Intelligent Algorithms for DNA Detection

Electronic and Information Engineering MEng

Imperial College London

Andy Sungjun Han

Supervisor: Dr Jesus Rodriguez-Manzano

2nd Marker: Professor Timothy Constandinou

CONTENTS

I. Introduction

I. Introduction

Transformer Model for DNA Sequence Classification from
Loop-mediated isothermal amplification (LAMP)
Intelligent Algorithms for DNA Detection

1. Project Motivation

■ Project Background

Loop-Mediated Isothermal Amplification (LAMP) is a popular DNA amplification and detection technique. It utilizes fluorescence intensity plotted against time cycles to generate an amplification curve, enabling the identification of target DNA sequences. However, implementing "multiplex LAMP," which tests multiple DNA primers in a single sample, presents a notable challenge as it becomes difficult to differentiate between various amplification reactions.

Project's ultimate objective

The final goal of the project was to develop a deep learning algorithm capable of efficiently and accurately classifying amplification curves and melting curves into respective DNA sequence classes.
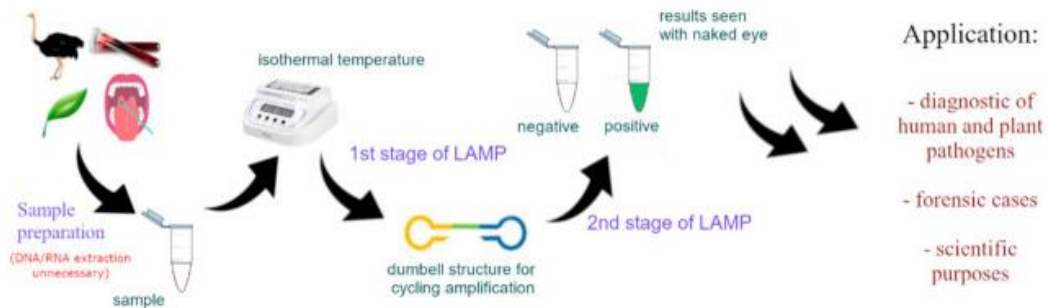
I. Introduction

2. Background Research

■ DNA Amplification and LAMP

- DNA Amplification
- LAMP – a simple, cost effective, isothermal reaction to amplify DNA sequences
- Efficient in diagnostic or detection
- Typical procedure as follows:

■ Multiplex LAMP

- Multiplexing – multiple target primer sets to test for multiple targets in one experiment
- Multiplex LAMP less developed compared to multiplex PCR
- Could use amplification curve or melt curve as a method for multiplex LAMP
- Use deep learning algorithms to classify curves to target DNA sequences
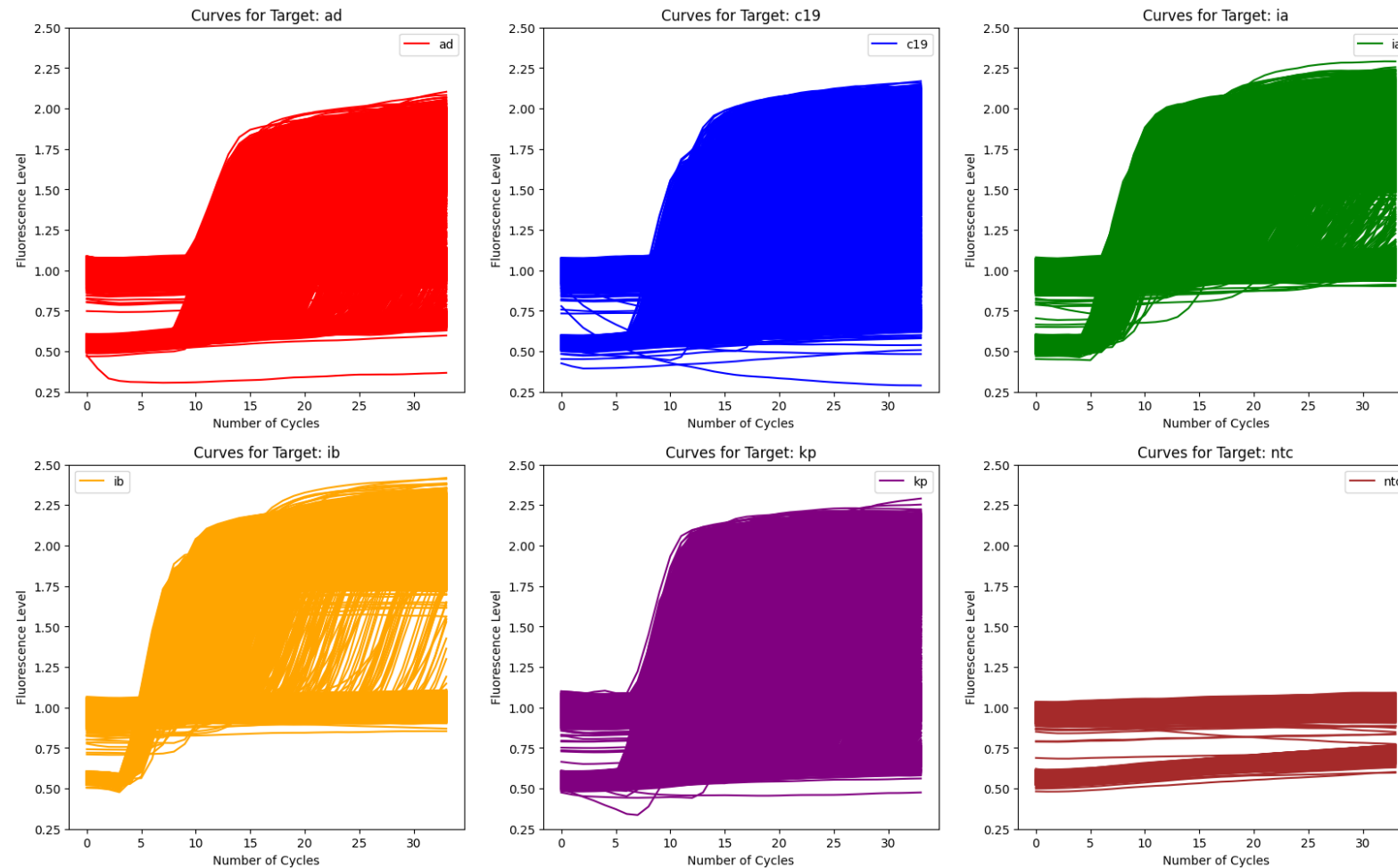
II. Implementation

II. Implementation

Intelligent Algorithms for DNA Detection

1. LAMP Dataset

■ Original Amplification Curve

Transformer Model for DNA Sequence Classification from
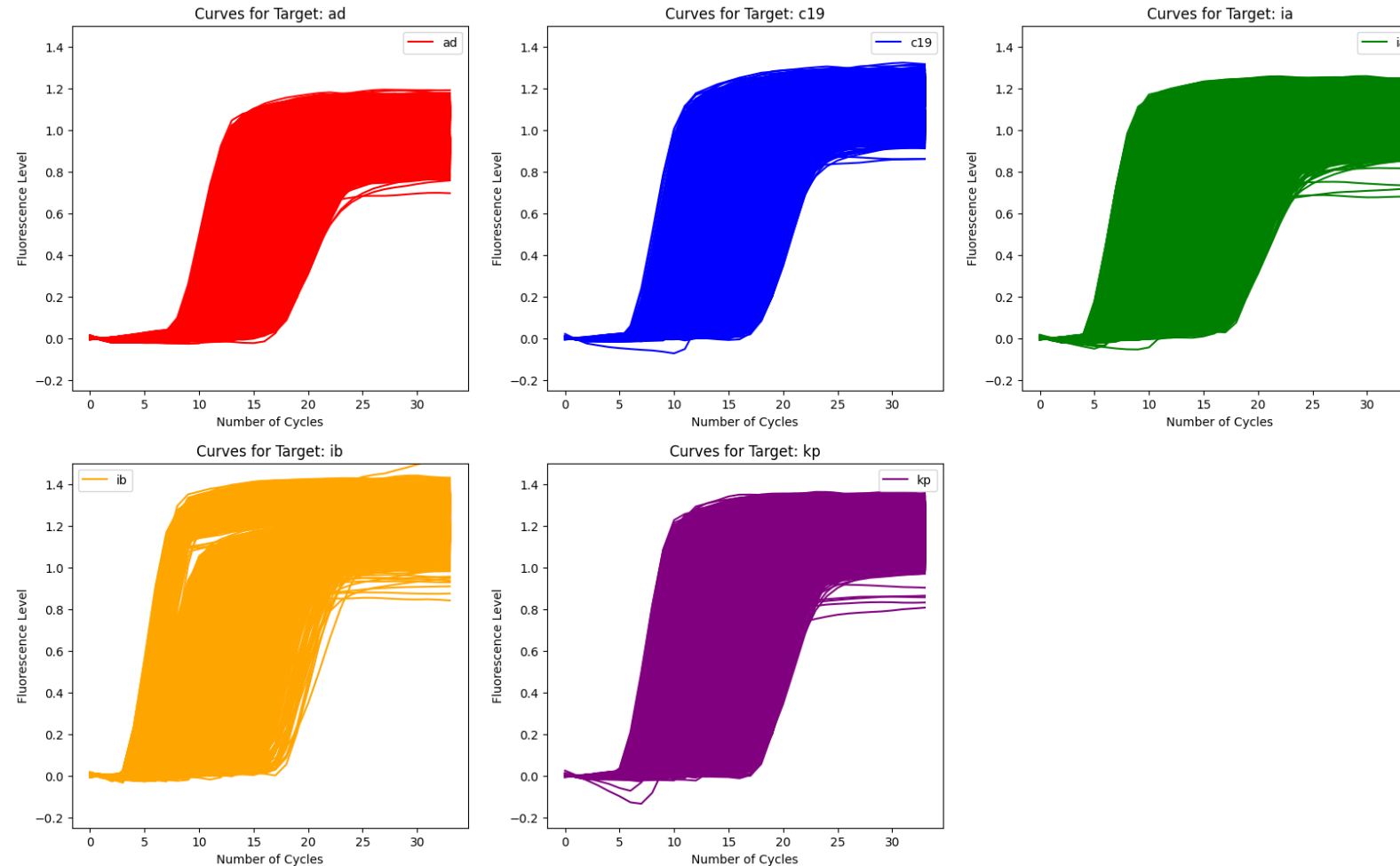Loop-mediated isothermal amplification (LAMP)
Intelligent Algorithms for DNA Detection
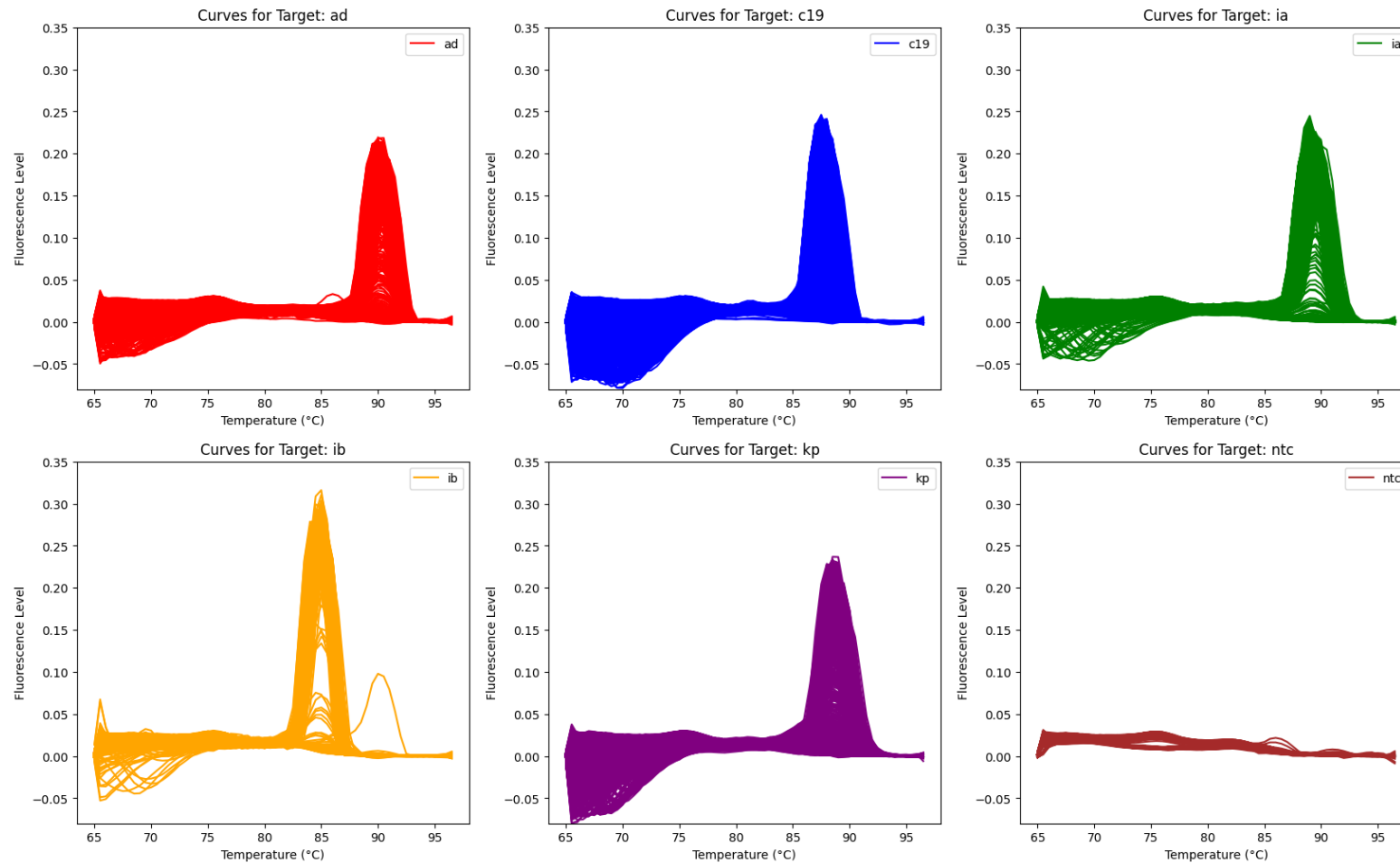
II. Implementation

1. LAMP Dataset

■ Pre-processed Amplification Curve

II. Implementation

1. LAMP Dataset

■ Original Melt Curve

II. Implementation

1. LAMP Dataset

■ Pre-processed Melt Curve

II. Implementation
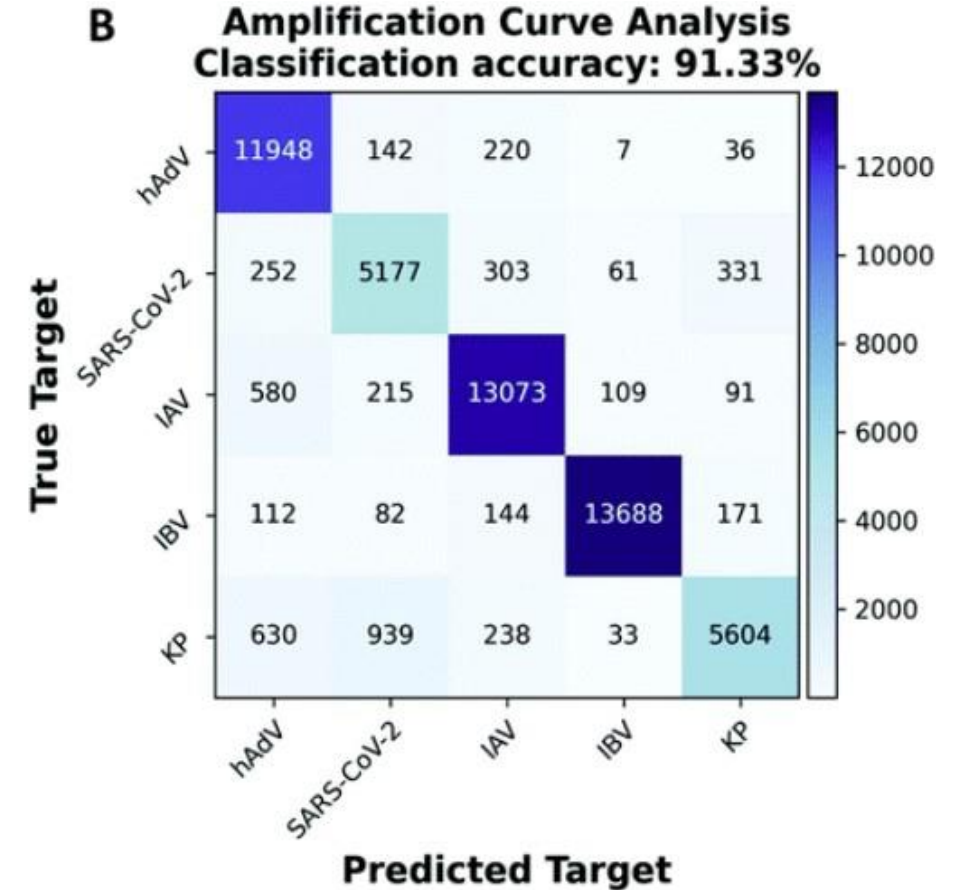
Intelligent Algorithms for DNA Detection

2. Baseline Models

■ "Single-channel digital LAMP multiplexing using amplification curve analysis" by Malpartida-Cardenas et al (2022)

- Dataset of five respiratory pathogens: human influenza A virus (IAV), human influenza B virus (IBV), severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), human adenovirus (hAdV), and Klebsiella pneumoniae (KP)
- Used K-NN Algorithm on Amplification Curve Data with parameter K set to 10
- Achieved classification accuracy of 91.33% ± 0.33% (mean ± std)

B Amplification Curve Analysis
Classification accuracy: 91.33%

Transformer Model for DNA Sequence Classification from
Loop-mediated isothermal amplification (LAMP)
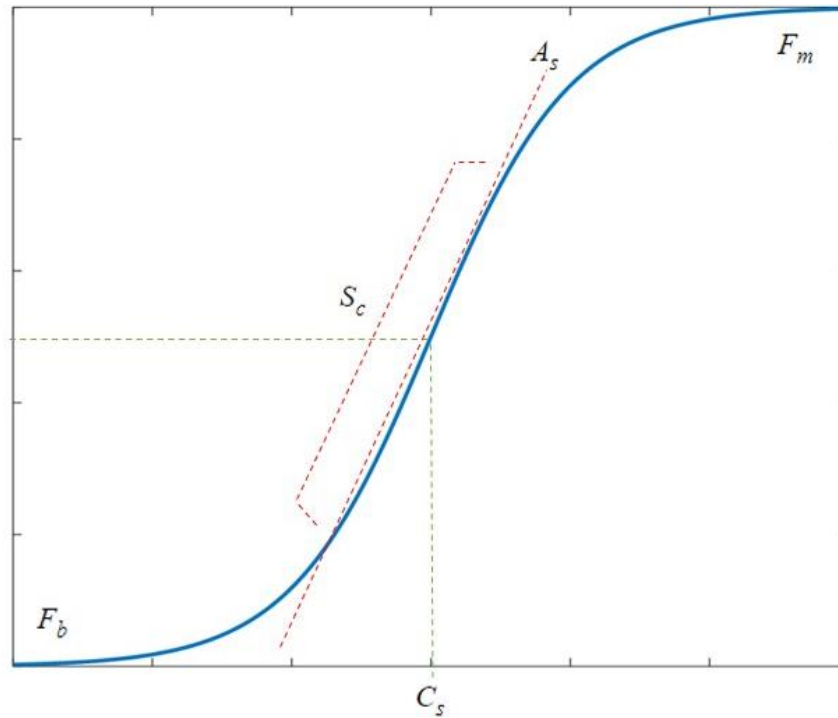Intelligent Algorithms for DNA Detection

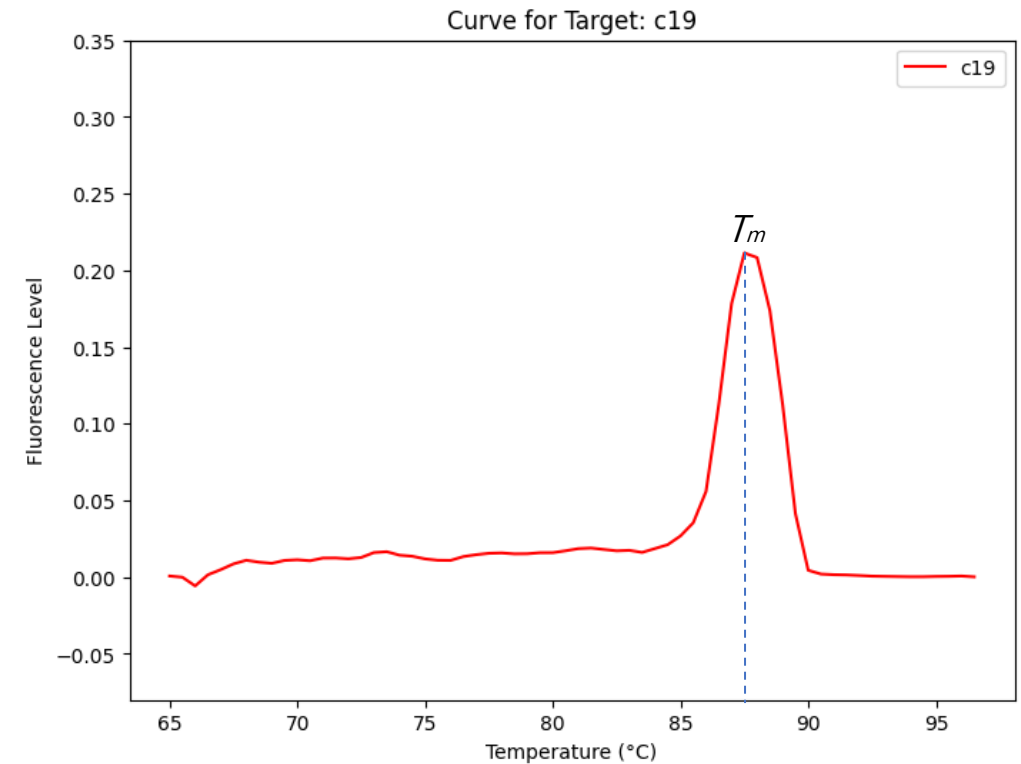# II. Implementation

2. Baseline Models

■ 5 Sigmoid Parameter Features from AC

$$f(t,p) = \frac{F_m}{\left(1 + e^{-S_c(t - C_s)}\right)^{A_s}} + F_b$$

■ Melt Curve Peak Feature from MC

II. Implementation
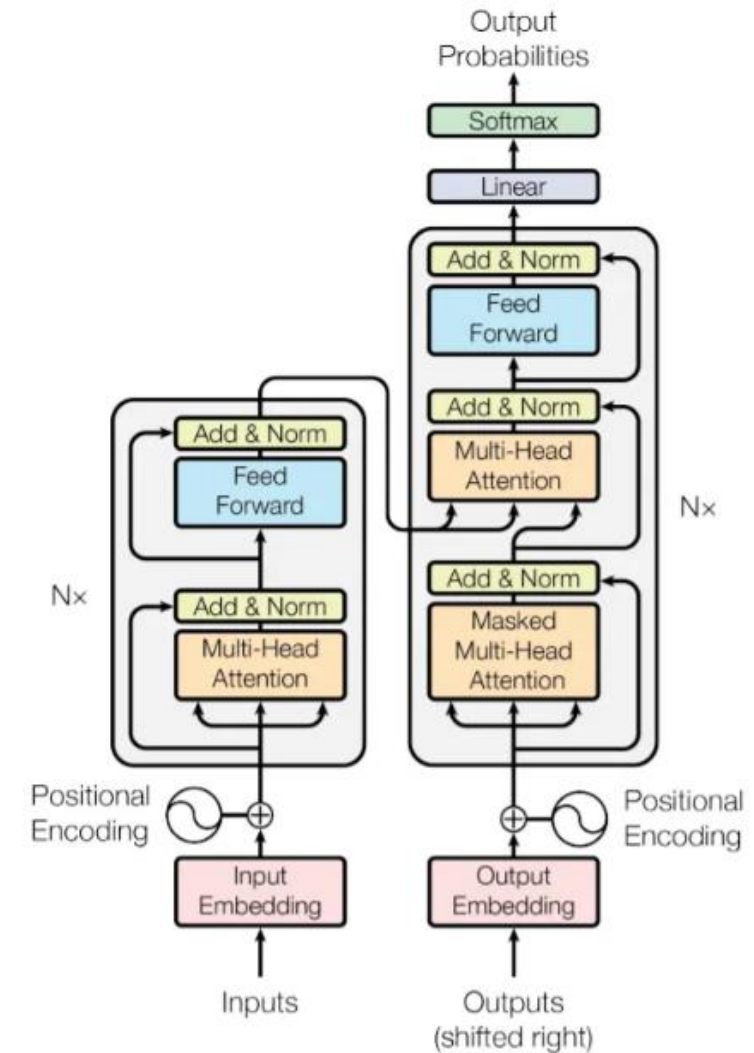
Intelligent Algorithms for DNA Detection

3. Transformer Model

■ Used Transformer Model as an alternative approach

- Task was to analyse time series data such as the amplification curve and melt curve.
- Utilised the effectiveness of the Transformer Model in handling time series classification tasks.
- Implemented positional encoding to incorporate the temporal order of the data
- Adjusted hyperparameters such as d_model, nhead, num_layers, lr, batch_size, and drop_out.

III. Results & Evaluation

III. Results

Intelligent Algorithms for DNA Detection

1. Baseline Models AC

■ Best Baseline Model (Random Forest)

Class	Precision	Recall	F1-Score	Support
AD	0.95	0.98	0.96	1691
IA	0.95	0.96	0.96	2272
IB	0.97	0.98	0.97	1634
KP	0.92	0.92	0.92	1778
C19	0.87	0.77	0.82	1043
Accuracy			0.94	
Macro Avg	0.93	0.92	0.93	8418
Weighted Avg	0.94	0.94	0.94	8418

Accuracy by Feature and Model

III. Results

Intelligent Algorithms for DNA Detection
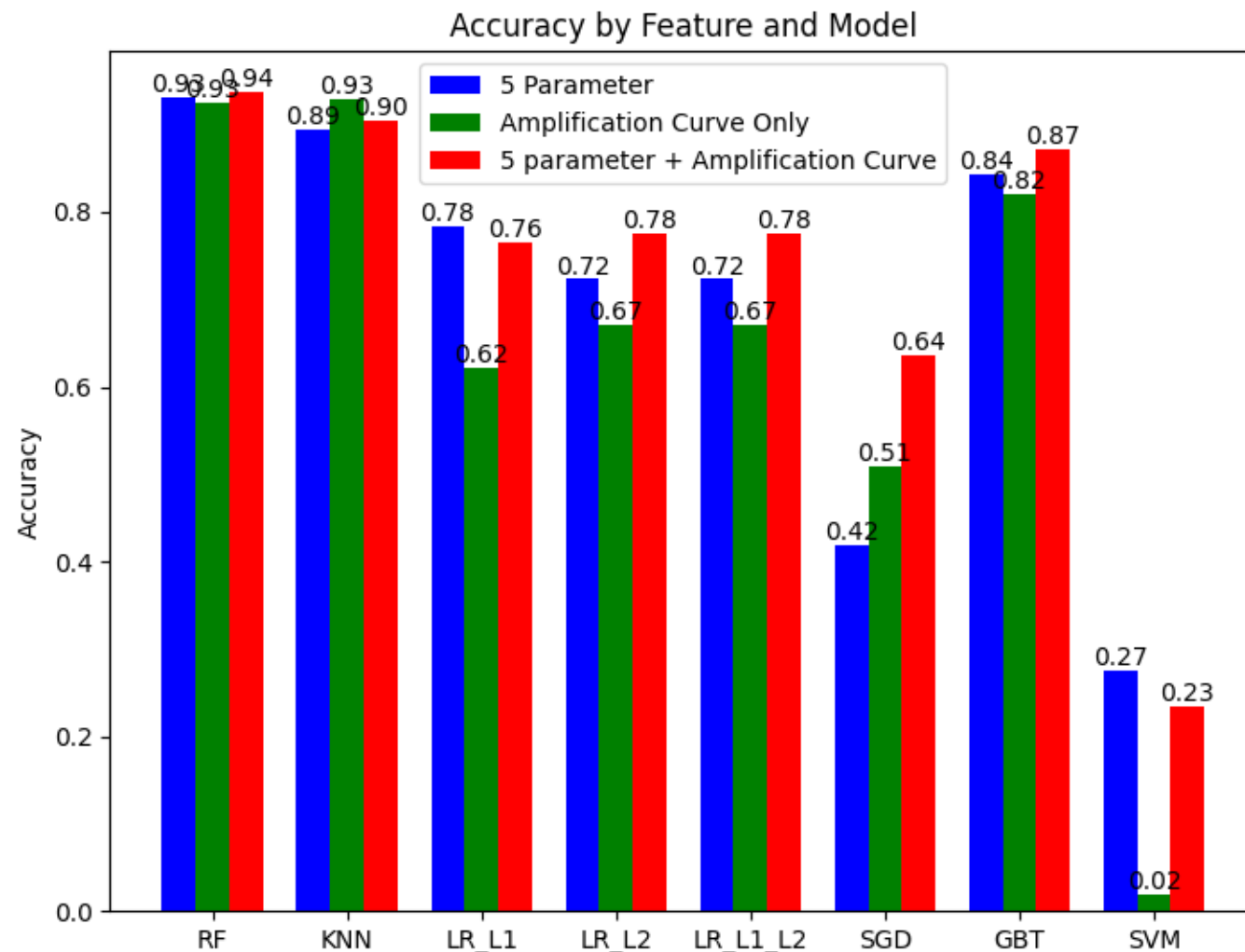
2. Transformer Model AC

■ Best Transformer Model

Class	Precision	Recall	F1-Score	Support
AD	0.95	0.98	0.97	2378
IA	0.95	0.93	0.94	2713
IB	0.98	0.99	0.99	2858
KP	0.94	0.89	0.91	1290
C19	0.80	0.85	0.82	853
Accuracy		0.95		
Macro Avg	0.93	0.93	0.93	10092
Weighted Avg	0.95	0.95	0.95	10092

Best Model Hyperparameters: d_model=128, nhead=4, num_layers=3, lr=0.001, batch_size=64, and drop_out=0.1

Confusion Matrix
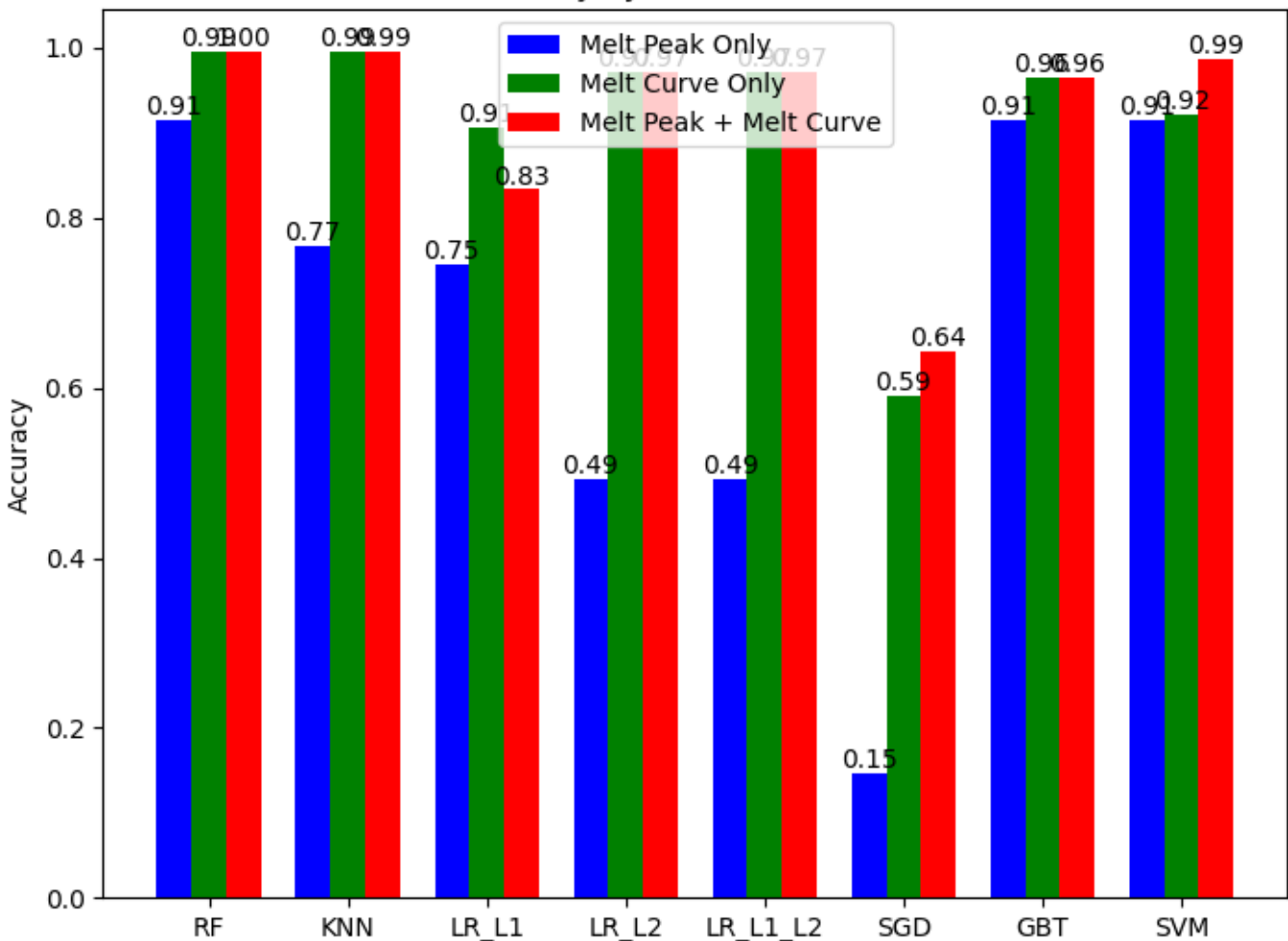
III. Results

Intelligent Algorithms for DNA Detection

3. Baseline Models MC

■ Best Baseline Model (Random Forest)

Class	Precision	Recall	F1-Score	Support
AD	0.99	1.00	0.99	1344
IA	0.99	0.99	0.99	2310
IB	1.00	1.00	1.00	3183
KP	1.00	0.97	0.99	1190
C19	1.00	1.00	1.00	1161
Accuracy		1.00		
Macro Avg	1.00	0.99	0.99	9188
Weighted Avg	1.00	1.00	1.00	9188

Accuracy by Feature and Model
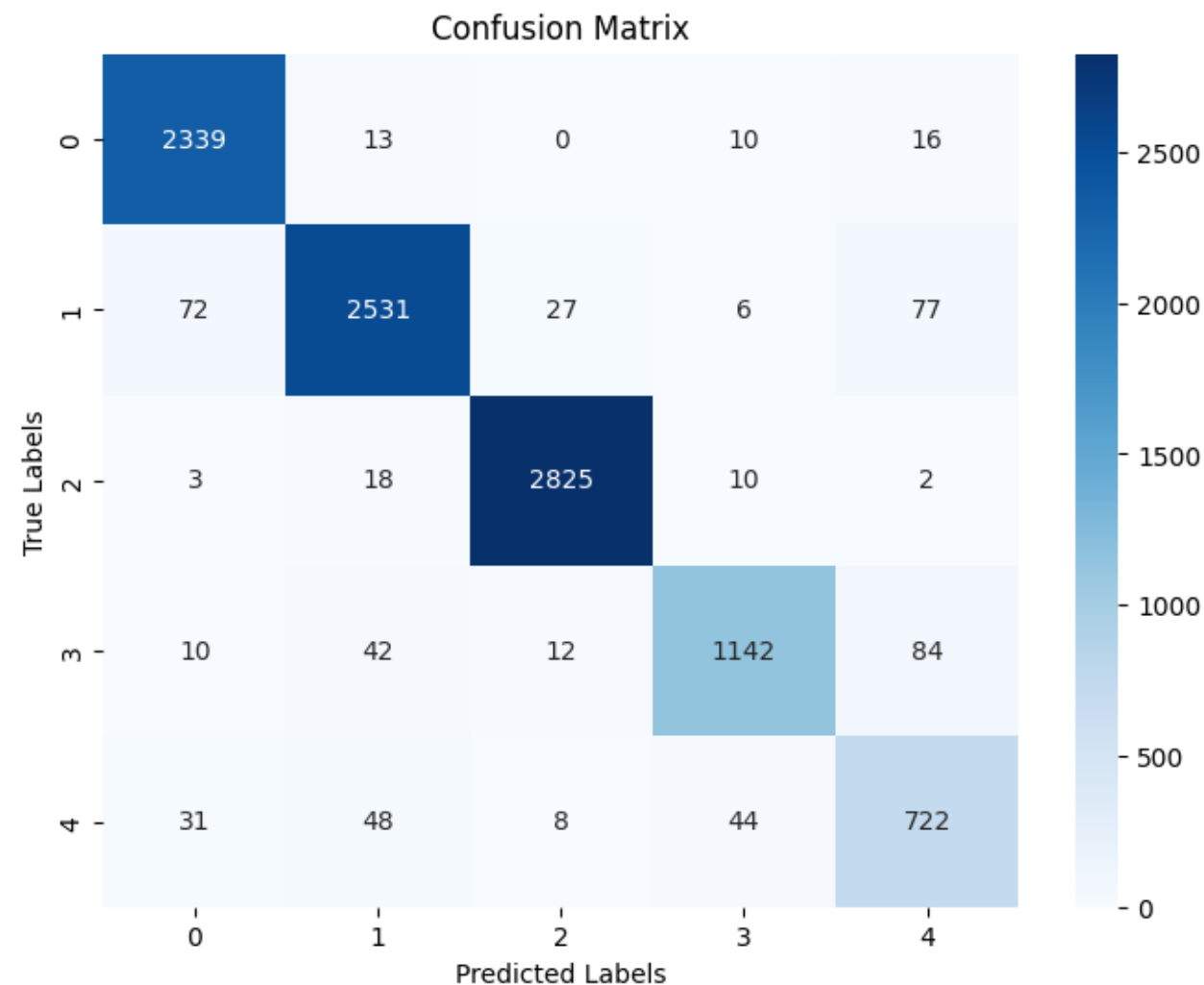
III. Results

Transformer Model for DNA Sequence Classification from
Loop-mediated isothermal amplification (LAMP)
Intelligent Algorithms for DNA Detection

4. Transformer Model MC

■ Best Transformer Model

Class	Precision	Recall	F1-Score	Support
AD	1.00	1.00	1.00	2378
IA	1.00	1.00	1.00	2713
IB	1.00	1.00	1.00	2858
KP	1.00	1.00	1.00	1290
C19	1.00	1.00	1.00	853
Accuracy			1.00	
Macro Avg	1.00	1.00	1.00	10092
Weighted Avg	1.00	1.00	1.00	10092

Best Model Hyperparameters: d_model=128, nhead=8,
num_layers=3, lr=0.0001, batch_size=64, and drop_out=0.1

Confusion Matrix

III. Evaluation
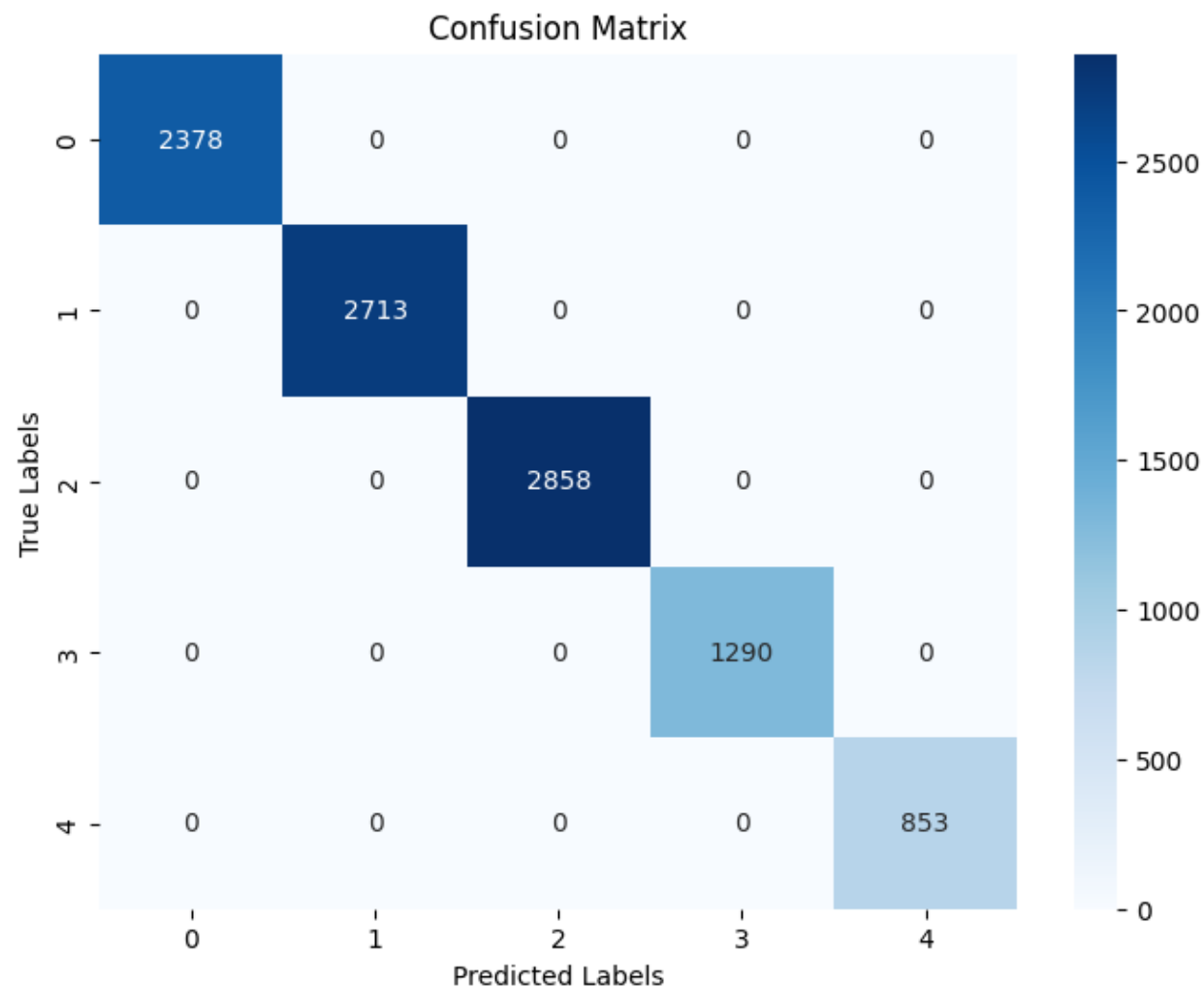
Transformer Model for DNA Sequence Classification from
Loop-mediated isothermal amplification (LAMP)
Intelligent Algorithms for DNA Detection

5. Evaluating Results

■ Performance of the Transformer Models

- Random Forest Model – amplification curve
  - Classification Accuracy: 93.59%
- Transformer Model – amplification curve
  - Classification Accuracy: 94.72%
- Random Forest Model – melt curve
  - Classification Accuracy: 99.52%
- Transformer Model – melt curve
  - Classification Accuracy: 100.00%

■ Biological Significance

- Obtaining melt curve vs amplification curve
- Costs in obtaining curves
- Additional factors for real life scenarios
- Melt curve models have higher accuracy compared to other models.
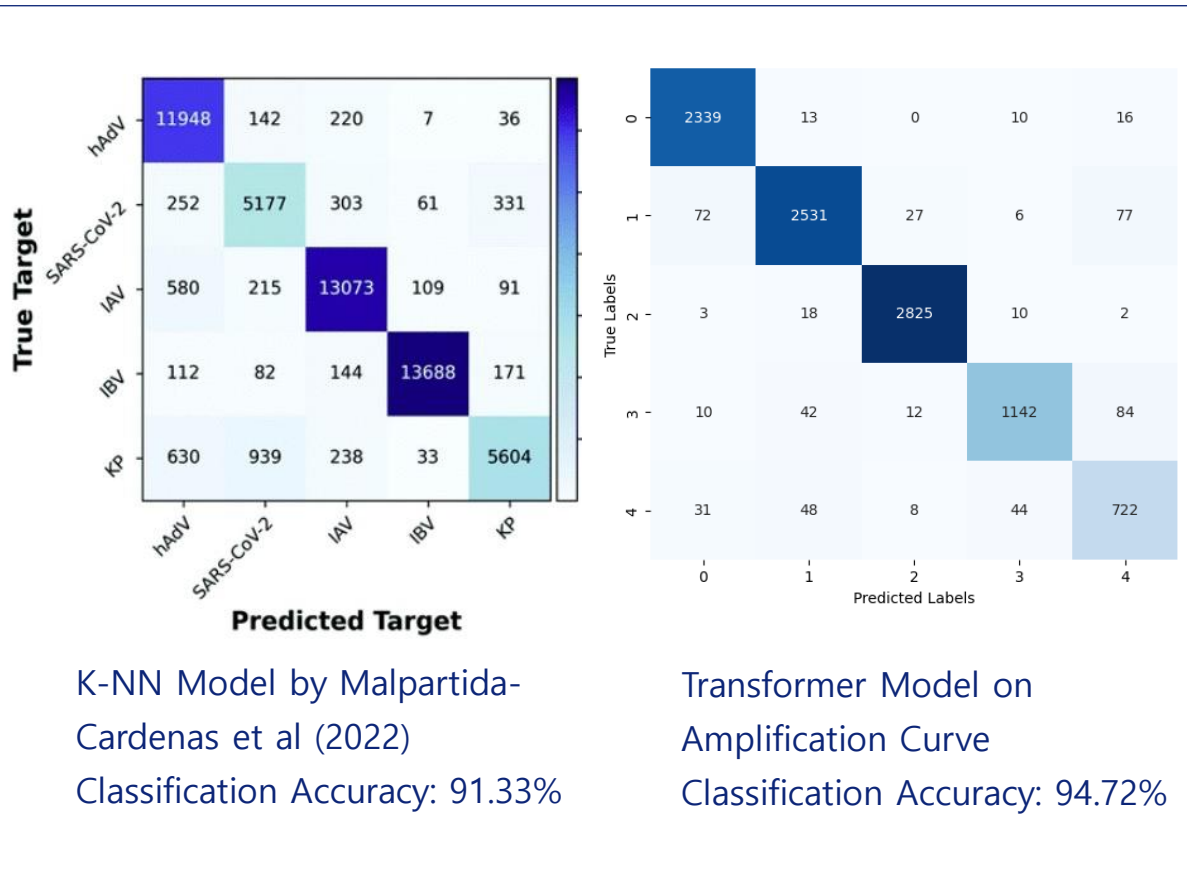- LAMP VS PCR – isothermal advantage

IV. Conclusion

IV. Conclusion

Intelligent Algorithms for DNA Detection

1. Comparison between models

K-NN Model by Malpartida-Cardenas et al (2022)
Classification Accuracy: 91.33%

Transformer Model on Amplification Curve
Classification Accuracy: 94.72%

2. Future Works

- Additional data for further training & testing of the model to enhance robustness
- Usage of transfer learning:
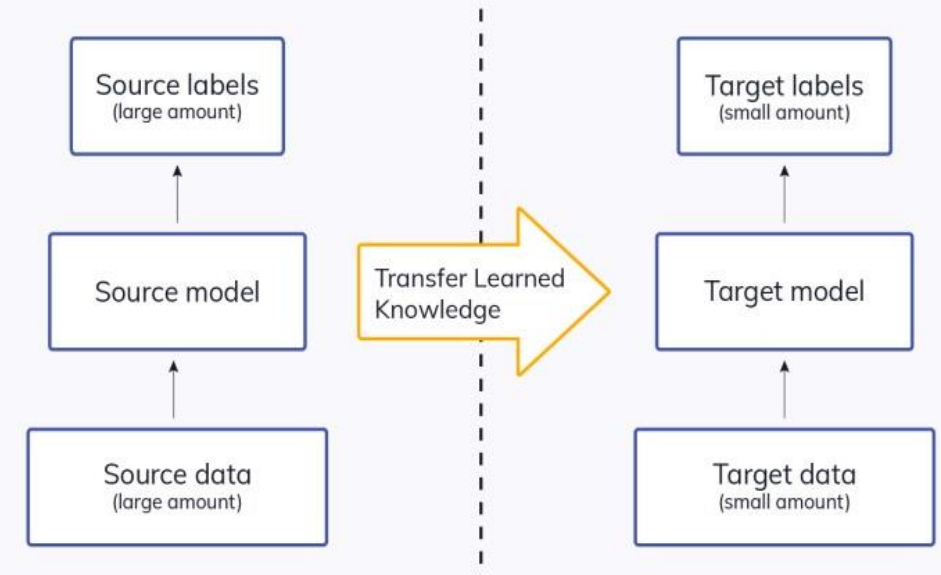
Thank you for listening

Any questions?

Imperial College London

Andy Sungjun Han

Supervisor: Dr Jesus Rodriguez-Manzano

2nd Marker: Professor Timothy Constandinou