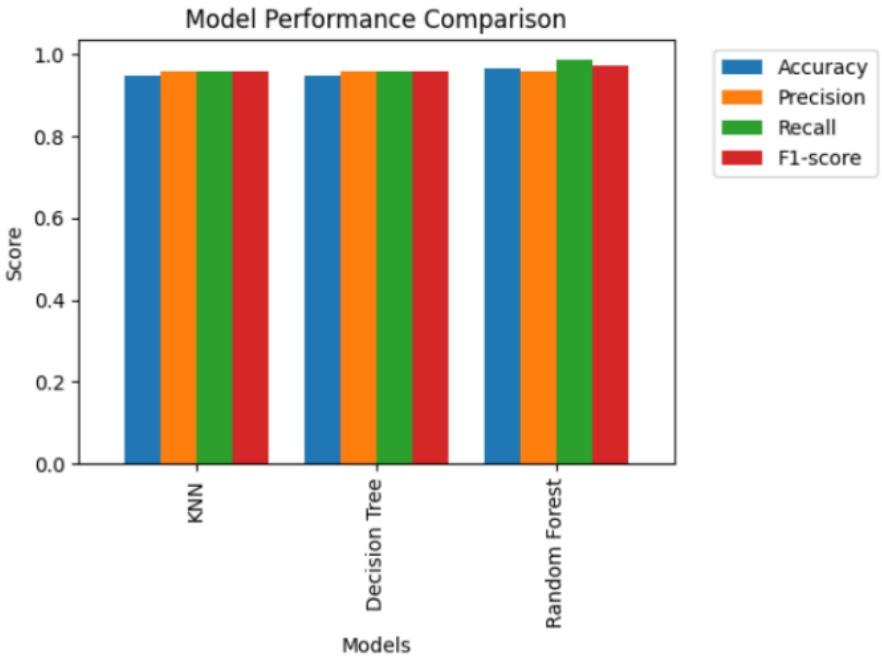


Figure 8: Model Performance Comparison Graph

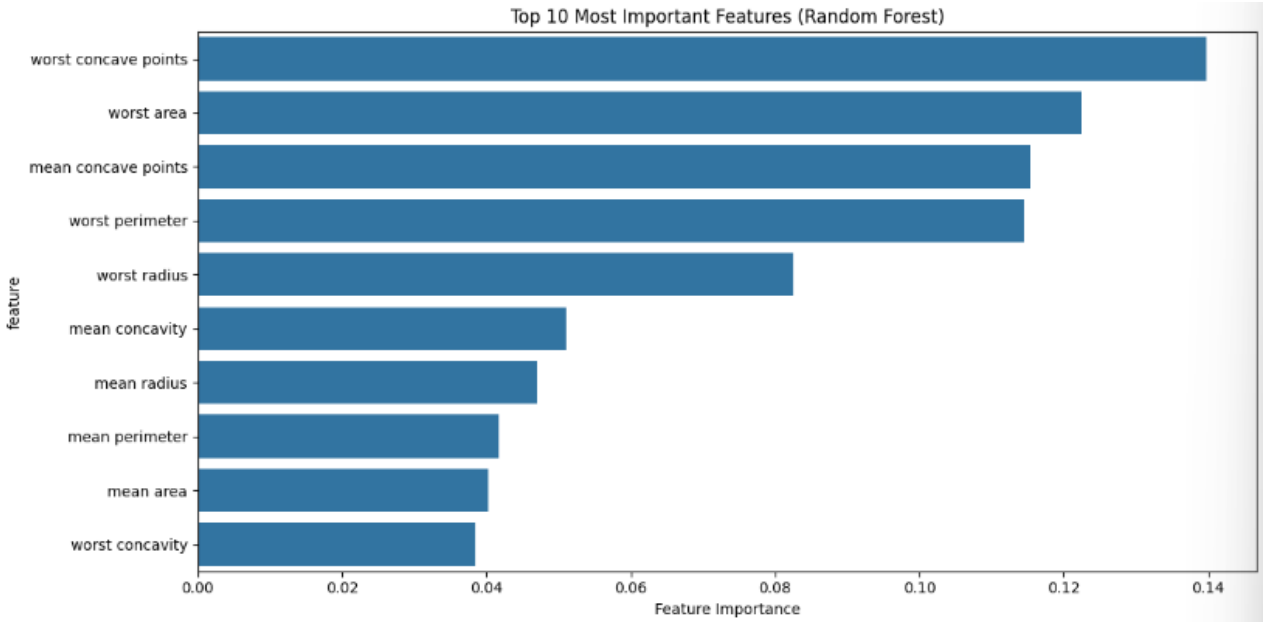


Figure 9: Feature Importance of Random Forest (Best Model)

Section 4: Conclusion

Through this project, I found that analyzing data with multiple models worked better than relying on a single approach. While all three methods performed well with over 94% Accuracy, and above 95% Precision, Recall, and F1-score, **Random Forest** proved to be the most trustworthy for this important medical task because it performed the best across all four key metrics. **Accuracy** (96.49%) shows how often my model was right overall - like getting 96 out of 100 diagnoses correct. **Precision** (95.89%) tells me how reliable the cancer predictions were - when my model said "This is cancer," it was right about 96% of the time, helping avoid unnecessary worry from false alarms. **Recall** (98.59%) was particularly important as it shows how well my model found actual cancer cases - if 40 patients had cancer, it would find 39-40 of them, rarely missing a real cancer case. Finally, the **F1-score** (97.22%) gives a balanced grade considering both precision and recall, showing that my model was good at both avoiding false alarms and catching actual cancer cases. These metrics work together to give a complete picture of how well my model performs in a real medical context.

Additionally, the graphs I generated support these conclusions (Figure 8), showing how each method improved with different settings and which features mattered most (Figure 9). As said in Section 3.6, I found that cells with more severe surface indentations ("worst concave points") and unusual size measurements ("worst area") were the strongest indicators of potential cancer, as cancerous cells typically have irregular shapes with more dents and are often larger than normal cells.

To put it all together, I not only identified **Random Forest** as the most effective model but also gained valuable insights into the physical features that best indicate cancerous cells. This combination of accurate prediction methods and understanding of key cellular features demonstrates the power of machine learning in supporting medical diagnostics. My findings suggest that automated systems using **Random Forest** classification could serve as a reliable tool to assist medical professionals in breast cancer diagnosis, potentially leading to earlier and more accurate detection of cancer cases.