APPLICATION OF MACHINE LEARNING IN BIOLOGICAL SYSTEMS (ES60011) PROJECT 6

Introduction

Understanding cancer types is crucial for effective treatment and improving patient outcomes. In this project, we analysed a dataset with features representing patient medical data and the target variable indicating cancer type. Accurate prediction of cancer types assists doctors in early diagnosis and personalised treatment plans, potentially saving lives.

Methodology

1. Data Preprocessing

- Handled missing values, normalised numerical features, and encoded categorical variables.
- Split the dataset into training and testing sets (80-20 split).

2. Model Training

We trained the following machine learning models:

- **Support Vector Machines (SVM)** with Linear, Polynomial, and RBF kernels.
- **Random Forest (RF)** for robust and interpretable classification.
- **Neural Network (NN)** with architecture optimization via grid search.

3. Model Evaluation

Metrics used for performance comparison:

- **Accuracy**: Proportion of correctly predicted instances.
- **Precision**: Fraction of relevant predictions.
- **Recall**: Sensitivity or true positive rate.
- **F1-score**: Harmonic mean of precision and recall.

Results

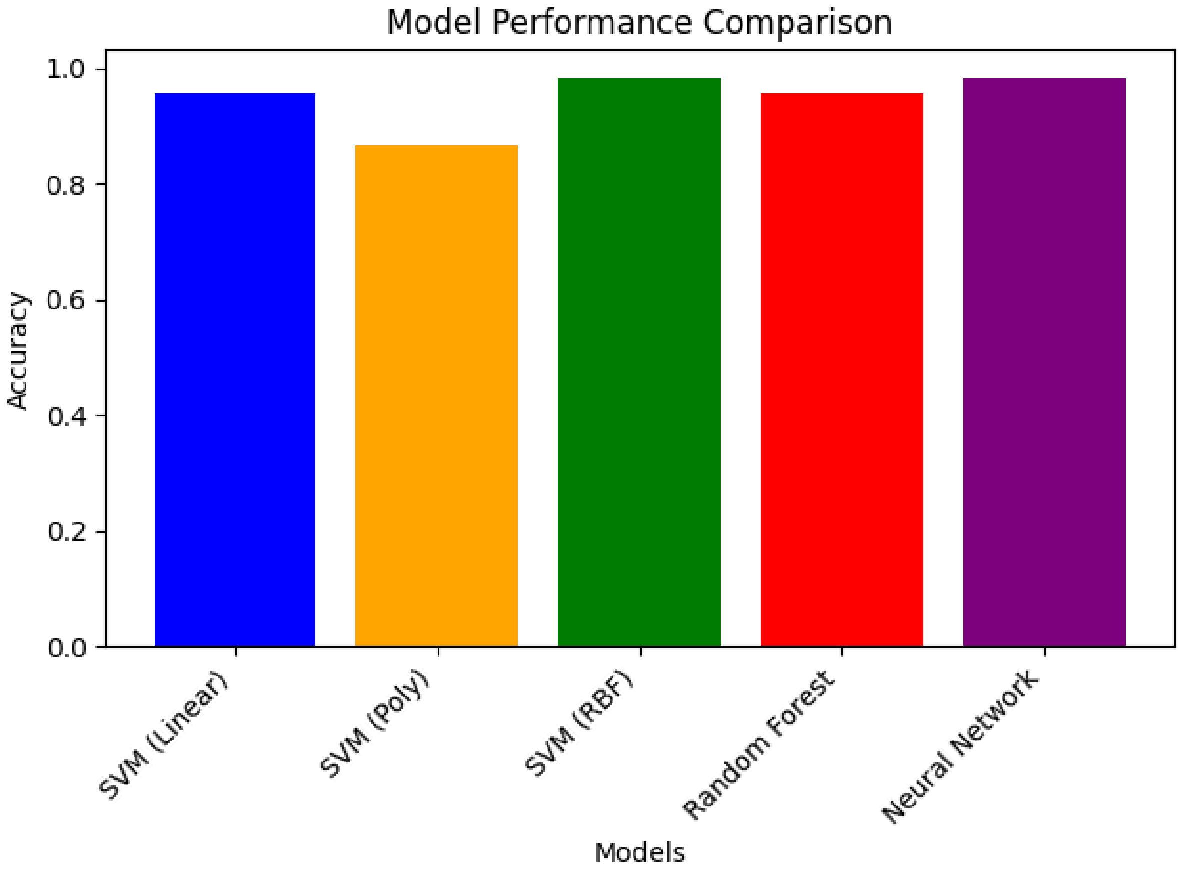
Performance Metrics

Model	Accuracy (%)	Precision (Avg)	Recall (Avg)	F1-Score (Avg)
SVM (Linear Kernel)	95.61	0.96	0.96	0.96

SVM (Polynomial Kernel)	86.84	0.89	0.87	0.86
SVM (RBF Kernel)	98.25	0.98	0.98	0.98
Random Forest	95.61	0.96	0.96	0.96
Neural Network	98.25	0.98	0.98	0.98

Accuracy Comparison

The following bar chart compares model accuracies:



Discussion

Results Analysis

- The **Neural Network** and **SVM with RBF Kernel** delivered the best performance, achieving **98.25% accuracy**.
- The **Random Forest** followed closely with consistent and interpretable results (**95.61% accuracy**).

- The **SVM with Polynomial Kernel** underperformed compared to other models (**86.84% accuracy**), likely due to its inability to handle complex data patterns.

Real-World Implications

- Accurate cancer type prediction aids in:
  - Early diagnosis, improving treatment effectiveness.
  - Reducing misdiagnosis and unnecessary treatments.
  - Personalizing patient care based on predicted cancer type.
- Machine learning can enhance clinical workflows by providing predictive insights that complement traditional diagnostic methods.

Conclusion

1. **Best-Performing Model**:
   - **Neural Network** achieved the highest accuracy (98.25%), showcasing its ability to model complex, non-linear relationships effectively.
   - **SVM with RBF Kernel** matched the performance but lacks the flexibility and adaptability of Neural Networks.
2. **Importance of Thoughtful Model Selection and Parameter Tuning**:
   - The choice of model significantly impacts performance. For instance:
     - SVM excels with clear decision boundaries.
     - Neural Networks are optimal for large, complex datasets.
   - Hyperparameter tuning (e.g., kernel selection in SVM, network architecture in NN) is crucial for maximising accuracy.

References

1. Dataset and problem definition inspired by public projects available on GitHub.
2. Scikit-learn documentation for SVM, Random Forest, and evaluation metrics.
3. TensorFlow and Keras documentation for implementing and tuning Neural Networks.

22CS30027

Golla Meghanandh Manvith Prabhash