DEVELOPMENT OF AN AI-BASED FRAMEWORK FOR PATHOGEN CLASSIFICATION THROUGH ANALYSIS OF RT-PCR DATA

A Major Project Report Submitted To The Department Of Computer Applications, Bharathiar University In Partial Fulfilment Of The Requirements For The Award Of The Degree Of,

MASTER OF SCIENCE IN DATA ANALYTICS

Submitted By,

CHANDRU G
(REG.NO: 22CSEG03)

Under the Guidance of,

Prof. Dr. V. BHUVANESWARI, M.C.A., M.Phil., Ph.D.,

Department of Computer Applications.

DEPARTMENT OF COMPUTER APPLICATIONS
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
BHARATHIAR UNIVERSITY
COIMBATORE-641046
TAMIL NADU

APRIL – 2024

CERTIFICATE

CERTIFICATE

This is to certify that the project titled **"DEVELOPMENT OF AN AI-BASED FRAMEWORK FOR PATHOGEN CLASSIFICATION THROUGH ANALYSIS OF RT-PCR DATA."** submitted to Bharathiar University in partial fulfilment of the requirement for the award of the degree of the **Master of Science in Data Analytics** is a record of the originalwork done by **CHANDRU G** under my supervision and guidance and this project work has not formed the basis for the award of any Degree/Diploma/Associate ship/Fellowship or similar title to any candidate of any University.

Place: Coimbatore

Date:

Project Guide

Head of the Department

Submitted for the University Viva-voice Examination held on _____

Internal Examiner

External Examiner

Microbiological Laboratory Research and Services India Private Limited

(An ISO 13485:2016 Certified Company)
0422-2425312/8098701010
E-mail: sales@microserv.in
Webpage: www.microserv.in

21nd April 2024,
Coimbatore

To Whom it may concern

This is to certify that **Mr. Chandru** has been our internship participant from 23rd January 2024 to 21nd April 2024. During this period, he has worked on developing **" Development of An AI-Based Framework For Pathogen Classification Through Analysis Of RT-PCR Data.".** This project aimed to develop software capable of interpreting molecular assays for diagnostic purposes, representing the first approach of its kind. Due to the interdisciplinary nature of the project, Mr. Chandru G as a Data science specialist had to collaborate with clinicians and molecular biologists from different fields.

As required by the project, Mr. Chandru G also demonstrated a strong interest in learning rt-PCR and related analysis protocols, which provided the foundation for the development ofthe software for automated interpretation of molecular assays.

In summary, we would like to thank Mr. Chandru G for his contributions during his internshipat Microbiological Laboratory Research and Services India Private Limited. We wish him the best for his future endeavours.

Regards, Dr. Rohit Radhakrishnan
Ph.D. Director (Research and operations)

arch and Operatio

DECLARATION

DECLARATION

I hereby declare that this project work title **"DEVELOPMENT OF AN AI-BASED FRAMEWORK FOR PATHOGEN CLASSIFICATION THROUGH ANALYSIS OF RT-PCR DATA"** submitted to Department of Computer Applications, Bharathiar University is arecord of original work done by **CHANDRU G** under the supervision and guidance of **Prof. Dr. V. BHUVANESWARI, M.C.A., M.Phil.,** Ph.D., Professor, Department of Computer Applications, Bharathiar University and that this project work has not formed the basis for the award of any Degree/ Diploma/ Associateship/ Fellowship or similar title to any candidate of any University.

Place: Coimbatore Signature of the candidate

Date:

COUNTERSIGNED BY

ACKNOWLEDGEMENT

ACKNOWLEDGEMENT

Union is strength. It gives me great pleasure to acknowledge with gratitude the personalities without whose help the completion of this project work would not have been possible.

I express my respectful thanks to **Dr. M. PUNITHAVALLI, M.Sc., M.Phil., Ph.D.,** Professor and Head, Department of Computer Applications, Bharathiar University, Coimbatore, for permitting me to carry out my project work.

I consider it a special privilege to convey my prodigious and everlasting thanks to my guide **Prof. Dr. V. BHUVANESWARI, M.C.A., M.Phil., Ph.D.,** Professor, Department of Computer Applications, Bharathiar University, for his valuable guidance and suggestions for this project work.

I also extend my sincere thanks to **Mr. S. PALANISAMY, MCA., M.Phil.,** Assistant Professor, Department of Computer Applications, Bharathiar University, for her valuable guidance and suggestions for this project work.

Additionally, I extend my sincere thanks to **Dr. ROHIT RADHAKRISHNAN, Ph.D.,** Director (Research and Operations), Microbiological Laboratory Research and Services (I) PVT LTD, Coimbatore, for providing the opportunity to work on their R&D project.

Finally, I express my thanks to my dear parents and my dear friends for their support and encouragement for the successful completion of this project. I am highly obliged to those who have helped me directly and indirectly in making this project a successful one.

INTRODUCTION

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Laboratory Information Management Systems (LIMS) transformed conventional laboratory operations into digitally-enabled infrastructure operations to attain high productivity and efficiency. LIMS aid a clinical laboratory to create an ecosystem for automating workflows, integrate instruments, manage samples, data management, real-time collaboration, perform data analytics, check quality control and patient reporting in a secured, user-friendly and polarized environment (Fig. 1). Thus, the software is crucial not only in a clinical lab but also in wide laboratories ranging from academic research, chemical labs, and manufacturing to agricultural testing, forensics, etc.,

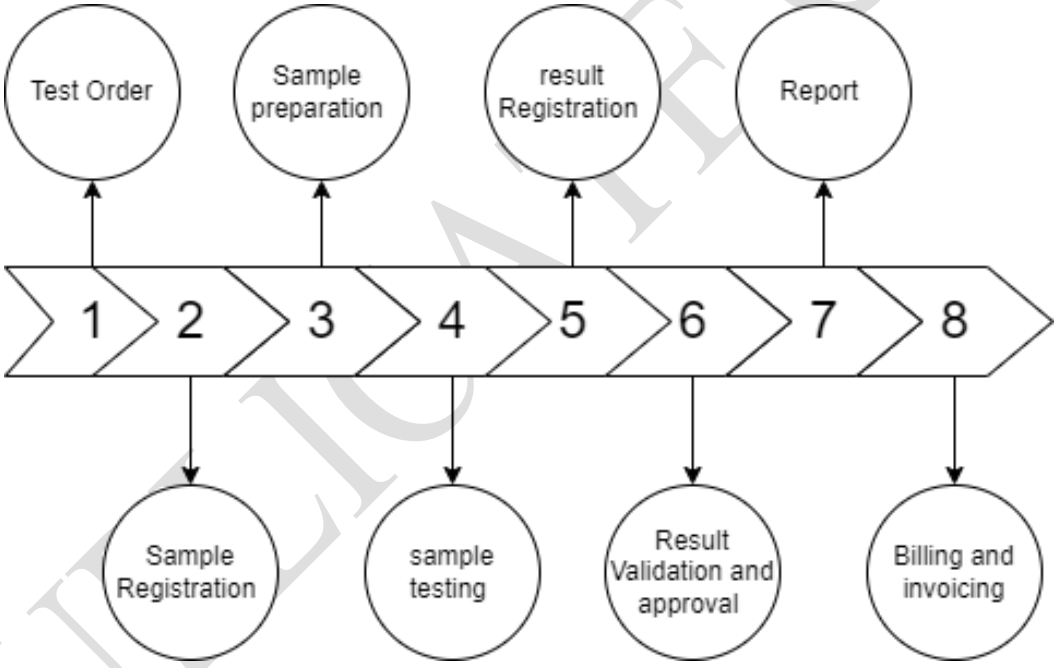


Figure 1: Capabilities of Laboratory Information Management System

Clinical laboratories are healthcare institutions that offer a variety of techniques to aid physicians with patient diagnostics, care, and management run by laboratory scientists. With the rapid advancements in hardware and software technology, all the specialized clinical laboratories are modernized with state-of-the-art laboratory machines and instruments for

testing and diagnosing with high quality and accurate results, therefrom quick analysis and reports made in respective software.

Shirts et al., stated the importance of analytics in clinical laboratories as 'Clinical laboratory analytics is the systematic evaluation and communication of clinical laboratory testing data to improve healthcare operations and patient outcomes. In looking at the system where the *analytics* part is the most demanding and includes complicated methods, such as analyzing and interpreting the test results and checking with quality control data to monitor the instrument's performance and accuracy. Along with these tasks, there are some limitations of data acquisition and integration phases in the analytics pipeline to integrateor acquire several other instruments with the respective data formats of their respective manufacturers' software/plugins in a general LIMS, taking into account large-sized labs. However, there are a few manufacturers/vendors that address this shortage by providing a solution for configuring instruments in the workflow, but this software might be expensive for mid to small-sized laboratories, or custom-built software can be a solution.

Clinical laboratory analytics should focus on improving*decision support* (i.e., the use of tools or systems to provide clinicians with relevant information, recommendations, and guidelines at the point of care when ordering and interpreting laboratory tests) during test ordering and result interpretation. This approach requires developing a strong decision support infrastructure that embodies both rule-based and*machine learning*-based algorithms into the clinical workflow. In addition to that the importance of using "offline" clinical laboratory analytics to analyze and enhance test utilization (e.g., identifying variations in test ordering patterns between clinicians that cannot be explained by clinical factors) was also discussed.

Most of the third-party vendors provide generic LIMS, in which the analytics pipeline involves transferring data from the LIMS to the different environments/platforms for analyzing the data, thus making a quite complex situation for the technicians and clinicians to perform analysis and interpretation in intermediate-level and national reference labs with faster deliverance and accurate results, where they own different kinds of instruments.Besides, mid to large-sized intermediate-level labs have laborious and time-consuming processes for doing analytics without LIMS. In concern to the lab budget, the specific or generic LIMS developed by various PCR manufacturers or software vendors may charge additional fees for updates or require ongoing maintenance for continued support and updates.

Molecular diagnostics is a high complexity clinical laboratory, which amplifies the genetic level (DNA or RNA) of cells or pathogens to detect mutations, gene expression, or infectious agents at the molecular level using PCR. Quantitative polymerase chainreaction (i.e., real-time PCR or rt-PCR) data analysis is a highly significant process that includes many primarily different techniques such as experiment setup, data processing, normalization, amplification analysis, efficiency and performance of PCR reaction, visualization of results, this technique allows DNA amplification in real-time accumulation of fluorescence in reaction. High-Resolution Melt Analysis (HRMA) is an advanced technique of conventional Melt Curve Analysis (MCA) with a rt-PCR instrument or its specialized instrument to identifythe melting temperatures (T_m) of DNA, though melt curve analysis also gives reliable results however, HRMA gives more accurate results compared to MCA. These techniques are doneby the various commercialized PCR manufacturers' data analysis software/plugins. Though stipulated statistical analyses and mathematical algorithms which can be done easily by clinicians with this software then, after the analyses part, information and insights gained finally lead to interpreting the results require certain expertise in the field.

Wrong interpretations and analysis of PCR test data and post-PCR data, (i.e., DNA melting curve, melt peaks, and HRMA curve) lead to serious consequences both for the patients affected and for the laboratory itself. However, the interpretation of PCR test result data is manually done by the clinicians/microbiologists with the domain knowledge (e.g., DNA melting, amplification of DNA, high-resolution melt analysis, characteristics of molecular pathogens, and thermodynamics of PCR) and other major parameters (such as., components added in the PCR compound, primers, and so on) by visual inspections and gain detailed insights from the analysis software. This visual interpretation highly requires intense laborious and time-consuming processes for novice clinicians/researchers and for complex case data, hence it impacts the right time for results to be delivered.

Several clinical laboratories own different PCR manufacturers' instruments and their respective software/plugins for their unique features and accurate results to obtain and interpret the data is challenging for mid to large-sized intermediate-level labs, which run numerous PCR experiments on a day-to-day basis. In addition to that, various democratized software/plugins/web-based applications are available, and each of them has its advantages and limitations for doing analytics after the experiment is done but, this might or might not give accurate results, according to different types of analysis and techniques used in the software.

As defined by Shirts *et al*., this project aims to improve microbiologists'/clinicians' decision support and assist them by leveraging Artificial Intelligence and Machine Learning with the advent of HRMA data during the conventional analytical process to improve the performance of the interpretation of results. Additionally, to address the absence in theanalytics phase by developing an automated application to aid laboratorians/clinicians.

1.1 ORGANIZATION PROFILE

The **Microbiological Laboratory Research and Services India Private Limited (Microserv)** is an ISO 13485:2016 certified medical manufacturing facility and a research institute. Microserv develops and produces diagnostic kits such as ready-to-use microbiological culture media for clinical and industrial use, molecular assay reagents and MLRS-STaTAST. MLRS-STaTAST is a novel patented antibiotic sensitivity testing technology jointly developed with Anna University. Microserv also provides training in molecular diagnosis jointly with Bharathiar University.

The **Microbiological Laboratory, Coimbatore** is a leading NABL-accredited clinical laboratory in India. It is the first clinical laboratory which has several molecular assays under the NABL scope since 2007. Microbiological Laboratory has developed and patented HRMA based molecular assay for infectious diseases which is currently being used for patient diagnosis. Microbiological Laboratory operates over 50 branches throughout India that are connected to a central server system, enabling the consistent delivery of high-quality reports across all locations.

1.2 PROBLEM DEFINITION

High-Resolution Melting Analysis (HRMA) involves monitoring the disassociation characteristics of double-stranded DNA during denaturation heating. Mutations andsequence variations in the DNA cause changes in the melting temperature and curve shape,allowing for sensitive and rapid detection of genetic variations without the need for expensive

probes or post-PCR processing. HRMA is aided by commercially available thermal-cycler (PCR) machines and respective analysis plugins/software for generating the melting curve graphs based on the raw fluorescence data.

VISUAL INTERPRETATION

The interpretation of HRM data is crucial and it requires a clear understanding of the melting temperatures of every DNA target. HRMA software provide visualization (graphs) of melting signals against temperature, which comprises both perfect and imperfect (noisy) signals. The result interpretation usually involves visual observations and analysis by technicians. Experts who perform interpretation must focus on removing such noisy signals from their analysis by following some thresholding metrics. As the number of samples scales up, the interpretation also requires scaling, and doing it manually is challenging and time-consuming. Typically, experts would have much experience, and the perception they have in interpretation is huge and impeccable. In practice, experts alone cannot perform interpretation at all times, and several other beginners and junior technicians are also often required to perform interpretation, considering the productivity.

VERSATILE SOFTWARE

Various PCR instrument manufacturers have their proprietary data analysis software and algorithms, which calculate and process the HRMA data with stipulated steps. Hence, different software can produce different melting curves for the same sample in HRMA analysis. This can occur due to differences in the algorithms used for data analysis and curve fitting, variations in the baseline correction and normalization methods, and other factors related to data processing and interpretation. To ensure accurate and reliable results, it is important to use standardized melting curve plotting protocol and software validated for the specific HRMA application.

PREDICTIVE ANALYSIS

Most of the analysis software of HRMA uses the fluorescence response vs temperature data for the samples tested to plot the melting curve. The software uses statistical techniques and mathematical algorithms to analyze the raw fluorescence signals to melt signals. This software has various features such as identifying mutations of the pathogens, genotyping and detecting SNPs. Predictive analysis techniques are unavailable in this software for studying the unique features (e.g., Melting temperature (T_m), melting peak height, curve shape, curve width,

inflection point, and area under the curve) of melt signals which can be used for creating digital signatures unique for each target.

1.3 BACKGROUND AND NEED

The analysis, interpretation, and reporting of HRMA in the context of enhancing the decision support of the clinicians with the interoperable customized predictive analysis and reporting will decrease the dependency on laborious visual interpretation of such complex data. With the existing software available, such have their limitations of producing melt curve with the standard statistical and mathematical methods, and to the extent there are few commercial software which utilizes some of the machine learning algorithms such as principal component analysis and k-means for dimensionality reduction and clustering the data for other applications such as genotyping and mutation scanning. However, this software is outdated with support only for Windows 7 platform.

Currently, HRMA data analysis and interpretation require technical expertise, which can result in variability and errors in the results. An AI-based framework can standardize the interpretation and analysis of HRMA data, leading to more accurate and reliable results. This framework can provide automated data management, making the process more efficient and less time-consuming. With the integration of machine learning algorithms, the framework can learn from past data and adapt to new data sets, improving its accuracy and efficiency over time. The AI-based framework can also reduce human error and increase the speed of analysis, making it possible to analyze more clinical samples in a shorter time frame. The predictive analysis of HRMA data allows for rapid identification of the pathogen causing the disease.The implementation of an AI-based framework for HRMA data management, interpretation, and reporting can also facilitate the sharing of data between laboratories and clinics, improving collaboration and accelerating the development of new diagnostic tools. Theframework can also be used to track the evolution of genetic variations in pathogens, enablingearly detection of emerging pathogens and their drug resistance patterns.

Overall, an AI-based framework for HRMA data management, interpretation, and reporting has the potential to revolutionize the clinical diagnosis of infectious diseases, making it more accurate, reliable, and efficient. It can provide a standardized approach to HRMA data analysis, allowing for better comparison of results across different laboratories, and can ultimately lead to the development of more effective diagnostic tools and treatment strategies.

1.4 PURPOSE OF THE STUDY

The purpose of the study aims to develop and implement an AI framework to analyze and interpret High-Resolution DNA melt data with faster deliverance of accurate results and reports to aid clinicians/laboratorians in an intermediate-level laboratory.

Interpretation of HRMA data is handled by clinicians with keen visual observations and inspections with high domain expertise, which is a time-consuming and laborious process, and the lack of data acquisition and extraction pipeline makes a bit tangled situation at intermediate-level and national reference laboratories. These factors influence the speed and accuracy of deliverance and providing responsive reports to physicians and patients. Harnessing state-of-the-art AI and ML techniques and algorithms to analyze, interpret, and report without human intervention in a web-based platform.

To encounter the interpretation of HRMA data in concern with rtPCR data analysis techniques, the team explored the HRMA data of various pathogens by performing pre-processing, and feature engineering, with appropriate statistical analyses to gain insights. Implemented some of the existing methodologies beginning from the bare method of examining the images of DNA melt signal, to the approach of unravelling the co-ordinates of raw-fluorescence signal, DNA melt signal, and lastly the combination of images and co- ordinates of DNA melt signals. Upon the researched methodologies, ultimately the team devised a feasible and practical solution of developing a Python-based library with custom- trained Deep Learning models for the interpretation, and finally measured and validated the results with the expert clinicians. Along with these research methods, the team developed automated software for data extraction from the PCR data analysis plugin/software.

The goal of the study is to implement the web-based AI framework for analysis, interpretation, and reporting to assist technicians and clinicians. Another goal of the study is to develop automated software for data extraction.

1.5 DEFINITIONS

This section provides a major list of concepts in detail such as PCR reaction, the processes behind the reaction, DNA melting behavior, amplification analysis, and so on.

1.5.1 POLYMERASE CHAIN REACTION

The Polymerase Chain Reaction (PCR) is a molecular technique used for DNA quantification, biomarker identification, genotyping, and mutation detection. The technique is based on the amplification of a specific segment of DNA into several copies, using a DNA polymerase enzyme (Fig. 2). PCR involves using short synthetic DNA fragments called *primers* designed based on sequences specific to each target. The segment of the DNA complementing the sequences of the primer will be amplified for multiple PCR cycles until it reaches the limit of detection.

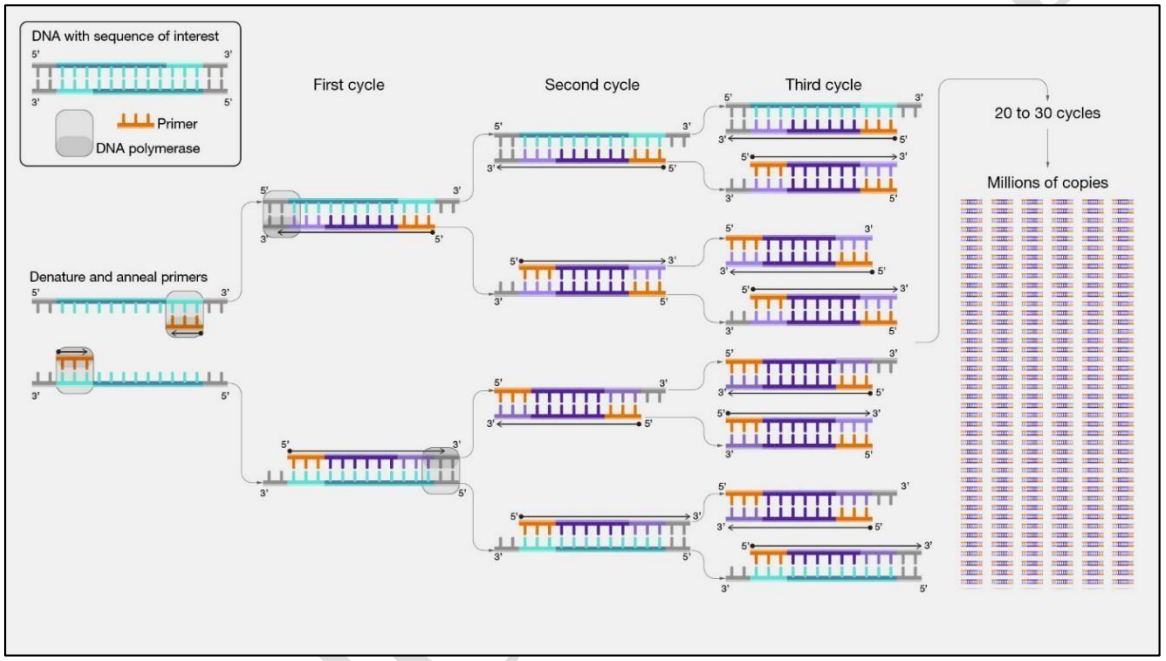


Figure 2: Amplification of DNA segments
Source: National Human Genome Research Institute

Traditional PCR demand that the product be examined following the completion of the reaction, this procedure is frequently referred to as *"end point"* analysis [7]. Due to the improvements in hardware and software, real-time PCR (also known as quantitative PCR or qPCR or rt-PCR) occurred as a new variation that constantly accumulates fluorescent signals from several polymerase reactions and permits to detect the DNA amplification at the right moment.

Real-time PCR uses commercially available fluorescence-detecting thermocyclers to amplify specific nucleic-acid sequences tagged with different types of fluorescent dyes (probes and SYBR™ green dye) and measure their concentration simultaneously. Target sequences are amplified and quantified simultaneously in the same PCR machine. Hence, the PCR amplification of the target sequence can be monitored in real-time thus eliminating

quantification steps such as agarose gel electrophoresis. PCR was widely used during therecent COVID-19 outbreak to manage the epidemic across the world and remains the gold- standard method for COVID-19 diagnosis. The COVID-19 diagnosis is the latest application that has popularized this qPCR technique in recent times, and the application of PCR in the diagnosis of pathogens (targets) such as bacteria, viruses, fungi, and other non-culture biomarkers has been available for several years.

1.5.2 DNA MELTING

The dissolution of the double-stranded DNA (dsDNA) helix into single coils is referred to as DNA melting. It can be accomplished by simply heating double-stranded DNA. The temperature at which the DNA strands dissociate into single coils depends on the number of hydrogen bonds holding the complementary strands. The most commonly used method to determine the melting temperature of a PCR product is to subject the product to a temperature gradient in the presence of intercalating dye. The intercalating dyes are chemicals that only emit light when bound to double-stranded DNA.

In a typical melting experiment, a PCR product is mixed with an intercalating dye, and fluorescence emitted by this mix is monitored as the sample is slowly heated (subjected to a temperature gradient). The outcome of the analysis is a curve displaying fluorescence changes emitted by the sample over the range of temperatures that the sample was subjected to, commonly referred to as a melting profile (Fig. 3).

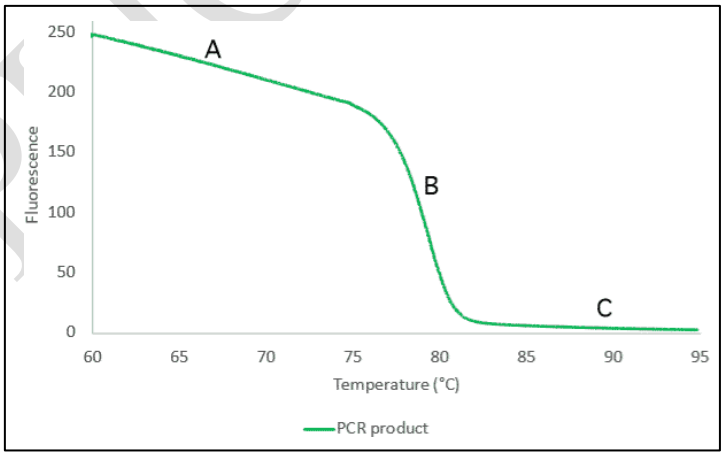


Figure 3: Melting profile of a PCR product

At the beginning of the experiment, the temperature is low and all PCR product in the sample is double-stranded. Thus, the fluorescence level is high in the sample (Fig. 3-A). Observing high levels of fluorescence, as the temperature increases up to the point, where all

hydrogen bonds within the PCR fragment are broken and the amount of double-stranded PCR product drastically decreases. Consequently, a sharp decrease in the detected fluorescence level (Fig. 3-B). At a high temperature, there is no double-stranded PCR product in the sample and the fluorescence levels are close to 0 (Fig. 3-C). The temperature at which the sharp drop in the fluorescence depends on the number of hydrogen bonds in the analyzed PCR product and hence is specific to the analyzed fragment.

1.5.3 MELT CURVE ANALYSIS

The Melt Curve is derived from the raw fluorescence data, by getting the first negative derivative ($-\ dF/dT$) of Fluorescence intensity and Temperature (fig. 4). In Melt curve analysis, the data comes as a result of HRM (in the case of specialized instrument used) being analyzed further, to determine the melting characteristics of several DNA in a more precise way. In this stage, the derivative of fluorescence intensity captured in real-time will be plotted against the temperature so, the temperature at which the dsDNA began to denature into ssDNA, the point at the melt peaks which resemble the melting temperature of the DNA. A threshold is manually set by the clinicians through keen visual inspections and observations with the helpof commercial PCR manufacturers' analysis software.

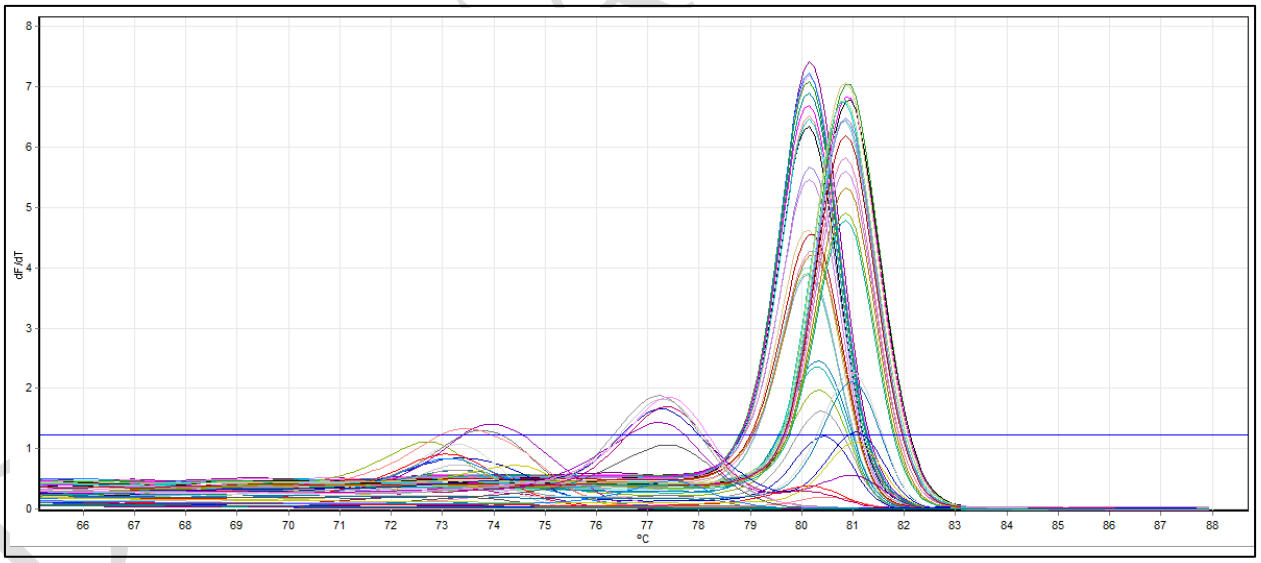


Figure 4: Negative derivative plot of the Melting curve

Source: QIAGEN's Q-Rex Software

In looking at Fig. 4, the characteristics of the Melt curves are observed and studied by concerning various features like:

- Peaks
- Shape of the curve
- Height
- Range
- Area under the curve

1.5.4 AMPLIFICATION CURVE ANALYSIS

The Amplification Curves are also known as 'growth curves' that display the graph of Cycle number vs Fluorescence (fig. 5), these data from real-time PCR are used to detect the presence of the PCR product (i.e., target DNA) and a threshold (C_t) line is to be set in between the exponential and linear phase, to identify which PCR product is amplified earlier in the PCR reaction cycles.

The expression levels of genes can be measured by either absolute or relative quantification. In absolute quantification, a calibration curve is used to relate the PCR signal to the input copy number, while relative quantification measures the relative change in mRNA expression levels. The accuracy of an absolute real-time rtPCR assay depends on the identical amplification efficiencies of both the native target and the calibration curve in the RT reaction and kinetic PCR. Relative quantification is a simpler method compared to absolute quantification since it does not require a calibration curve. It involves comparing the expression levels of a target gene to a reference gene and is sufficient for most investigations into changes n gene expression. The units used for relative quantification are unimportant and can be compared across multiple real-time RT-PCR experiments.

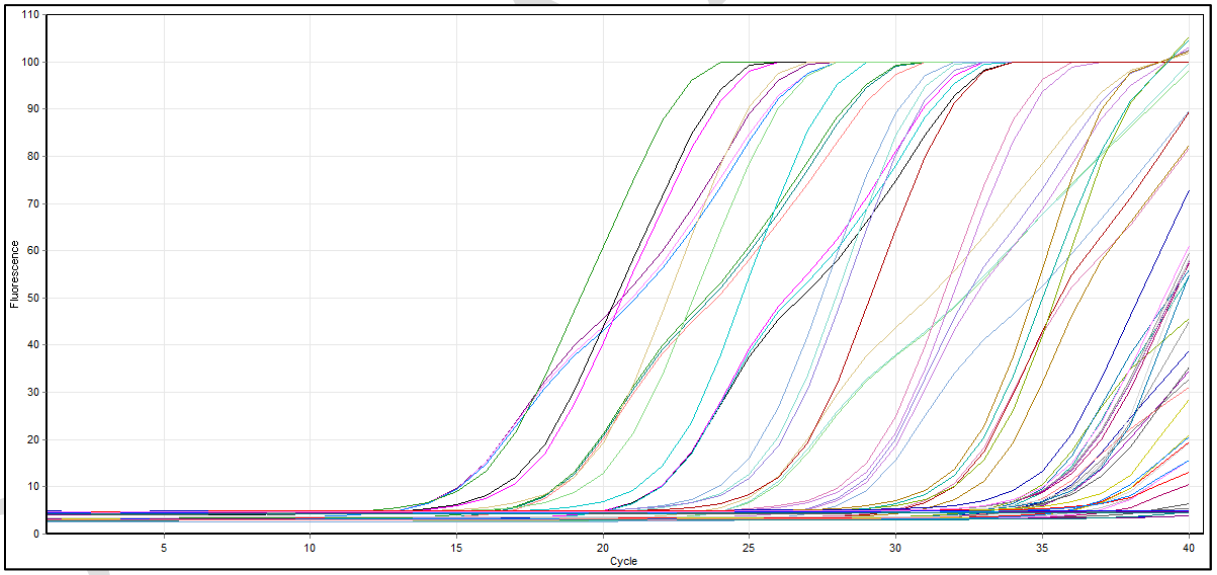


Figure 5: PCR Amplification Cure
Source: QIAGEN's Q-Rex Software

The three phases of PCR:

- Exponential phase
- Linear phase
- Plateau phase

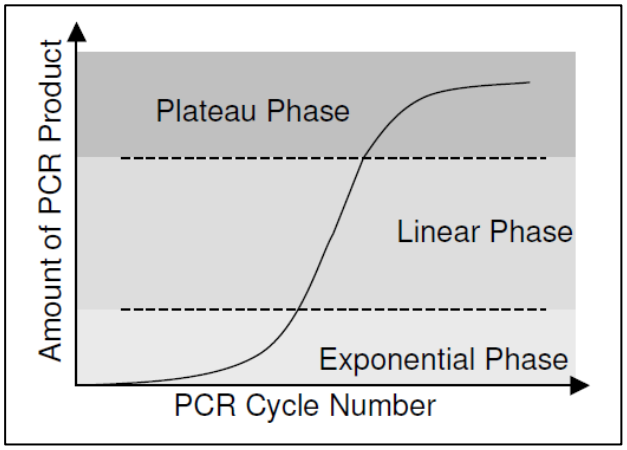


Figure 6: Theoretical plot of PCR, three phases

The PCR will eventually reach the plateau phase during later cycles and the amount of product will not change because some reagents become depleted. The exponential phase is the earliest segment of the PCR, in which the product increases exponentially because the reagents are not limited. the linear phase is characterized by a linear increase in the product as PCR reagents become limited. the PCR will eventually reach the plateau phase during later cycles.

1.5.5 HIGH-RESOLUTION DNA MELT ANALYSIS

A novel DNA analysis technique called High-Resolution Melting (HRM/HRMA) is a significant method that was developed in 2002 through a collaborative effort between the University of Utah, USA, and Idaho Technology Inc., USA. for analyzing geneticvariations such as SNPs (single nucleotide polymorphisms), mutations, and methylations in PCR amplicons. It is a homogeneous, close-tube, post-PCR technique that allows researchers to study the thermal denaturation of double-stranded DNA in greater detail than *traditional melting curve analysis*, resulting in higher information yield, with the advanced hardware.

By analyzing the disassociation (melting) behavior of nucleic acid samples, HRMA can differentiate between samples based on their sequence, length, guanine-cytosine (GC) content, or strand complementarity and it can even detect single base changes like SNPs. It is a powerful tool that enables the detection of unknown variations in PCR amplicons, making it a valuable alternative to sequencing and the range of applications including:

- Mutation discovery
- Screening for loss of heterozygosity
- DNA fingerprinting
- SNP genotyping
- DNA methylation analysis

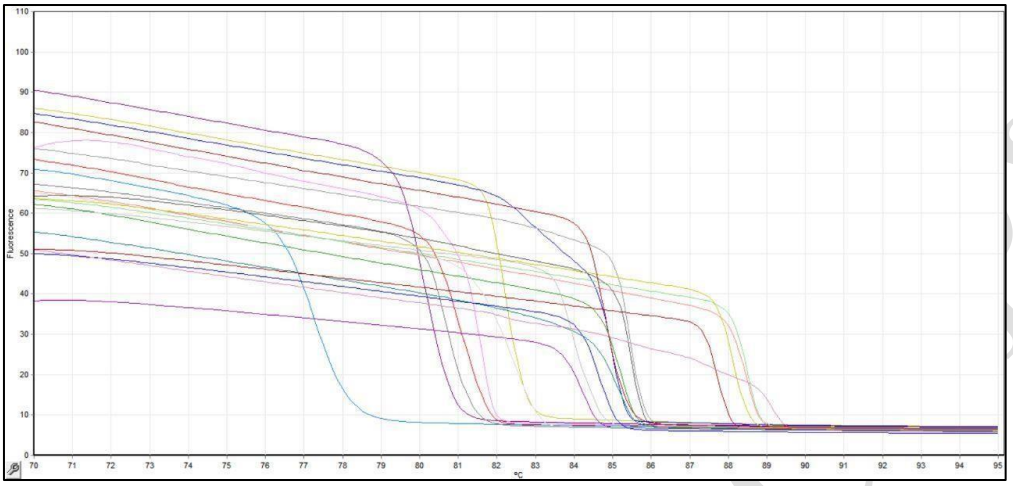


Figure 7: High-Resolution Melt Curve
Source: QIAGEN's Q-Rex Software

1.5.6 OVERVIEW OF DNA MELT SIGNAL INTERPRETATION

DNA melt signals are the important output of PCR experiments, as they provide information about the characteristics of the amplified DNA. As the temperature increases, the double-stranded DNA begins to denature into single-strand, and the DNA-binding dye will dissociate from the DNA, causing a decrease in fluorescence. The temperature at which half of the double-stranded DNA is denatured is called the melting temperature (T_m).

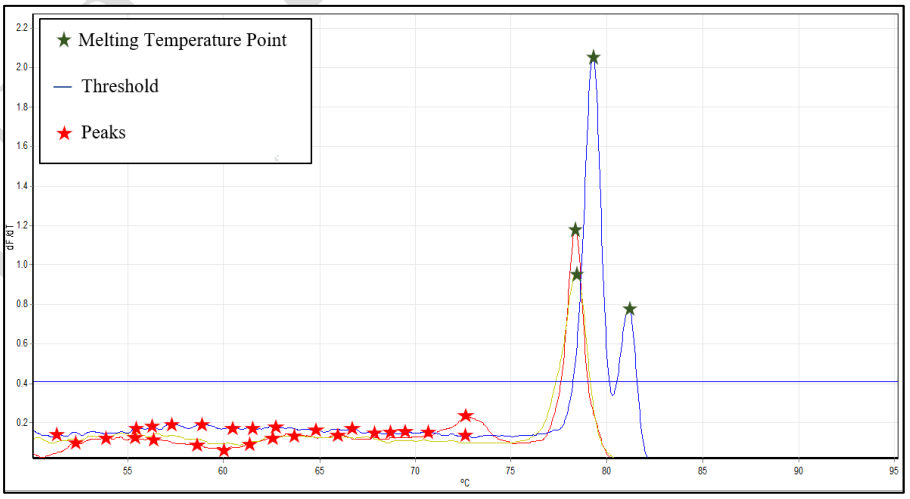


Figure 8: Reference DNA melting signal – Meningitis Panel

Each pathogen will be having different DNA melting temperatures and their respective signals will also possess different shapes and sizes. Usually, DNA melting signals are in bell-

shaped curve with peaks, which denotes the melting point or the melting temperature of the DNA. Interpretation will be made through visual inspection, performed on such signal's shape, peaks, and size. Fig. 8 shows the DNA melting signals of **Haemophilus influenzae (HI), Streptococcus pneumoniae (SP), and Neisseria meningitidis (NM).**

No.	Pathogen Name	Temperature of melt-Observed Tm	Temperature of melt-Expected Tm	Cycle of Threshold	Cycle of Threshold	Threshold
1	Haemophilus influenzae	78.35 ℃	77±1 ℃	27.25	26	0.4
2	Streptococcus pneumoniae	78.40 ℃	78±1 ℃	24.88	25	0.4
3	Neisseria meningitidis	79.27 ℃ /81.13 ℃	79/81±1 ℃	19.52	2	0.4

Table 1: DNA melting temperature standards for Meningitis Panel of the positive sample [fig. 8]

A. Melting Temperature

Pathogen	Target T_m
Acinetobacter baumanii	81ºC
Bacteriode fragillis	80 ºC
CONS-coagulase-negative *Staphylococci*.	80 ºC
Enterobacter spp.	84 ºC
Enterococccus feacalis	82 ºC
Enterococccus spp.	84 ºC
Group A *Streptococcus*-GAS	82 ºC
Serratia macessens	85 ºC
Staphylococcus spp. (Gram-positive)	77 ºC
Streptococcus agalactiae-GBSC	77 ºC
Streptococcus pneumoniae (Gram-positive)	78 ºC

Table 2: Sample DNA melting temperature standards for Sepsis Panel

There is a set of pre-defined DNA melting temperature standards recorded by clinicians and microbiologists for identifying and distinguishing melt signals during the interpretation process. Peaks found from such melt signals (Fig. 8) for a proposed pathogen, that seemed to be satisfying (±1) the standard, will be treated as 'Positive', and if not, will be treated as 'Negative' and vice-versa.

B. Thresholding

Thresholding, on the other hand, plays a crucial role in this process, which is being set manually during the analysis, for eliminating noisy signals and unwanted peaks. It is a simple numerical figure on the *y-axis* (i.e., derivative of fluorescence over temperature), where only those peaks will be considered on or above such numerical figure (Fig. 8 and Table 1).

Apart from Melting temperature and threshold, there are several other parameters that must also be put into consideration for interpreting signals, and they are,

- Temperature point at which the signal start rising.
- Temperature point at which the signal falls/saturated.
- Prominence of the signal.
- Area under the curve.

Such attributes of DNA melt signals will be considered as features, and relevant feature engineering techniques must be employed to bring the best out of them. The techniques and methodologies are briefly elaborated in the upcoming sections.

1.6 ROLE OF DATA IN REAL-TIME PCR

The post-PCR methods such as Melt Curve Analysis, Amplification Curve Analysis and HRMA are undergone using rt-PCR machines, followed by the PCR experiment using thermal-cycler machines. In common, most post-PCR data are collectively known as DNA melting curves, amplification curves, and HRMA.

- **Amplification curves:** PCR generates amplification curves that show the increase in fluorescence signal over time, indicating the amount of amplified DNA. This data can be used to determine the starting amount of target DNA and to assess the efficiency and sensitivity of the PCR reaction.
- **Melting curves:** PCR melting curves show the dissociation of double-stranded DNA into single strands as the temperature is increased. These curves can be used to determine the melting temperature (T_m) of the amplified DNA, which can help to identify specific DNA sequences and to detect mutations.
- **HRMA data:** HRMA provides information on the melting behavior of PCR products, which can be used to identify and distinguish different PCR amplicons based on their melting temperature (T_m). HRMA data can be used to detect mutations, SNPs, and other sequence variations in DNA samples. It can also be used to evaluate PCR performance, including specificity and sensitivity, and to optimize PCR conditions.

The interpretation of rt-PCR results data includingvarious numerical data (other than melting curves, amplification curves and HRMA) that grant the assessment of various analytical parameters such as linearity, accuracy, precision, specificity, and so on to determine the rt-PCR instrument's efficiency, specificity and otherresults. From the post-PCR data, the HRMA data (i.e., melting curves) is promising to do analyses and interpretations about the specificity and identity of the amplified product (i.e., target DNA) by employing the pioneering ML techniques. Overall, HRMA data plays an important role in real-time PCR by providing valuable information on PCR products and helping to improve the accuracy and reliability of PCR-based assays.

1.7 PREDICTIVE ANALYSIS IN DIAGNOSIS

Diagnostic is itself an analysis of "What happened?", "Why happened?" and "Where happened?". There are a lot of advancements are been introduced day by day in healthcare and some of them already exist. In such a way, predictive analysis of diagnostic data is not a new approach. There are a lot of provisions and support were already been introduced to aid many researchers and organizations in boosting their routine work.

Some examples are,

- Predicting heart disease using electronic data, medical data, and patient information.
- Predicting various health complexities using medical images like X-rays, CT scans, and MRIs.
- Predicting Cancer with data on tumours (benign or malignant).

In the field of molecular diagnosis, predicting targets may involve considering several factors. Melt curve analysis is one of the steps in diagnosing with PCR and it gives information more on the melting nature of a dsDNA. Clinicians/Microbiologists will study the melt curves and observe the distinct variations in the graphs thus determine the presence of a specific target DNA. It is important to note that only melting analysis would not suffice to confirm any presence of target DNA (pathogen) in a patient sample, and additional analysis may require such as

- Sequencing
- Phylogenetic analysis
- Multiplex PCR

- Specific primer/probe design
- Culture and isolation
- Serology

As a result, this project will cover predictive analysis on PCR diagnosis data, specifically HRM data, that gives Melt curves and peaks. With relevant feature extraction and feature engineering, finally, predictions will be made on classifying pathogen classes using various predictive analysis techniques.

MACHINE LEARNING IN DIAGNOSIS

Machine Learning is a great choice of predictive modelling, and it is a robust technology, which has been globally applied in various applications ranging from classifying spam mail to predicting diseases. Due to its high compatibility and sound algorithmic resource, many real-world problems can be solved using Machine Learning and diagnosis is not an exception for applying it.

MACHINE LEARNING OVER STATISTICAL MODELLING

Both tend to use similar predictive approaches like regression, classification, and clustering, but they differ in many ways.

Statistical modelling is a subset of mathematical modelling where it hugely involves assumptions, relationships between random and non-random variables, and estimating population with sample data. Choosing statistical modelling as a predictive analysis technique will require more understanding of the variables involved in the data and the respective relationships between them. Once these perceptions are satisfied, a sensible assumption has to be made, to explain the relationships between variables and the resulting prediction. This is why statistical modelling is highly preferred when proper interpretation and explanations are demanded. On the other hand, Machine Learning is also another predictive analysis technique, which is a branch of computer science and artificial intelligence, that mainly works on the principle of pattern analysis and often introduces challenges in interpreting their learning pattern. Compared to statistical modeling ML models can work with large data sets, and it rejects the chances of making assumptions on the given data, as it learns from the pattern through a weight-based approach. As a result, predictions by these models are powerful and more accurate.

In the context of performing predictive analysis with HRM data, typically it's biological data which is complicated due to its complexity and high variability in nature because every experiment is done and influenced under the clinical environment. So, the Machine Learning approach is preferable over statistical modeling, owing to its pattern learning process which gives highly accurate results while statistical modeling is best if the characteristics of the data should not vary for formulating the hypothesis. Concurrently, several statistical techniques were also used to find insights during the development of this project. Besides both techniques have their advantages and limitations.

1.8 OVERALL RESEARCH AIM AND OBJECTIVES

The overall aim of the project is to create an AI-based framework for analyzing and interpreting the HRM data without involving any human assistance. The scope of the project starts from the necessary data extraction/acquisition to the end report presentation.

The objectives can be enumerated as:

- To setup data acquisition pipelines for extracting and acquiring all the data, necessary for the analysis and interpretation.
- To develop data pre-processing modules for cleaning and transforming acquired data.
- To conduct research on the domain and the given problem to formulate the best and most suitable solution.
- To conduct relative research on previously undergone research and solutions made for problems in the same regard.
- To formulate an effective approach to applying Machine Learning algorithms to the problem.
- To evaluate and validate results with domain expertise and look for further improvisation.
- To setup modules and components for effective data storing and accessing.
- To develop supporting software components to aid the workflow.
- To augment all the components into a single apex system.

CHAPTER 2

TRADITIONAL EXISTING SYSTEM

The High-Resolution Melting Analysis are being done with the help of specificsoftware and existing hardware components. In practice, without these components, HRM cannot be done. However, doing manually is more complex and prone to error. Since the term**"High-Resolution"** itself depicts the technology of capturing fluorescence in **"High- Resolution"** which is thereby demanding sophisticated and engineered technical components. There are already many commercially available instruments and plugin tools used for runningPCR tests, and most of them are engineered with cutting-edge technology.

2.1. DATA ANALYSIS SOFTWARE

2.1.1 QIAGEN's ROTOR-GENE

Rotor-Gene Q series are commercial thermal cycle instruments used in many laboratories for running PCR tests and analyzing the results using their respective versions of plugin software.

Rotor-Gene offers several individual components for analysis like,

- *Melt Analysis*
- *HRM Analysis*
- *Screen Clust Analysis*

The components come with various names and versions, and among them, QIAGEN's '**Rotor-Gene Q-Rex**' is a default software tool that comes with every Rotor-Gene instrument for analyzing the run files of completed PCR tests. Q-Rex offers both *Melt* and *HRM* analysis in a single package.

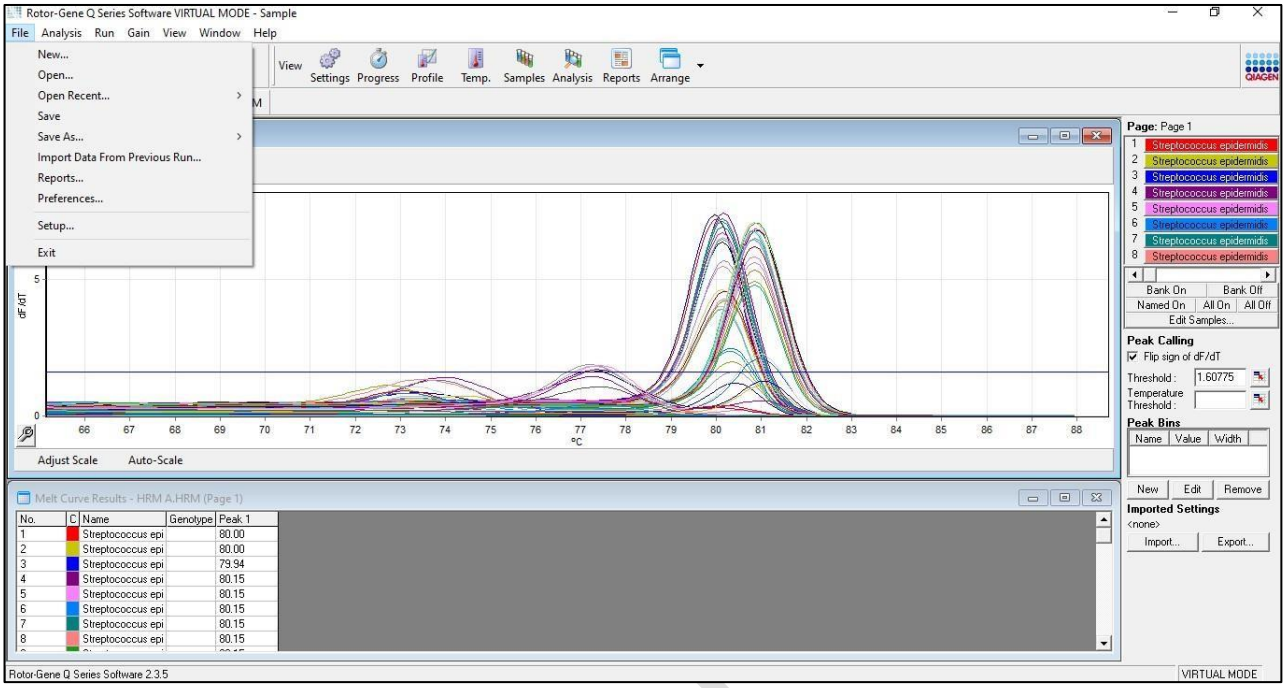
2.1.1 (A) ROTOR-GENE Q- SERIES SOFTWARE

MELT ANALYSIS

The "Melt Curve Analysis" function of the Rotor-Gene Q software can be used for checking the specificity of a reaction, genotyping, and measuring protein stability with differential scanning fluorimetry. The function analyses the first derivative (dF/dT) of the raw

melting data and identifies peaks in the selected temperature and fluorescence range. The data can be used for genotyping by defining "Peak Bins" based on peak characteristics of known genotypes.

A



B

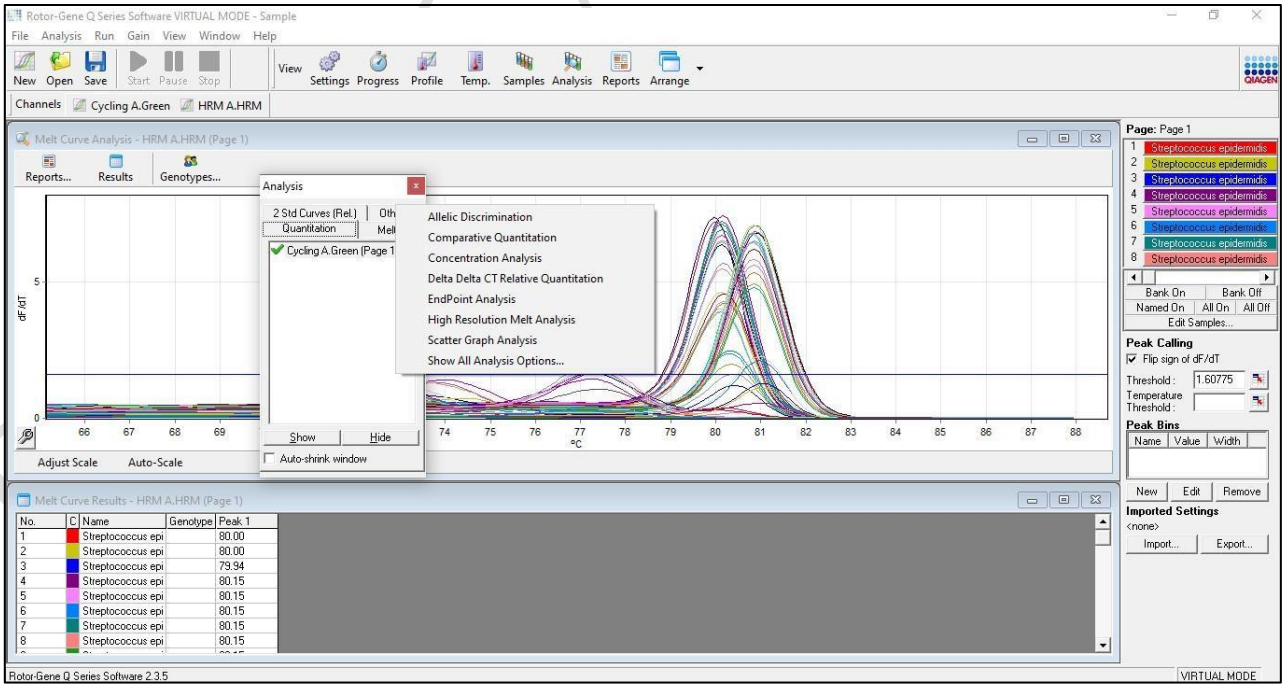


Fig 9, A and B: Rotor-Gene Q-Rex Interface
Source: QIAGEN's Q-Rex Software

HRM ANALYSIS

High-resolution melting (HRM) analysis using Rotor-Gene Q software involves selecting a data set and defining normalization regions to compensate for variations between samples. Genotype names are assigned, and control samples are selected. Results are displayed in the “HRM Results” table with automatic identification results for each genotype. Confidence levels are assigned to each sample, and a threshold value for the "Confidence Percentage" can be defined. The “HRM Normalized Graph” plot displays different curves relative to a selected genotype in a “Difference Graph” to emphasize differences between samples. Once the process is complete, tables and graphs can be exported.

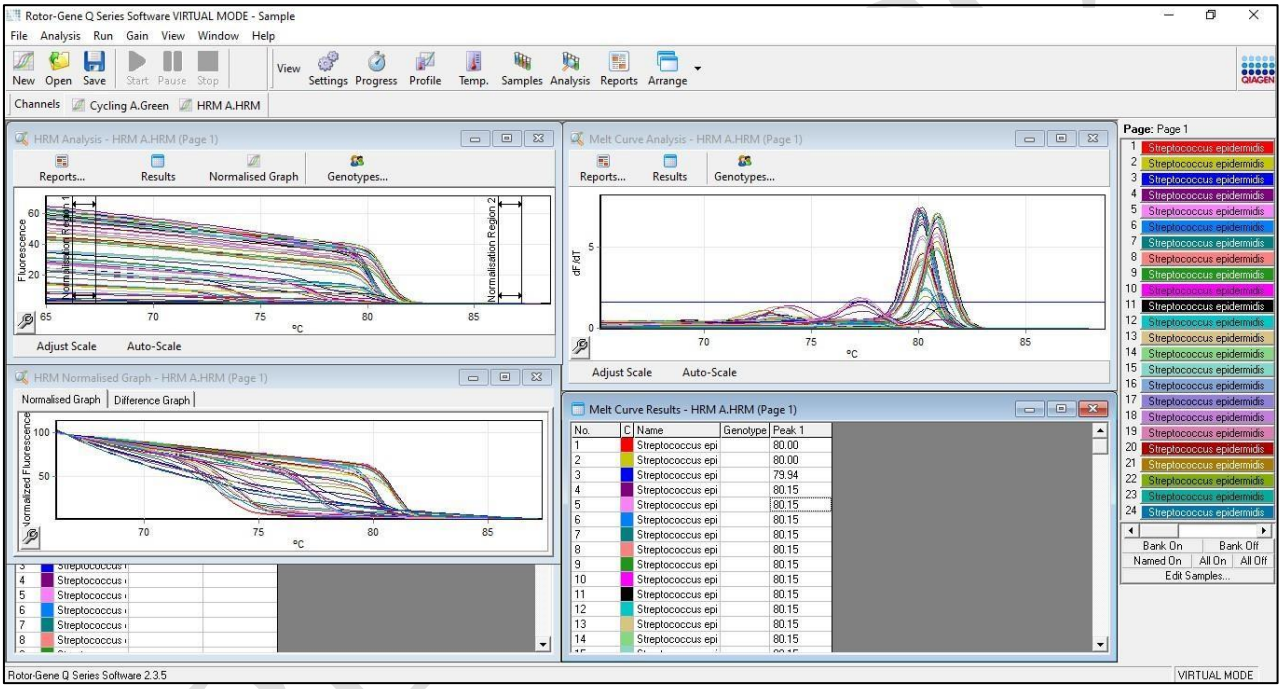


Fig 10: Rotor-Gene Q Rex Interface
Source: QIAGEN’s Q-Rex Software

2.1.2 (B) ROTOR-GENE SCREENCLUST HRM SOFTWARE

Rotor-Gene ScreenClust HRM Software is a powerful tool for the analysis of high-resolution melting (HRM) data from the Rotor-Gene Q or Rotor-Gene 6000 cyclers. By grouping samples into clusters based on their dissociation (melting) curve characteristics, Rotor-Gene ScreenClust HRM Software enables applications such as genotyping and mutation scanning. The number of clusters can either be defined by the user, if they have known controls

for each genotype (supervised mode) or the software can aid the user in determining the number of clusters in a sample set (unsupervised mode).

Rotor-Gene ScreenClust HRM Software provides:

- Innovative mathematical approach to HRM analysis.
- Highly accurate identification of genotypes in supervised mode.
- Automatic detection of new mutations in unsupervised mode.
- Robust statistics for classifying and interpreting HRM data.
- Minimal effort and standardized processes for data interpretation.

HRM analysis on a Rotor-Gene cycler produces raw data that can be further analyzed using Rotor-Gene ScreenClust HRM Software. Rotor-Gene ScreenClust HRM Software analyses HRM data in 4 steps:

1. Normalization
2. Generation of a residual plot
3. Principal component analysis
4. Clustering

HRM curves can have different starting points, therefore the scale of each melt is different Rotor-Gene ScreenClust HRM Software only compares samples that are on the same scale, which is achieved by normalization. Raw data are normalized by applying curve scaling to a line of best fit so that the highest fluorescence value is equal to 100and the lowest is equal to zero. Next, the curves are differentiated and a composite median curve is constructed using the median fluorescence of all samples. The melttraces for each sample are subtracted from this composite median curve to draw a residual plot. Consecutively, the individual sample characteristics are extracted by principal component analysis from the residual plot. The principal component analysis is a well- established method of data analysis.

However, Rotor-Gene ScreenClust HRM Software is the first software application to apply principal component analysis to HRM data. The principal component analysis highlights similarities and differences in the data and is used to create a cluster plot.

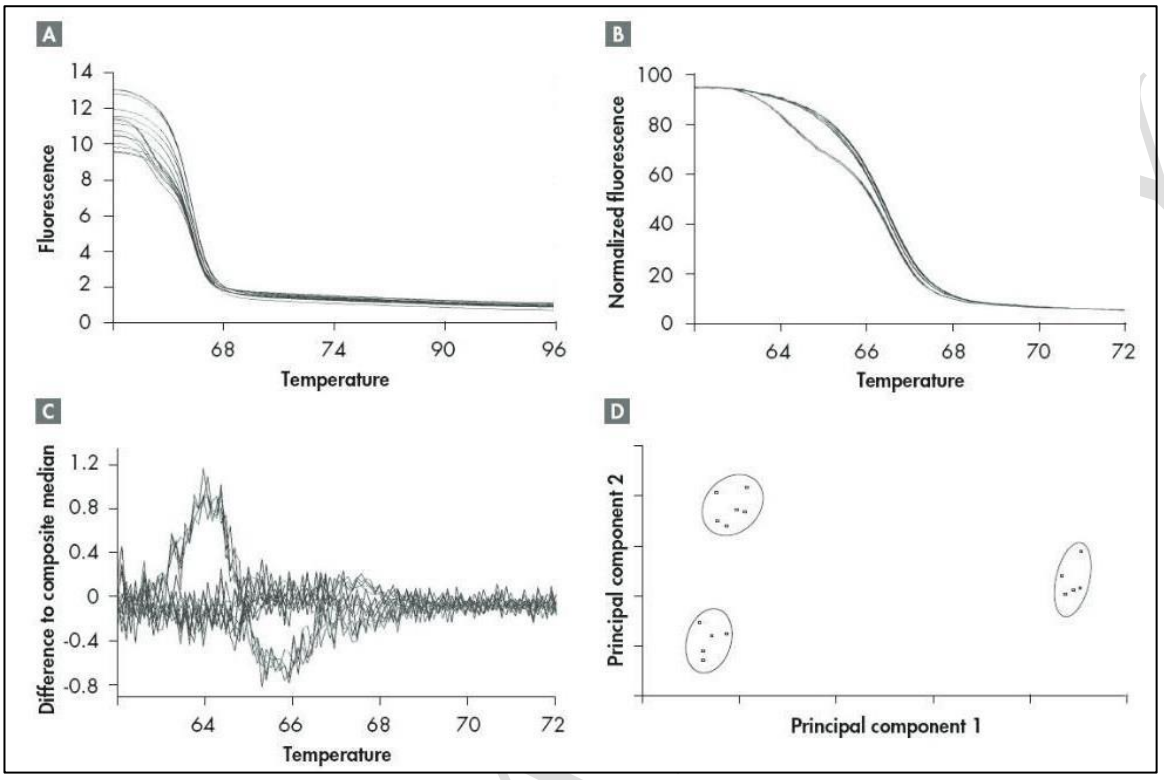


Figure 11: Data analysis performed in ScreenClust HRM Software (adapted from ScreenClust Software user manual)

Rotor-Gene ScreenClust HRM Software performs clustering (grouping) of data according to allele in either supervised or unsupervised mode. Supervised mode is often used for SNP genotyping, where the genotypes are known. In supervised mode, the user assigns one or more control samples for each cluster and the software classifies (auto calls) all unknown samples to clusters according to their characteristics. The unsupervised mode is used when there is no or only partial prior knowledge of the genotypes present in the samples. In unsupervised mode, the software calculates the optimal number of clusters by itself. This feature is an excellent tool for the discovery of new polymorphisms. In addition to the easy-to-interpret cluster plot, Rotor-Gene ScreenClust HRM Software provides statistical probabilities and typicality in a results table to allow easy comparison of results from different experiments.

2.1.3 BIO-RAD'S CFX SERIES

2.1.2 (A) CFX MANAGER

The software plots the relative fluorescence unit (RFU) data collected during a melt curve as a function of temperature. To analyze melt peak data, the software assigns a beginning and ending temperature to each peak by moving the threshold bar. The floor of the peak area is specified by the position of the melt threshold bar. A valid peak must have a minimum height relative to the distance between the threshold bar and the height of the highest peak.

- Melt Curve: Viewing the real-time data for each fluorophore as RFUs per temperature for each well.
- Melt Peak: Viewing the negative regression of the RFU data per temperature for each well.
- Well selector: Wells to show or hide the data.
- Peak spreadsheet*:* Viewing as a spreadsheet of the data collected in the selected well.

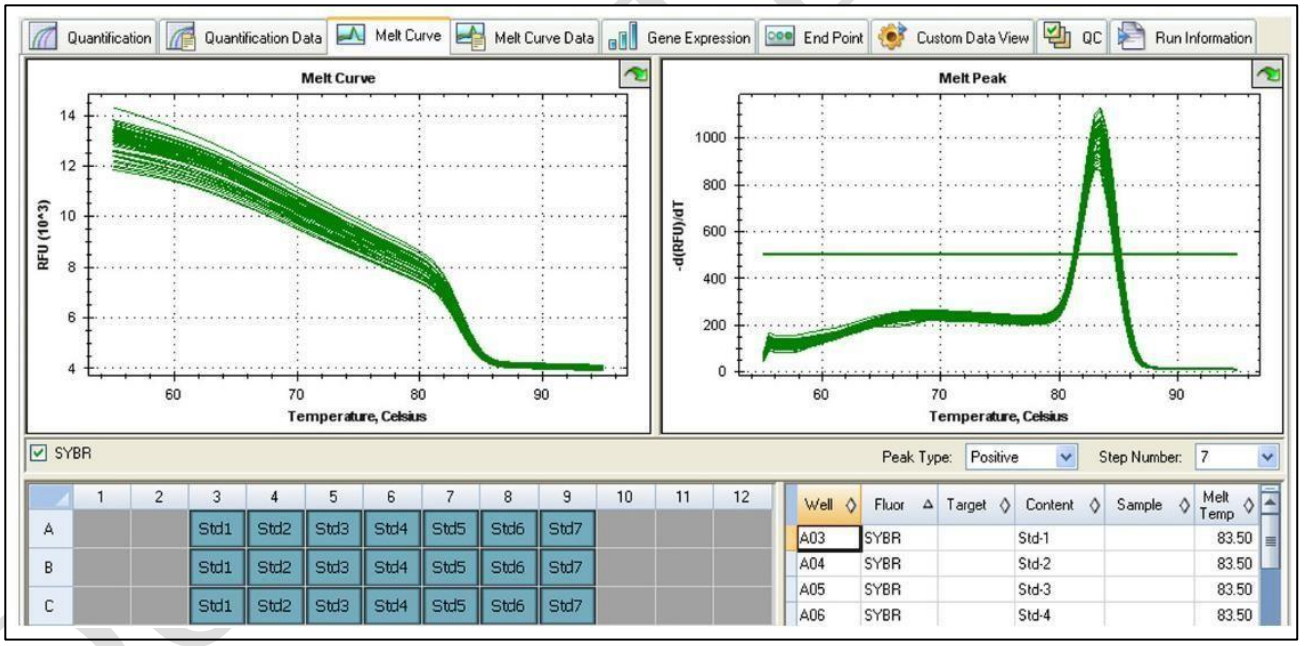
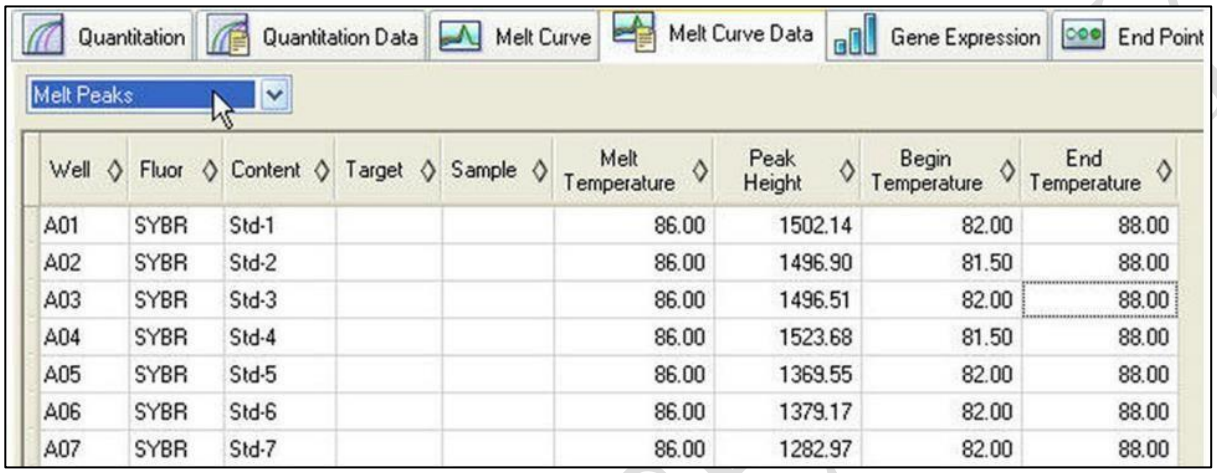


Figure 12: Interface of BIO-RAD CFX Manager

The Melt Curve Data shows the data from the Melt Curve in multiple spreadsheets, including all the melt peaks for each trace. Select one of these four options to show the melt curve data in different spreadsheets:

- Melt Peaks: Listing all the data, including all the melt peaks, for each trace
- Plate: Listing a view of the data and contents of each well in the plate
- RFU: Listing the RFU quantities at each temperature for each well
- -d(RFU)/dT: Listing the negative rate of change in RFU as the temperature (T) changes. This is the first regression plot for each well in the plate



Well	Fluor	Content	Target	Sample	Melt Temperature	Peak Height	Begin Temperature	End Temperature
A01	SYBR	Std-1			86.00	1502.14	82.00	88.00
A02	SYBR	Std-2			86.00	1496.90	81.50	88.00
A03	SYBR	Std-3			86.00	1496.51	82.00	88.00
A04	SYBR	Std-4			86.00	1523.68	81.50	88.00
A05	SYBR	Std-5			86.00	1369.55	82.00	88.00
A06	SYBR	Std-6			86.00	1379.17	82.00	88.00
A07	SYBR	Std-7			86.00	1282.97	82.00	88.00

Figure 13: Melt Peak Spreadsheet – CFX Manager

2.1.3 BIO MOLECULAR SYSTEMS – MIC

2.1.3 (A) MICPCR SOFTWARE

The micPCR software offers a Melt Analysis option that enables the determination of the peak dissociation temperature (T_m) of a sample from the melt data. This feature is useful in detecting non-specific amplicons like primer dimers, thereby serving as a measure of analytical specificity for an assay. Melt Analysis can also be applied for genotyping using chemistries such as dual hybridization probes. The software displays a graph of the first derivative curve plotted as dF/dT (y-axis) against temperature (°C, x-axis) for the first target selected in the Assays list. Users can set the melt curve threshold to any value and adjust other melting parameters available for genotyping.

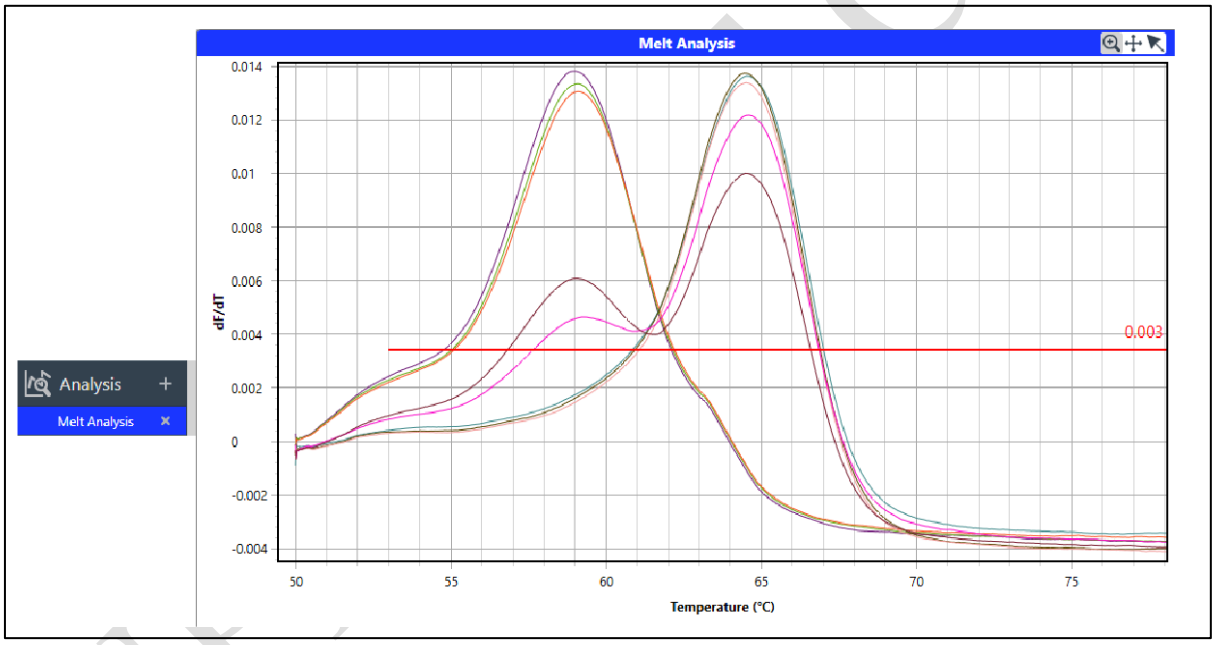


Figure 14: Interface of melt curve graph in Bio Molecular System micPCR software

2.1.4 THERMO FISHER – QUANTSTUDIO

2.1.4 (A) QUANTSTUDIO DESIGN & ANALYSIS SOFTWARE

The Melt Curve experiment is used in Thermo Fisher PCR reactions with SYBR Green dye to determine the melting temperature (T_m) of the amplification products. Tm is the temperature at which 50% of the DNA is double-stranded and 50% is dissociated into single-stranded DNA. Melt Curve analysis is included in the default run method for any experiment type that uses SYBR Green reagents. Multiple peaks in a melt curve indicate additional amplification products, often due to non-specific amplification or primer-dimer formation.

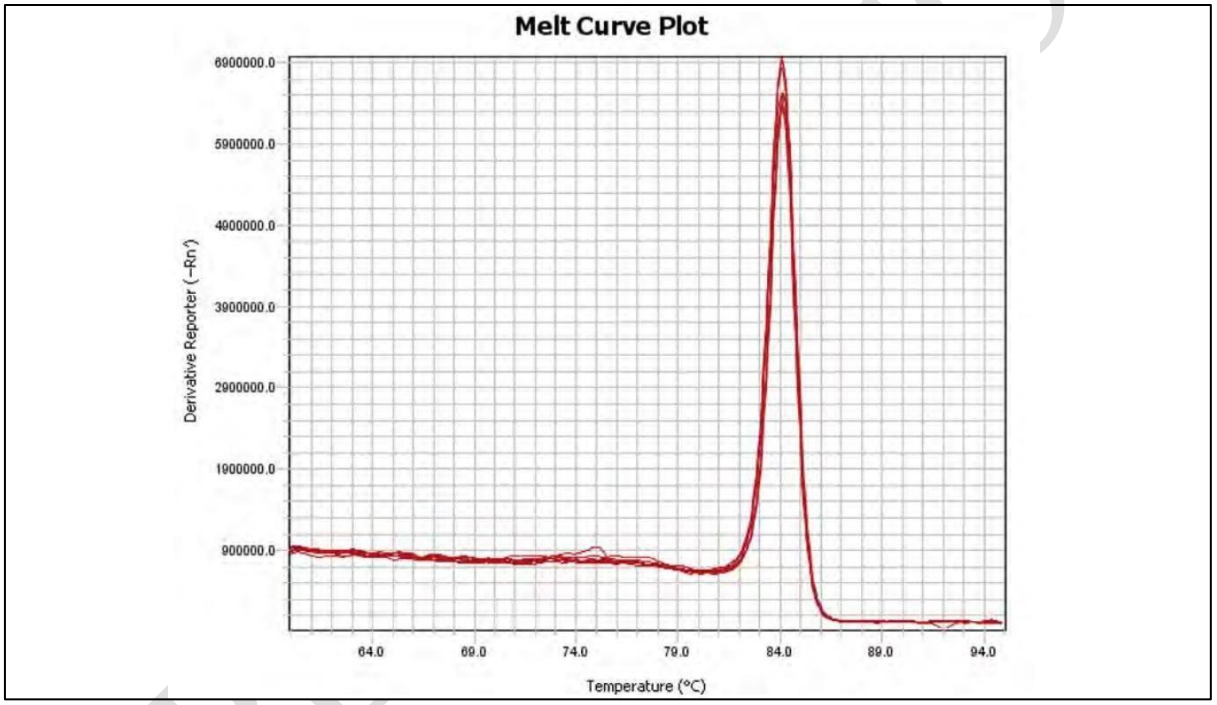


Figure 15: Interface of Melt curve graph in ThermoFisher QuantStudio software

2.1.5 ROCHE – LIGHTCYCLER SERIES

2.1.5 (A) LIGHTCYCLER SOFTWARE

The LightCycler uses fluorescence measurements to perform melting temperature analysis, which determines the melting temperature (T_m) of each sample. The analysis produces a Melting Curves chart that shows the downward curve in fluorescence as samples melt and a Melting Peaks chart that plots the first negative derivative of sample fluorescent curves to display the melting temperature of each sample as a peak. This allows for easier comparison between samples.

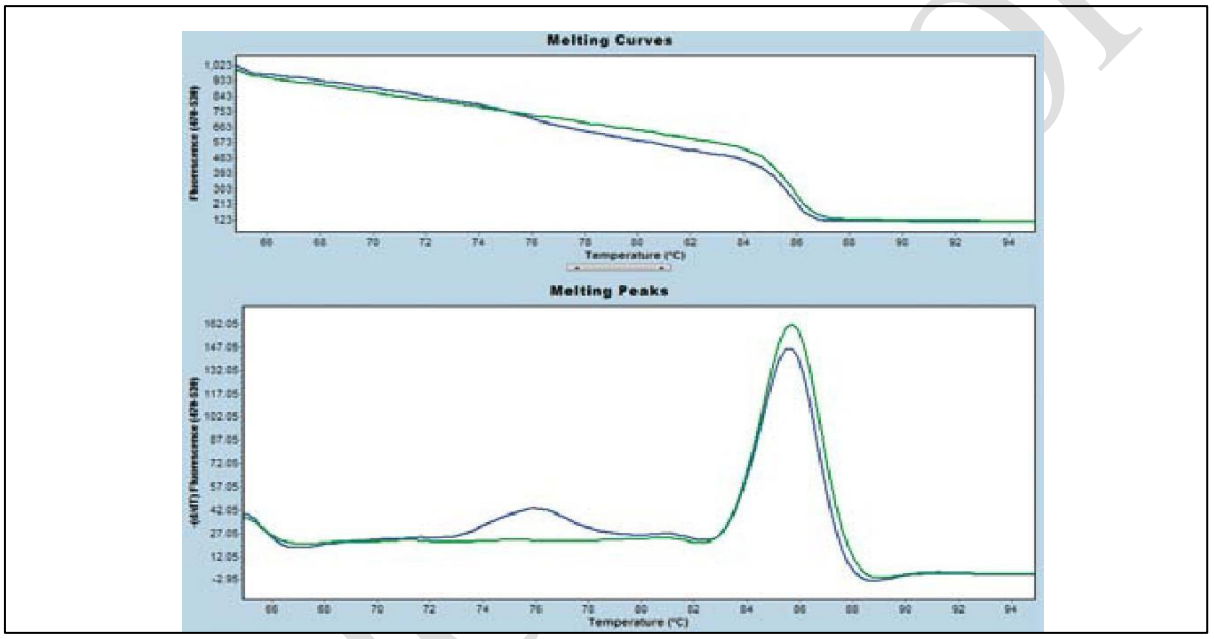


Figure 16: Interface of Roche's LightCycler

2.2 DEMOCRATIZED SOFTWARE

2.2.1 DNA-UTAH

2.2.1 (a) uANALYZE

uAnalyze is a web-based tool that analyses high-resolution melting data of PCR products. It uses recursive nearest-neighbor thermodynamic calculations to predict a melting curve. The tool accepts unprocessed melting data from *LightScanner-96*, *LS32*, or *HR-1* data files or via a generic format for other instruments. A fluorescence discriminator identifies low-intensity samples, and the background is removed either as an exponential or by linear baseline extrapolation. The precision and accuracy of experimental melting curves are quantified, and a temperature overlay is provided to focus on the curve shape.

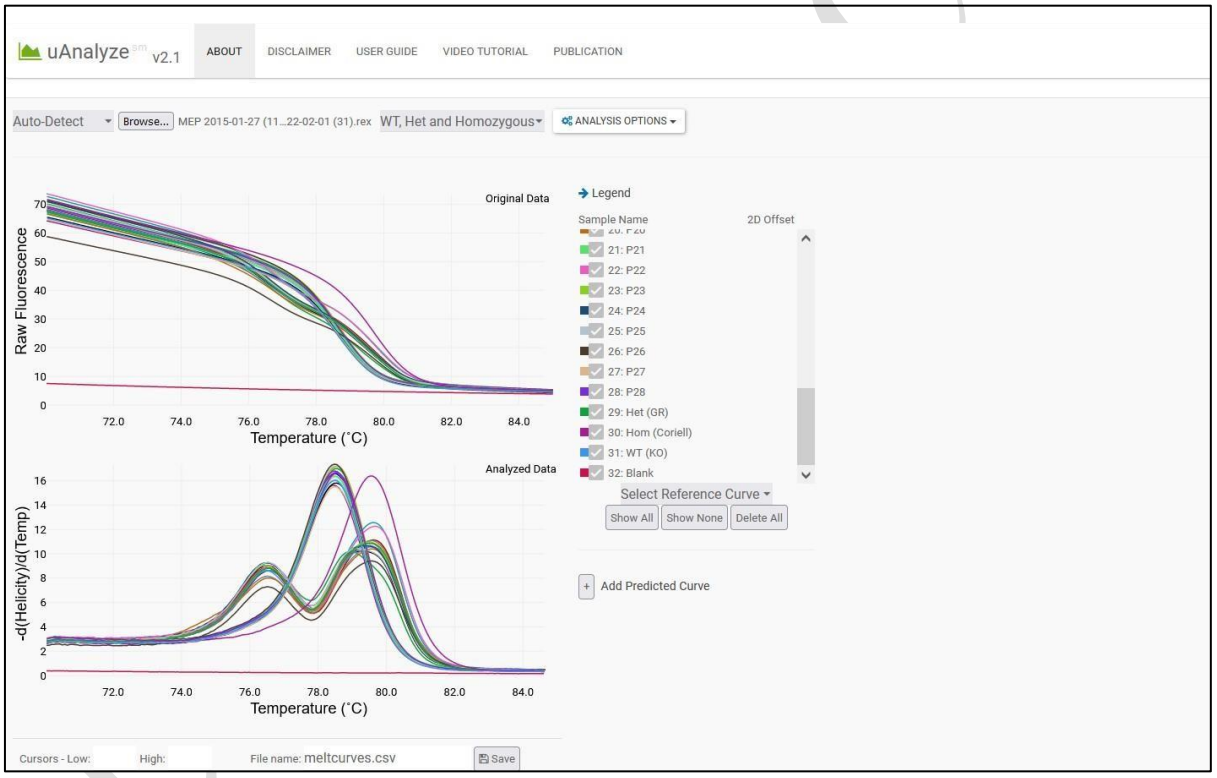


Figure 17: uANALYZE Interface

2.1.4 b) uMELT

uMelt is a web-based tool used for predicting the DNA melting curves and denaturation profiles of PCR products. The user inputs an amplicon sequence and defines thermodynamic and experimental parameters including nearest neighbour stacking energies, loop entropy effects, and cation concentrations. Using an accelerated partition function algorithm, uMelt calculates and visualizes the mean helicity and dissociation probability at each sequence position within a temperature range. The predicted curves and profiles display stability and loss of helicity with increasing temperature. Results from fluorescent high-resolution melting experiments match the predicted melting domains and their relative temperatures, but the absolute melting temperatures may vary. uMelt provides a convenient platform for the simulation and design of high-resolution melting assays.

Figure 18: uMELT Interface

CHAPTER 3

LITERATURE REVIEW

Untergasser *et al.* 2021, made the analyses of amplification and melting curves to provide valuable information on the quality of individual reactions in quantitative PCR (qPCR) experiments and result in more reliable and reproducible quantitative results. The new web-based LinRegPCR web application provides visualization and analysis of a single qPCR run, displaying the analysis results on the amplification curve and melting curve analysis in tables and graphs. It also provides a stand-alone back-end RDML (Real-time PCR Data Markup Language) Python library and several companion applications for data visualization, analysis, and interactive access. The use of the RDML data standard enables machine-independent storage and exchange of qPCR data, and the RDML tools assist with importing the data from the files exported by the qPCR instrument.

Moniri *et al.* (2020) demonstrates that the large volume of raw data obtained from real-time PCR instruments can be exploited to perform data-driven multiplexing in a single channel using machine learning methods. This approach, referred to as Amplification Curve Analysis (ACA), was used to multiplex 3 carbapenem-resistant genes in the presence of single targets, resulting in an accuracy of 99.1% (N = 16188). To support the analysis, a formula was derived to estimate co-amplification occurrence in PCR based on multi-variate Poisson statistics. Combining this method with probe-based assays will increase multiplexing capabilities.

Wisittipanit *et al.* 2020 used a modified high-resolution DNA melting curve analysis (m-HRMa) to classify Salmonella spp. into clusters and a machine learning (dynamic time warping) algorithm (DTW) to create a phylogeny tree of Salmonella strains (n = 40) collected from homes, farms, and slaughterhouses in northern Thailand. DTW and ms-HRMa clustering analyses were able to generate molecular signatures of the Salmonella isolates, resulting in 25 ms-HRM and 28 DTW clusters compared to 14 clusters from a standard HRM analysis. The new Salmonella sub-typing protocol identified five S. Weltevraden subtypes with S.Weltevreden subtype DTW4-M1 being predominant. This suggests that transmission of salmonellosis in northern Thailand is likely to be farm-to-farm through contaminated chicken stool.

Athamanolap *et al.* 2014 made an automated HRM curve classification based on machine learning methods and learned tolerance for reaction condition deviations enables reliable, scalable, and automated HRM genotyping analysis with broad potential clinical and epidemiological applications.

Roediger *et al.* 2013. implemented the MBmca package with R, for DNA Melting Curve Analysis on microbead surfaces. Particularly, for the use of the second derivative melting peaks as an additional parameter to characterize the melting behaviour of DNA duplexes.

Dwight *et al.* 2011 created a web-based tool called "uMeltSM" for predicting DNA melting curves and denaturation proles of PCR products. It uses an accelerated partition function algorithm to calculate and visualize the mean helicity and dissociation probability at each sequence position at different temperatures. Results from fluorescent melting experiments match the number of predicted domains and their relative temperatures, but current libraries do not account for the rapid melting rates and helix-stabilizing dyes used in experiments.

Smith *et al.* 2009 defines Methylation of DNA as a common mechanism for silencing genes and is increasingly being implicated in many diseases. They describe and validate a rapid, in-tube method to quantitate DNA methylation using the melt data obtained following the amplification of bisulphite-modified DNA in a real-time thermocycler. The parameters derived provide an objective description and quantitation of the methylation in a specimen andcan be used for statistical comparisons of methylation between specimens

CHAPTER 4

PROPOSED METHODOLOGIES

Creating an AI-based framework for interpreting and reporting PCR tests requires extensive study and a major focus on understanding significant result analysis techniques like as melt curve analysis and cycle threshold analysis. As a result, completing proper research is critical for developing a strong solution to the challenge at present.

4.1 CORE COMPONENTS

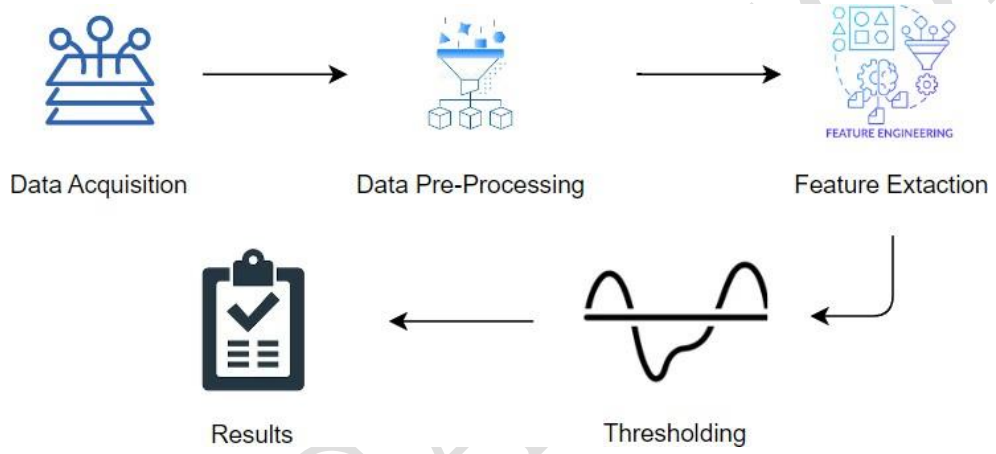


Figure:19 Process of workflow

In this project, the primary focus lies in the AI-based interpretation of Melt and Cycle thresholds, where obtaining accurate and reliable information from raw data plays a vital role. Given this objective, significant attention should be directed towards data acquisition and feature engineering. It will serve as the foundation for setting thresholds and conducting further analysis.

Interpreting the results of PCR test samples relies heavily on melt and cycle thresholds as predominant key measures. Each pathogen exhibits unique peaks in the melt curve, aiding in identifying their presence in test samples, while cycle threshold take-off points assist in understanding the severity of infection in patients. This analysis is facilitated by extracting features through feature engineering. Subsequently, upon detecting the presence of pathogens based on their nature and characteristics, LLM-based reports will be generated by prompting the required instructions to the Large Language Model.

4.2 PROPOSED METHODOLOGIES

In this project, three methodologies were proposed to analyze key measures such as Melt and Amplification curves. Each of them yielded notable results.

- One approach involved utilizing existing software.
- Another approach focused on the extracting the of co-ordinates of raw fluorescence and cycle threshold using GUI-based Automated Human-Computer Interaction.
- The third approach centered on file parsing to extract raw fluorescence and cycle threshold data.

These methodologies were implemented as part of our data pre-processing and feature engineering process. However, not all of them were successful, as each had its limitations and disadvantages. Subsequent approaches aimed to address the limitations of previous methods and yielded observable improvements.

CHAPTER 5

APPROACH ON UTILIZING EXISTING SOFTWARES

5.1 INTRODUCTION

Data acquisition refers to extracting required information from the raw data. It's super important for analyzing data. This is where you gather information from various sources like databases, files etc. Sometimes data can be messy, so you need to tidy it up by removing errors or inconsistencies. It's crucial to structure the data in a way that makes it easy to work with during analysis. This involves transforming the data into a format suitable for analysis, like converting it into tables or charts. As this project focuses on interpreting the Melt and Amplification curve, there is a need to extract required information from raw data.

5.2 RAW DATA

Polymerase Chain Reaction (PCR) is a crucial tool used in labs to check patient samples for different germs and genetic traits. To do this, labs use various PCR machines like Roto Gene Q Rex software, Bio-Rad's CFX series, and Thermo-Fisher's QuantStudio. When the Roto Gene Q Rex software is used, it creates .rex files as output after the test samples are inputted. Up until now, the only way to analyze these samples has been through this software.

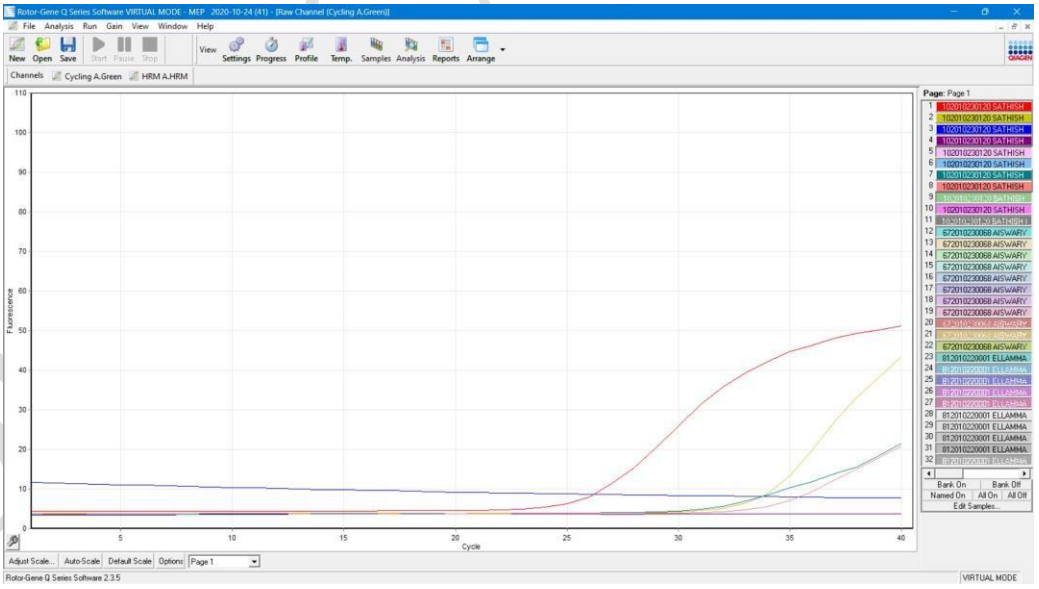


Figure:20 Raw Fluorescence signal data
Source: Qigen's Rotor Gene Q-rex software

Within the .rex files multiple test data measures are present. The focus is to extract two key measures from raw data which includes High Resolution Melt and Cycle Threshold. Before this, to extract the information (.xls file) from the raw data (.rex file) is only through Rotor Gene Q rex software. To finding the pattern in Melt and Amplification Curve, it requires lots of test samples data (.rex). Initially the information has been extracted manually through the Rotor Gene Q res software using its software plugins.

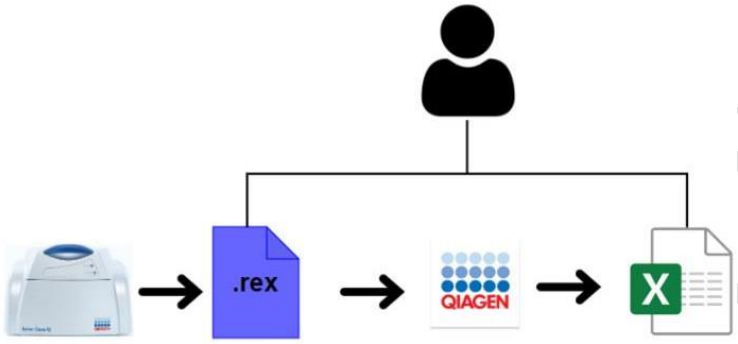


Figure:21 Manual Conversion

The limitations of Qiagen's Rotor Gene Q-rex Software in extracting the information from the raw data are given as below:

- Since it is a proprietary software, it's not possible to extract the necessary information from the raw data without the aid Rotor Gene Q-rex Software.
- Manually extracting information from many files is time-consuming.

5.3 EXTRACTOR

To overcome the above-mentioned disadvantages, the previous project work found a solution to extracting the information from the raw data using GUI based software called Extractor. It is a lightweight simple GUI-based application that extracts '.rex' files from the Qiagen's Rotor-Gene Q Software to the necessary '.xls' file. It's built for the users such as laboratory technicians and clinicians who handle and run PCR experiments especially in Qiagen's Rotor-Gene Q thermal cycler machine. Extractor automated the user role by simply put the raw data file directory and desired directory to which the excel files are stored in your system, which saves time and not to burned out from this repetitive task.

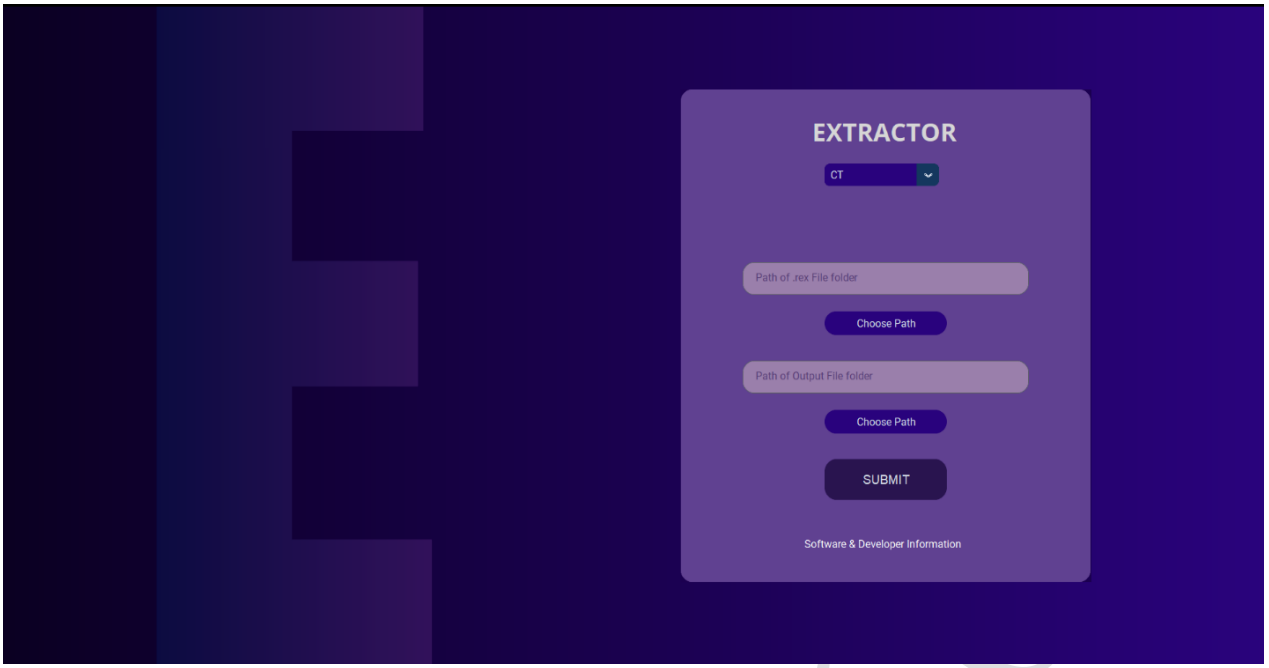


Figure:22 User Interface of Extractor

ADVANTAGES

- It removes the need for the user manually extract information from raw data one by one.
- It reduces the time required to extract information with the aid of Rotor Gene Q Rex software by 40%.
-

DISADVANTAGES

- The software utilizes snipped images of the required software options to search for patterns on the screen. When it detects a matching pattern, it automatically clicks in that region. This is the operational mechanism of the software. However, the challenge arises because the pixel values of the snipped images vary for each system. As a result, the algorithm encounters difficulties when operating on different computers.
- The Extractor works with the only aid of Rotor Gene Q rex software. As it is a proprietary software (Qiagen's Rotor Gene Q rex), this software is not available to all the user who wants to extract information from raw data.
- Since the Extractor is a GUI-based software, it doesn't function in a multi-threading environment. This means that until the processing is completed, the user is unable to access any other application while it's still running.

5.4 PyHRM

PyHRM is a first Python-based library for processing High Resolution Melting (HRM) data to extract features like Melting Temperatures, Take-off, and Touch-down points of melting signal Temperature at which peak start rising and temperature at which peak falls, Peak prominences, and Area Under the curve. Additionally, the library offers vision-based filtering to eliminate noisy signals from the data and provides only genuine peaks with all the above-mentioned features.

Figure: A library for processing DNA Melting signal with feature extraction

In this project, by using PyHRM some difficulties were raised in different stages. It was listed as follows

- When installing the PyHRM library for the first time, there are difficulties in utilizing its features for further analysis due to variable conflicts, which were identified as bugs. Although these conflicts were resolved on local machines, the fixes were not implemented in the library's repository.
- PyHRM relies on dependencies such as Extractor or Rotor Gene Q-rex software, which only accept (.xlsx) files as inputs. It does not support direct run files (.rex) as input files.
- PyHRM has its focus on High Resolution Melt to extract features from the transformed data of .rex file format which is excel file format.
- It does not focus on Cycle threshold data to extract the feature where Cycle threshold plays a vital role in interpreting the results
- Though it helps to extract features from the HRM data, it is unable to classify test samples from different patients within the same input file.
- PyHRM utilizes a vision-based approach to remove noise signals from the melt data. However, due to the insufficient noisy signal images for each pathogen, it misclassifies genuine signals as noisy signals. This raises doubts about the reliability of the results obtained from the library.

5.5 MELT CURVE INTERPRETER

The Melt Curve Interpreter is a web-based application designed for analyzing and interpreting the results generated by the Extractor and PyHRM library. This module includes various files and folders, such as .py, .html, .css, and .h5, which are used for the final classification and interpretation of the Meningitidis panel data.

Figure:23 Web based application for analyzing and interpreting results

MCI WORKING DIFFICULTIES

Although the Melt Curve Interpreter (MCI) has two sections for uploading High-resolution melt data and Cycle Threshold data, it cannot interpret both datasets simultaneously. The foundation for MCI is based on the Python library PyHRM and Extractor software. Therefore, the limitations encountered with these components will also apply to the Melt Curve Interpreter.

RESULT AND DISCUSSION

Interpreting High-Resolution Melt and Cycle Threshold data using existing software like Extractor, Melt Curve Interpreter, and the Python library PyHRM faces several limitations across data acquisition, preprocessing, and feature engineering stages. Extractor software extracts data from .rex files with a digital filter applied, which may not provide accurate features from test samples. While PyHRM library serves as a foundation for extracting Melt Curve features, interpreting PCR test samples requires more than just Melt features. The vision-based approach demands a greater number of 'single-peak', 'double-peak', and noisy signals for effective classification. Insufficient data of these images hinders efficient signal classification.

CONCLUSION

Raw data input is essential for precise feature extraction, especially in classifying melt signals as 'single-peak', 'double-peak', or 'noisy signals'. Finding an alternative method for this classification is imperative. Addressing the limitations of Extractor, PyHRM, and Melt Curve Interpreter is crucial for enhancing the software. Upgrading the software to integrate all these functionalities into one comprehensive tool will streamline the analysis process.

CHAPTER 6
GUI-BASED AUTOMATED EXTRACTION OF RAW FLUORESCENCE AND CYCLE THRESHOLD COORDINATES

The identified limitations from the previous approach have led us to a dead end, hindering further analysis. Therefore, we are transitioning to a new approach focused on the Automated Extraction of raw fluorescence and cycle threshold coordinates. This shift aims to overcome previous constraints and enable more comprehensive analysis. The main drawback of Extractor lies in its image-based approach, where it compares predefined snipped images of software options on the screen. The system faces difficulty in detecting these patterns due to pixel variations across different systems, rendering it unable to accurately locate the snipped images on the screen. A generalized GUI-based Automated Human Computer Interaction is needed to address these above limitations.

6.1 KEYBOARD AND MOUSE BASED APPROACH

A unique keyboard and mouse navigation pattern has been identified for extracting key features such as HRM and CT. This approach leverages the consistency in keyboard and mouse operations across different systems, allowing for more reliable feature extraction with the help of Rotor Gene Q-rex Software. Utilizing the keyboard and mouse-based navigation approach, High Resolution Melt (HRM) and Cycle Threshold data have been extracted. Using the previously found solution in PyHRM, features have been extracted. However, the extracted feature is not reliable due to its vision-based approach, as it misclassifies genuine signals as noisy signals. As an alternative, there is a possibility of neglecting noisy signals without the need for a vision-based approach (Image Classification). By considering the top two peak prominences, genuine peaks are not misclassified, thus aiding in the removal of noisy signals. Once the genuinely extracted features are obtained, the next crucial step is to threshold the samples of the pathogen using the values provided by clinicians for the test results.

During the process, while working with a SEPSIS panel that tests for 30+ pathogens. Samples were collected across 60+ branches, and during the sample collection process, each pathogen was named differently from one branch to another. This discrepancy created a significant problem for the project. To overcome this issue, we need to establish generic naming conventions that are commonly followed by all branches.

6.2 GENERIC NAMING CONVENTIONS

Naming conventions play a vital role in data preprocessing and analysis. Inconsistent naming of pathogens within the same dataset can lead to confusion during analysis, treating the same pathogen under different names as separate entities. This not only affects the accuracy of the analysis but also creates problems during data storage, increasing the risk of **data inconsistency** and duplicates. Moreover, inconsistent naming conventions can lead to difficulties in data retrieval in the future.

The main issues we face are incorrect identification of pathogen names and the presence of spaces in the names. This makes it challenging to split the names during analysis, which in turn makes it difficult to apply thresholding to the pathogen results using the values provided by clinicians. Therefore, establishing clear and standardized naming conventions is crucial for ensuring data integrity and smooth data analysis processes. The formulated naming conventions are as follows:

Format: ***SAMPLE_ID: BAR_CODE NAME PATHOGEN***

Sample ID	Numerical Character	Sample Identification during the tests
Bar Code	Numerical Character	Unique Identification for patient
Patient Name	Text Character	Patient name (Spaces avoided)
Pathogen Name	Text Character	Use generic name (Short name)

Table:3 Naming Convention

Example: ***11: 01234567890 Patient_name Pathogen_name***

Before Naming Convention

After Naming Convention

Figure:24 Pathogen Name was changed into Generic Short Name

6.3 EXAMINING FEATURE EXTRACTION

From the obtained naming pattern, a manual data cleaning process across the sepsis panel was done, in which around 30+ pathogen names were changed to a generic short name provided by the laboratory. After the cleaning process, samples were tested with the threshold values provided by the clinicians to find the results. After finding the results, it was compared to the original reports provided by the clinicians previously, but most of the reports were mismatched. One of the example files in which Panbacteria family test are as follows:

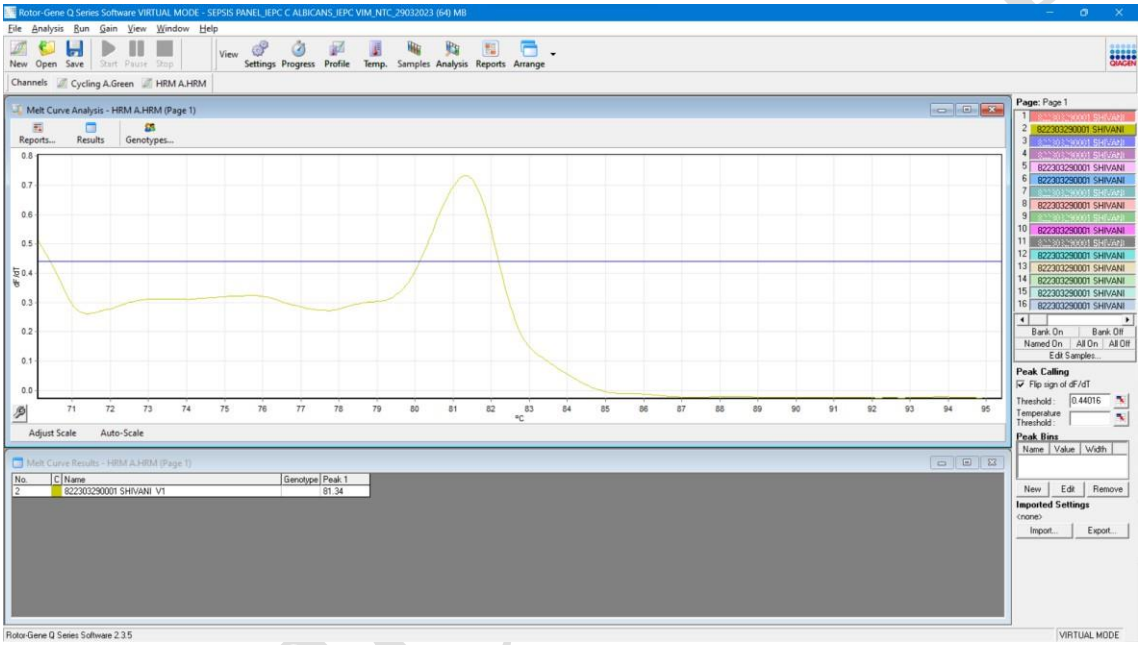


Figure:25 Rotor-Gene Visuals V1 (81-86)

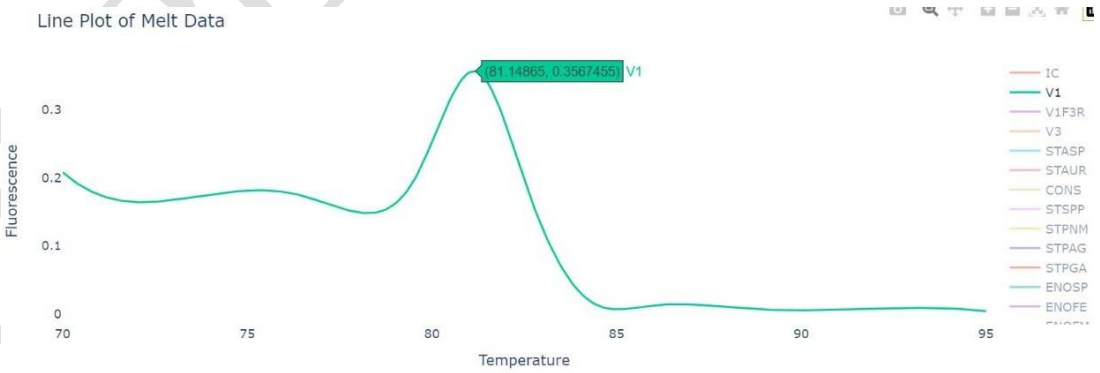


Figure:26 Extracted Data V1 (81-86)

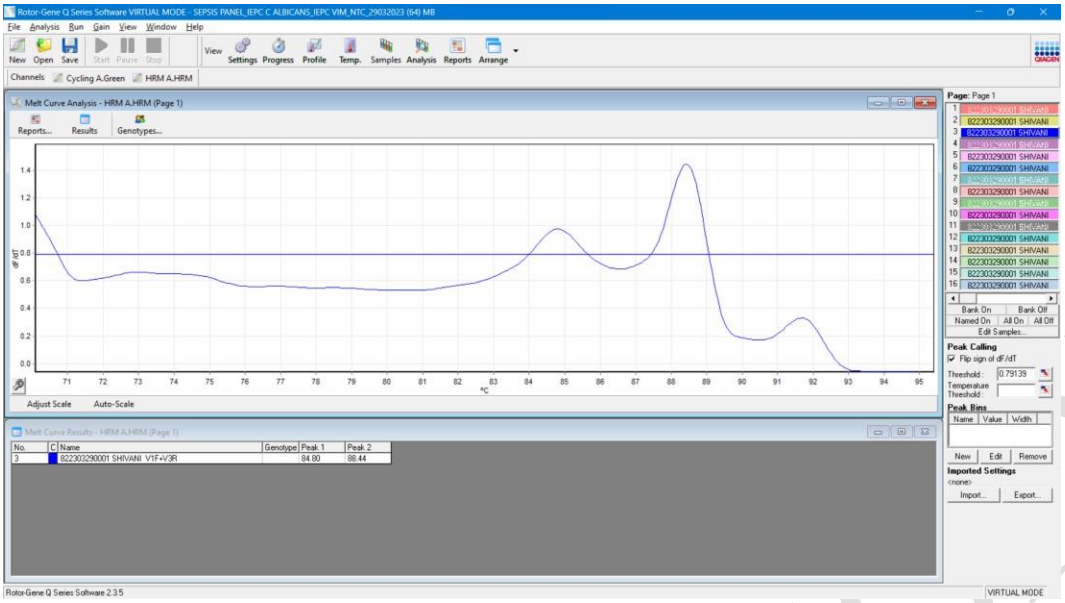


Figure:27 Rotor-Gene Visuals V1F3R (81-88)

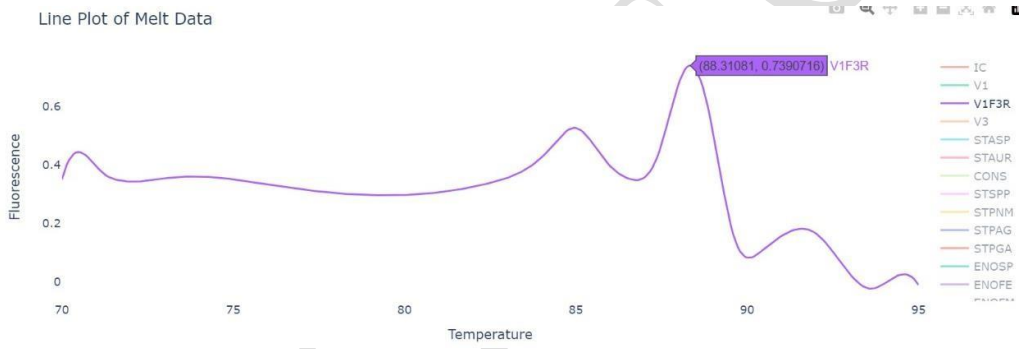


Figure:28 Extracted Data V1F3R (81-88)

From the above visuals, it is clear that V1 and V1F3R are positive, indicating they belong to the panbacterial family. The features extracted from Rotor Gene Q-rex Software are mostly the same as the extracted features. However, upon comparing these results with the original reports, it was found that the reports were unmatched.

Spec.type	Test Name	Results	Previous Results (Date)	units	Reference Ranges/Methods
	Pan-bacteria DNA #	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity

Figure:29 Final Report

RESULT AND DISCUSSION

In this approach, the main reason for the misclassification are the inputs are extracted from a GUI-Based Automated tool where the data undergo digital filtering, making it challenging to access the raw data. During the melt conversion process to the already filtered data, PyHRM applies another filter, affecting the small differences between the extracted features and the data from the Rotor Gene Q-rex software. This difference may be a reason for discrepancies in the reports. Another significant factor could be the lack of consideration for the amplification curve, also known as the cycle threshold, which helps interpret the severity of pathogen infection in patients and plays a crucial role in filtering out noisy signals.

Not only do these mentioned factors play a role, but other factors such as the patient's medical history and symptoms also contribute significantly to determining the results of the SEPSIS panel. Unfortunately, there is no available data related to symptoms and the patient's medical history in the given input files. Additionally, many factors can affect the test results, so the results cannot be taken as the final outcome for interpretation.

CONCLUSION

In conclusion, the main reasons for errors in the test results are due to how the data is processed and filtered, making it difficult to get accurate readings. The method used to convert the data may introduce additional filtering, leading to discrepancies compared to other software. Also, important information like the severity of infection and patient symptoms are not considered in the analysis, which can affect the results. Since various factors can influence the outcome, the results should not be solely relied upon for interpretation.

CHAPTER 7

AN APPROACH CENTERED TO PARSING OF ROTOR EXPERIMENT FILE WITH LOGICAL THRESHOLDING

In the previous chapter, the data transformation of Rotor-Gene experiment files into Excel file format was done using a GUI-based tool with dependencies on Rotor-Gene Q-rex software. High-resolution melt and cycle threshold were extracted separately from the Rex files. However, this approach caused several issues during the feature extraction and interpretation of results from multiple patients available in the experiment files. To overcome the above problem, we must use the *Rex file* as the input file. In this chapter, an alternative method was approached for data extraction by parsing the required elements from the Rex file as follows

7.1 FILE PARSING

These Rotor-Gene experiment (.rex) files are more complex repositories of valuable information, containing multiple channels that contain critical elements such as High-Resolution Melt (HRM), Amplification Curve data, and detailed Sample information. Within these files, we encounter various channels, each of which plays an important role in the data extraction process. Amplification Curve channel, also known as Cycling A. Green, is one of the most significant. This particular element within the raw channels contains critical attributes such as startX, stepX, and MinX points, which are useful in determining the determine the trend and intervals of the x-axis values, similar to what one would see in the Rotor Gene Q software interface.

Following that, within this same element, another critical aspect, namely the 'reading' values. These 'reading' values are connected according to the previously identified x-axis values and are concatenated accordingly, forming a list of information known as Master CT.

The parsing process does not end with the Amplification Curve; rather, it continues to include other important elements found in .rex files, such as HRM data and sample details. Similar exact parsing techniques are used to extract relevant information from these elements, resulting in a thorough and accurate data compilation.

ALGORITHM:1 ALGORITHM TO EXTRACT RAW FLUROCENSE FROM RUN FILES

Input: *Rotor Gene Experiment files*
Output: *Excel files*
Import required libraries*: pandas, xml.etree.ElementTree, os*
Initialization of objects*: Read the input .rex file and Create variables for HRM and CT*
Amplification:
Parse for amplification data from amplification tag
Extract and process amplification data
Return processed amplification data
High Resolution Melt:
Parse for HRM data from HRM tag
Extract and process HRM data
Return processed HRM data
Sample details:
Parse for sample details for both HRM and Amplification
Extract and process sample details
Return processed sample details
Define HRM Data function:
Get HRM and sample details with respect to startX and stepX points
Create data frames for HRM data
Concatenate data frames and save to Excel
Define CT Cycle function:
Get Amplification and sample details with respect to startX and stepX points
Create data frames for CT Cycle data
Concatenate data frames and save to Excel
End

By using this improved approach to data extraction and parsing, we can avoid the problems that affected previous methodologies. With this technique, we can extract the exact raw files from the .rex files without applying any digital filters to the data. This helps us get much closer to the upcoming results compared to the results obtained from the past approach. This improved method not only allows for a more efficient extraction process but also ensures accurate interpretation and analysis of results. This is especially important when dealing with data from multiple patients within experiment files.

Text	X	Y	Text	X	Y	Text	X	Y
1: 12207328 SHIMNA EV	70	57.87376429	2: 12207328 SHIMNA HSV	70	28.31804497	3: 12207328 SHIMNA VZV	70	7.204840631
1: 12207328 SHIMNA EV	70.2	56.95175606	2: 12207328 SHIMNA HSV	70.2	28.14892805	3: 12207328 SHIMNA VZV	70.2	6.955531689
1: 12207328 SHIMNA EV	70.4	56.37153175	2: 12207328 SHIMNA HSV	70.4	27.83736063	3: 12207328 SHIMNA VZV	70.4	6.743088935
1: 12207328 SHIMNA EV	70.6	55.59940708	2: 12207328 SHIMNA HSV	70.6	27.63735845	3: 12207328 SHIMNA VZV	70.6	6.609801378
1: 12207328 SHIMNA EV	70.8	55.20988785	2: 12207328 SHIMNA HSV	70.8	27.40778562	3: 12207328 SHIMNA VZV	70.8	6.524321791
1: 12207328 SHIMNA EV	71	54.420608	2: 12207328 SHIMNA HSV	71	27.25551232	3: 12207328 SHIMNA VZV	71	6.440833143
1: 12207328 SHIMNA EV	71.2	53.93817749	2: 12207328 SHIMNA HSV	71.2	27.06492714	3: 12207328 SHIMNA VZV	71.2	6.364283716
1: 12207328 SHIMNA EV	71.4	53.28381087	2: 12207328 SHIMNA HSV	71.4	26.86778058	3: 12207328 SHIMNA VZV	71.4	6.320922662
1: 12207328 SHIMNA EV	71.6	52.65796902	2: 12207328 SHIMNA HSV	71.6	26.6271022	3: 12207328 SHIMNA VZV	71.6	6.245009028
1: 12207328 SHIMNA EV	71.8	52.220442	2: 12207328 SHIMNA HSV	71.8	26.40944315	3: 12207328 SHIMNA VZV	71.8	6.226116906
1: 12207328 SHIMNA EV	72	51.41669119	2: 12207328 SHIMNA HSV	72	26.16551679	3: 12207328 SHIMNA VZV	72	6.187342641
1: 12207328 SHIMNA EV	72.2	50.93634445	2: 12207328 SHIMNA HSV	72.2	25.95025559	3: 12207328 SHIMNA VZV	72.2	6.130966005
1: 12207328 SHIMNA EV	72.4	50.11662979	2: 12207328 SHIMNA HSV	72.4	25.66435238	3: 12207328 SHIMNA VZV	72.4	6.10265507
1: 12207328 SHIMNA EV	72.6	49.31323502	2: 12207328 SHIMNA HSV	72.6	25.48875994	3: 12207328 SHIMNA VZV	72.6	6.085679408
1: 12207328 SHIMNA EV	72.8	47.97785262	2: 12207328 SHIMNA HSV	72.8	25.26392658	3: 12207328 SHIMNA VZV	72.8	6.025133789
1: 12207328 SHIMNA EV	73	47.18793833	2: 12207328 SHIMNA HSV	73	25.04768444	3: 12207328 SHIMNA VZV	73	6.032008733
1: 12207328 SHIMNA EV	73.2	46.40984559	2: 12207328 SHIMNA HSV	73.2	24.83687519	3: 12207328 SHIMNA VZV	73.2	5.998226139
1: 12207328 SHIMNA EV	73.4	45.538262	2: 12207328 SHIMNA HSV	73.4	24.57907988	3: 12207328 SHIMNA VZV	73.4	5.974924341
1: 12207328 SHIMNA EV	73.6	44.62511853	2: 12207328 SHIMNA HSV	73.6	24.37803137	3: 12207328 SHIMNA VZV	73.6	5.929773286

Figure:30 Extracted raw flurocense form Rotor-Gene

Text	X	Y	Text	X	Y	Text	X	Y
1: 12207328 SHIMNA EV	70	57.87376429	2: 12207328 SHIMNA HSV	70	28.31804497	3: 12207328 SHIMNA VZV	70	7.204840631
1: 12207328 SHIMNA EV	70.2	56.95175606	2: 12207328 SHIMNA HSV	70.2	28.14892805	3: 12207328 SHIMNA VZV	70.2	6.955531689
1: 12207328 SHIMNA EV	70.4	56.37153175	2: 12207328 SHIMNA HSV	70.4	27.83736063	3: 12207328 SHIMNA VZV	70.4	6.743088935
1: 12207328 SHIMNA EV	70.6	55.59940708	2: 12207328 SHIMNA HSV	70.6	27.63735845	3: 12207328 SHIMNA VZV	70.6	6.609801378
1: 12207328 SHIMNA EV	70.8	55.20988785	2: 12207328 SHIMNA HSV	70.8	27.40778562	3: 12207328 SHIMNA VZV	70.8	6.524321791
1: 12207328 SHIMNA EV	71	54.420608	2: 12207328 SHIMNA HSV	71	27.25551232	3: 12207328 SHIMNA VZV	71	6.440833143
1: 12207328 SHIMNA EV	71.2	53.93817749	2: 12207328 SHIMNA HSV	71.2	27.06492714	3: 12207328 SHIMNA VZV	71.2	6.364283716
1: 12207328 SHIMNA EV	71.4	53.28381087	2: 12207328 SHIMNA HSV	71.4	26.86778058	3: 12207328 SHIMNA VZV	71.4	6.320922662
1: 12207328 SHIMNA EV	71.6	52.65796902	2: 12207328 SHIMNA HSV	71.6	26.6271022	3: 12207328 SHIMNA VZV	71.6	6.245009028
1: 12207328 SHIMNA EV	71.8	52.220442	2: 12207328 SHIMNA HSV	71.8	26.40944315	3: 12207328 SHIMNA VZV	71.8	6.226116906
1: 12207328 SHIMNA EV	72	51.41669119	2: 12207328 SHIMNA HSV	72	26.16551679	3: 12207328 SHIMNA VZV	72	6.187342641
1: 12207328 SHIMNA EV	72.2	50.93634445	2: 12207328 SHIMNA HSV	72.2	25.95025559	3: 12207328 SHIMNA VZV	72.2	6.130966005
1: 12207328 SHIMNA EV	72.4	50.11662979	2: 12207328 SHIMNA HSV	72.4	25.66435238	3: 12207328 SHIMNA VZV	72.4	6.10265507
1: 12207328 SHIMNA EV	72.6	49.31323502	2: 12207328 SHIMNA HSV	72.6	25.48875994	3: 12207328 SHIMNA VZV	72.6	6.085679408
1: 12207328 SHIMNA EV	72.8	47.97785262	2: 12207328 SHIMNA HSV	72.8	25.26392658	3: 12207328 SHIMNA VZV	72.8	6.025133789
1: 12207328 SHIMNA EV	73	47.18793833	2: 12207328 SHIMNA HSV	73	25.04768444	3: 12207328 SHIMNA VZV	73	6.032008733
1: 12207328 SHIMNA EV	73.2	46.40984559	2: 12207328 SHIMNA HSV	73.2	24.83687519	3: 12207328 SHIMNA VZV	73.2	5.998226139
1: 12207328 SHIMNA EV	73.4	45.538262	2: 12207328 SHIMNA HSV	73.4	24.57907988	3: 12207328 SHIMNA VZV	73.4	5.974924341
1: 12207328 SHIMNA EV	73.6	44.62511853	2: 12207328 SHIMNA HSV	73.6	24.37803137	3: 12207328 SHIMNA VZV	73.6	5.929773286

Figure:31 Extracted raw flurocense through File parsing

This file parsing method don't need any dependency (Rotor Gene Software). This novel algorithm to extracted raw flurocense data from .rex file format.

7.2 SMOOTHENING HIGH RESOLUTION MELT

The extracted HRM data through file parsing is in the raw format. Since it's a raw format data, couldn't directly use to for further analysis, because there are lot of unwanted noisy data. Need to convert raw florescence data to digital filter applied data, so that, must use smoothening approach to resolve the noisy data. If raw florescence data converted derivative melt plot without smoothening, will look like below,

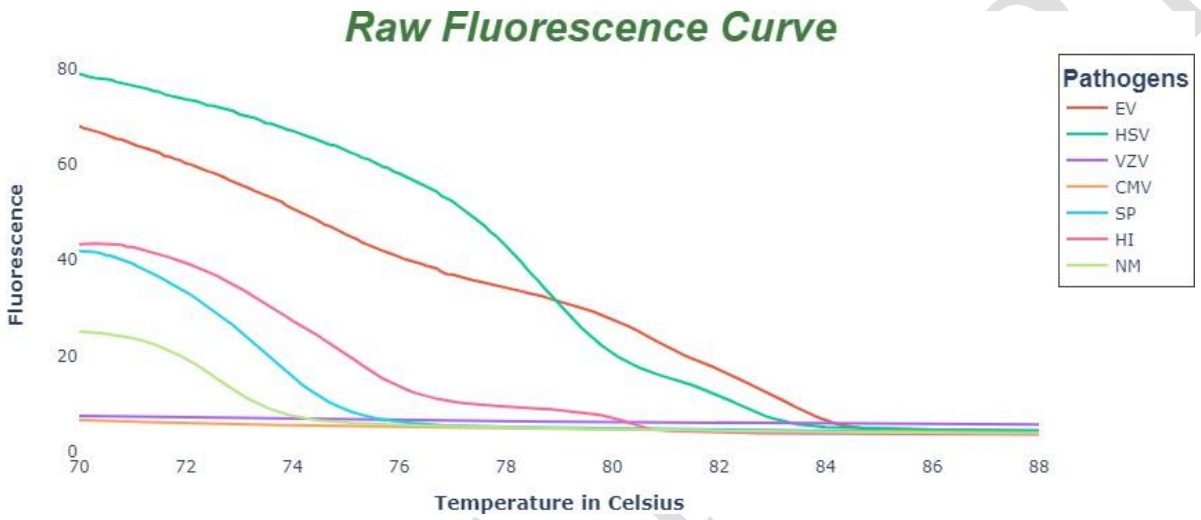


Figure:32 Before smoothening

7.3 SAVITZKY-GOLAY SMOOTHENING

Savitzky-Golay filter has emerged as a powerful tool for smoothing and differentiating noisy data. The Savitzky-Golay filter operates by fitting a polynomial to a window of adjacent data points. The width of this window and the degree of the polynomial are adjustable parameters that allow users to fine-tune the filtering process based on the characteristics of the data under analysis. Through this polynomial fitting, the filter effectively smooths the data, reducing noise and preserving essential features such as peak heights and widths. It helps to obtain a clearer representation of underlying trends and patterns.

The key advantages of employing the Savitzky-Golay filter include its ability to preserve important data features while effectively reducing noise. Unlike some traditional smoothing techniques that may over smooth or distort the data, the Savitzky-Golay filter strikes a balance between noise reduction and feature retention. Moreover, it is robust against outliers and performs well even with data of varying sampling intervals.

1. Smoothing Operation:

- Given data points y_i for $i = 1,2,3 \ldots . N$, where N is the total number of data points.
- We want to find the smoothed value $\hat{y}_i$ at a point y_i using a window of size m (an odd number) and a polynomial of degree n.

Equation

$$\hat{y} = \sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}} c_i y_{i+j}$$

  - c_j are the filter coefficients determined using least squares fitting.
  - These coefficients depend on the window size m and the degree of the polynomial n.

2. Coefficient Calculation:

- To compute the filter coefficients c_j, we solve the system: $Y = XC$

  Where**,**

  Y is a column vector of data points within the window.

  X is a matrix with rows representing powers of indices.

  C is a column vector of filter coefficients.

$$C = (X^T X)^{-1} X^T Y$$

3. Differentiation Operation:

- The Savitzky-Golay filter can also perform numerical differentiation.
- The derivative at a point y_i is computed as:

$$\frac{d\hat{y}_i}{dx} = \sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}} c_i \frac{d}{dx} y_{i+j}$$

Where**,**

$\frac{d\hat{y}_i}{dx}$ -This represents the derivative of the data point y_i with respect to x, which is what we want to compute

$\sum_{j=-\frac{m-1}{2}}^{\frac{m-1}{2}}$ -This is a summation over the window of size m centered around the data point yi. It iterates through all the neighbouring data poAhints within this window

$\boldsymbol{cj}$	-These are the filter coefficients determined using the Savitzky-Golay filter, which are used to weight the contributions of each neighbouring data point to the derivative calculation.
$\frac{\boldsymbol{d}}{\boldsymbol{dx}}\boldsymbol{y_{i+j}}$	-This represents the derivative of the data point y_{i+j} with respect to x. It's the derivative of each neighbouring data point within the window.

After utilizing *savogal filter* the raw flurocense data converted derivative melt plot. Now it is easy to interpret the melt signals and extract the melt features. The melt plot looks like below after savogal filter used,

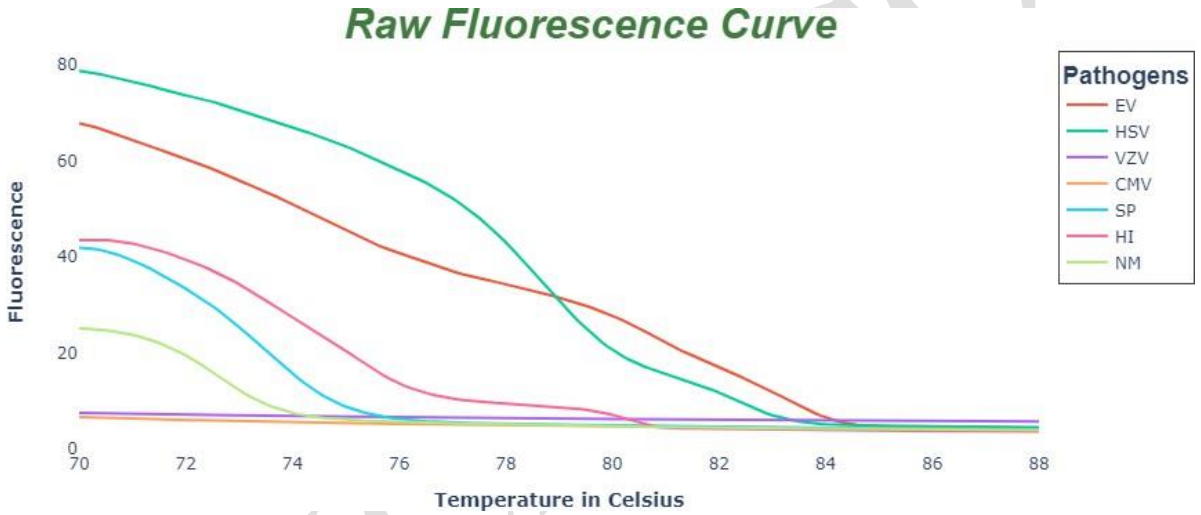


Figure:33 After Smoothening by Savitzky-Golay Filter

7.4 MELT CONVERSION

Melt conversion serves as a crucial analysis aspect in identifying the peak Tm of pathogens. To conduct this analysis, HRM data undergo differentiation. Within the derivative melt plot, each pathogen exhibits various peaks, including 'single peak,' 'double peak,' and 'multiple peaks' (noise). If a pathogen reaches its predetermined Tm threshold range, it indicates the presence of the pathogen in patients' test samples. The formula for converting to derivative melt plot,

$$M(T) = \frac{dT}{dF}$$

Where,

- $M(T)$ = melt curve

- $\frac{dT}{dF}$ = rate of change of fluorescence intensity with respect to temperature

7.4 B-SPLINE INTERPOLATION

B-Spline interpolation is a widely used technique in signal processing for approximating and reconstructing continuous signals from a set of discrete data points. B-Splines, short for basis splines, are piecewise polynomial functions defined by their degree and knots. The degree of a B-Spline determines the order of continuity, with higher degrees yielding smoother curves. The knots are the points where the polynomial pieces connect, defining the boundaries of the spline's segments. In signal processing, B-Spline interpolation is advantageous due to its ability to provide a flexible and smooth reconstruction of signals, particularly useful when dealing with noisy or irregularly sampled data. By adjusting the degree of the B-Spline and the spacing of the knots, engineers can tailor the interpolation to suit various signal processing applications, including image processing, audio signal reconstruction, and data smoothing.

One notable property of B-Spline interpolation is its local support, meaning that the influence of each data point on the interpolated result is confined to a limited region determined by the degree of the spline. This property allows B-Spline interpolation to avoid the global oscillations that may occur with other interpolation methods, making it particularly suitable for preserving the characteristics of the original signal while minimizing artifacts. Additionally, B-Spline interpolation is computationally efficient, with algorithms available for both uniform and non-uniform knot distributions, enabling its widespread application in real-time and offline signal processing tasks. Overall, B-Spline interpolation stands as a powerful tool in the signal processing toolbox, offering versatility, accuracy, and efficiency in reconstructing continuous signals from discrete data points.

7.5 MELT CONVERSION PATTERN

There is some another process to convert fluorescence data to derivative melt plot. There is some pattern existing Rotor Gene Q rex software while converting fluorescence data to melt signals. The x-coordinates values are interpolated with some patterns after the melt conversion in Rotor-Gene Q series software. By manually inspecting melt data file of Rotor Gene, the logic to interpolate the x-coordinates of HRM data, here is logic,

- Calculate the difference between the value in the first row and the value in the second row of the HRM_data.
- Set the start_value attribute to be the value in the first row plus half of the calculated difference.
- Set the end_value attribute to be the value in the last row minus half of the calculated difference.
- Set the space attribute to be one-third of the calculated difference.

ALGORITHM:2 ALGORITHM TO INTERPOLATE X-COORDINATES OF HRM

Input: *X-Coordinates of fluorescence data*
Output: *Interpolated X-Coordinates*
Initialization of objects: *Create empty list object "x-coordinates"*
FUNCTION *interpolation (column)*

> ***Difference*** ← *difference between first and second values of HRM x-coordinates*
> ***Start value*** ← *adding mid-point of difference to the first value of HRM x-coordinate*
> ***End value*** ← *subtracting mid-point of difference to the end value of HRM x-coordinate*
> ***Interval*** ← *divide difference value by 3*
>
> **WHILE** *Start value <= End value* **DO**
>
> ***x-coordinates*** ← *append the Start value till the condition is unsatisfied*
> ***Start value*** ← *add Interval value till the condition is unsatisfied*
>
> **END WHILE**

RETURN ***x-coordinates***

FOR *i = 0* TO *len (dataframe.columns)* **DO**

> ***Interpolated Signals*** ← *Interpolate the signal values with the corresponding interpolated temperature values using scipy.interpolate.splrep() with a 's' value*
> ***Smoothened Melt Signals*** ← *append each signal columns of raw fluorescence to its corresponding columns of Melting signal data frame.*

END FOR

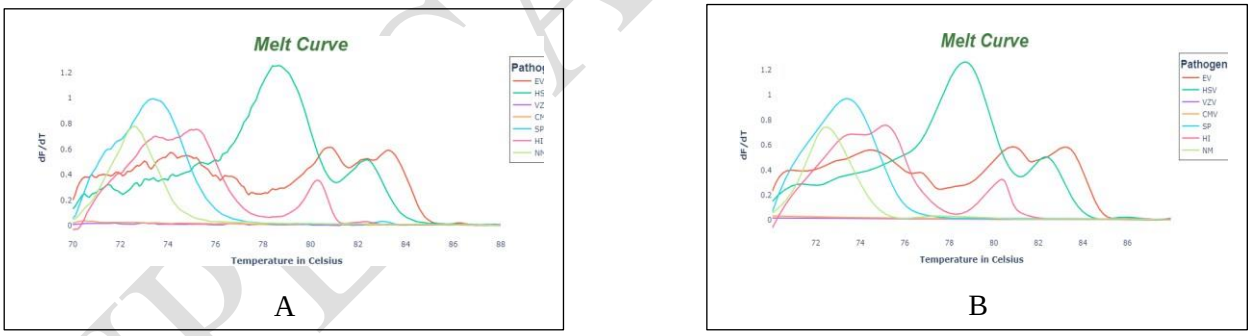
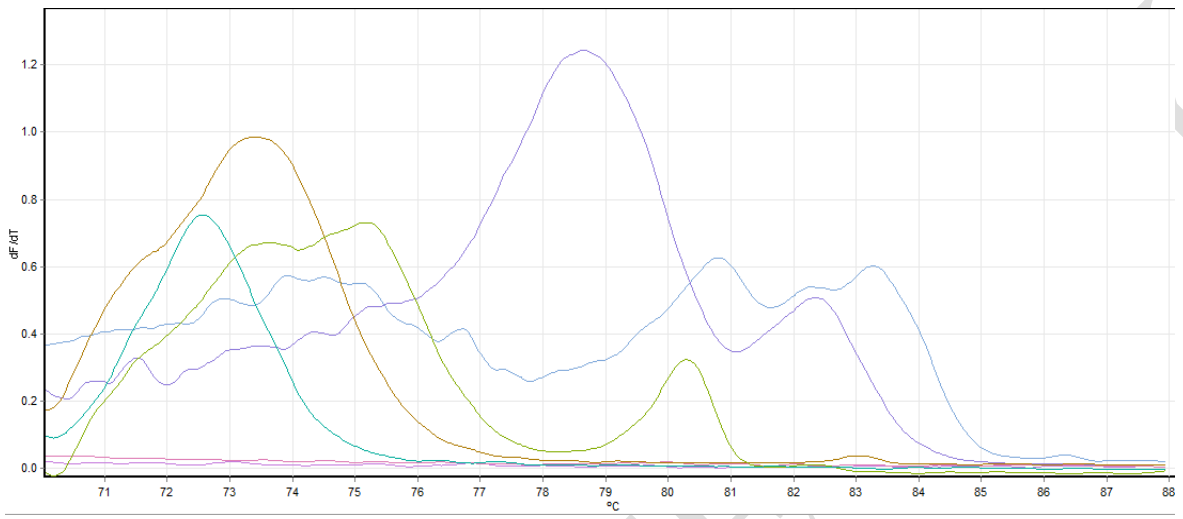


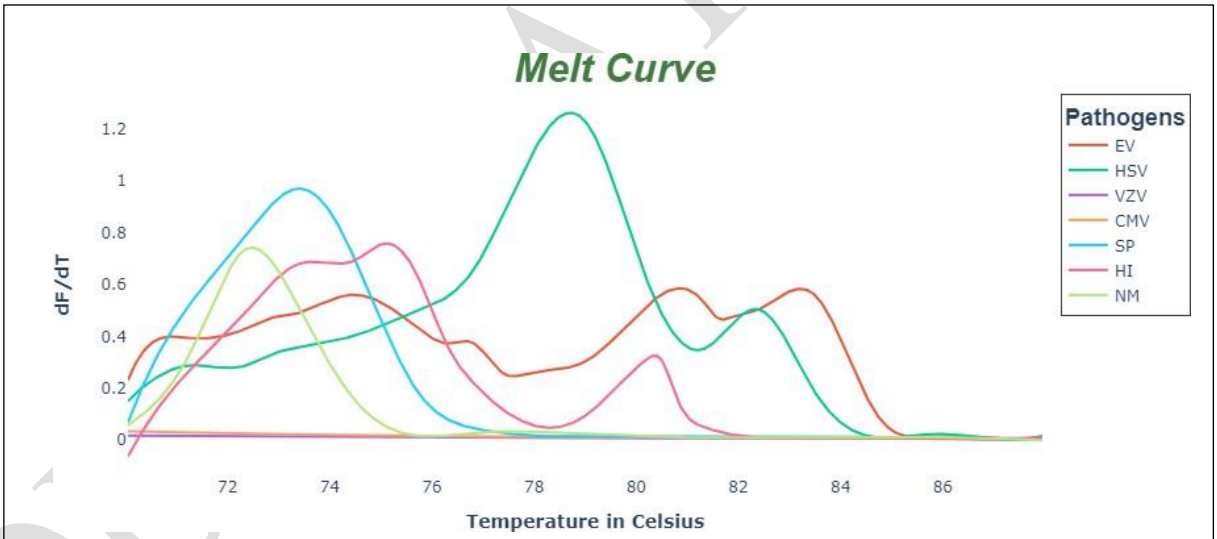
Figure:34 Melting Signal without Interpolation (A) and Melting Signal with Interpolation (B)

After applying interpolation to melt signals, it looks fine and there is no noise in Figure (B) compare to Figure (A).

After performing B-spline interpolation, it is necessary to check the similarity between manually plotted melt signals and machine-generated melt signals. Only by doing so, it will be easier to extract features from the melt signals and compare the peak melting temperatures (Tm) of machine-generated and manually generated melt signals.



Melt Signals Plot (A)
Generated by Rotor Gene Q-rex Software



Melt Signals Plot (B)
Generated using Python code

There is no significant difference between these melt signal plots. Which ensures that now it is ready to extract melt signal features.

7.6 SIGNAL PROCESSING FOR EXTRACTING MELT FEATURES

Peak: A peak is a point on a graph or signal where the value reaches a local maximum. In data analysis, peaks often represent significant features or events, such as the highest intensity in a spectrum or the maximum value in a signal.

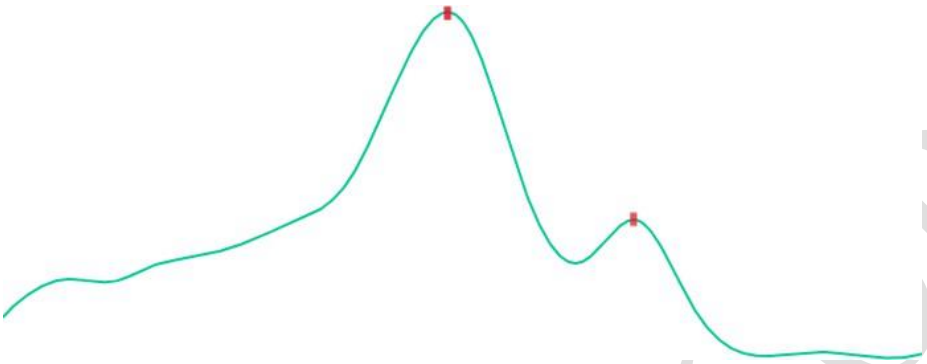


Figure:35 Peak identification

Finding peaks in signals presents a typical challenge in signal processing, aiming to detect and isolate significant features such as peaks or troughs within a given signal. A peak represents a local maximum or the highest point in the signal, while a valley indicates a local minimum or the lowest point in the signal. These algorithms are essential across various domains like image processing, speech recognition, and data analysis, often serving as a crucial initial step in analyzing data to unveil vital insights into the underlying signal generation process. Multiple techniques exist for peak detection, ranging from straightforward threshold-based methods to more advanced approaches like wavelet transforms and machine learning.

These methodologies vary in terms of accuracy, computational complexity, and resilience to noise and other signal irregularities. One prevalent method involves identifying local maxima through a sliding window in what is known as the "local maxima" approach. Another popular method, the "derivative-based" approach, entails computing the signal's derivative and detecting peaks where the derivative changes its sign. The peak-finding algorithm function operates on a 1-D array, identifying all local maxima through straightforward comparison of neighbouring values.

Peak Prominence: Peak prominence is a measure of how much a peak stands out from the surrounding baseline of a signal. It is defined as the vertical distance between the peak's highest point and the lowest point of the curve that connects the peak to the nearest higher peak. Essentially, it quantifies how much a peak rise above its surroundings.

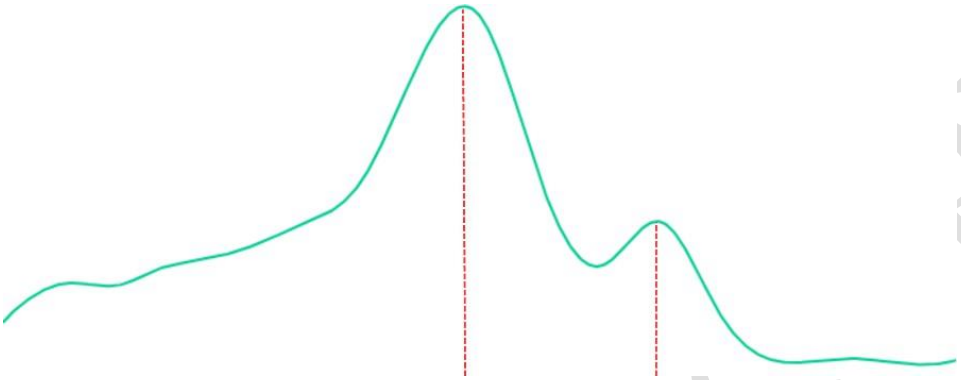


Figure:36 Peak Prominence

Signal peak prominence represents the relative height of a peak within a signal in relation to its neighbouring peaks, offering valuable insights for signal processing and analysis. This metric indicates the significance of a peak within the entire signal, defined as the vertical distance between the peak and the lowest point on the curve connecting adjacent higher peaks. Essentially, it denotes the minimum height needed to descend from the peak to a higher neighbouring peak or the signal's baseline. Widely utilized in peak detection algorithms, signal peak prominence aids in filtering out noise and pinpointing the most noteworthy peaks.

Peaks with greater prominence are typically more pertinent to the underlying signal-generating process than those with lower prominence. Moreover, prominence serves to compare and quantify variations between peaks across different signals or datasets. For instance, it enables comparison of peak amplitudes in EEG signals originating from distinct brain regions or assessment of peak intensities in spectra derived from diverse chemical compounds.

Peak Width: Peak width refers to the horizontal distance between the points where the signal reaches half of its maximum amplitude on either side of the peak. It provides information about the spread or width of the peak in the domain of the signal.

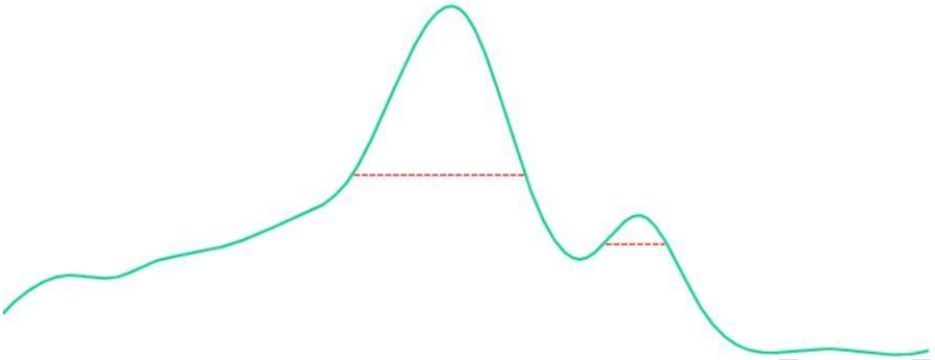


Figure:37 Peak width

Signal peak width denotes the extent of a peak within a signal along the x-axis, typically representing time or frequency. This parameter holds significance in signal processing and analysis as it offers insights into the duration or dispersion of a specific phenomenon depicted by the peak. Peak width plays a crucial role in peak detection algorithms, particularly for discerning closely positioned or overlapping peaks. In such scenarios, peak width aids in distinguishing between individual peaks and noise or artifacts within the signal.

Various methods exist for quantifying peak width, depending on the signal's characteristics and the intended application. One commonly used measure is the full width at half maximum (FWHM), representing the peak's width at half of its maximum amplitude. Another prevalent measure is the peak width at the baseline, indicating the width of the peak at a specific fraction of its maximum amplitude.

By utilizing the peak width, one can compute the take-off and touch-down points by referencing the start and end positions of the width. Additionally, the peak's width is determined based on its relative height (Peak Prominence), which is typically set to 100%.

Peak Take-off Point: The peak take-off point is the point on the rising edge of a peak where the signal begins to deviate noticeably from the baseline. It marks the onset of the peak's ascent and is often used to define the start of a peak's measurement.

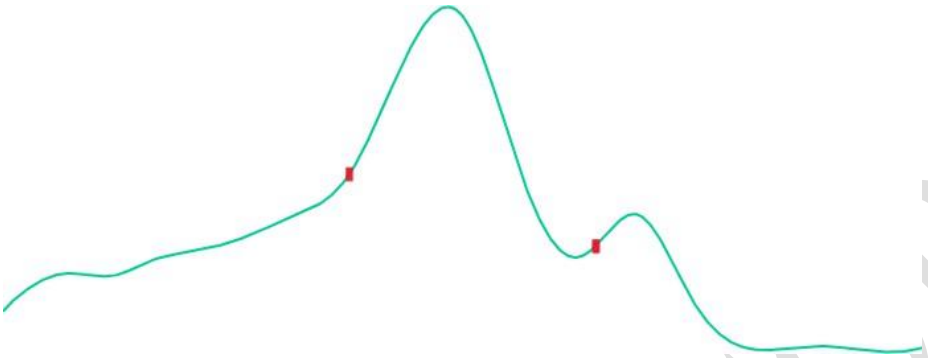


Figure:38 Take off point

Peak Takedown Point: The peak takedown point is the point on the falling edge of a peak where the signal returns to the baseline level after reaching its maximum value. It marks the end of the peak's measurement and is often used to define the peak's duration.

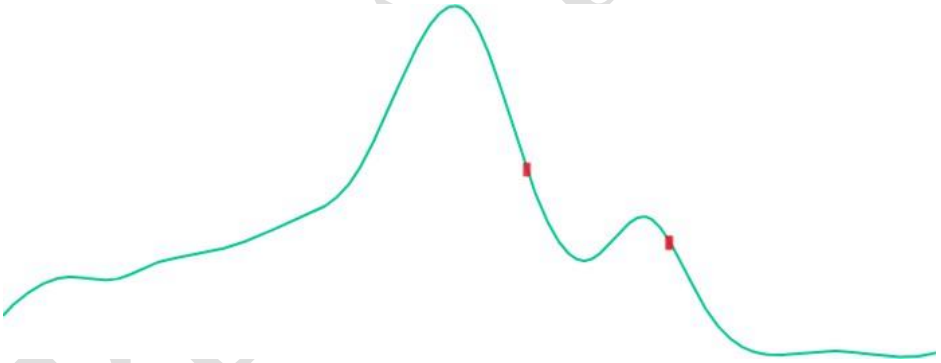


Figure:39 Take down point

Peak Area Under the Curve: The peak area under the curve is the integral of the signal within the boundaries of a peak. It represents the total quantity or magnitude of the phenomenon that the peak signifies. Integrating the signal over the peak's width provides a measure of the total signal intensity or concentration within that peak.

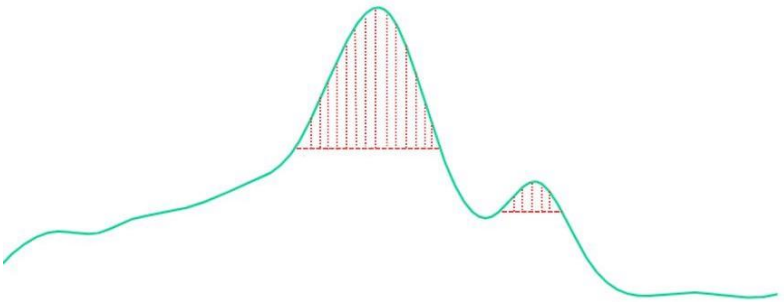


Figure:40 Area Under the Curve

The concept of the area under the curve in signal processing and analysis is foundational, representing the cumulative energy or power of a signal across a defined time frame. This principle finds widespread application in diverse fields like physics, engineering, biology, and finance. Calculating this area yields crucial insights into the signal, including its mean, variance, and distribution, while also serving as a valuable tool for detecting anomalies and patterns. Such insights have practical implications, enabling tasks like signal classification, denoising, and compression. Various methods exist for computing the area under the curve, ranging from numerical integration to techniques like Simpson's rule and Monte Carlo integration.

Simpson's Rule is a numerical integration method employed to estimate the area beneath a curve. It involves approximating the curve with a series of parabolic functions and subsequently computing the area of each resulting parabola. The formula for Simpson's Rule approximates the integral of a function over an interval [a, b] as follows:

$$\int [a,b] f(x)dx \approx \frac{b-a}{6} * [f(a) + 4f(a+b)2 + f(b)]$$

Here, f(x) represents the function to be integrated, [a, b] denotes the integration interval, and (a+b)/2 is the midpoint of the interval. Simpson's Rule essentially divides the interval [a, b] into subintervals of equal width, approximates the curve over each subinterval using a parabolic function, and then sums the areas under these parabolas to approximate the total area under the curve. Simpson's Rule is particularly renowned for its accuracy, especially when dealing with smooth functions with relatively simple shapes. However, its effectiveness may diminish when applied to functions exhibiting sharp changes or irregularities in their shape.

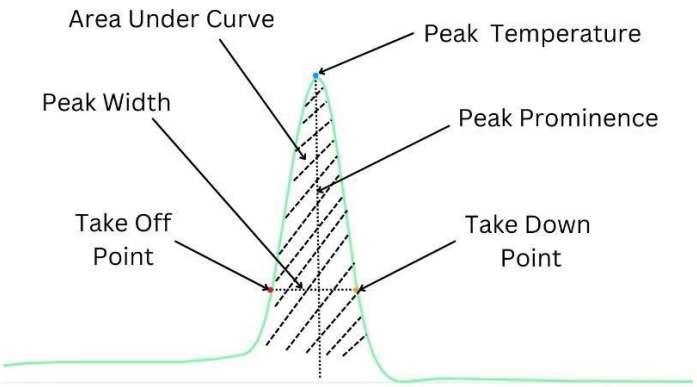


Figure:41 Calculate All features from DNA Melt Signal

In terms of peak detection and feature extraction through signal processing techniques, notable outcomes have been achieved. This approach effectively identifies various features within a DNA melting signal, such as peak prominence, width, and area under the curve. Nonetheless, it's worth mentioning that the peak detection process operates without employing any thresholding. Consequently, all peaks present in the signal are captured, regardless of whether they are relevant for analysis. This underscores the importance of thresholding in eliminating unwanted signals, a critical aspect in this context.

Figure:42 Noisy Signals detected peaks

Since no DNA melting is seen, the negative signal in Figure should not be taken into consideration. However, using signal processing techniques has resulted in the minute noises being incorrectly caught as peaks. It is necessary to apply appropriate thresholding algorithms to these signals to eliminate undesirable and noisy peaks before analyzing them. In the current setting, thresholding has been carried out visually and manually by configuring the prominences of DNA melting signals. A sufficient degree of prominence will be established, meaning that only peaks above this threshold will be considered, any other peaks will be discarded as undesired or noisy signals.

Since it is an AI-based framework designed to operate independently of human intervention, appropriate logic must be used to determine self-thresholding levels, ensuring that only necessary peaks are captured for analysis.

7.7 NOISE SIGNALS REMOVAL

In the existing method, our seniors used a vision-based approach to remove unnecessary noisy signals. However, this method requires many noisy images to retrain the convolutional model, making it inefficient for classifying noisy signals. It classifies the genuine peak and double-peak signals as noise. Dut to this issue, there is some positive case pathogens are misclassified as noisy signals. To overcome this issue, need to find the efficient solution in feasible way. As previously said, thresholding logic needs to be very appropriate and designed to identify only real signals. Peaks with sound prominences in most of the melting signals will be selected for examination. All other factors, such as melting temperature, are occasionally considered, even little peaks with lesser prominences.

There could be "Single-peak," "Double-peak," or "Noise signals" in each melt signal. The threshold for peak prominences must be established to only consider necessary peaks. Only the peaks that are 20% and higher than the maximum peak height in the melt signal will be regarded as true peaks if there are several peaks. There are occasions when the second peak is considered in High Resolution Melt analysis. Thus, it is imperative to hold onto the required peaks. The necessary peaks feature only extracted after this thresholding logic.

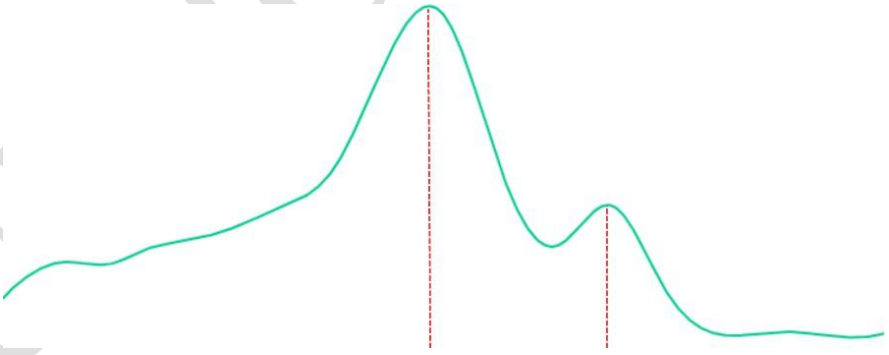


Figure:43 Measuring Prominences

In a prior project, the prominence value was determined at the beginning of the take-off point. Because of this, the actual height of the peak could not be determined from the ground line. Thus, it is necessary to derive the peak's prominence value from the ground line. Only top 2 highest peaks are considered for further analysis. After that, set the threshold to retain the necessary peaks.

7.8 LOGICAL BASED RESULT

laboratory technician uses Peak Tm values to identify pathogens in High Resolution Melt Analysis. There are matching peak Tm value ranges for each disease. The presence of the pathogen in the patient's body is indicated if its peaks fall within its associated Tm range. Based on the threshold data, the result is interpreted through logical conditions. The corresponding Tm ranges for each pathogen listed below,

Pathogen	Short Name	Lower bound	Upper bound
Enterovirus	EV	83	85
Herpes Simplex Virus-1	HSV-1	82	84
Herpes Simplex Virus-2	HSV-2	85	87
Varicella Zoster Virus	VZV	75	77
Cytomegalovirus	CMV	83	85
Streptococcus pneumoniae	SP	77	79
Haemophiles influenzae	HI	77	79
Neisseria meningitidis	NM	78	82

Table:4 Tm Threshold for Meningoencephalitis pathogens

ALGORITHM:3 LOGICAL CONDITIONS FOR DETECT THE RESULT OF PATHOGEN

***Input*:** *Features data, Threshold data*
***Output*:** *Result data frame*
***Initialization of objects*:** *Initialize an empty dictionary called '**result**' to store the detection results.*
FOR *each set of Tm1, Tm2, and column in the features data:*
  FOR *each set of min_threshold, max_threshold, and short_name in the threshold data:*
    IF *the **column** matches the **short_name:***
      IF *short_name is "**HSV**":*
        IF *Tm1 or Tm2 falls within the HSV thresholds:*
          *If yes, set the detection result to "**Detected**".*
          If no, check other HSV thresholds:
            *If any of the other HSV thresholds are met, set the result to "**Detected**".*
            *If not, set the result to "**Need Manual Interpretation (Check Tm Value)**".*
        ELSE:
          "Not Detected"
      IF *short_name is "**NM**":*
        IF *Tm1 or Tm2 falls within the NM thresholds:*
          *If yes, set the detection result to "**Detected**".*
          If not, check other NM thresholds:
            *If any of the other NM thresholds are met, set the result to "**Detected**".*
            *If not, set the result to "**Need Manual Interpretation (Check Tm Value)**"*
        ELSE:
          "Not Detected"
      IF *other short_names:*
        IF *Tm1 or Tm2 falls within the specified thresholds:*
          *If yes, set the detection result to "**Detected**".*
          If not, check if they are close to the thresholds:
            *If yes, set the result to "**Need Manual Interpretation (Check Tm Value)**".*
            *If not, set the result to "**Not Detected**".*
        ELSE:
          "Not Detected"
    ELSE
      Continue
  END FOR
END FOR

melt_result ← *Create a data frame using **result** dictionary*
RETURN melt_result

Pathogens are classified as *'Detected'* if they fall within the designated Tm range, otherwise marked as *'Not Detected'*. Results as *'Need Manual Interpretation (Check Tm value)'* when the Tm range is within one unit before or after the specified threshold.

RESULT AND DISCUSSION

The logical-based approach works based only the Tm range. The Laborations also consider the peak width, peak prominence of the melt signal. But there is no specified threshold for those parameters, they only assess those parameters by visually. Now the logical approach solely works on Tm range, so it doesn't consider the other features. Other features also play crucial role while interpreting the Melt Signal.

CONCLUSION

The decision based on the Tm range does not yield reliable results alone. Other features also play a role in determining the outcomes. Utilizing these additional features is necessary to accurately predict pathogen results. Future methodologies will address this need, ensuring more accurate predictions of pathogen outcomes.

CHAPTER 8
AMPLIFICATION CURVE

Amplification curve is a sigmoidal shape with three visually distinct apparent phases or regions. The first phase is near the baseline with a slow upward trend in the line. The second phase is a strong upward swing in the line. Finally, the third phase, is a plateau where the amplification signal tapers off and ceases to grow. Two points on the curve are of particular interest to us. One is the CT value, The crossing of this statistical noise threshold is the basis for calling a sample positive in a qualitative assay, and the cycle number at which it occurs is the basis for generation of a standard curve and quantitation of starting template in a quantitative PCR.

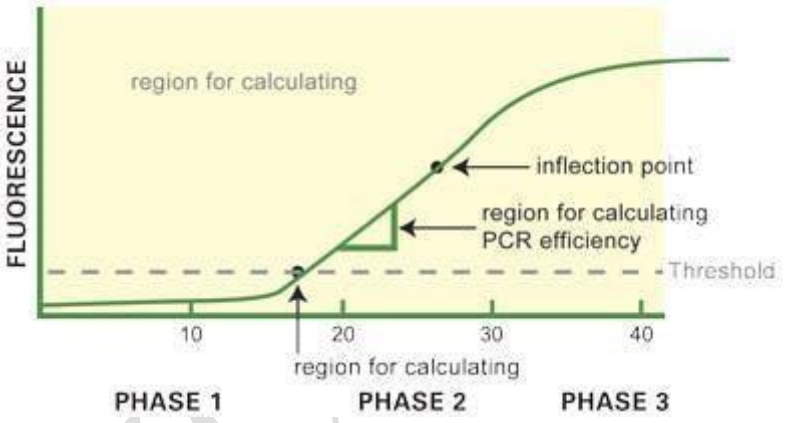


Figure:44 qPCR amplification curve shape

8.1 METHODOLOGIES

- Smoothening
  - Moving Average
  - Savitzky-Golay
- Baseline Subtraction
- Extracting the continuous positive values from the Rate of change
- Linear regression

8.2 MOVING AVERAGE

The moving average is a fundamental statistical tool used to analyse time-series data. It plays a crucial role in smoothing out short-term fluctuations, thereby revealing underlying trends and patterns that might otherwise be obscured by noise. By computing the average of neighbouring data points, it creates a more stable representation of the data, making it easier to identify long-term trends and make informed decisions based on reliable information. There are different variants of the moving average each with its own characteristics and applications. The choice of which moving average to use depends on the specific analysis and the desired emphasis on recent data points versus historical data. Here we used ***Three-point moving average***. This type of moving average calculates the average of three consecutive data points, where each data point is equally weighted. This moving average is often used in smoothing time series data to reduce noise and identify trends by creating a smoother representation of the data compared to the individual data points. The mathematical expression is as follows:

$$X_i = \frac{x_{i-1} + x_i + x_{i+1}}{3}$$

Where:

X_i - smoothed value at i^{th} cycle and

x_i - value at i^{th} cycle before smoothing.

8.3 SAVITZKY-GOLAY FILTERING

The filter effectively smooths the data, reducing noise and preserving essential features such as peak heights and widths. It helps to obtain a clearer representation of underlying trends and patterns.

The key advantages of employing the Savitzky-Golay filter include its ability to preserve important data features while effectively reducing noise. Unlike some traditional smoothing techniques that may over smooth or distort the data, the Savitzky-Golay filter strikes a balance between noise reduction and feature retention. Moreover, it is robust against outliers and performs well even with data of varying sampling intervals.

A

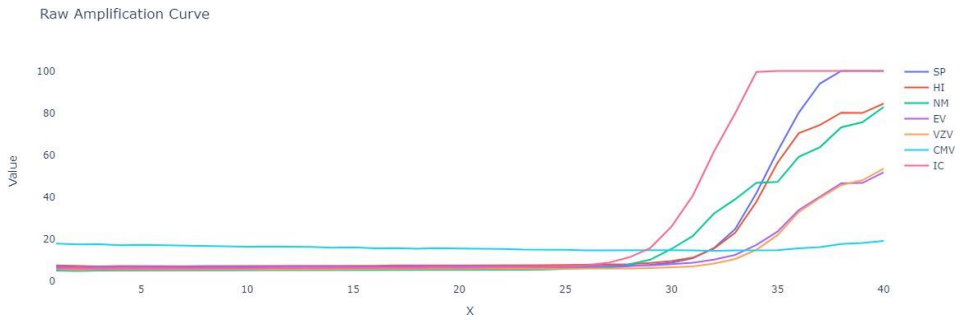


Figure:45 Raw Amplification curve

B

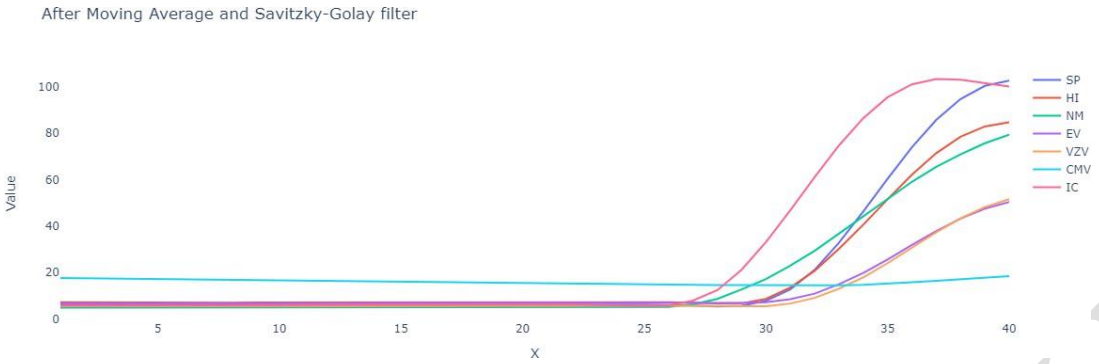


Figure:46 Digital Filter Applied Data

The above image-**B** represents the curve obtained from raw data after applying the mentioned moving average method and the digital filter called Savitzky-Golay Filtering for the smoothing process. This process increases the accuracy of the rising points.

ALGORITHM:4 ALGORITHM TO PRE-PROCESS DATA

Input: *Amplification curve data (data frame)*
Output: *Pre-processed amplification curve data (ct)*
Import necessary libraries: *NumPy, pandas, scipy.signal*
Initialize data frame object*: ct to store pre-processed data*
FUNCTION *pre-process (raw data)*
    FOR *each pathogen in data.iloc[:, 0::3].iloc[0]:*
        Split pathogen string and extract the last word
        Append extracted pathogen to the pathogens list
    FOR *each column, pathogen in zip (data.iloc[:, 2::3].columns, pathogens):*
        Rename column in data to pathogen
        Drop columns from data.iloc[:, 0::3] and data.iloc[:, 4::3]
        Rename the X column to "X" in the data frame:
        Rename column x column name in data to "X"
RETURN *Pre-processed data*
FUNCTION *moving average (Pre-processed data):*
    Initialize an array moving average with zeros of length equal to data
    Calculate moving average using the moving average formula
    FOR *each column in columns [1:]:*
        Calculate moving average of the column data
        Normalize the moving average data
        Update the column data in the data frame with the normalized data
RETURN *smoothened data*
FUNCTION *savogal (smoothened data):*
    FOR *each column in columns [1:]:*
        Apply Savitzky-Golay filter
        Update the column data in the data frame with the smoothed data
RETURN *the pre-processed CT data frame*

8.4 BASELINE SUBTRACTION

The baseline is the noise level in early cycles, typically measured between cycles 3 and 15, where there is no detectable increase in fluorescence due to amplification products. The number of cycles used to calculate the baseline can be changed and should be reduced if high template amounts are used. Set the baseline so that growth of the amplification plot begins at a cycle number greater than the highest baseline cycle number. Baseline subtraction involves establishing a baseline signal level in the early cycles of the PCR reaction, typically before significant amplification occurs. This baseline signal is then subtracted from the fluorescent signal obtained in subsequent cycles. The purpose of this subtraction is to eliminate background noise and non-specific signals, thereby improving the accuracy and reliability of CT values. The mathematical expression is as follows:

$$X_i = x_i - x_{min}$$

Where:

X_i - value at i^{th}cycle after baseline subtraction,

x_i - value at i^{th}cycle before baseline subtraction and

x_{min} - minimal value in the cell through the whole PCR run.

ALGORITHM:5 ALGORITHM TO BASELINE SUBTRACTION

***Input**: Digital filter applied Amplification data*
***Output**: Baseline subtracted data*
FOR *column IN ct.columns[1:]:*
    mean value = mean of data [0:15]
    normalized data = (ct[column] / mean value - 1) / 10
    ct[column] = normalized data
RETURN *Baseline subtracted data*

THE SIGNIFICANCE OF BASELINE SUBTRACTION CAN BE SUMMARIZED AS FOLLOWS:

- **Noise Reduction**: By subtracting the baseline signal, you remove background noise and non-specific fluorescence, leading to more accurate CT values. This is crucial for distinguishing true amplification signals from background fluctuations.
- **Improved Sensitivity**: Baseline subtraction enhances the sensitivity of the assay by allowing the detection of low-level target nucleic acids that may be obscured by noise without this correction.
- **Data Consistency**: Consistent baseline subtraction ensures that CT values obtained from different samples or experiments are comparable and reliable. It standardizes the analysis process, making your results more robust and reproducible.

- **Enhanced Interpretation**: Clear baseline-subtracted amplification curves provide a visual representation of target amplification dynamics, aiding in the interpretation of results and the identification of any issues such as inhibition or poor amplification efficiency.

The following figure represents the output after applying moving average and Savitzky-Golay filtering. The figure exhibits noisy signals in the baseline, which can adversely affect the accuracy of the rising points in the amplification curve.

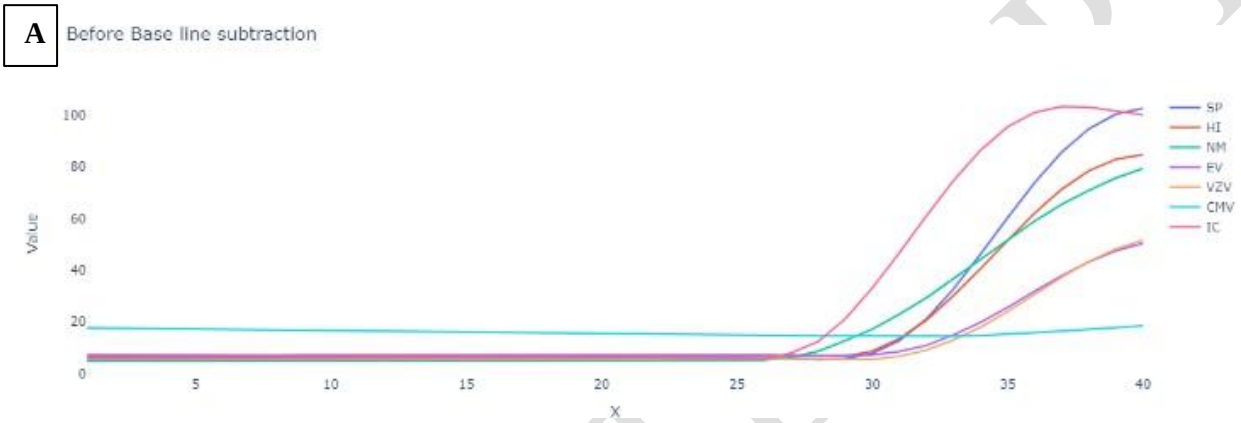


Figure:47 Before Baseline subtraction

The inability to extract precise CT features from the data processed by the digital filter is due to the fact that phase 1 of the amplification curve does not maintain a consistent level. To address this challenge, baseline subtraction was implemented on the CT data after applying the digital filter. This approach aids in normalizing phase 1 of the amplification curve, thereby enhancing the accuracy of the results.

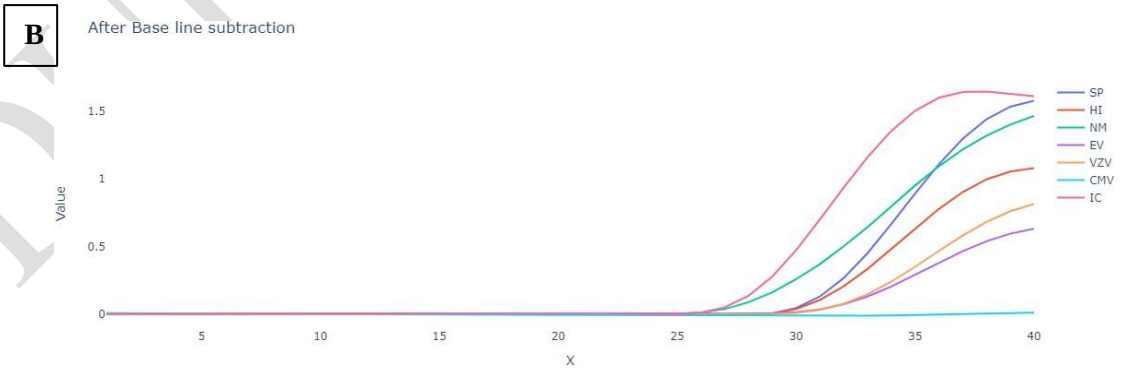


Figure:48 After Baseline subtraction

After baseline subtraction, all curves in phase 1 were normalized to obtain accurate results for the CT features. This normalization process aids in extracting exact values, similar to those obtained using the Rotor Gene Q-rex software.

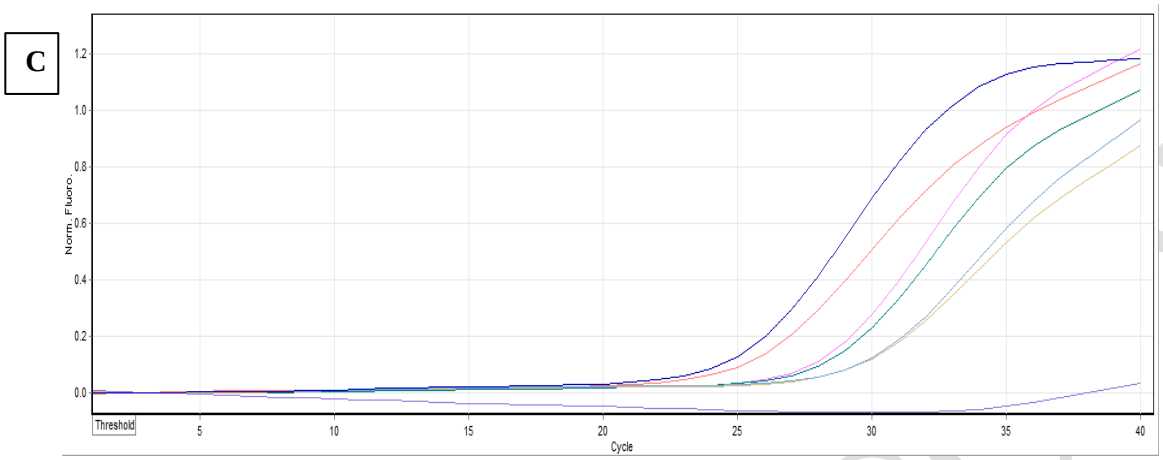


Figure:49 The original Amplification curve obtained from the Rotor-Gene Q Rex Software

The above figure serves as a reference for the curve obtained from proprietary software, demonstrating its similarity to our previously obtained results.

8.5 AMPLIFICATION CURVE AND DIFFERENTIATED VALUES

An amplification curve typically exhibits an initial phase characterized by low fluorescence levels, followed by an exponential growth phase where fluorescence increases rapidly, eventually plateauing in a saturation phase. Differentiating the amplification curve yields a series of differentiated values, reflecting the rate of change in fluorescence between successive cycles. Positive values in this series correspond to an increase in signal, indicating ongoing target amplification, while negative or zero values suggest that arising points are not reached. The objective of identifying continuous positive values within the differentiated values series is to pinpoint segments of the amplification curve where target amplification has its take of points

8.6 APPROACH

- To find the rate of change of processed Amplification Curves
- Find out the continuous positive values from the rate of change values

ALGORITHM:6 ALGORITHM TO FIND CONTIOUS POSITIVE VALUES

***Input**: differentiated values (list of values)*
***Output**: continuous positive values (list of tuples representing continuous positive sequences)*
FUNCTION *find continuous positive values (differentiated values):*
    ***Initialize** continuous positive values as an empty list*
    ***Initialize** current sequence length as 0*
    ***Initialize** start index as None*
    FOR *i FROM 0 TO len (differentiated values) - 1:*
        IF *differentiated values[i] > 0 THEN*
            IF *current sequence length == 0 THEN*
                Set start index to i
                Increment current sequence length by 1
        ELSE*:*
            IF *current sequence length > 0 THEN*
                IF *current sequence length >= min sequence length*
                    Append *tuple (start index, i - 1) to continuous positive*
                Set *current sequence length to 0*
                Set *start index to None*
    END FOR

    IF *current sequence length >= min sequence length THEN*
        Append tuple (start index, len (differentiated values) - 1) to continuous positive values
RETURN *continuous positive values*

STEPS

- Initialize variables to track the current continuous positive sequence length current sequence length.
- Iterate through the differentiated values series, examining each value to determine if it indicates positive amplification. If a positive value is encountered, update the current sequence length and start index accordingly.
- Upon encountering a non-positive value (zero or negative), evaluate if a continuous positive sequence exists based on the minimum sequence length criteria. If so, record the start and end indices of the sequence in `continuous positive values`.
- After processing all values, check if a continuous positive sequence extends to the end of the series. If present and meeting the minimum length requirement, include it in the final list of continuous positive values.

CONCLUSION

The algorithmic approach presented provides a systematic and efficient method for identifying continuous positive values in RT-PCR amplification curves. By delineating segments of sustained target amplification, this approach enhances the interpretability and utility of RT-PCR data in diverse biological and diagnostic contexts. The accurate identification of continuous positive values contributes significantly to the reliability and precision of RT-PCR-based analyses, underscoring its importance in molecular biology research and clinical diagnostics.

CHAPTER 9

APPROACH BASED ON FIXING THE REGRESSION LINE TO FIND RISING POINT

The previous approach, which involves identifying continuous positive values in the rate of change to find the rising point, faces several limitations. One major issue is the possibility of fluctuations within the amplification curve. These fluctuations can arise due to various factors such as noise in the data or minor variations in the underlying signal. As a result of these fluctuations, the calculations for the rising point can become inaccurate.

The approach relies on identifying the longest continuous positive values from the rate of change in the amplification curve. However, this method may not effectively filter out small changes or noise that occur within the rate of change. For instance, if there are minor fluctuations or noise present in the data, they may be mistakenly included in the calculations for the rising point. This inclusion of noisy signals can significantly impact the accuracy of the analysis, leading to misleading results.

Therefore, the limitations of this approach stem from its inability to effectively separate noisy signals from the actual amplification curve. This can result in inaccurate calculations and a less reliable determination of the rising point in the data. To overcome from these issues an alternative approach is used to find the rising point from the amplification curve using linear regression line.

9.1 STEPS

- Fix the Linear regression line in the actual values of the amplification curve
- Below the linear regression line of the actual curve find the maximum distance from the x-axis and fitted linear regression line

ALGORITHM:7 ALGORITHM TO EXTRACT FEATURES OF CT DATA

***Input**: CT Pre-processed data (data frame with columns)*
***Output**: CT Features (data frame with Pathogen, Take-off Point, Y-Coordinate, Status)*

FUNCTION *take off (CT data):*
Initialize empty lists: pathogens, take off points, y coordinates, status
Initialize data frame
FOR *column IN CT Pre-processed data.columns[1:]:*
X = Extract column 0 values from CT Pre-processed data as X (reshaped as (-1, 1))
y = Extract current column values from CT Pre-processed data as y
Initialize Linear Regression model
Fit the model with X and y
Predict y values using the model
Calculate difference as predicted value – actual value
Find take off index as index of max difference
Find positive values as indices where difference > 0
IF *y [take off index] >= 0 THEN*
Append column to pathogens list
Append X [take off index] to take-off points list
Append y[take off index] to y-coordinates list
IF y [-1] < 0.043 THEN
Append "Noise" to status list
ELSE:
Append "Normal" to status list

ELSE *IF y[take off index] < 0 THEN*
FOR idx FROM take-off index TO positive values [-1]:
IF y[idx] > 0 THEN
Append column to pathogens list
Append X[idx] to take off points list
Append y[idx] to y-coordinates list
IF y [-1] < 0.043 THEN
Append "Noise" to status list
ELSE:
Append "Normal" to status list
BREAK
***Create** data frame with columns 'Pathogen', 'Take of Point', 'Y-Coordinate', 'Status'*
RETURN *CT Features*

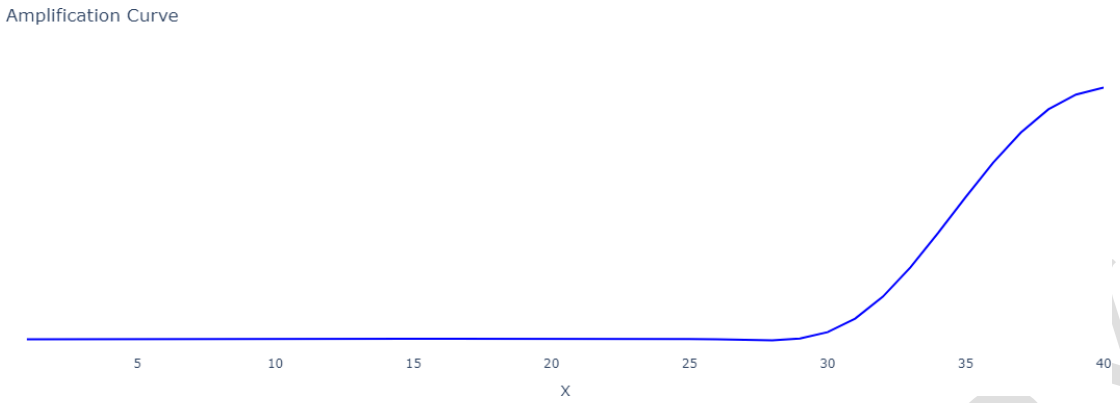


Figure:50 Amplification curve

This methodology operates on CT data, which is a output of Pre-processed CT data, obtained from the previous methodology. Within this approach, a loop iterates over each column of the CT data, excluding the first column, which typically holds sample identifiers. For each pathogen or target represented by a column, the function performs linear regression using the CT values as independent variables (X) and the corresponding amplification data as dependent variables (y). This regression models the amplification trend, helping to identify deviations from the expected behaviour.

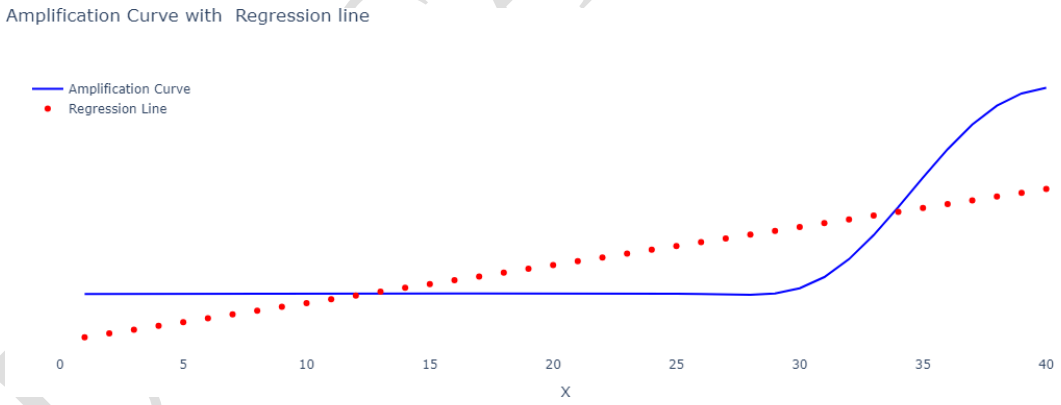


Figure:51 Amplification curve with linear regression line

After fitting the linear regression model, the function calculates predicted amplification values and computes the difference between predicted and actual values. The "take off index" is determined as the point with the largest positive difference, signifying a notable increase in amplification. If the amplification value at this index is positive, indicating a genuine amplification event, relevant information such as pathogen name, take-off point (cycle

number), actual amplification level (Y-coordinate), and a status label (Normal or Noise) are recorded. The status label depends on whether the final amplification value falls below a certain threshold (0.043 -here used), distinguishing between normal amplification and noise.

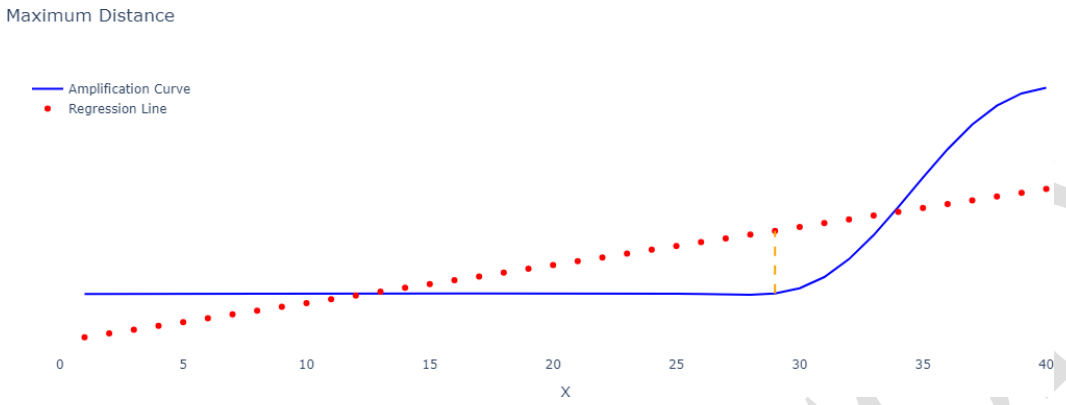


Figure:52 Maximum Difference between Actual and predicted line

Finally, the extracted information organized into a structured data frame ("CT result") with columns for Pathogen, Take-off Point, Y-Coordinate, and Status. This structured output provides a concise summary of significant events or deviations observed in the amplification curves of different pathogens, aiding researchers in interpreting RT-PCR test results from a data analytics perspective.

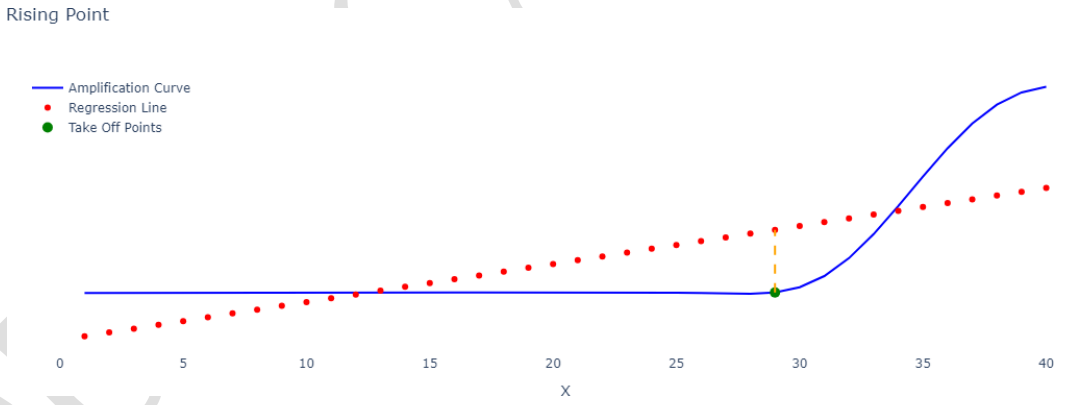


Figure:53 Amplification curve with Rising point

CONCLUSION

This method provides more closure value of the rising point of the Amplification curve which is more similar to the Rotor gene Q rex software which will more reliable value to make further analysis

CHAPTER 10

AN APPROACH ON MACHINE LEARNING MODEL TO PREDICT THE RESULTS OF PATHOGENS

In the previous approach, the logical conditions failed to produce reliable results. Low prominence peaks were erroneously identified as positive using this method. However, the prominence and width of the curves play a critical role in interpreting melt signals to predict pathogen results. Despite this, the quantity of positive data available for training machine learning models is significantly lower compared to negative data. Naturally, there are more instances of negative data since not all patients exhibit positive results. To train the Machine learning model the “Detected” and “Not Detected” classes should be balanced. But it’s not like that,

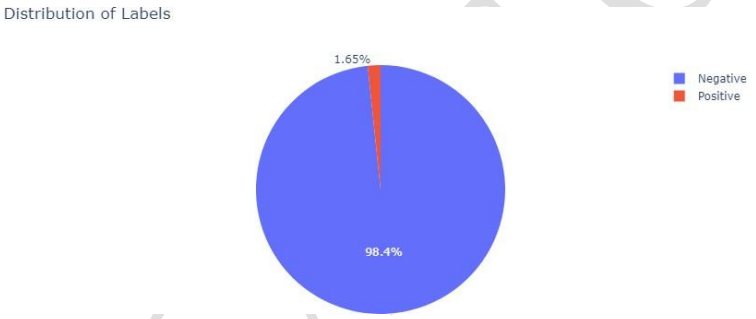


Figure:54 Distribution of the class

The figure above illustrates a highly imbalanced class distribution. To develop the machine learning model effectively, it is imperative to equalize the class labels. Upon manual inspection of the positive results, a specific pattern was discerned within the positive results dataset.

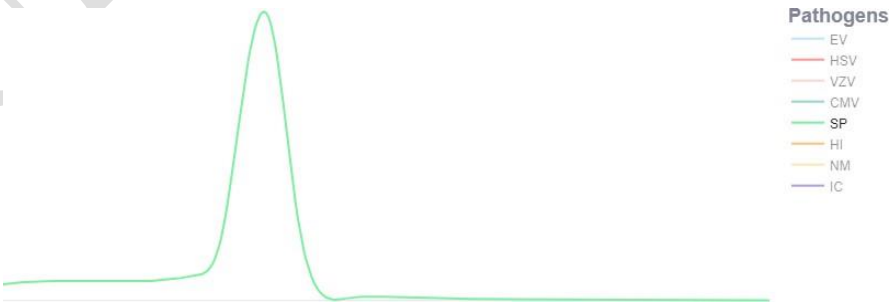


Figure:55 Streptococcus Pneumoniae DNA Melt Signal (Detected Case)

The above figure depicts the DNA melt signal of Streptococcus pneumoniae. This represents a detected case. The following are the features detected from the DNA melt signals,

Target	Temperature	Width	Prominence	Take off point	Take down point	Area under the Curve
SP	78.5667	34.9585	2.5855	77.2333	79.5667	4.2513

Table:5 Detected Features of Streptococcus Pneumoniae

In the provided table, the width of the DNA melt signals appears notably narrow, with the take-off and take-down points exhibiting a deviation of approximately 1 unit from the temperature. Through manual inspection of the positive result data, these specific patterns were identified.

Target	Temperature	Width	Prominence	Take off point	Take down point	Area under the Curve
HI	78.78333333	41.59756227	0.688485542	78.01666667	79.41666667	0.681286207
HSV	84.98333333	42.10716818	1.187007584	84.15	85.55	1.184505043
EV	82.71666667	147.3239213	0.993240299	78.51666667	83.41666667	2.616116744
CMV	84.45	32.12423826	4.600685537	83.88333333	84.95	3.452585678
CMV	84.56666667	26.57131188	1.188226922	83.7	85.43333333	1.492474833
CMV	84.43333333	26.92940362	1.485060318	83.5	85.3	1.91573981
CMV	84.96666667	26.80166254	1.99740973	84.03333333	85.83333333	2.496809773
EV	83.71666667	83.16210356	1.63088407	81.78333333	84.55	2.89543765
SP	78.85	73.31449402	0.216611822	77.45	79.91666667	0.389402059
EV	84.45	53.51726184	2.500981155	83.31666667	85.11666667	2.910686213
VZV	77.11666667	55.2398961	1.721596981	76.08333333	77.91666667	2.150378891
SP	78.71666667	60.67807177	2.525000625	77.65	79.68333333	3.510470322
CMV	85.55	31.73922926	3.814785915	84.98333333	86.05	2.811323563
CMV	85.41666667	46.71831045	2.461361635	84.51666667	86.08333333	2.631859611

Table:6 Detected Features of Positive result

The specific patterns identified through manual inspection are evident in the positive results data. In contrast, the negative results data do not exhibit such distribution. Specifically, the width of the negative data significantly surpasses that of the positive result data. Additionally, the take-off and take-down points display considerable deviation from the peak temperature.

10.1 SYNTHETIC DATA GENERATION

Due to the lack of positive data, couldn't able to predict the results of the pathogens, Machine Learning models required balanced class labels, then only it can learn the patterns of the class labels without any bias. Already the positive data patterns are found, so it's easy to synthesize the data for positive cases.

ALGORITHM:8 TO GENERATE SYNTHETIC DATA

Input: *Target name, number of samples to generate, Targets_threshold*

Output: *Synthetic data for positive class*

Initialization of Objects: *Create an empty List object as "list"*

Import necessary libraries: *numpy*

FUNCTION *Synthetic_data_generation (Target name, samples count to generate)*

    FOR *i=0 to range(samples count to generate)*

        Tm ← *numpy.random.uniform(Targets_threshold[Target name])*

        width ← *numpy.random.uniform(observed range from positive data)*

        prominence ← *numpy.random.uniform(observed range from positive data)*

        Take_off ← *numpy.random.uniform(observed range from positive data)*

        Take_down ← *numpy.random.uniform(observed range from positive data)*

        Auc ← *numpy.random.uniform(observed range from positive data)*

        Ct_value ←*numpy.random.randint(values within 13 to 30)*

        Result ← *"Positive"*

        list. append ([Tm, width, prominence, Take_off, Take_down, Auc, Ct_value, Result])

    END FOR

RETURN *list*

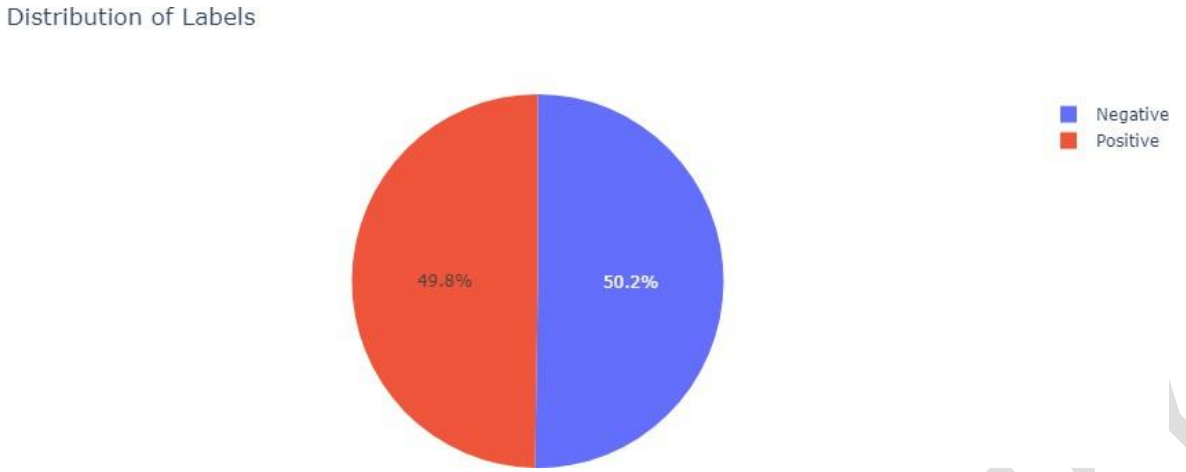


Figure:56 After Balancing the Class labels

Based on the specific pattern inspected, synthetic data has been generated and the class imbalance issue has been rectified. The generated data is provided below.

Target	Temperature	Width	Prominence	Take_off_point	Take_down_point	AUC	Ct_value	Result
SP	78.91666667	60.63850807	2.507138602	77.81666667	79.81666667	3.458802005	27	P
SP	78.68333333	65.97513349	3.353447639	77.51666667	79.71666667	4.938072589	21	P
CMV	85.31666667	32.05976313	5.159021721	84.75	85.81666667	3.825338694	25	P
HI	78.75	32.89237954	1.246639209	78.15	79.25	0.94799408	33	P
CMV	84.95833333	26.12081787	2.657524939	83.875	86.04166667	4.244418344	24	P
CMV	83.08333333	31.84287686	3.071800526	82.51666667	83.58333333	2.282632794	22	P
CMV	84.58333333	33.53513486	6.962858227	83.98333333	85.08333333	5.334156876	22	P
HSV	83.43333333	31.79183229	2.339806781	82.23333333	84.36666667	3.474065497	30	P
CMV	84.16666667	26.60683187	2.599883311	83.23333333	85.03333333	3.353508678	31	P
VZV	76.63333333	32.25901125	0.946064588	75.5	77.7	1.409512886	29	P
SP	78.3	32.2020761	1.53191781	77.1	79.23333333	2.293590372	20	P
CMV	84.58333333	30.12428188	6.573009739	84.05	85.05	4.627905851	18	P
SP	78.15	78.27849818	0.877041858	76.65	79.25	1.558190658	29	P
SP	78.41666667	64.30379939	1.110390743	77.35	79.48333333	1.621284018	16	P
EV	84.35	46.48476779	3.11140523	83.41666667	84.95	3.260646574	26	P
EV	84.87050185	57.29055231	1.227362981	83.8810333	85.13680567	3.756226278	22	P
EV	83.38238153	45.65947208	0.526650994	83.27788517	84.34513402	6.878385958	21	P
EV	84.94436094	81.83529236	2.024385495	83.65026517	85.96152707	2.324705213	25	P
EV	84.9174778	75.32501007	1.387664677	84.38745339	85.30153445	3.224888066	25	P
EV	83.16353043	69.79225273	3.896468984	82.54749111	84.73624859	2.121918265	28	P

Table:7 Generated positive data

10.2 RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

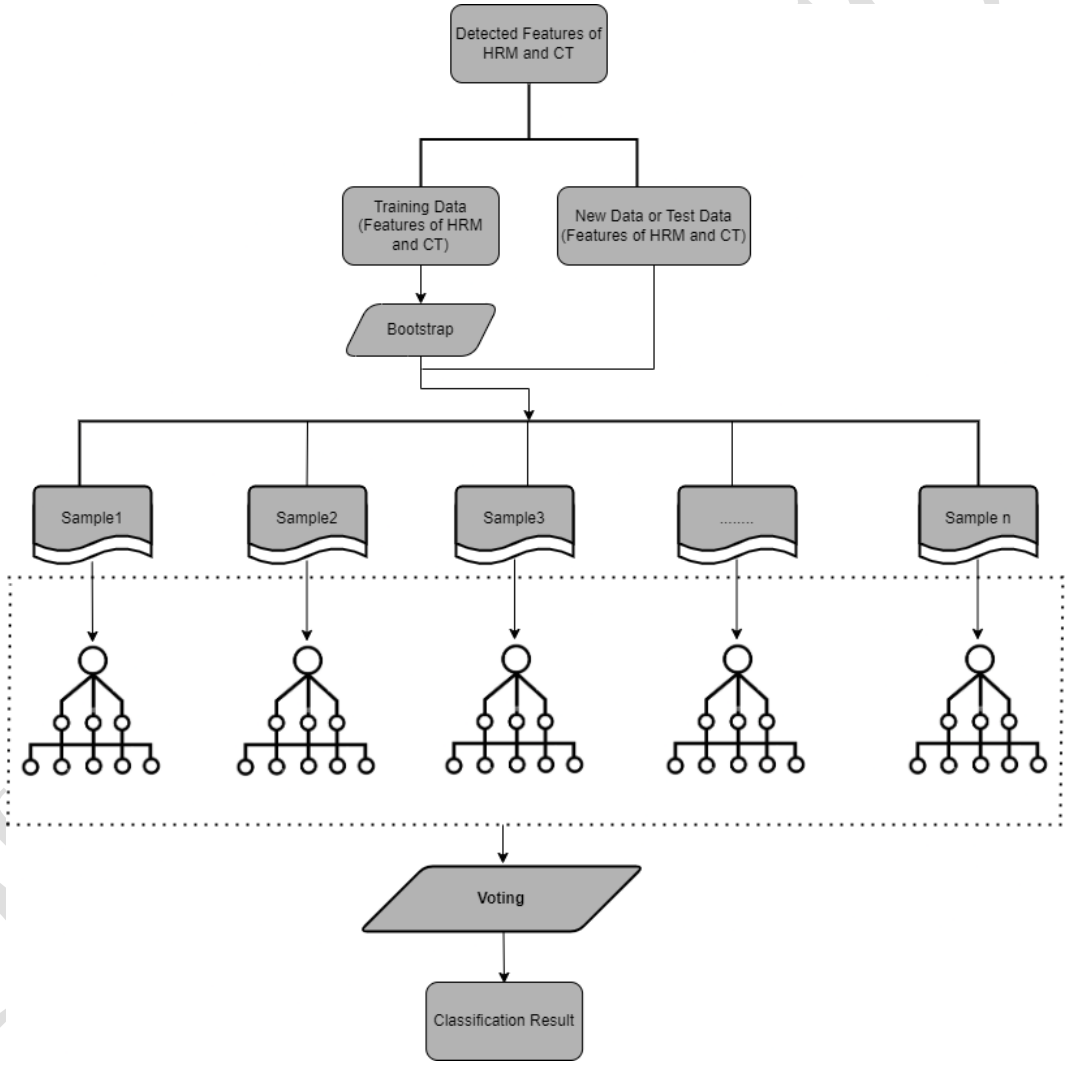


Figure:56 Random Forest Architecture

10.3 MODEL ACCURACY AND CLASSIFICATION REPORT

	Precision	Recall	F1-Score	Support
Negative	0.99	0.95	0.97	405
Positive	0.99	0.99	0.97	403
Accuracy			0.97	808
Macro Average	0.97	0.97	0.97	808
Weighted Average	0.97	0.97	0.97	808

Table:8 Classification report of Random Forest Model

The above classification report shows that the model predicts the unknown data with 97% accuracy.

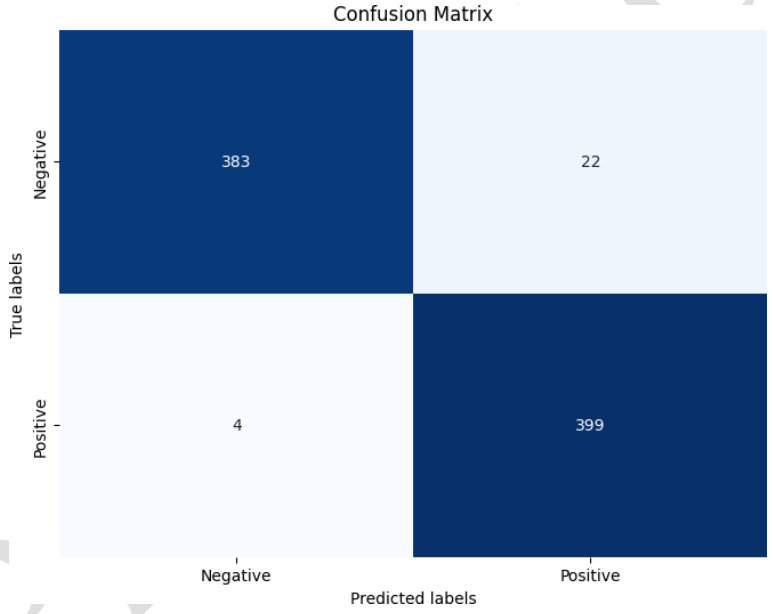


Figure:57 Confusion Matrix for Test data

The Above Confusion matrix show that the model classifies the pathogens with less false-positive rates.

RESULT AND DISCUSSION

The machine learning model-based approach provides reliable results with higher true-positive rates and accuracy. The ML model learns patterns from the given data, making it easy to classify pathogen results. However, since synthetic data is used, the exact patterns of positive data may not be learned by the machine learning model. It's necessary to obtain the original positive data patterns. There are some grey areas when interpreting pathogen results, so incorporating options to store exceptional patterns in the database is essential.

CONCLUSION

The machine learning model accurately predicts the pathogens' results with lower false-positive rates. The provided results are reliable when compared to the original ones. Thus, it's evident that predicting pathogen results without manual interpretation is feasible. This model facilitates result interpretation within seconds, thereby reducing laboratory time consumption.

CHAPTER 11
SYSTEM DESIGN & DEVELOPMENTS

11.1 COMPONENTS

For developing the AI-based framework for automated analysis, interpretation and data management for the HRM data, which is generalized and optimized, three major components were developed by the team are

- REXTRACTOR (Tool Data Extraction)
- PyMLRS (Python library for Feature engineering)
- PATHOGEN DETECTOR (Tool for Predicting the results)

The team has developed three major components for developing an AI-based framework for automated analysis, interpretation, and data management of High-Resolution Melting (HRM) data. These components include the data extraction tool called "REXTRACTOR," the feature engineering library called "PyMLRS," and the prediction tool called "Pathogen Detector".

The REXTRACTOR tool is used to extract data from the raw Rotor gene experiment files, while the PyHRM library used for feature engineering, which involves extracting relevant features from the Rotor Gene Experiment files. Finally, the Melt curve Interpreter tool Pathogen Detector uses predictive analytics and Machine learning models to interpret the extracted features and predict the presence of the intended molecular target in a clinical sample tested.

These three components work together to develop an AI-based framework that can automate the analysis and interpretation of HRM data, allowing for faster and more accurate diagnosis of infectious diseases. With this framework, clinicians can make more informed decisions andplan the course of treatment for their patients.

AI -FRAMEWORK

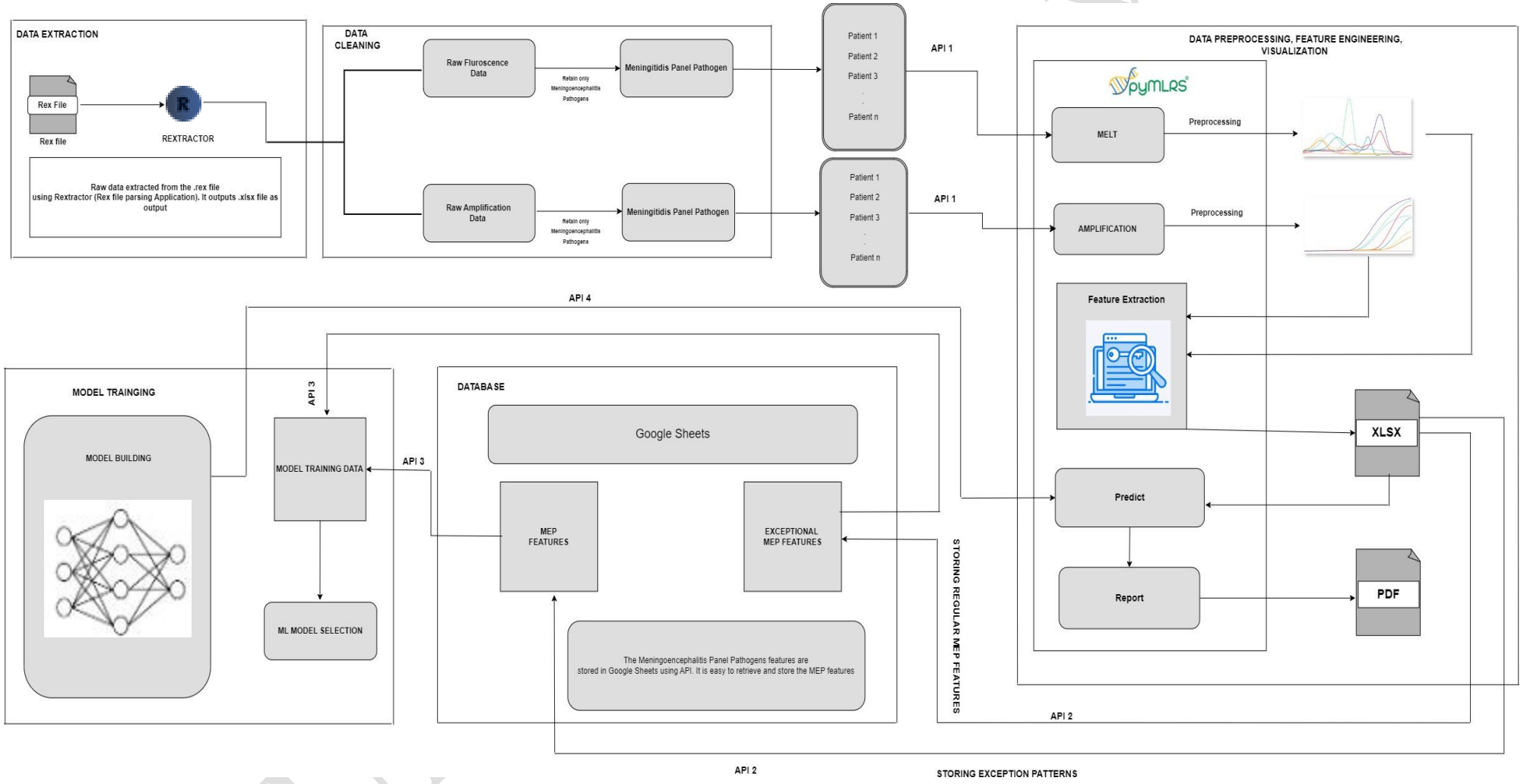


Figure:58 Work Flow for AI-Based Framework

11.2 REXTRACTOR

Rextractor is a dedicated application crafted to aid users of the Rotor-Gene PCR cycler in the extraction of raw data from ".rex" files generated by the Rotor-Gene Q-Rex Software Series. Rextractor utilizes a novel algorithm for extracting HRM and Amplification data from .rex files. This is notable because, thus far, there has been a lack of standalone software capable of converting .rex file formats into Excel file formats without the dependency of proprietary software (Rotor Gene Q-rex Software).

Figure:59 User interface of Rextractor

After the successful experiment ran in *Rotor-Gene Q* cycler, it produces the raw data and the users which can be only opened and analyzed via **Qiagen's Q-Rex Software**. If a specific run file (raw data) has to be exported into desired formats such as *text*(.txt), *HTML Table*(.html), *XML*(.xml), *excel*(.xls) given by the **Qiagen Rotor-Gene Q-Rex Software**. Here we automated the user role by our R**EXTRACTOR** Software, by which you simply put the raw data file directory and desired directory to which the excel files are stored in your system, which saves time and not to burned out from this repetitive task.

FRAMEWORK

The frame work for the Tool Rextractor is given as follows

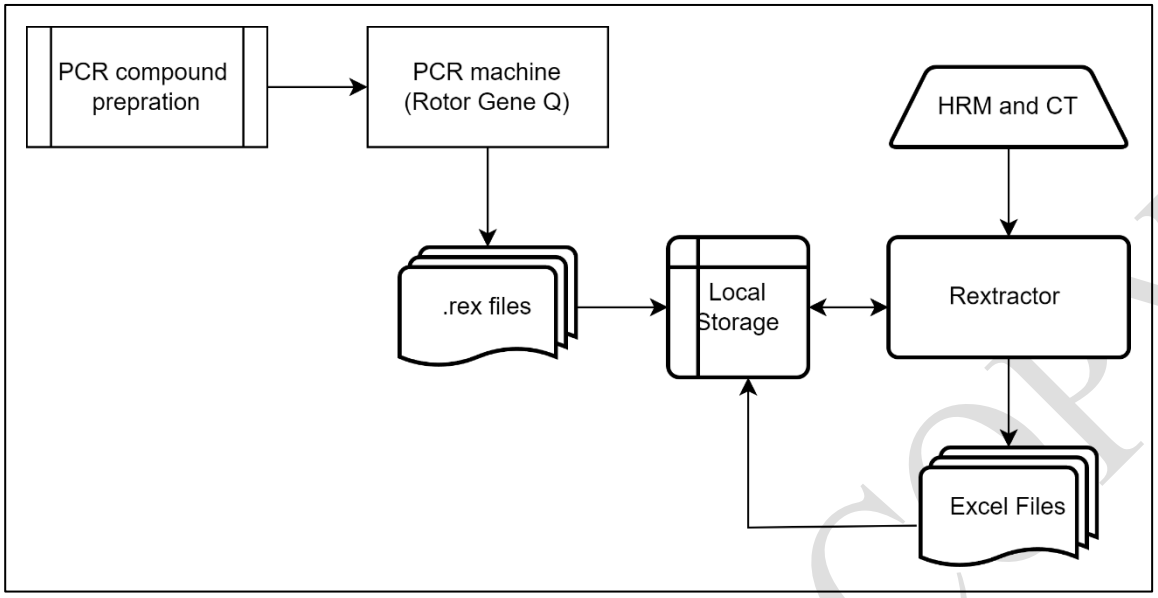


Figure:60 Framework of Rextractor

MANUAL CONVERSION

It is a time-consuming process and often introduce frustration and inconsistency, while converting bulks of raw data by manual process.

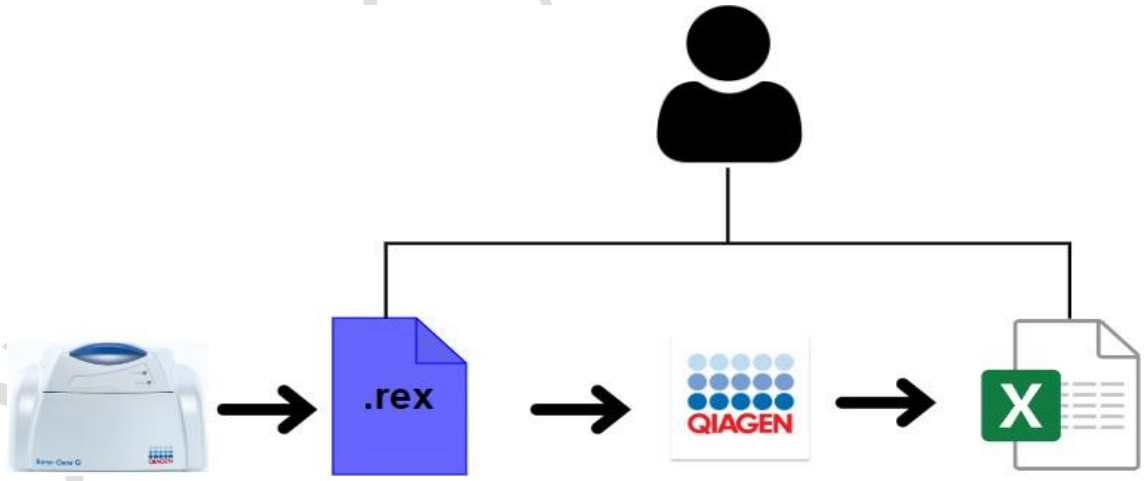


Figure:61 Tradition method to covert rex to xlsx

CONVERSION THROUGH REXTRACTOR

Rextractor efficiently extracts raw data from .rex files to .xlsx format in less than a second. It processes hundreds of .rex files in less than one minute

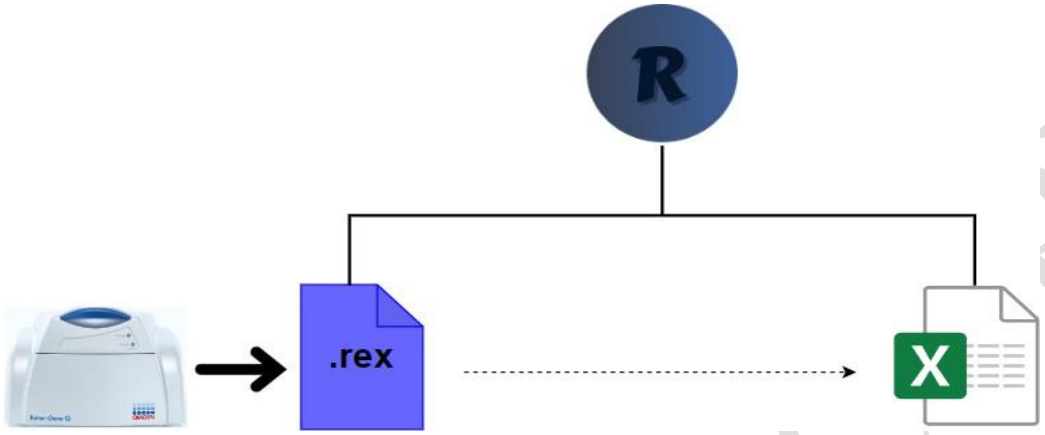


Figure:62 File Parsing method to convert rex to xlsx

TYPES OF DATA TO EXTRACTS

Select the type of data from the drop-down menu and enter the respected fields below and finally enter submit (figure).

- Amplification Curve
- High Resolution Melt

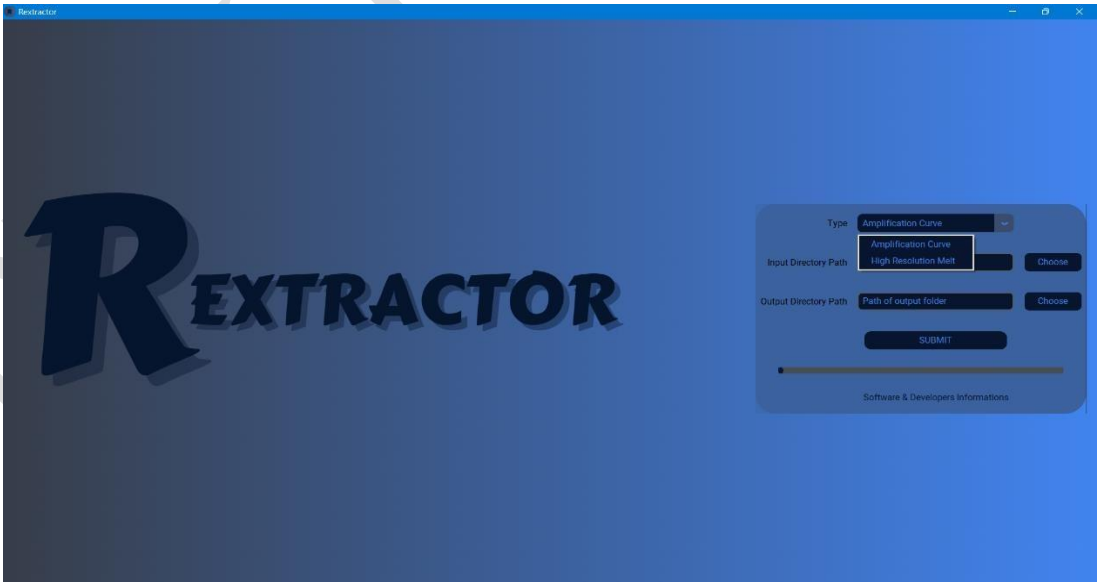


Figure:63 Features in Rextractor

11.3 PyMLRS

PyMLRS is python library which is mainly used for the transformation of raw data into the pre-processed data and helps in feature extraction where processing High-Resolution Melt (HRM) data, especially, DNA melting signals to extract features like Peak Temperatures, Take-off and Touch-down points of melting signal (Temperature at which peak start rising and temperature at which peak falls down), Peak prominences, and Area Under the curve. Additionally, the library offers interactive visualization for DNA melting signal

INSTALLING FROM PIP

The PyMLRS library can be installed with using pip command

python -m pip install PyMLRS or pip3 install PyMLRS

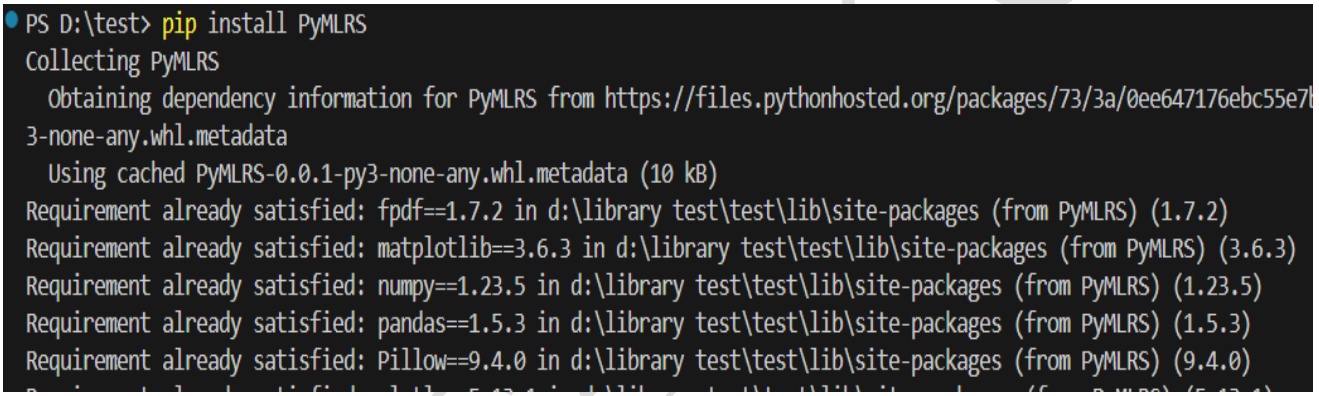


Figure:64 Installing PyMLRS

DEPENDENCIES OF PYMLRS

The following are the necessary dependencies for the PyMLRS.

- fpdf==1.7.2
- matplotlib==3.6.3
- numpy==1.23.5
- pandas==1.5.3
- Pillow==9.4.0
- plotly==5.13.1
- scikit_learn==1.4.1.post1
- scipy==1.13.0
- openpyxl==3.1.2

FILE STACK OF THE LIBRARY

The file stack of the PyMLRS library consists of various files including .py and .pkl files which are basement for feature detection of HRM data**.**

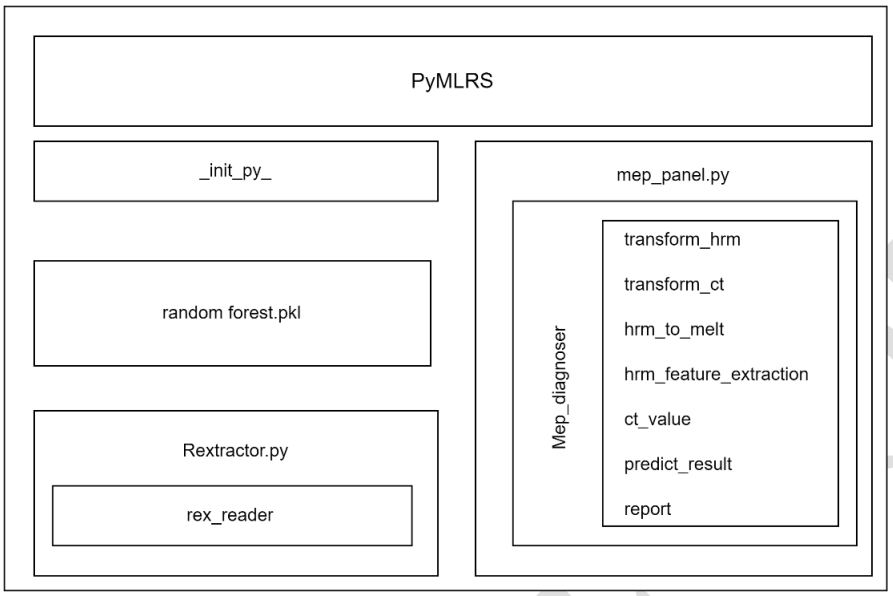


Figure:65 File Stack of PyMLRS

FEATURES

(a) Rextractor
   1. Extract the data from Rotor Gene Experiment(.rex) files.
      i. High Resolution Melt (HRM)
      ii. Amplification Curve - Cycle Time (CT)
   2. Processing Data
      i. Filter Only MEP Pathogens
      ii. Separate by patients

(b) MEP panel
   1. Feature Extraction
      i. Target – Pathogen Name
      ii. Temperature (Peak of the Melt Curve)
      iii. Width
      iv. Prominence
      v. Take of Point
      vi. Take down Point
      vii. Area Under the curve
   2. Interactive Visualization.
      i. High Resolution Melt
      ii. Melt Curve
      iii. Amplification Curve
   3. Result
      i. Interpreting the results of the pathogens
      ii. Generate report for test Results

INPUT DATA FORMAT

The input file for the library is the run file from the Rotor-Gene machine, which is in the format of a Rotor-Gene Experiment (.rex) file, and we can directly provide the .rex file without any preprocessing.

WORKING OF LIBRARY

The library has been imported in the any notebook IDEs, by with the following commands in figure. Next to create a class instance for the mep_diagnoser and rextractor module present in the PyMLRS

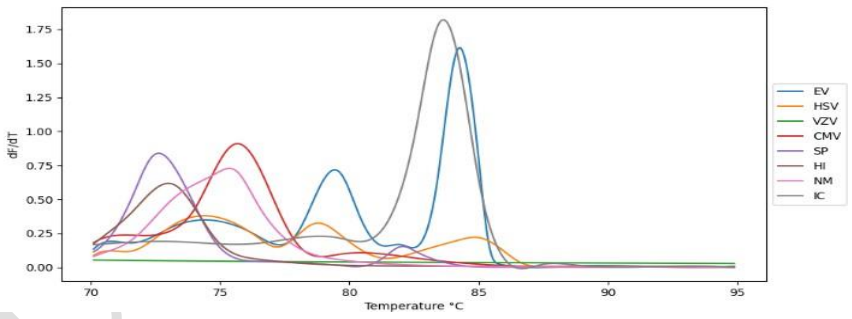
```
from Rextractor import rex_reader
from Mep_panel import Mep_diagnoser

patient_id,hrm,ct = rex_reader("D:\MEP+ MEP VZV MUMPS RT PCR RUN FILE (58 RXNS) BATCH-02 05-04-2024.rex")
for id in patient_id:
    mep = Mep_diagnoser()
    mep.read_hrm(hrm[id])
    mep.read_ct(ct[id])
    mep.hrm_to_melt()
    mep.hrm_feature_extraction()
    mep.ct_value()
    mep.Tm_threshold()
    mep.predict_result()
    mep.report(output_fie_path=f"{id}.pdf",patient_id=559)
```

Figure:66 Sample code for working process

SAMPLE OUTPUT

Melt Curve



Features

Target	Tm1	Width1	Prom1	Top1	Tdp1	auc1	Tm2	Width2	Prom2	Top2	Tdp2	auc2
EV	84.3	28.13	1.61	83.23	85.1	2.14	79.43	37.85	0.72	78.03	80.57	1.34
HSV	74.37	73.26	0.38	72.03	76.9	1.44	78.77	33.39	0.33	77.7	79.9	0.59
VZV	0	0	0	0	0	0	0	0	0	0	0	0
CMV	75.7	54.52	0.91	73.7	77.37	2.46	0	0.0	0	0.0	0.0	0
SP	72.63	51.75	0.84	70.97	74.43	2.08	0	0.0	0	0.0	0.0	0
HI	72.97	54.84	0.62	70.83	74.5	1.73	0	0	0	0	0	0
NM	75.37	82.53	0.73	71.77	77.3	2.86	0	0.0	0	0.0	0.0	0
IC	83.63	45.82	1.82	81.97	85.03	3.9	0	0.0	0	0.0	0.0	0

11.4 PATHOGEN DETECTOR

Pathogen Detector is a user-friendly web-based application which mainly combines the Rextractor and PyMLRS library. This module consists of various files and folders such as .py, .pkl, .xls file for the final classification and interpretation of Meningitidis panel

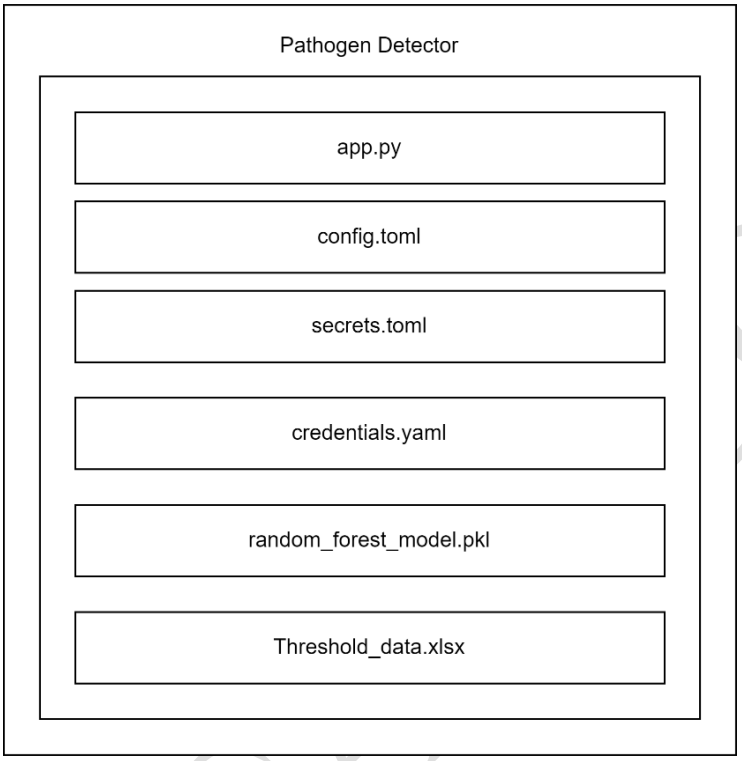


Figure:67 File Stack of Pathogen Detector

Pathogen detector have log-in page where unauthorized persons cannot access the web application without any credentials

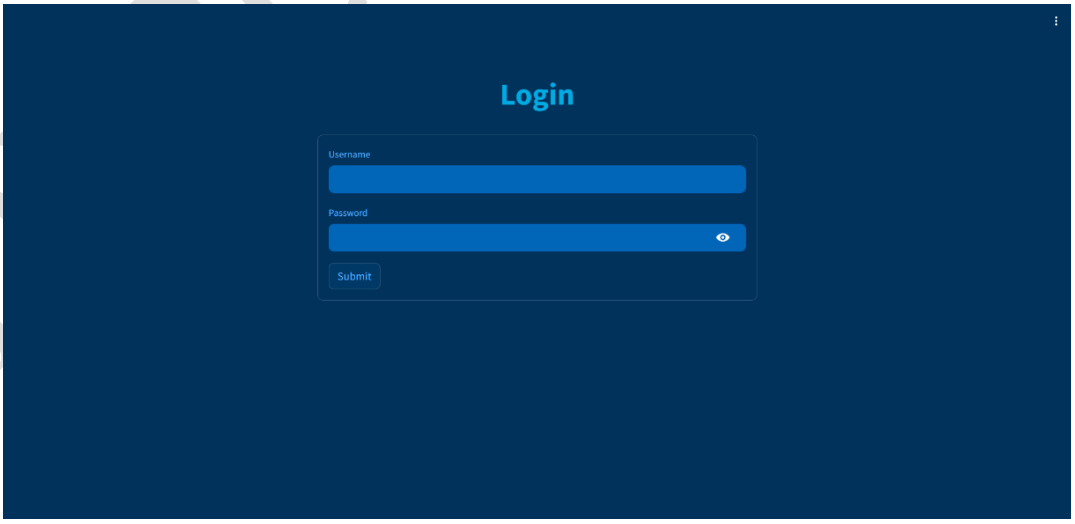


Figure:68 Login page

The file can be simply drag and drop from the local file directory and raw fluroscence can be visually plot according to the need. The multiple tabs in user interface will simply navigate us to the multiple functions as follows

Figure:69 User Interface of Pathogen Detector

(Source: Pathogen Detector)

The above images are snippet taken from the web application to generate reports by simply uploading the .rex files in web application

CHAPTER 12

TEST RESULT

12.1 TEST DATA

The following unknown patient sample data has been used to test the Pathogen Detector's Machine Learning Model of pathogen classification.

```
<?xml version="1.0" encoding="utf-8" ?>

<Experiment>
<RexHeader>REX 3.15</RexHeader>
<Notes></Notes>
<Operator>MICROLAB</Operator>
<Rotor>3</Rotor>
<RunId></RunId>
<StartTime>2024-04-05T12:44:49</StartTime>
<FinishTime>2024-04-05T14:43:43</FinishTime>
<TemplateFilename>C:\Users\Operator\Desktop\RT- PCR Protocols\HRMA PROTOCOL\MEP  04-0
<HighSpeedRotor>True</HighSpeedRotor>
<SpikeSuppression>True</SpikeSuppression>
<ReactionVolume>20</ReactionVolume>
<OilLayerVolume>0</OilLayerVolume>
<AmbientCompensation>True</AmbientCompensation>
<Signature>d1be5a07046628e2a636b5faeb61d6e0</Signature>
<TemplateSignatureState>2</TemplateSignatureState>
<Samples>
```

MEP+ MEP VZV MUMPS RT PCR RUN FILE (58 R... 4/8/2024 2:33 PM Rotor-Gene Experime...

Figure:70 Rex file (Run File of Meningoencephalitis)

The above file fed as input for Pathogen Detector Web application, it will give the HRM, CT Analysis and Predict the result of all pathogens in meningoencephalitis.

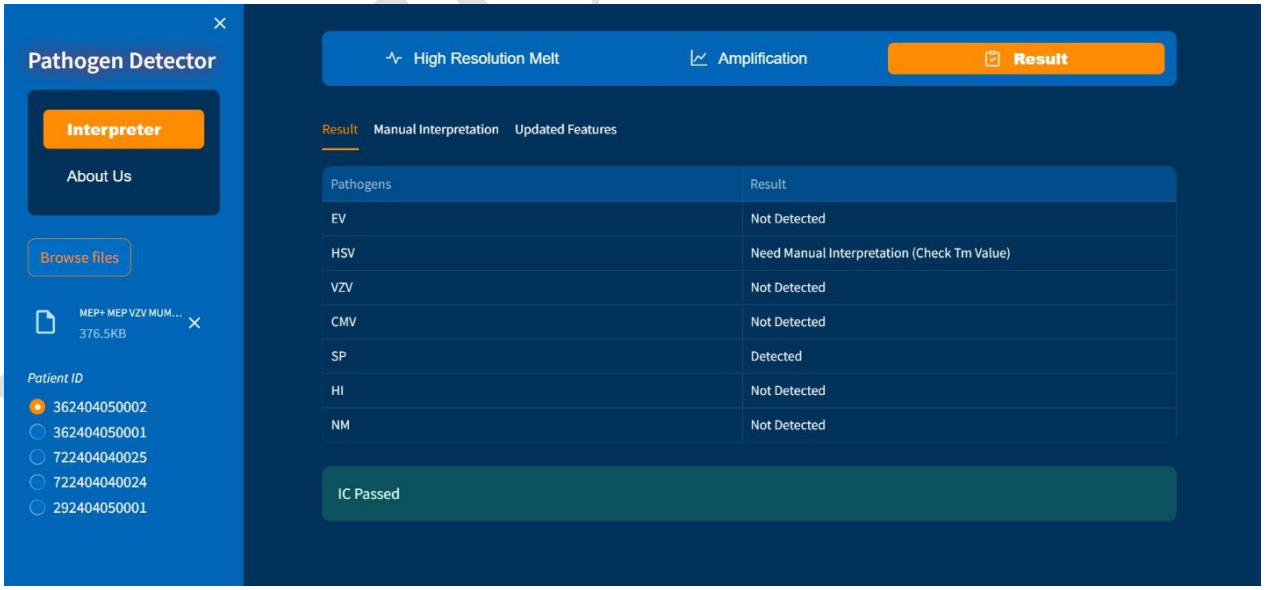
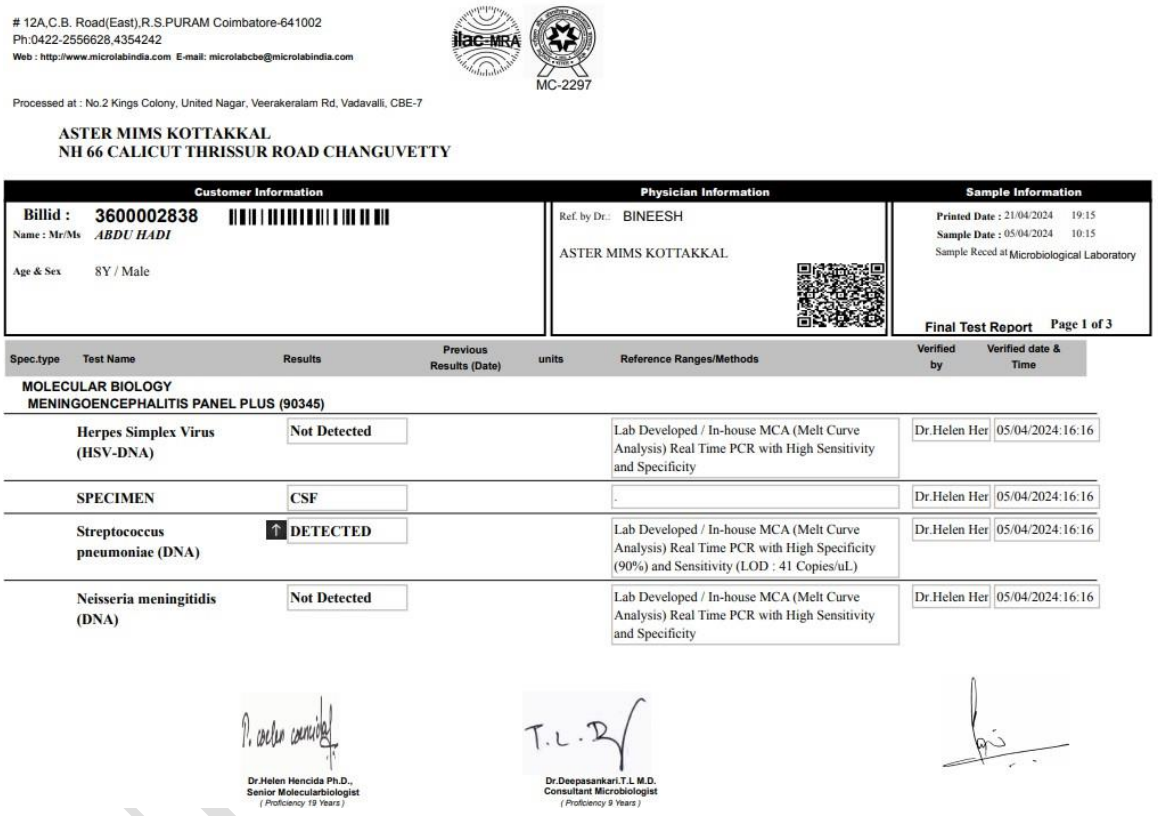


Figure:71 Result of the Input file

The figure above illustrates that the model predicts Streptococcus pneumoniae as **"Detected"** while the rest of the pathogens are classified as **"Not Detected"** or **"Need Manual Interpretation (Check Tm value)."** If the Tm range falls within 1 unit before or after the associated Tm range, it should be classified as **"Need Manual Interpretation (Check Tm value)."** Similarly, if the Tm range falls within the associated range and the Ct value lies between 28 to 30, it should be classified as **"Need Manual Interpretation (Check Ct value)."** Thus, the result of the Pathogen detector combines logical and machine learning-based approaches to interpret the results.

To prove the model result is correct, the original result of the patient sample is given below,



12A,C.B. Road(East),R.S.PURAM Coimbatore-641002
Ph:0422-2556628,4354242
Web : http://www.microlabindia.com E-mail: microlabcbe@microlabindia.com

MC-2297

Processed at : No.2 Kings Colony, United Nagar, Veerakeralam Rd, Vadavalli, CBE-7

ASTER MIMS KOTTAKKAL
NH 66 CALICUT THRISSUR ROAD CHANGUVETTY

Customer Information	Physician Information	Sample Information
Billid : 3600002838 Name : Mr/Ms *ABDU HADI* Age & Sex 8Y / Male	Ref. by Dr.: BINEESH ASTER MIMS KOTTAKKAL	Printed Date : 21/04/2024 19:15 Sample Date : 05/04/2024 10:15 Sample Reced at Microbiological Laboratory

Final Test Report Page 1 of 3

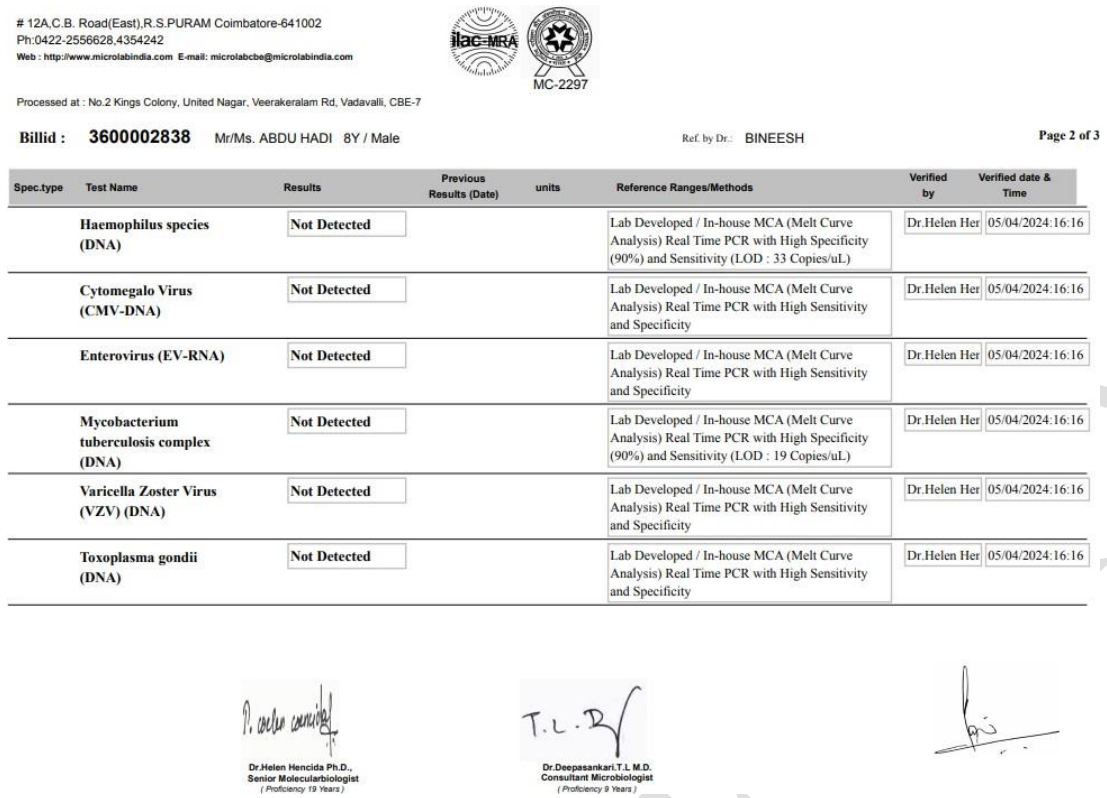
Spec.type	Test Name	Results	Previous Results (Date)	units	Reference Ranges/Methods	Verified by	Verified date & Time
MOLECULAR BIOLOGY MENINGOENCEPHALITIS PANEL PLUS (90345)							
	Herpes Simplex Virus (HSV-DNA)	Not Detected			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity	Dr.Helen Her	05/04/2024:16:16
	SPECIMEN	CSF			.	Dr.Helen Her	05/04/2024:16:16
	Streptococcus pneumoniae (DNA)	↑ DETECTED			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Specificity (90%) and Sensitivity (LOD : 41 Copies/uL)	Dr.Helen Her	05/04/2024:16:16
	Neisseria meningitidis (DNA)	Not Detected			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity	Dr.Helen Her	05/04/2024:16:16

Dr.Helen Hencida Ph.D.,
Senior Molecularbiologist
(Proficiency 19 Years)

Dr.Deepasankari.T.L M.D.
Consultant Microbiologist
(Proficiency 9 Years)

Figure:72 Original Report

(Source: Microbiological Laboratory)



12A,C.B. Road(East),R.S.PURAM Coimbatore-641002
Ph:0422-2556628,4354242
Web : http://www.microlabindia.com E-mail: microlabcbe@microlabindia.com

MC-2297

Processed at : No.2 Kings Colony, United Nagar, Veerakeralam Rd, Vadavalli, CBE-7

Billid : 3600002838 Mr/Ms. ABDU HADI 8Y / Male Ref. by Dr.: BINEESH **Page 2 of 3**

Spec.type	Test Name	Results	Previous Results (Date)	units	Reference Ranges/Methods	Verified by	Verified date & Time
	Haemophilus species (DNA)	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Specificity (90%) and Sensitivity (LOD : 33 Copies/uL)	Dr.Helen Her	05/04/2024:16:16
	Cytomegalo Virus (CMV-DNA)	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity	Dr.Helen Her	05/04/2024:16:16
	Enterovirus (EV-RNA)	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity	Dr.Helen Her	05/04/2024:16:16
	Mycobacterium tuberculosis complex (DNA)	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Specificity (90%) and Sensitivity (LOD : 19 Copies/uL)	Dr.Helen Her	05/04/2024:16:16
	Varicella Zoster Virus (VZV) (DNA)	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity	Dr.Helen Her	05/04/2024:16:16
	Toxoplasma gondii (DNA)	**Not Detected**			Lab Developed / In-house MCA (Melt Curve Analysis) Real Time PCR with High Sensitivity and Specificity	Dr.Helen Her	05/04/2024:16:16

Dr.Helen Hencida Ph.D.,
Senior Molecularbiologist
(Proficiency 19 Years)

Dr.Deepasankari.T.L M.D.
Consultant Microbiologist
(Proficiency 9 Years)

Figure:73 Original Report

(Source: Microbiological Laboratory)

The Figure indicates that the model report aligns with the original report, ensuring that our machine learning model accurately predicted the presence of the intended pathogens in the patient sample data.

CONCLUSION

In the current scenario, the diagnosis of infectious diseases is rapidly shifting towards molecular assays, with several major biotechnology companies developing ready-to-use molecular kits. However, the reporting of these molecular assays largely depends on visual interpretation and analysis by technicians, significantly affecting the acceptability of these assays in commercial diagnostic setups such as clinical laboratories and hospitals.

This project aims to lay the foundation for developing a framework for automated analysis of molecular assays, which is the first of its kind. We have successfully shown that by using predictive analytics and deep learning models on High-Resolution Melting data and Amplification Curve, several distinct features can be extracted that can be used to develop an algorithm to indicate the presence of the intended molecular target in a clinical sample tested.

This project can be further developed into a full-fledged software that can aid clinicians in diagnosing several diseases and planning the course of treatment. This software has the potential to revolutionize the molecular diagnosis field and improve the digital compatibility of molecular assay interpretation with the existing laboratory information management system.

REFERENCES

[1] Create a Google Sheets Data Entry Form with Python & Streamlit | Quick & Easy Tutorial. (n.d.). [Video]. Retrieved April 22, 2024, from https://www.youtube.com/watch?v=_G5f7og_Dpo

[2] M. (n.d.). GitHub - mkhorasani/Streamlit-Authenticator: A secure authentication module to validate user credentials in a Streamlit application. GitHub. https://github.com/mkhorasani/Streamlit-Authenticator

[3] PCR/qPCR Data Analysis. (n.d.). https://www.sigmaaldrich.com/IN/en/technical-documents/technical-article/genomics/qpcr/data-analysis

[4] Commonly used terms in PCR. (n.d.). https://www.qiagen.com/us/knowledge-and-support/knowledge-hub/bench-guide/pcr/commonly-used-terms-in-pcr/commonly-used-terms-in-pcr

[5] Access NCBI through the World Wide Web (WWW). (1995, February). Molecular Biotechnology, 3(1), 75–75. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3258155/

[6] Staff, M. (2016, April 12). Interpretation of qPCR curve shapes. Medical Laboratory Observer. https://www.mlo-online.com/home/article/13008268/interpretation-of-qpcr-curve-shapes

[7] Data Analysis on the ABI PRISM 7700 Sequence Detection System: Setting Baselines and Thresholds. Overview. Data Analysis Tutorial - PDF Free Download. (n.d.). https://docplayer.net/14940629-Data-analysis-on-the-abi-prism-7700-sequence-detection-system-setting-baselines-and-thresholds-overview-data-analysis-tutorial.html(Data Analysis on the ABI PRISM 7700 Sequence Detection System: Setting Baselines and Thresholds. Overview. Data Analysis Tutorial - PDF Free Download, n.d.)

[8] Ruijter, J. M., Ramakers, C., Hoogaars, W. M., Karlen, Y., Bakker, O., Van Den Hoff, M. J., & Moorman, A. F. (2009, February 22). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Research. https://doi.org/10.1093/nar/gkp045

[9] F. (n.d.). GitHub - FEUSION/PyHRM: A library for processing DNA Melting signal with feature extraction and automatic thresholding. GitHub. https://github.com/FEUSION/PyHRM

[10] Technologies, L. B. (2022, November 14). qPCR troubleshooting: interpreting amplification curves and diagnosing problems. https://www.linkedin.com/pulse/qpcr-troubleshooting-interpreting-amplification-/