Predicting Cancer Diagnoses Based on Patient miRNA Data

Ellora Devulapally and **Riana Doctor**
{eldevulapally,ridoctor}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

1 Introduction

Cancer is a complex disease with many different types, each with its own genetic signature. It's one of the main causes of death globally, and one of the challenges in modern cancer research is figuring out how to sort different cancer types accurately based on biological data. This is where tools like microRNA (miRNA) analysis come in. MicroRNAs are small molecules that help control which genes get turned on and off, and they tend to show different patterns depending on the cancer (Aristeidis G Telonis 2017). In this project, we developed and experimented with machine learning models to classify cancer types using miRNA data from patients samples sourced from the Cancer Genome Atlas (TCGA). The dataset includes six distinct cancer types, each stored in separate folders, which required pre-processing to combine patient files and assign accurate labels. Once the dataset was complete, we trained two primary models: a Support Vector Machine (SVM) classifier and a Random Forest classifier, along with an additional Neural Network model for practice.The SVM classifier achieved a maximum accuracy of 0.95, the Neural Network model performed better with an accuracy of 0.967, and the Random Forest classifier achieved the highest accuracy at 0.969. These finding suggest that miRNA-based models have a strong potential for multi-class cancer classification, when the data is carefully organized and built to suit the complexity of the task.

2 Exploring the Dataset

We first examined the distribution and structure of the miRNA data provided by TCGA. Each cancer type was stored in its own directory, and each patient's sample was saved as an individual text file containing miRNA names and associated read counts. We found that the number of patient samples varied between cancer types, creating a noticeable class imbalance. This imbalance may influence model performance by making it easier for the classifier to learn patterns from more common cancer types and harder to accurately predict the rarer ones.

We also verified that all patient files contained the same set of miRNA identifiers. Across all samples, the read count values were numeric and consistently reported as normalized counts, reducing the risk of data type issues during modeling. We then proceeded to cleaning and organizing the data for input into a classification model. The following sections, Data Description, Data Aggregation, and Data Visualization, provide a more detailed overview of these steps and the structure of the dataset.

3 Data Pre-processing

Data Description

The dataset comprises samples from six cancer types:

- Breast invasive carcinoma
- Kidney renal clear cell carcinoma
- Lung adenocarcinoma
- Lung squamous cell carcinoma
- Pancreatic adenocarcinoma
- Uveal melanoma

Each sample is represented by a .txt file containing expression levels miRNAS, and each cancer type has its own directory, serving as the class label.

Data Aggregation

To prepare the raw miRNA data, we wrote a python script to collect and structure patient data across the six cancer types stored in separate folders. For each folder, we located all .mirbase21.mirnas.quantification.txt files, and extracted the columns containing the miRNA identifier and its normalized expression value reads_per_million_miRNA_mapped. After gathering all of the files into a single dataset, we reshaped it into a wide format, where each row corresponds to one patient and each column corresponds to a specific miRNA. Any missing values were filled with zeros, and the cleaned dataset was saved as a CSV file.

Data Visualization

Figure 1 shows the distribution of patient samples across the six cancer types included in the dataset. The visualization reveals a clear imbalance. Breast Invasive Carcinoma has a much larger number of samples compared to the other categories, while Pancreatic Adenocarcinoma and Uveal Melanoma have notably fewer. We found that this is important because a classifier could become biased toward predicting the more frequent cancer types and underperform on the underrepresented ones. Recognizing this early in preprocessing influenced our later modeling decisions, such as

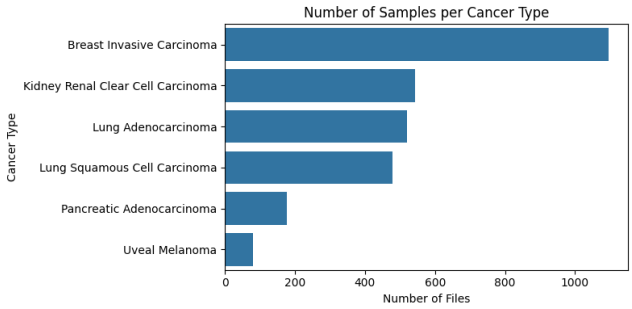

Figure 1: Number of Samples per Cancer Type

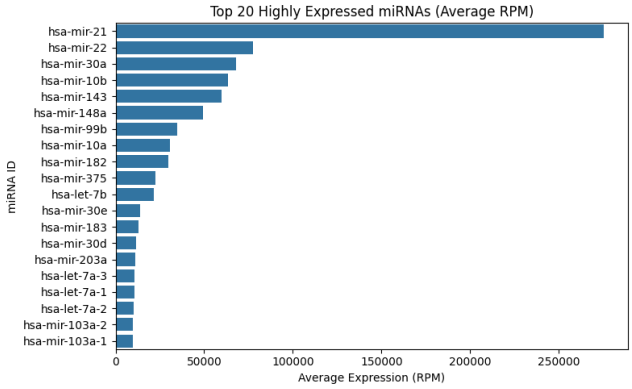

Figure 2: Top 20 Highly Expressed miRNAs

applying class weighting to ensure fair performance across all cancer types.

Figure 2 displays the top 20 most highly expressed miRNAs across all cancer types in the dataset. Among these, hsa-mir-21 stands out as the most expressed by a large margin. We found this important because it suggests that hsa-mir-21 plays a significant role across multiple cancers. Identifying strong outliers early in our analysis helped us understand biological trends and guided future decisions, such as normalization. This helped ensure that the high expression of hsa-mir-21 wouldn't negatively influence the model training.

4 Model Implementation

We evaluated the effectiveness of different machine learning approaches in classifying cancer types based on miRNA expression data. We implemented three models: a Support Vector Machine (SVM), a Neural Network model, AND A Random Forest classifier. Each model was trained and tested using the same preprocessed dataset. The following subsections describe our implementation details, parameter choices, and evaluation results for each model.

Support Vector Machine Classifier

In this section, we trained a Support Vector Machine (SVM) classifier using the wide-format miRNA CSV file created during data pre-processing. We began by loading the CSV

```
Test Accuracy: 0.91

Classification Report:
                                precision    recall  f1-score   support

        Breast_InvasiveICarcinoma       0.85      0.99      0.92       219
Kidney Renal Clear Cell Carcinoma       1.00      0.90      0.95       109
             Lung Adenocarcinoma       0.95      0.89      0.92       104
      Lung Squamous Cell Carcinoma       0.94      0.81      0.87        96
         Pancreatic Adenocarcinoma       0.94      0.86      0.90        35
                   Uveal Melanoma       1.00      0.81      0.90        16

                         accuracy                           0.91       579
                        macro avg       0.95      0.88      0.91       579
                     weighted avg       0.92      0.91      0.91       579
```

Figure 3: SVM First Classification Report

```
Test Accuracy: 0.92

Classification Report:
                                precision    recall  f1-score   support

       Breast Invasive Carcinoma       0.89      0.99      0.94       219
Kidney Renal Clear Cell Carcinoma       1.00      0.89      0.94       109
             Lung Adenocarcinoma       0.95      0.89      0.92       104
      Lung Squamous Cell Carcinoma       0.89      0.89      0.89        96
         Pancreatic Adenocarcinoma       0.94      0.86      0.90        35
                   Uveal Melanoma       1.00      0.88      0.93        16

                         accuracy                           0.92       579
                        macro avg       0.95      0.90      0.92       579
                     weighted avg       0.93      0.92      0.92       579
```

Figure 4: SVM Class Weight Change Classification Report

file, selecting all miRNA expression columns as features, and excluding the file_id and cancer_type identifiers. The dataset was then split into training and testing subsets using an 80/20 ratio, and features were standardized using StandardScaler to ensure that all miRNAs contributed equally to the model's decision boundary. The initial SVM was trained with a radial basis function (RBF) kernel, C=1.0, and default settings. This baseline model achieved an accuracy of .91. Figure 3 shows the classification report for this model.

Next, to address potential class imbalance in the dataset, we added the parameter class_weight='balanced' to the SVM. This adjustment automatically scales the penalty associated with misclassifying samples from minority classes. This ensures that each cancer type contributes proportionally to the model's training objective. As a result, the accuracy improved to .92, and we observed more consistent performance across classes. More specifically, minority categories such as Uveal Melanoma and Pancreatic Adenocarcinoma achieved higher recall scores of 0.88 and 0.86, indicating that the model became better at correctly identifying samples from these underrepresented groups. Additionally, Breast Invasive Carcinoma maintained a strong recall of 0.99, showing that balancing the weights did not compromise performance on the larger classes. Figure 4 shows the updated classification report for this model configuration.

Lastly, we applied a logarithmic transformation to the input features using X = np.log1p(X). Since, miRNA expression data often includes extreme values and wide differences across samples, this step helped even things out. Taking the log of one plus each value compressed very large numbers and reduced the effect of outliers. This made the data more balanced and easier for the model to interpret. After applying this transformation, the SVM's accuracy improved to .95. Figure 5 shows the classification report after these

Test Accuracy: 0.95

Classification Report:

	precision	recall	f1-score	support
Breast Invasive Carcinoma	0.96	0.99	0.98	219
Kidney Renal Clear Cell Carcinoma	1.00	0.94	0.97	109
Lung Adenocarcinoma	0.96	0.90	0.93	104
Lung Squamous Cell Carcinoma	0.87	0.97	0.92	96
Pancreatic Adenocarcinoma	0.97	0.86	0.91	35
Uveal Melanoma	1.00	0.94	0.97	16
accuracy			0.95	579
macro avg	0.96	0.93	0.94	579
weighted avg	0.95	0.95	0.95	579

Figure 5: SVM Final Classification Report (After Logarithmic Transformation)

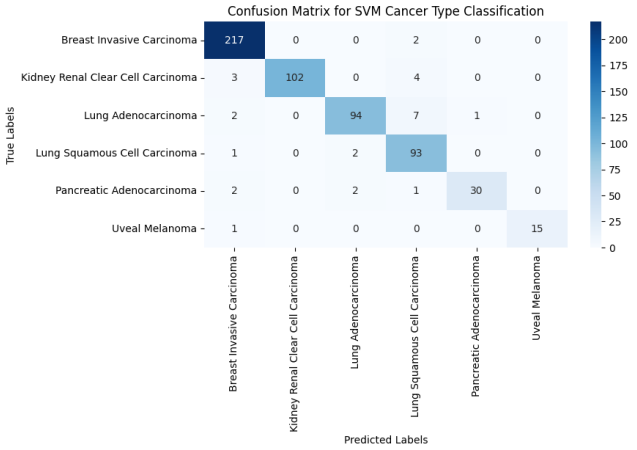

Figure 6: SVM Confusion Matrix

improvements. Furthermore, figure 6 shows the confusion matrix for the final SVM model. It displays strong diagonal dominance and minimal misclassficiations across all six cancer types. Overall, the final SVM model performed the best with an accuracy of 0.95.

Random Forest Classifier

In this section, we started with a very minimal notebook for training our RandomForestClassifier, which we trained on the wide-forest miRNA CSV we had created in data pre-processing (Pedregosa et al. 2025b). First, we loaded the CSV, picked the feature columns (which were all the columns other than the patient id/fileid and cancertype identifier), split the data into train/test, trained a baseline Random Forest, and then printed several classification reports and showed a simple feature-importance plot. The feature importance view was helpful because it showed the importance of the top 25 miRNA in identifying cancer diagnoses. For the raw Random Forest model, we experimented with several different hyperparameters to see how the edits affected accuracy. The hyperparameters that we played around with were n estimators and max depth We began by setting the number of trees in the forest, or n estimators, to 300. We knew that more trees would lead to smoother results, but be significantly slower. However, if there were too few trees,

Test Accuracy: 0.97

Classification Report:

	precision	recall	f1-score	support
Kidney Renal Clear Cell Carcinoma	1.00	1.00	1.00	109
Lung Adenocarcinoma	0.90	1.00	0.95	104
Lung Squamous Cell Carcinoma	1.00	0.90	0.95	96
Pancreatic Adenocarcinoma	1.00	0.97	0.99	35
Uveal Melanoma	1.00	1.00	1.00	16
accuracy			0.97	360
macro avg	0.98	0.97	0.98	360
weighted avg	0.97	0.97	0.97	360

Figure 7: Classification Report (Standard Random Forest)

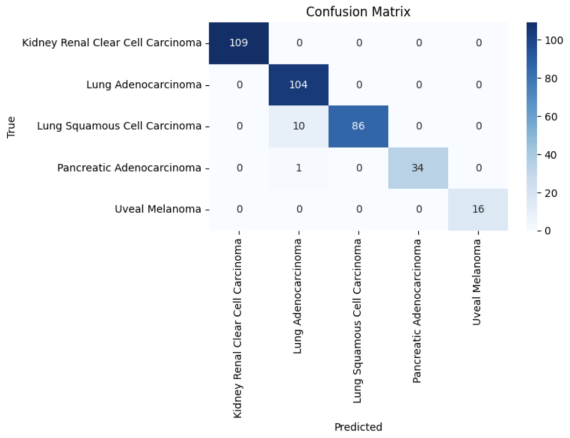

Figure 8: Confusion Matrix (Standard Random Forest)

the model would underfit the data. We saw that at a value of 300 for n estimators and a value of None for max depth, the accuracy was reported at .97. When we adjusted max depth to 10 and n estimators to 100, we saw the accuracy dropped to 0.964. At a value of 500, and for 1000 for n estimators and a value of None for max depth, we saw that the accuracy seemed to plateau, staying at .969. Even with a lower max depth of 10, we saw that the accuracy stayed the same. Figure 3 is the classification report and Figure 4 is the confusion matrix for the standard Random Forest model.

We noticed that in the original paper, the authors decided to binarize the data – basically, rather than creating measurements between 0 and 1, they determined the classifiers between either a 0 and 1. Thus, we decided to do the same thing and note whether it led to a more accurate model. In our Random Forest Binarized file, for each sample, we marked the top X percent most expressed miRNAs as 1 and the rest as 0. Then, we trained the same basic RandomForestClassifier on the binary vector. We started with a TOP PCT value of 0.20, meaning that for the top 20 percent per sample, we'd mark them as 1, and else, they'd be marked 0. This gave us an accuracy of 0.961. When we moved the TOP PCT down to 0.10, the accuracy went up to 0.93. When we moved it down to TOP PCT to 0.05, we saw that the accuracy plateaued at 0.93. Figure 5 is the classification report and Figure 6 is the confusion matrix for the binarized Random Forest model. Based on the Confusion Matrices

Test Accuracy: 0.93

Classification Report:

	precision	recall	f1-score	support
Breast Invasive Carcinoma	0.91	0.99	0.95	219
Kidney Renal Clear Cell Carcinoma	0.99	0.96	0.98	109
Lung Adenocarcinoma	0.90	0.83	0.86	104
Lung Squamous Cell Carcinoma	0.94	0.89	0.91	96
Pancreatic Adenocarcinoma	0.97	0.91	0.94	35
Uveal Melanoma	1.00	1.00	1.00	16
accuracy			0.93	579
macro avg	0.95	0.93	0.94	579
weighted avg	0.94	0.93	0.93	579

Figure 9: Classification Report (Binarized Random Forest)

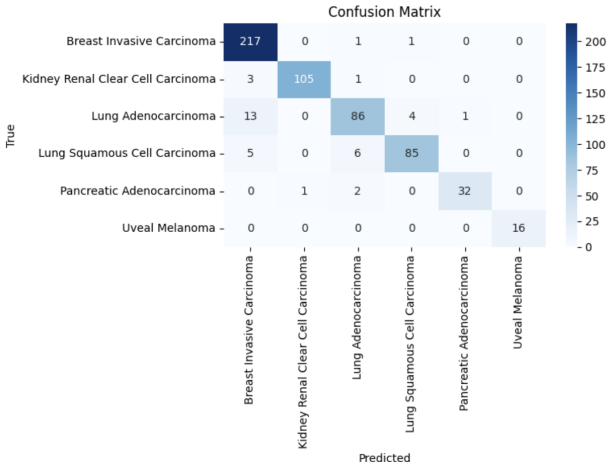

Figure 10: Confusion Matrix (Binarized Random Forest)

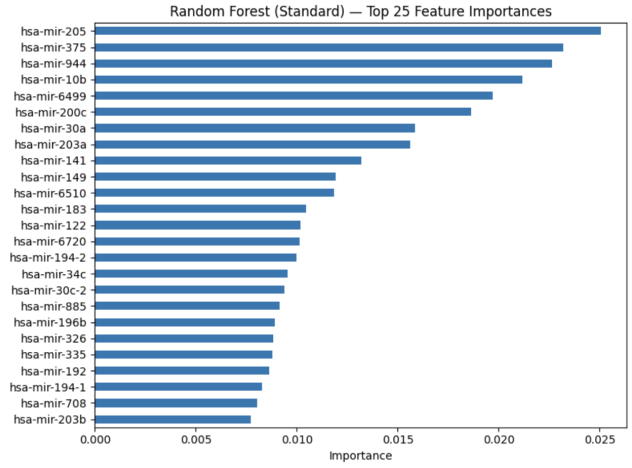

Figure 11: Top 20 Highly Expressed miRNAs (Standard Random Forest)

(Figures 4 and 6) we can see that both the standard random forest and binarized random forest models performed extremely well across all six cancer types. However, it seems that the binarized RF seems even cleaner in some categories such as less errors for Lung Adenocarcinoma and Pancreatic Adenocarcinoma, for example.

Beyond our baseline models, we also wanted to examine the feature importance stability. We can see the feature importance breakdown in figures 7 and 8, and it seems that the top-ranked miRNAs are largely the same between the Standard and the Binarized models (for example, hsa-mir-205, hsa-mir-375, hsa-mir-944, hsa-mir-141, and hsa-mir-200c). There are a few small shifts in order such as hsa-mir-200b, thus it seems that binarization does not change which features are informative, but perhaps how their magnitude contributes to the decision boundary.

The overall stability of both models demonstrate strong dominance and that Random Forest can effectively capture the distinct miRNA expression signatures of each cancer type.

Neural Net

In addition to our Random Forest experiments, which were very successful, we wanted to practice implementing a simple neural network using the MLPClassifier from scikit-learn (Pedregosa et al. 2025a). The goal was not to create a fully optimized deep learning model and have the highest possible accuracy rate. Instead, we wanted to see whether a lightweight appraoch could achieve comparable performance on the same data.

The network was built using the same pre-processed wide-format dataset, where each sample represented one patient and each feature corresponded to an individual miRNA expression level. We standardized all input features using a StandardScaler, encoded the six cancer types as integer labels, and then trained an MLP with two hidden layers (256 and 128 neurons) and ReLU activations. Early stopping was

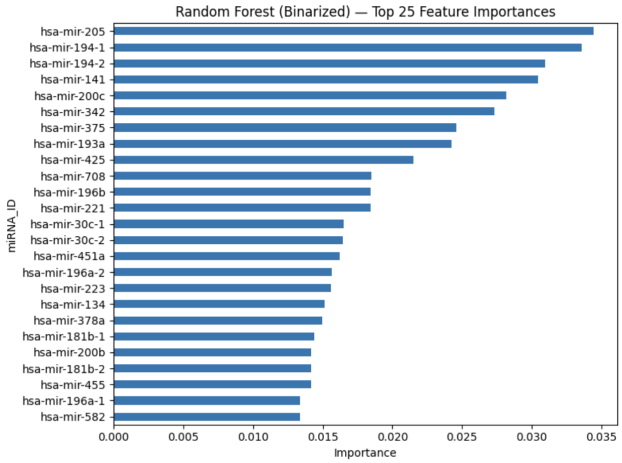

Figure 12: Top 25 Highly Expressed miRNAs (Binarized Random Forest)

Test Accuracy: 0.967

	precision	recall	f1-score	support
Breast Invasive Carcinoma	0.98	0.99	0.99	219
Kidney Renal Clear Cell Carcinoma	1.00	0.97	0.99	109
Lung Adenocarcinoma	0.95	0.92	0.94	104
Lung Squamous Cell Carcinoma	0.91	0.96	0.93	96
Pancreatic Adenocarcinoma	0.97	0.97	0.97	35
Uveal Melanoma	1.00	0.94	0.97	16
accuracy			0.97	579
macro avg	0.97	0.96	0.96	579
weighted avg	0.97	0.97	0.97	579

Figure 13: Classification Report (Neural Net)

enabled to prevent overfitting, and all other hyperparameters were kept at their default scikit-learn settings. Despite its simplicity, the model achieved a test accuracy of 0.967, which is essentially on par with the Random Forest results. Figure shows the confusion matrix, where we can see strong diagonal dominance and minimal misclassification across all six cancer types. This indicates that even a shallow neural architecture is capable of capturing the key nonlinear relationships in the miRNA feature space. Overall, this experiment served primarily as a quick comparison baseline. Because the MLP required more careful tuning (e.g., scaling, learning-rate selection, and layer size) and offered less interpretability than the Random Forest models, we chose not to pursue deeper neural network variations for this assignment.

5 Broader Impacts

Using machine learning to analyze biological data like miRNA expression data could really change how cancer is diagnosed. Models such as the ones developed in this project show that it's possible to identify cancer types with high accuracy, which could help doctors make quicker and more precise decisions. If applied carefully, then tools like these could help early detection and lead to more personalized treatments, which could improve outcomes for patients who might otherwise face delayed diagnoses.

However, there are also risks that can come with relying

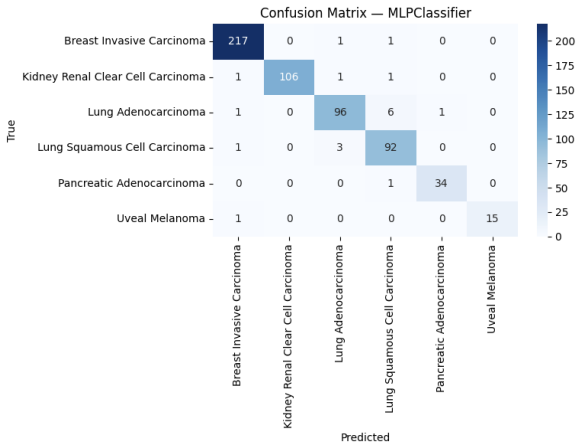

Figure 14: Confusion Matrix (Neural Net)

on these systems. One of the biggest concerns is when the model predicts incorrectly. For example, cases where it may predict that someone doesn't have cancer when they actually do could be detrimental. A mistake like this could mean a person misses out on early treatment, which could be life-threatening. Conversely, false positives could lead to unnecessary stress and medical procedures. This is why it's critical that these models are always used alongside human expertise, not as a replacement for it.

Furthermore, there can also be the issue of bias in the data itself. TCGA may not equally represent all populations or cancer subtypes. If a model learns from an unbalanced dataset, then it could end up performing worse for certain groups, which would deepen existing inequalities in healthcare. To solve this, future research should focus on making datasets more inclusive and testing models for fairness, so that they can benefit everyone equally. Overall, if handled responsibly, machine learning has the potential to make cancer diagnosis faster, more accurate, and more equitable, as long as we stay aware of the risks that come with it.

6 Conclusions

In this project, we trained and compared three models to classify cancer types using patient miRNA data from TCGA. The SVM achieved an accuracy of 0.95, the Neural Network performed slightly better at 0,967, and the Random Forest reached the highest accuracy at 0.969. All three models showed strong performance across classes, though some imbalance between cancer types may have influenced the results. These results show that miRNA data holds clear patterns that can be used to differentiate between cancer types, even with fairly simple models. Future work could focus on tuning hyperparameters, testing on a wider range of cancers, or examining more closely on how specific features connect to biological processes.

References

Aristeidis G Telonis, Rogan Magee, P. L. I. C. E. L. I. R. 2017. Knowledge about the presence or absence of mirna

isoforms (isomirs) can successfully discriminate amongst 32 tcga cancer types. In *Nucleic Acids Research*, 2973–2985.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, 2025a. Mlpclassifier — scikit-learn 1.7.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Retrieved on Nov. 12, 2025.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, 2025b. Randomforestclassifier — scikit-learn 1.7.2 documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Retrieved on Nov. 12, 2025.