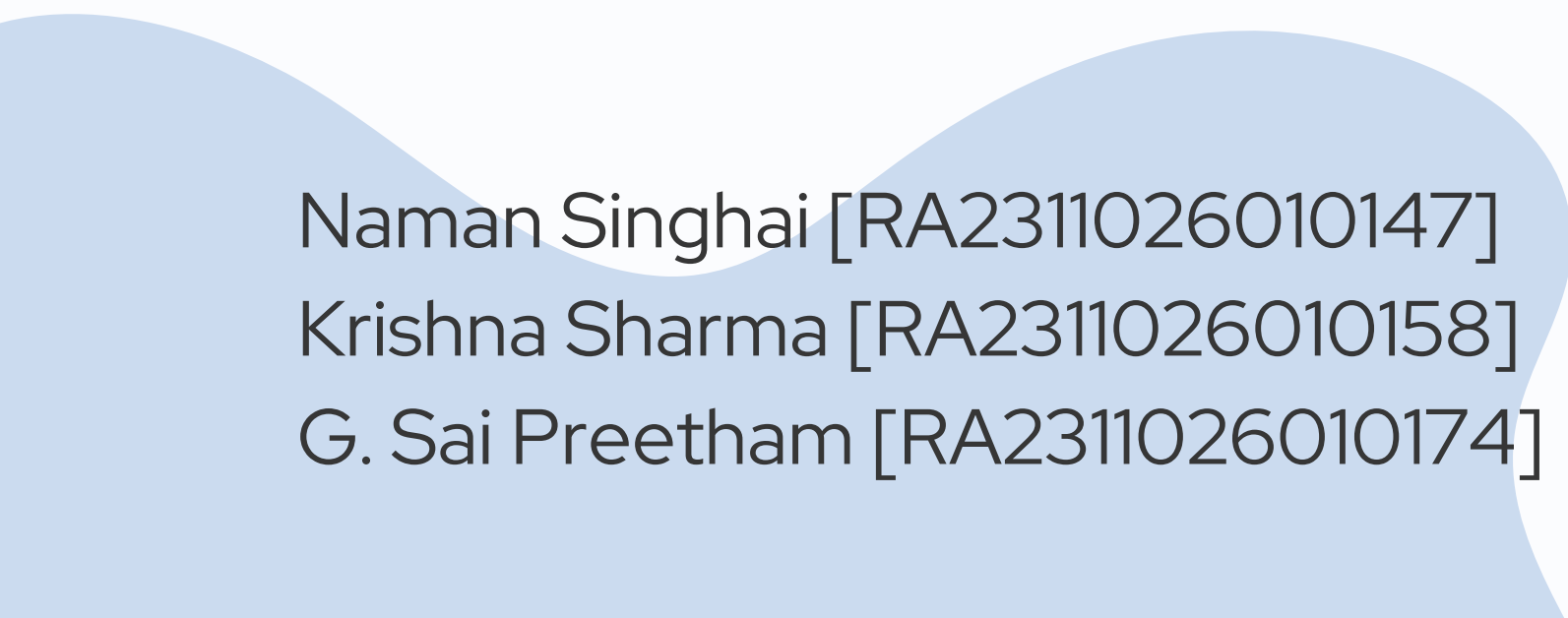




DNA Classification Using Machine Learning



Naman Singhai [RA2311026010147]
Krishna Sharma [RA2311026010158]
G. Sai Preetham [RA2311026010174]

Introduction

DNA (Deoxyribonucleic Acid) sequences are fundamental building blocks that store genetic information in all living organisms. Classification of these sequences is critical in biological research, especially for tasks like gene identification, mutation detection, and disease prediction.

With the increasing amount of genetic data, manual classification is no longer viable. Machine learning offers automated, accurate, and scalable solutions for recognizing patterns in DNA sequences and predicting anomalies.

This project leverages supervised learning techniques to classify DNA promoter sequences using a dataset from the UCI Machine Learning Repository.

Problem Statement

Manual analysis of DNA sequences is time-consuming, error-prone, and inefficient, especially with large-scale genomic datasets.

The primary challenge is to accurately identify and classify DNA sequences (such as promoter regions) to support early diagnosis of genetic disorders or to understand biological functions.

Key Problems:

- 01 Identify Market Trends
- 02 Examine Regulatory Changes
- 03 Assess Technological Advancements



Literature Survey – Study_1

Title: ML Using Intrinsic Genomic Signatures for Rapid Pathogen Classification – COVID-19 Case Study (PLOS ONE, 2020)

Authors: Gurjit S. Randhawa, Maximillian P. M. Soltysiak, Hadi El Roz, Camila P. E. de Souza, Kathleen A. Hill, Lila Kari

Method:

- Used MLDSP-GUI: combines Chaos Game Representation (k=7), DFT & Pearson correlation
- Classifiers: SVM (linear, quadratic), KNN, Subspace models (10-fold CV)
- Dataset: 5000+ viral genomes, including COVID-19

Key Outcomes:

- Achieved 100% accuracy in classifying COVID-19 at all taxonomic levels (Realm to Sub-genus)
- Alignment-free & fast genomic pattern recognition

Limitations:

- Lacks biological interpretability & mutation-level insights
- K-mer sensitive (k=7)
- High computational cost & no clinical validation

Literature Survey – Study 2

Title: ML-Based Detection of Early-Stage Colorectal Cancer via cfDNA Sequencing (BMC Cancer, 2019)

Authors: Nathan Wan, David Weinberg, Tzu-Yu Liu, Katherine Niehaus, Eric A. Ariazi, Daniel Delubac, Ajay Kannan, Brandon White, et al. (Affiliated with Freenome, South San Francisco, CA, USA)

Method:

- Applied ML (SVM, Logistic Regression) on WGS data from plasma cfDNA (546 CRC patients, 271 controls)
- Features: Gene-level read counts, normalized + tumor fraction (TF) via IchorCNA
- Confounder-aware validation (age, batch, etc.)

Key Outcomes:

- AUC 0.92 (standard CV); 85% sensitivity at 85% specificity
- AUC 0.83 (confounder-controlled); sensitivity 71%
- Sensitivity increased with tumor stage & TF

Limitations:

- Confounding factors required complex CV design
- TF via CNVs less sensitive than mutation-based approaches
- Retrospective study with potential sample bias
- Limited interpretability; uncertain signal origin

Literature Survey – Study 3

Title: DNA Sequence Classification with Deep Learning – A Survey (Menoufia J. Engg. Res., 2021)

Authors: Samia M. Abd-Alhalem, El-Sayed M. El-Rabaie, Naglaa F. Soliman, Salah Eldin S. E. Abdulrahman, Nabil A. Ismail, Fathi E. Abd El-samie

Overview:

- Reviewed DNA classification methods: Alignment-based, Alignment-free, ML-DSP, and Deep Learning (DL)
- DL models discussed: ANN, CNN, RNN with preprocessing, training strategies & tools (TensorFlow, Keras, PyTorch)
- Practical example: CNN applied on 2D-encoded bacterial DNA

Key Insights:

- DL offers improved accuracy and automation in genomic pattern recognition
- Tools and architectures well-suited for large-scale DNA datasets

Limitations:

- Encoding DNA remains a challenge
- DL models are resource-intensive and prone to overfitting
- Models lack biological interpretability
- Small datasets limit real-world generalization

Literature Survey – Study 4

Title: ML in DNA Microarray Analysis for Cancer Classification (APBC Conf., 2003)

Authors: Sung-Bae Cho & Hong-Hee Won (Yonsei Univ., Korea)

Method:

- Compared 7 feature selection methods + 4 ML classifiers (MLP, KNN, SVM, SASOM) & ensemble voting
- Datasets: Leukemia, Colon, Lymphoma (high-dimensional gene expression data)
- Used top 25 genes per method for classification

Key Findings:

- KNN and MLP performed robustly across datasets
- Feature selection significantly improved classifier accuracy
- Ensemble voting showed potential despite basic implementation

Limitations:

- Small sample size vs high-dimensional gene space → risk of overfitting
- SVM performance inconsistent
- Feature redundancy impacted ensemble effectiveness
- No biological validation of selected genes

Literature Survey – Study 5

Title: Classifying Cancer Patients Based on DNA Sequences Using ML

Authors: Fahad Hussain, Umair Saeed, Ghulam Muhammad, Noman Islam, Ghazala Shafi Sheikh

Method:

- Used supervised ML classifiers (ANN, KNN, DT, Naïve Bayes, RF, SVM, Fuzzy Classifier)
- Dataset: RNA-Seq (291 samples, 5 cancer types)
- Tools: Implemented using KNIME; evaluated via accuracy, precision, recall, specificity, ROC

Key Findings:

- Demonstrated feasibility of ML in DNA-based cancer classification
- Comparative analysis showed Random Forest and SVM as top performers

Limitations:

- Small dataset size and high dimensionality → overfitting risk
- No feature selection or DL comparison
- Lacked clinical validation or deployment perspective

Findings In The Existing System

Limited Scalability of Traditional Techniques

Traditional approaches like motif matching or regular expressions fail to scale with large genomic datasets and cannot generalize across species or contexts.

Inconsistent Accuracy of ML Models

Several existing ML models struggle with performance inconsistency due to the high dimensional and complex nature of DNA sequences.

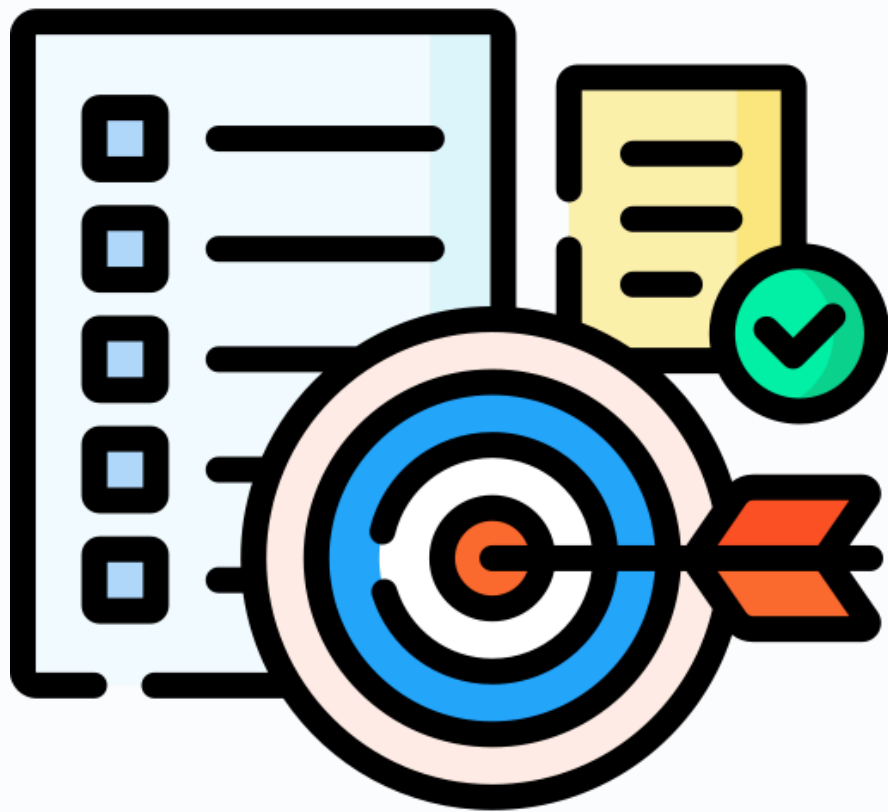
Lack of Biological Interpretability

While some models achieve decent accuracy, they offer minimal insight into the biological reasoning behind classifications.

High Resource Demand of Deep Learning

Deep learning approaches require large datasets and significant computational power, which may not be practical for small-scale or resource-constrained projects.

Objectives



01

To preprocess raw DNA promoter sequences into a format suitable for machine learning models.

02

To train and evaluate various machine learning classifiers for DNA classification.

03

To identify the most effective model based on accuracy and performance.

04

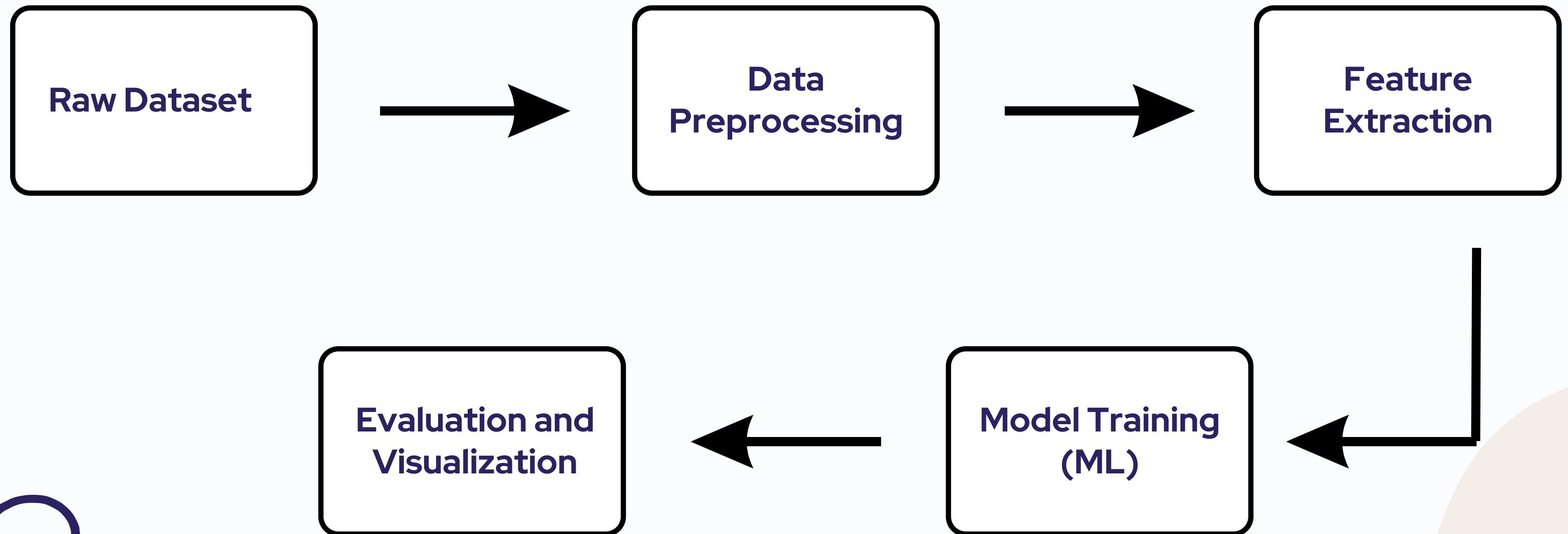
To visualize and interpret results using graphs.

05

To provide a lightweight, reproducible pipeline for DNA sequence classification.

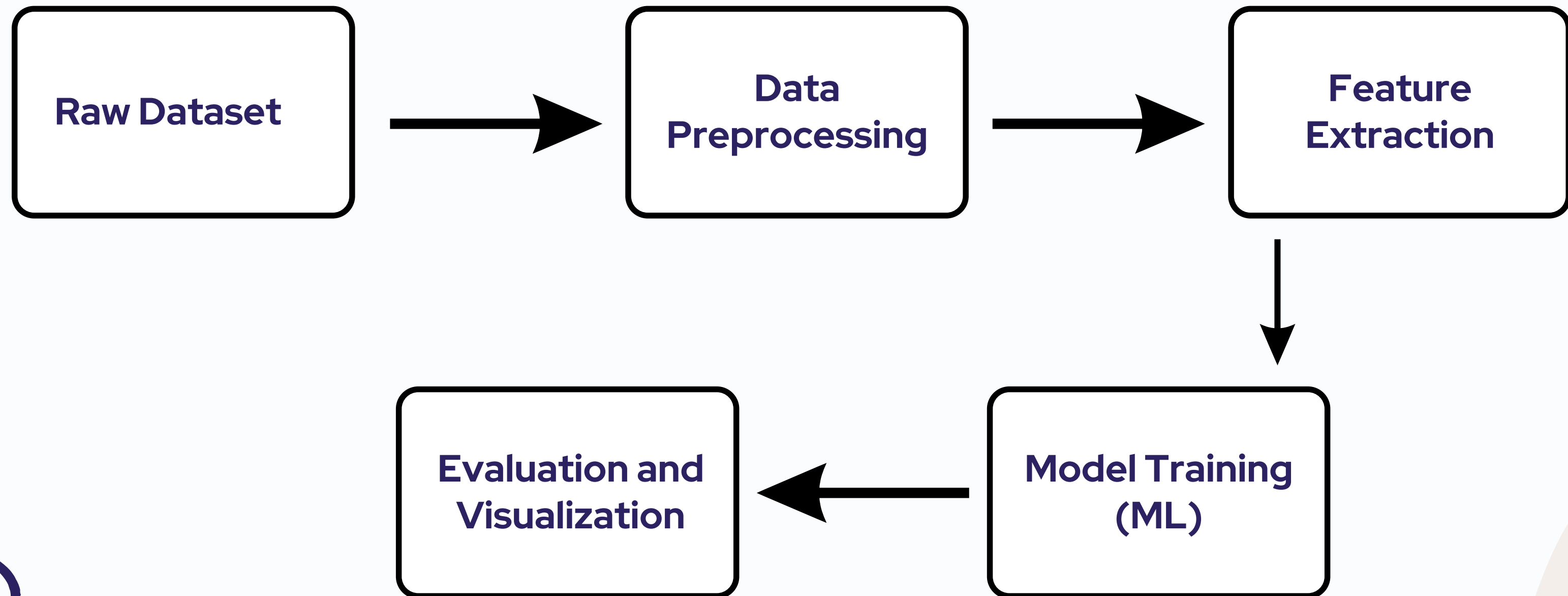
Proposed Methodology.

Architecture Diagram



Proposed Methodology.

Architecture Diagram





Preprocessing Steps & Feature Engineering



Step 1: Load Dataset

The .data file contains sequences along with labels (+1 for promoter, -1 for non-promoter).

Step 2: Split Sequences

Each DNA string is split into individual characters (nucleotides): A, T, G, C.

Step 3: Label Encoding

Target labels are encoded into numerical values:

+1 → 1 (Promoter)

-1 → 0 (Non-promoter)

Step 4: One-Hot Encoding of Nucleotides

Each nucleotide is converted into a binary vector:

A = [1, 0, 0, 0]

T = [0, 1, 0, 0]

G = [0, 0, 1, 0]

C = [0, 0, 0, 1]

Feature Engineering:

- Each DNA sequence (57 characters) becomes a matrix of size (57×4) using one-hot encoding.
- This results in 228 features per sample (57 nucleotides \times 4 values each).
- Final dataset = numerical matrix that machine learning models can process.

Machine Learning Models and Evaluation

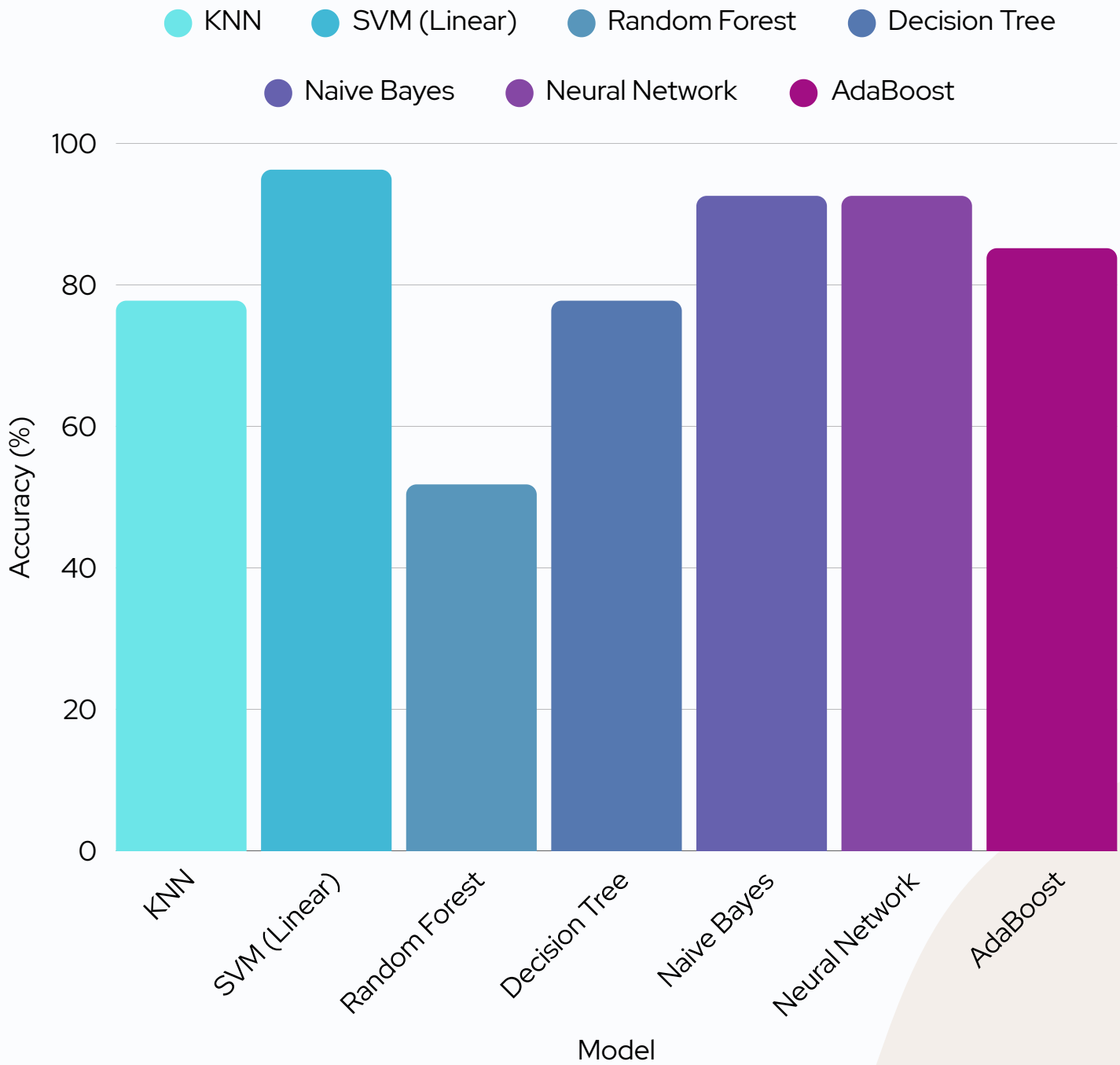
ML Model	Description
K-Nearest Neighbors (KNN):	Classifies based on closest DNA sequences in feature space.
Support Vector Machines (SVM):	Trained using linear, RBF, and sigmoid kernels. Tries to find a hyperplane to best separate promoter vs. non-promoter sequences.
Decision Tree Classifier:	Constructs a flowchart-like structure of decisions based on features.
Random Forest Classifier:	An ensemble of multiple decision trees, improves accuracy and reduces overfitting.

ML Model	Description
Neural Network (MLPClassifier):	Uses one or more hidden layers to learn complex patterns in DNA sequences.
Naive Bayes:	Based on Bayes' Theorem assuming independence between features.
AdaBoost Classifier:	Boosting technique that combines weak learners to form a strong classifier.
Gaussian Process Classifier:	Probabilistic model that learns a distribution over functions to classify DNA

Results and Discussions

Accuracy Scores:

Model	Accuracy (%)
KNN	77.77
SVM (Linear)	96.29
Random Forest	51.8
Decision Tree	77.77
Naive Bayes	92.59
Neural Network	92.59
AdaBoost	85.19



Results and Discussions

Model Evaluation Results:

```
K Nearest Neighbors
,0.7777777777777778
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.65      0.79         17
,      1          0.62      1.00      0.77         10
,
,      accuracy                0.78         27
,      macro avg      0.81      0.82      0.78         27
,weighted avg      0.86      0.78      0.78         27
,
,Gaussian Process
,0.8888888888888888
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.82      0.90         17
,      1          0.77      1.00      0.87         10
,
,      accuracy                0.89         27
,      macro avg      0.88      0.91      0.89         27
,weighted avg      0.91      0.89      0.89         27
,
,Decision Tree
,0.7777777777777778
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.65      0.79         17
,      1          0.62      1.00      0.77         10
,
,      accuracy                0.78         27
,      macro avg      0.81      0.82      0.78         27
,weighted avg      0.86      0.78      0.78         27
,
```

```
,Random Forest
,0.5185185185185185
,
,      precision    recall  f1-score   support
,
,      0          0.70      0.41      0.52         17
,      1          0.41      0.70      0.52         10
,
,      accuracy                0.52         27
,      macro avg      0.56      0.56      0.52         27
,weighted avg      0.59      0.52      0.52         27
,
,Neural Net
,0.9259259259259259
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.88      0.94         17
,      1          0.83      1.00      0.91         10
,
,      accuracy                0.93         27
,      macro avg      0.92      0.94      0.92         27
,weighted avg      0.94      0.93      0.93         27
,
,AddaBoost
,0.8518518518518519
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.76      0.87         17
,      1          0.71      1.00      0.83         10
,
,      accuracy                0.85         27
,      macro avg      0.86      0.88      0.85         27
,weighted avg      0.89      0.85      0.85         27
,
```

```
,Naive Bayes
,0.9259259259259259
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.88      0.94         17
,      1          0.83      1.00      0.91         10
,
,      accuracy                0.93         27
,      macro avg      0.92      0.94      0.92         27
,weighted avg      0.94      0.93      0.93         27
,
,SVM Linear
,0.9629629629629629
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.94      0.97         17
,      1          0.91      1.00      0.95         10
,
,      accuracy                0.96         27
,      macro avg      0.95      0.97      0.96         27
,weighted avg      0.97      0.96      0.96         27
,
,SVM RBF
,0.7777777777777778
,
,      precision    recall  f1-score   support
,
,      0          1.00      0.65      0.79         17
,      1          0.62      1.00      0.77         10
,
,      accuracy                0.78         27
,      macro avg      0.81      0.82      0.78         27
,weighted avg      0.86      0.78      0.78         27
,
```

- **Best Model Performance:**

- The SVM (Linear) classifier achieved the highest accuracy of 96.29%, making it the most effective model for this task.

- **Naive Bayes and Neural Networks:**

- Naive Bayes and Neural Networks also performed well, with accuracies around 92.59%.
- Both models demonstrated good ability to capture complex patterns in DNA sequences.

- **Random Forest:**

- Random Forest showed the lowest accuracy (51.8%), likely due to its assumption of feature independence, which doesn't hold in DNA sequence data.

- **Importance of Preprocessing:**

- One-hot encoding of nucleotide sequences was essential for converting the data into a format suitable for machine learning models.

- **Class Imbalance Consideration:**

- The dataset may have had a class imbalance, which could have influenced model performance. Random Forest was particularly resilient to this imbalance.

- **Precision vs. Recall:**

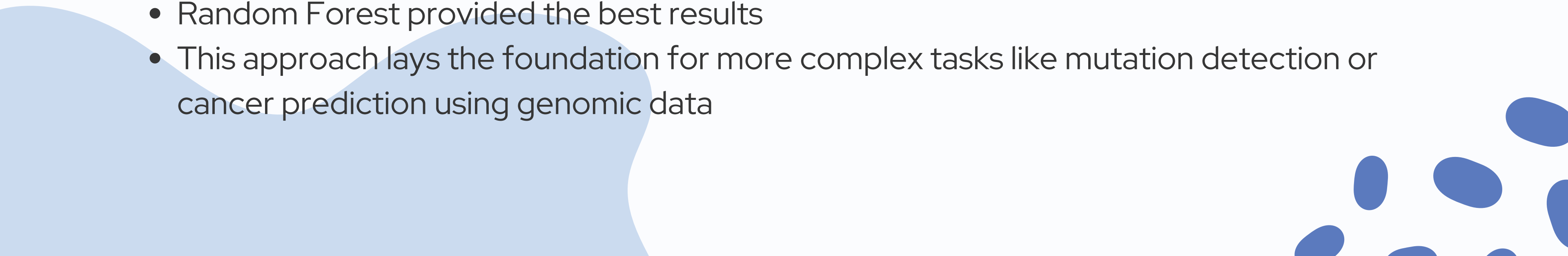
- In genetic classification tasks, recall (identifying true promoter sequences) is more critical than precision, and Support Vector Machine (Linear) provided a good balance.

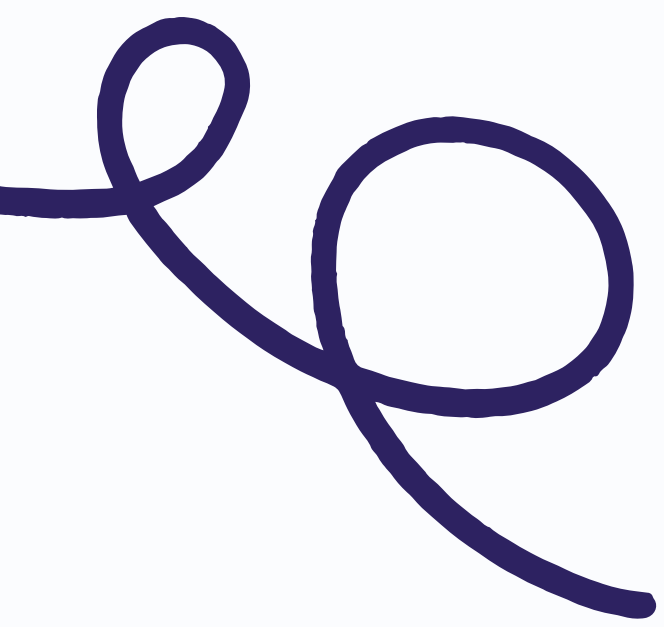


Conclusion

This project successfully demonstrated the use of machine learning in DNA sequence classification.

It provides a lightweight and efficient framework for promoter recognition, an essential task in genetic diagnostics and biological research.

- Multiple models were compared
 - Preprocessing techniques such as one-hot encoding proved effective
 - Random Forest provided the best results
 - This approach lays the foundation for more complex tasks like mutation detection or cancer prediction using genomic data
- 



Thank You!

