Predictive Modeling for Defective Laser Identification

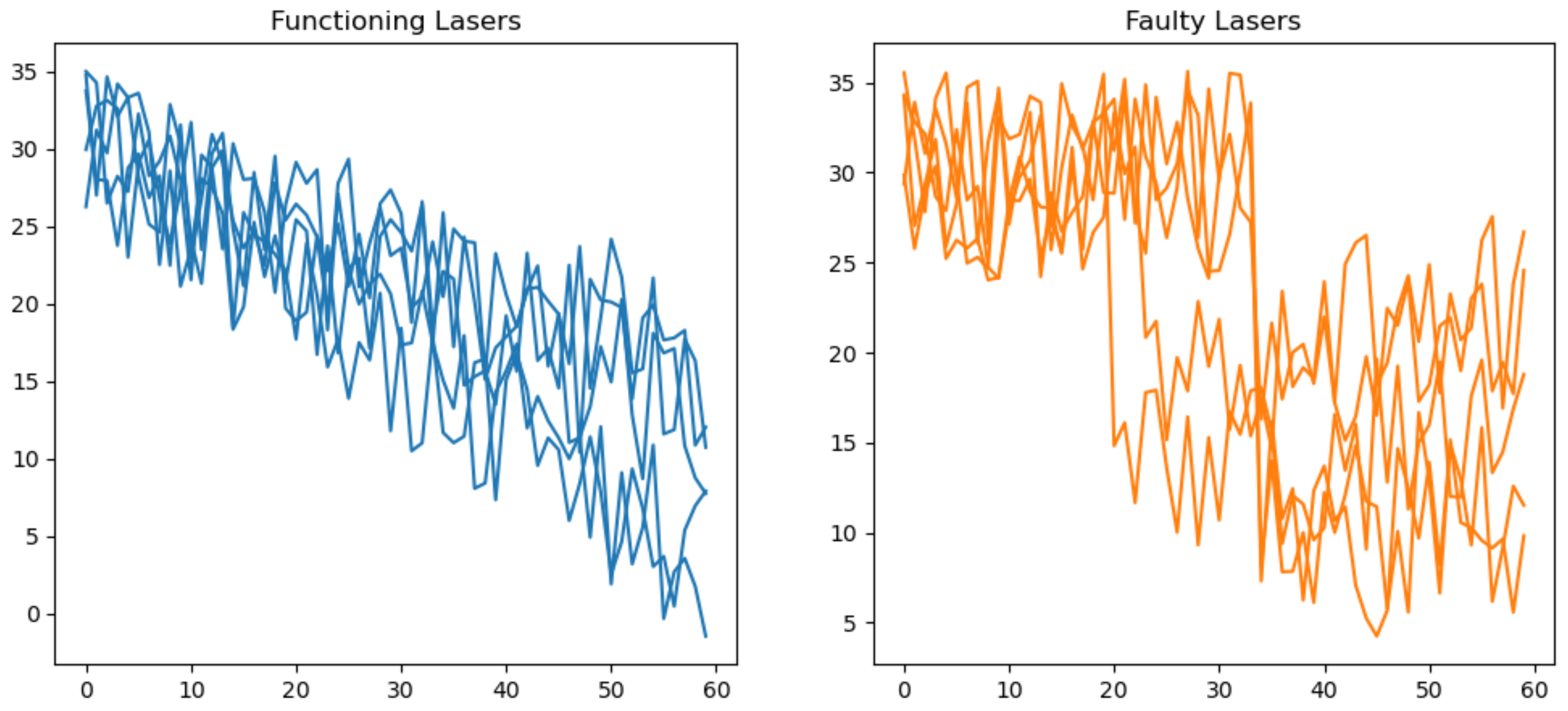
University of Potsdam
Iana Arefeva
7 October 2024

Problem Definition

- **Task:** Predict if a laser is defective based on its <u>intensity measurements</u>.
- **Type of the problem:** <u>Binary Classification</u>.
- **Data characteristics:**
  - **Given nput:** 60 Intensity Measurements over one minute per laser.
  - **Output: class label:** 1 for functioning, -1 for faulty.

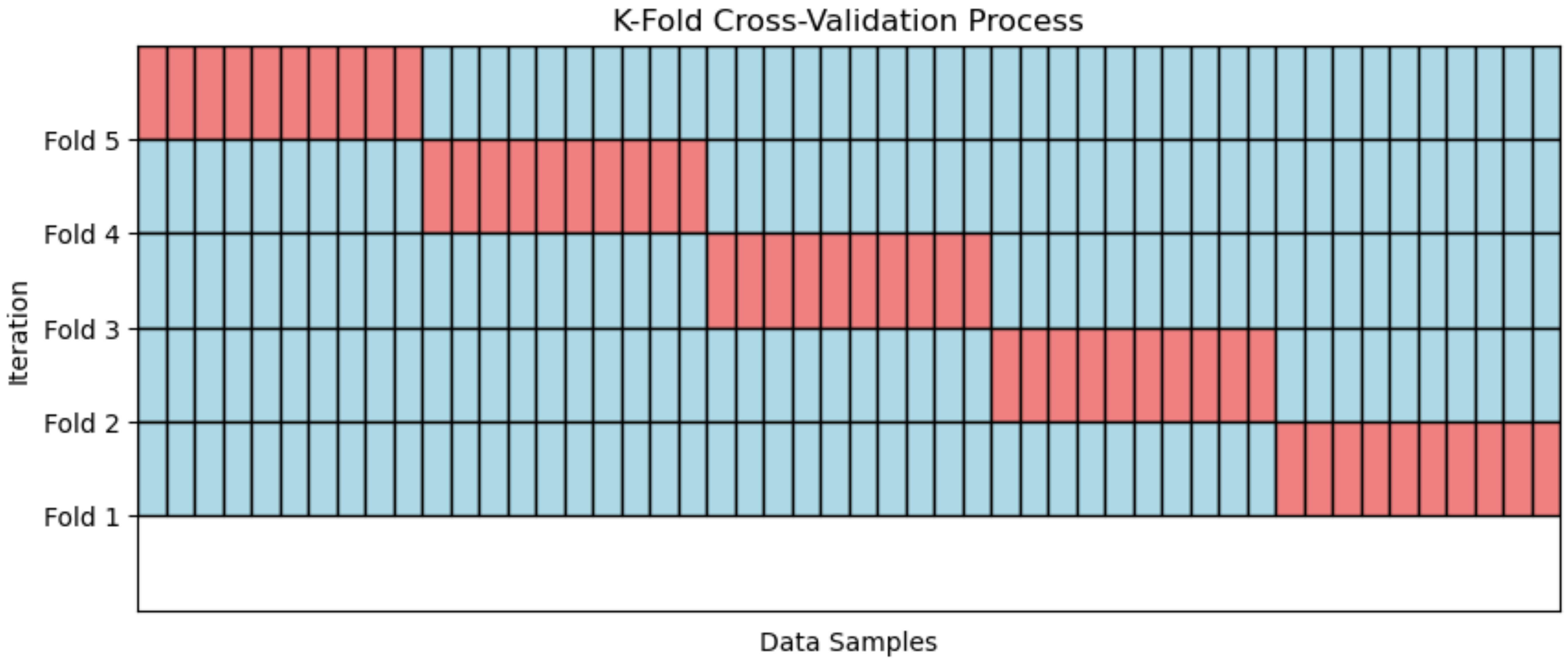
Dataset overview

- **Size:** 200 samples, each with 60 intensity measurements.
- **Class Balance:** 100 functioning lasers, 100 faulty lasers (balanced dataset - 50/50%).
- **Feature types:** time series data representing intensity over time



Evaluation protocol and metrics

- **Evaluation Protocol:**
  - Stratified 5-Fold Cross-Validation - ensures balanced class representation in each fold for better generalisation.
  - Hold-out Test Set - for final performance validation.
- **Metrics Used:**
  - Accuracy, Precision, Recall, F1 Score, ROC AUC.



Metric	Definition
Precision	TP / (TP + FP)
Recall	TP / (TP + FN)
F1-Score	2 * (Precision * Recall) / (Precision + Recall)
Accuracy	(TP + TN) / (TP + TN + FP + FN)

Model selection

Models Used:

- **Logistic Regression:** Baseline linear model.
- **Support Vector Machine (SVM):** Effective for binary classification, uses kernel trick for non linearity.
- **Random Forest:** Ensemble of decision trees, good for non-linear relationships and feature importances.
- **Neural Network (MLP):** captures complex relationships, good for non-linear patterns.

Hyperparameter Tuning and Cross-Validation

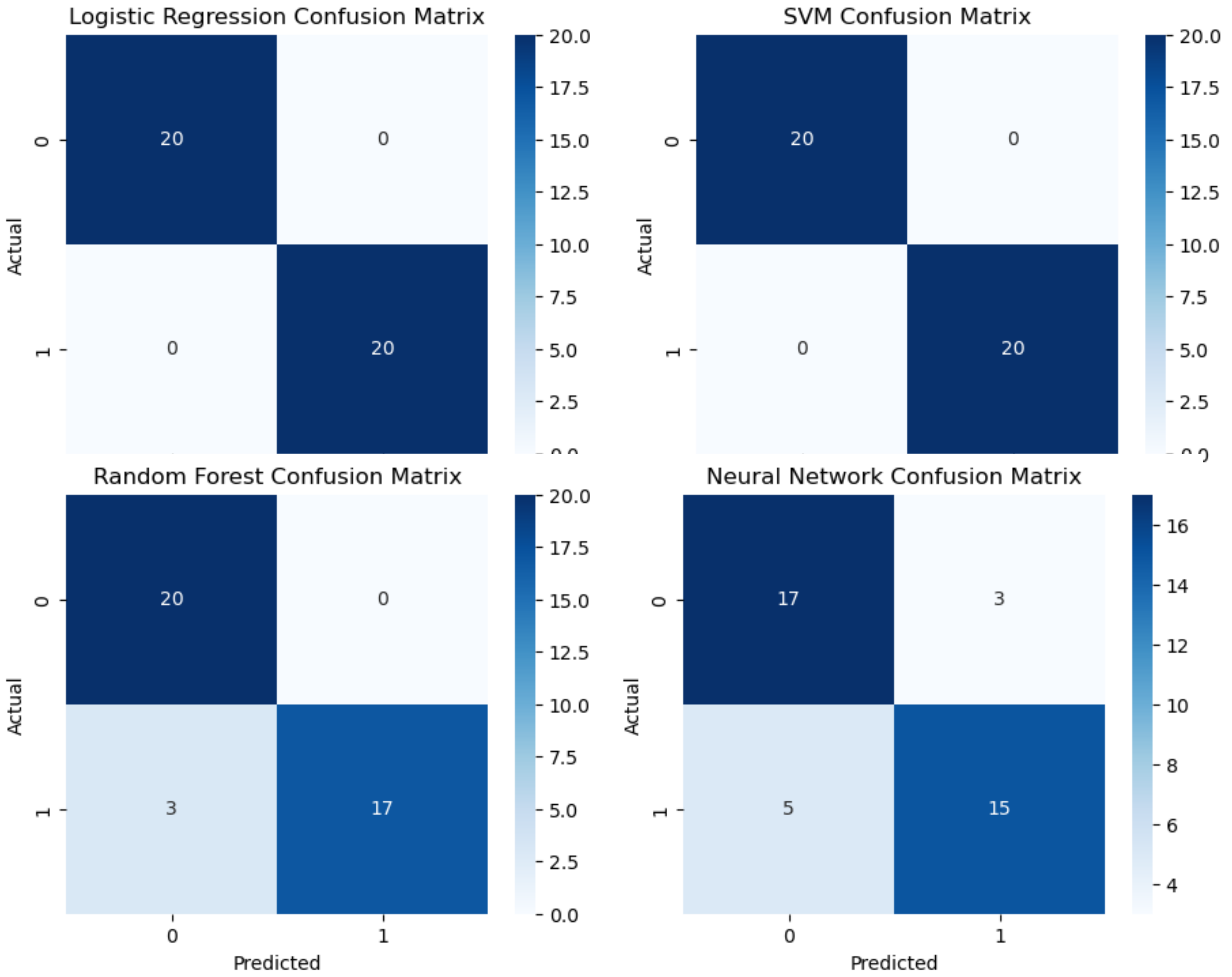
Tuning approach: used F1-score as the primary evaluation metric for hyperparameter tuning

```
Best parameters for Logistic Regression: {'C': 0.01}
Best cross-validation F1 score: 0.9681

Training SVM...
Best parameters for SVM: {'C': 10, 'kernel': 'rbf'}
Best cross-validation F1 score: 0.9935

Training Random Forest...
Best parameters for Random Forest: {'max_depth': 5, 'n_estimators': 50}
Best cross-validation F1 score: 0.9677

Training Neural Network...
Best parameters for Neural Network: {'activation': 'tanh', 'hidden_layer_sizes': (128,)}
Best cross-validation F1 score: 0.9552
```

Model evaluation metrics

Metrics Used:

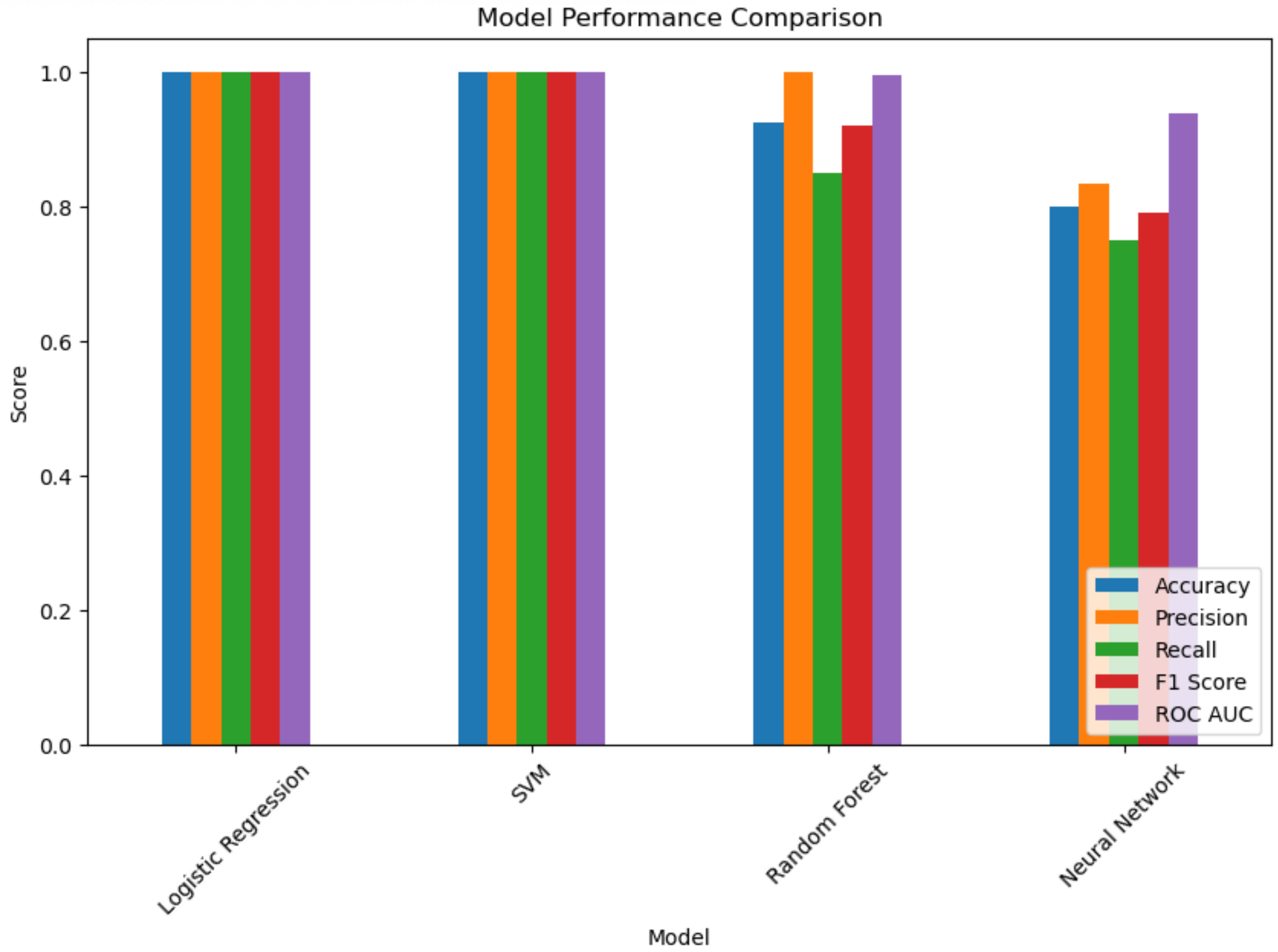
- **Accuracy**: Overall correctness.
- **Precision**: Proportion of predicted positives that are correct.
- **Recall**: Proportion of actual positives that are correctly predicted.
- **F1-Score**: Balance between Precision and Recall.
- **ROC AUC Score**: Trade-off between True Positive Rate (TPR) and False Positive Rate (FPR).



Model evaluation and results

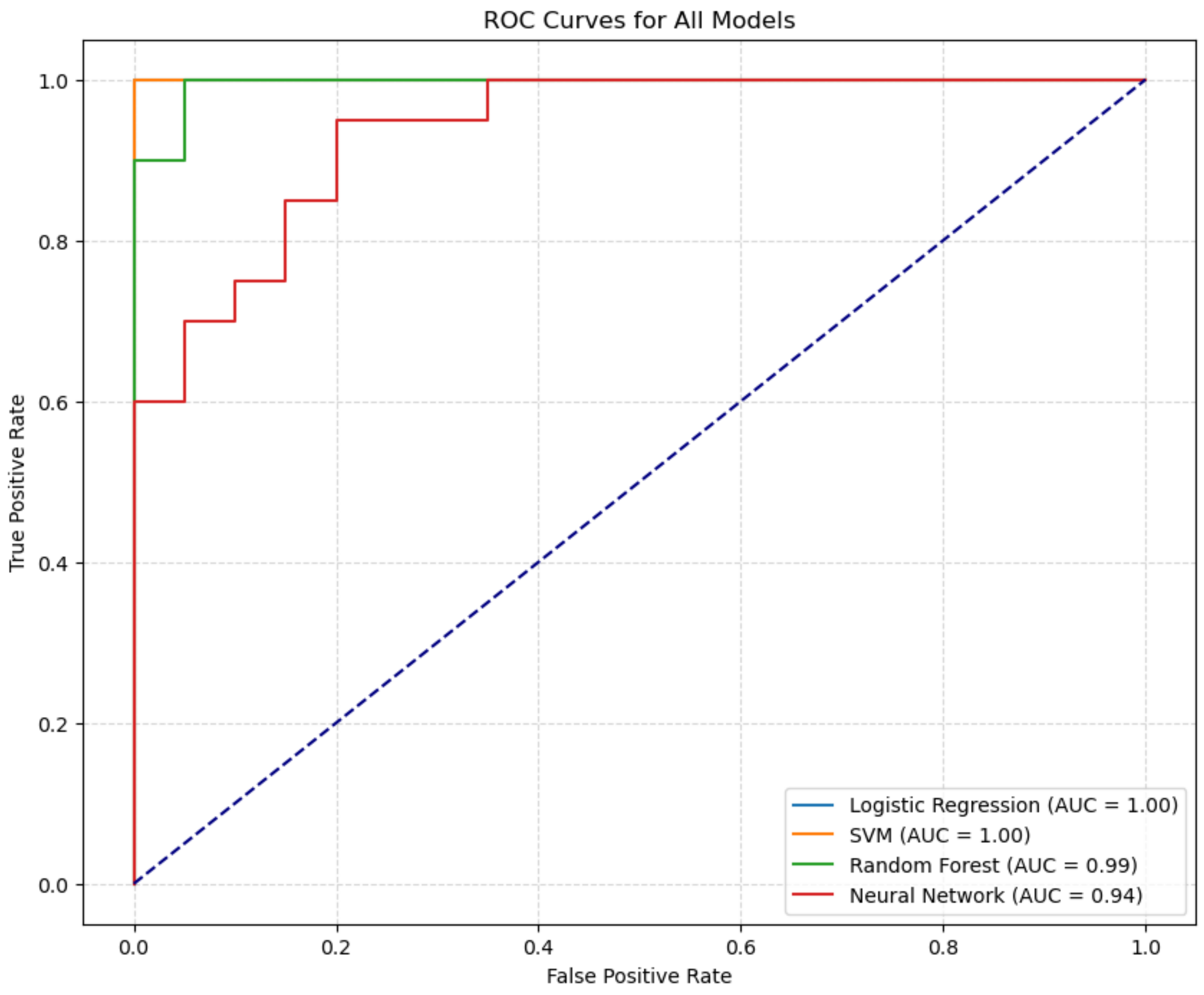
Model Performance Comparison:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	Logistic Regression	1.000	1.000000	1.00	1.000000	1.0000
1	SVM	1.000	1.000000	1.00	1.000000	1.0000
2	Random Forest	0.925	1.000000	0.85	0.918919	0.9950
3	Neural Network	0.800	0.833333	0.75	0.789474	0.9375



ROC Curve Analysis

ROC Curve: Represents the trade-off between TPR and FPR for each model.



Conclusion

- **Best Model:** Neural Network achieved the best metrics but needs further verification to avoid overfitting
- **Robust Choice:** SVM performed very well without signs of overfitting.

Model	Strengths	Limitations
Logistic Regression	- Easy to interpret	- Limited to linear relationships
	- Computationally efficient	- Sensitive to outliers
SVM	- Effective in high dimensions	- High training complexity
	- Powerful with the kernel trick	- Needs careful hyperparameter tuning
Random Forest	- Handles non-linear relationships	- Difficult to interpret
	- Provides feature importance insights	- High memory usage
Neural Network (MLP)	- Captures complex, non-linear relationships	- Prone to overfitting
	- Adaptable through hidden layers	- Requires significant computational power