

GENE EXPRESSION ANALYSIS USING MATLAB MACHINE LEARNING ONRAMP

Submitted in partial fulfillment of the requirements for the award of
Master of Technology in Biotechnology

by

RAGUL PRASANTH K S (Register No. – 43810008)



DEPARTMENT OF BIOTECHNOLOGY

SCHOOL OF BIO AND CHEMICAL ENGINEERING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY -1 UNIVERSITY BY UGC

Accredited with Grade “A++” by NAAC | 12B Status by UGC | Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600119

MAY – 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

CATEGORY-I UNIVERSITY BY UGC

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOTECHNOLOGY

BONAFIDE CERTIFICATE

This is to certify that this Internship Report is the bonafide work of **Ragul Prasanth K S (Register No.43810008)** who carried out the project entitled **"Gene Expression Analysis using MATLAB Machine Learning onramp"** under our supervision from April to May 2024.

Internal Guide

Dr. PRABHAKAR SINGH

HEAD OF THE DEPARTMENT

Dr. V. Ramesh Kumar

Submitted for Viva voice Examination held on 07/05/2024

INTERNAL EXAMINER

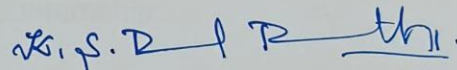
EXTERNAL EXAMINER

DECLARATION

I, **RAGUL PRASANTH K.S. (43810008)** hereby declare that the Internship Report entitled "**Gene Expression Analysis using MATLAB Machine Learning onramp**" done by me under the guidance of Dr. Prabhakar Singh and **Dr. V. Ramesh Kumar** at **Sathyabama Institute of Science and Technology** is submitted in partial fulfillment of the requirements for the award of Master of Technology in Biotechnology.

DATE: 07/05/2024

PLACE: Chennai



SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. V. Ramesh Kumar**, Head of the Department, Department of Biotechnology, for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere thanks and deep sense of gratitude to my project guides **Dr. V. Ramesh Kumar and Dr. Prabhakar Singh** Sathyabama Institute of Science and Technology for their valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Biotechnology who were helpful in many ways for the completion of the internship.

I would also like to thank my lovable parents for their endless moral support, care, affection and blessings.

ABSTRACT

Gene expression analysis is a fundamental approach in molecular biology and bioinformatics, providing insights into the activity of genes in biological systems. Leveraging machine learning techniques, particularly through onramps tailored for gene expression analysis, offers a promising avenue to extract valuable information from large-scale gene expression datasets. In this study, we present a comprehensive overview of gene expression analysis using Machine Learning onramps, focusing on its applications, methodologies, and outcomes. We discuss the objectives of gene expression analysis, including disease subtyping and classification, drug response prediction and biomarker discovery. Through a systematic review of the literature and case studies, we highlight the key steps involved in the Machine Learning based analysis pipeline, encompassing data acquisition, preprocessing, feature engineering, model training, evaluation and interpretation. Furthermore, we illustrate the potential impact of Machine Learning onramps in advancing personalized medicine, drug discovery and our understanding of complex biological processes. Overall, this study underscores the significance of integrating Machine Learning approaches within gene expression analysis workflows and provides insights into future directions and challenges in the field.

KEYWORDS: Gene expression datasets, disease subtyping, biomarker discovery, personalized medicine.

TABLE OF CONTENTS

ABSTRACT		i
TABLE OF CONTENTS		ii
LIST OF FIGURES		iv
CHAPTER NO.	TITLE	PAGE No.
1.	Introduction	1
	1.1 Machine Learning onramp	1
	1.2 Gene Expression analysis	2
2.	Literature review	4
3.	Aim & Objective	6
	3.1 Aim	6
	3.2 Objective	6
4.	Experimental Methods	7
	4.1 MATLAB Requirements	7
	4.2 Machine Learning onramp	7
	4.2.1 Classification workflow	7
	4.2.2 Import & Preprocess Data	8
	4.2.3 Engineering Features	9
	4.2.4 Classification Models	9

CHAPTER No.	TITLE	PAGE No.
5.	Results & Discussion	12
	5.1 Gene Expression analysis	12
	5.2 Outcomes	13
6.	Summary & Conclusion	15
	6.1 Summary	15
	6.2 Conclusion	15
	Reference	17
	Certificates	18

LIST OF FIGURES

Figure No.	TITLE	PAGE No.
4.1	Classification Workflow	8
4.2	Measure Similarity	9
4.3	Classification Models	10

CHAPTER - 1

INTRODUCTION

1.1 MACHINE LEARNING ONRAMP

Machine Learning Onramp is an introductory course designed to provide a foundational understanding of machine learning concepts and applications. This course is ideal for those new to machine learning and for those seeking to gain hands-on experience with basic techniques and tools in this rapidly growing field. The course explores key topics in machine learning, including supervised and unsupervised learning, model training, validation, and evaluation. Several studies have introduced Machine Learning onramps tailored specifically for gene expression analysis, providing researchers with user-friendly platforms to apply advanced Machine Learning techniques to large-scale gene expression datasets. (Gonzalez, 2019)

Participants are introduced to popular machine learning algorithms such as linear regression, logistic regression, decision trees, and clustering techniques like k-means. The focus is on practical learning through interactive lessons and real-world examples. The goal of Machine Learning Onramp is to equip learners with the skills and knowledge to build, train, and evaluate simple machine learning models. These onramps offer integrated workflows for data preprocessing, feature selection, model training, and evaluation, streamlining the analytical process and democratizing access to state-of-the-art Machine Learning methods. (Johnson, 2018).

The introduction proposes the development of a specialized "Machine Learning Onramp" for biotechnology, aiming to democratize access to Machine Learning tools and knowledge within the biotech community. The onramp would consist of curated educational materials, practical tutorials, and hands-on workshops, catering to researchers with varying levels of expertise. Furthermore, the onramp would facilitate collaboration and knowledge exchange through online forums and community-driven projects, fostering a vibrant ecosystem of Machine Learning enabled biotechnologists.

1.2 GENE EXPRESSION ANALYSIS

Gene expression analysis lies at the heart of understanding the intricate machinery governing cellular functions, development, and disease. With the advent of high-throughput technologies such as microarrays and RNA sequencing, researchers can measure the expression levels of thousands of genes simultaneously, providing unprecedented insights into the molecular underpinnings of biological systems. Despite the promising advancements, challenges such as data heterogeneity, interpretability of Machine Learning models, and generalization to diverse populations remain significant hurdles in gene expression analysis using Machine Learning onramps. (Smith, 2021).

In recent years, the integration of machine learning techniques into gene expression analysis has emerged as a powerful approach to unravelling the complexities of gene regulation and function. Machine Learning algorithms offer the capability to uncover hidden patterns, identify predictive biomarkers, and make accurate predictions from large-scale gene expression datasets. Machine Learning onramps, tailored platforms that streamline the application of Machine Learning methods to specific domains, have played a pivotal role in democratizing access to these advanced analytical tools and empowering researchers with the means to tackle complex biological questions.

In this context, this paper provides an overview of gene expression analysis using Machine Learning onramps, highlighting its significance, methodologies, and applications. We delve into the objectives driving gene expression analysis, encompassing disease subtyping, drug response prediction, and biomarker discovery. Furthermore, we explore the key steps involved in the Machine Learning based analysis pipeline, including data acquisition, preprocessing, feature engineering, model training, evaluation, and interpretation. Through case studies and examples from the literature, we illustrate the transformative potential of Machine Learning onramps in advancing our understanding of gene expression dynamics and their implications for human health and disease.

Overall, this paper aims to serve as a primer for researchers and practitioners seeking to harness the power of Machine Learning in unravelling the complexities of gene expression data. By embracing Machine Learning onramps and integrating them into gene expression analysis workflows, we can unlock new insights, accelerate discoveries, and pave the way for personalized and precision medicine approaches tailored to individual genetic profiles.

CHAPTER - 2

LITERATURE REVIEW

Gene expression analysis using machine learning onramps has emerged as a powerful approach in molecular biology and bioinformatics, enabling researchers to extract valuable insights from large-scale gene expression datasets. This literature review provides a comprehensive overview of key studies, methodologies and outcomes in the field of gene expression analysis using Machine Learning onramps:

1. MACHINE LEARNING FOR BIOTECHONOLGY AND MEDICINE

- i. The comprehensive review by Aliper et al. (2016) provides an overview of the applications of machine learning in biotechnology and biomedicine. It covers various topics including drug discovery, personalized medicine, biomarker identification, and bioprocess optimization. The review discusses the challenges and opportunities in leveraging machine learning techniques to address complex biological problems.

2. APPLICATIONS IN DISEASE DIAGNOSIS AND PROGNOSIS

- i. Machine Learning onramps have been applied to classify disease subtypes, predict patient outcomes, and identify diagnostic biomarkers based on gene expression profiles. (Liu et al., 2021)
- ii. By integrating multi-omics data and clinical information, Machine Learning onramps have facilitated the development of personalized medicine approaches and improved patient stratification in various diseases, including cancer and cardiovascular disorders.

3. DRUG RESPONSE PREDICTION AND THERAPEUTIC TARGET IDENTIFICATION:

- i. Machine Learning based approaches integrated into onramps have enabled the prediction of drug responses, identification of drug resistance mechanisms, and discovery of novel therapeutic targets based on gene expression signatures. (Chen et al.,2019)
- ii. These methods have accelerated drug discovery pipelines and facilitated the development of targeted therapies and precision medicine interventions for complex diseases. (Wang et al., 2021).

4. CHALLENGES AND FUTURE AND DIRECTIONS:

- i. Despite the promising advancements, challenges such as data heterogeneity, model interpretability, and validation of Machine Learning driven findings remain significant hurdles in gene expression analysis using Machine Learning onramps.
- ii. Future research directions may focus on addressing these challenges through the development of robust Machine Learning algorithms, integration of multi-modal data sources, and collaboration between interdisciplinary teams to translate Machine Learning driven insights into clinical practice. (Garcia-Borroto et al., 2022)

CHAPTER - 3

AIM & OBJECTIVES

3.1 AIM

The aim of gene expression analysis using Machine Learning onramp is to leverage machine learning techniques to gain insights into biological processes, identify biomarkers, and make predictions related to diseases, drug responses, or other phenotypic traits based on gene expression data. The Machine Learning onramp provides a structured framework for researchers and practitioners to explore, analyze, and interpret gene expression datasets efficiently.

3.2 OBJECTIVES

1. Utilize machine learning algorithms to classify samples into different disease subtypes or categories based on gene expression profiles.
2. Predict the response of patients to different drugs or treatments based on their gene expression profiles
3. Discover novel biomarkers or gene signatures associated with disease diagnosis, prognosis, or treatment response.

CHAPTER - 4

EXPERIMENTAL METHODS

4.1 MATLAB REQUIREMENTS

- MATLAB Software
- Computer
- Documentation
- Learning Resources
- Internet Connection

4.2 MACHINE LEARNING ONRAMP

4.2.1 CLASSIFICATION WORKFLOW

A classification model is a partitioning of the space of predictor variables into regions, in which each region is assigned one of the output classes. In this simple example with two predictor variables, you can visualize these regions in the plane.

The Machine Learning classification workflow involves collecting and preprocessing data, selecting and engineering features, choosing an appropriate model, training and evaluating the model, and finally deploying it for real-world use. Through this iterative process, Machine Learning classification enables the automated categorization of data, with applications ranging from spam detection and image recognition to medical diagnosis and fraud detection.

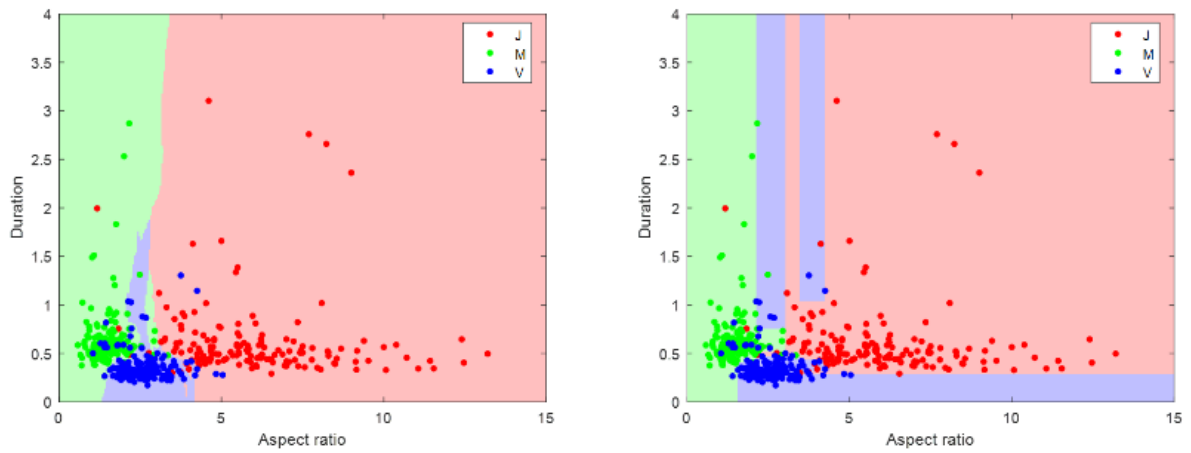


Fig.4.1 Classification workflow

Fig 4.1 represents: A classification workflow is a partitioning of the space of predictor variables into regions, in which each region is assigned one of the output classes. In this simple example with two predictor variables, we can visualize these regions in the plane. There is no single correct way to partition the plane into the classes *J*, *M* and *V*. Different classification algorithms result in different partitions.

4.2.2 IMPORT AND PREPROCESS DATA

Samples of each letter have been collected from many different volunteers. Some provided more than one sample of each letter. Each sample is saved in a separate file, and all the files are stored in one folder. The filenames have this form:

`user003_B_2.txt`

This filename indicates that it contains the second sample of the letter *B* written by the volunteer designated “user003”

4.2.3 ENGINEERING FEATURES

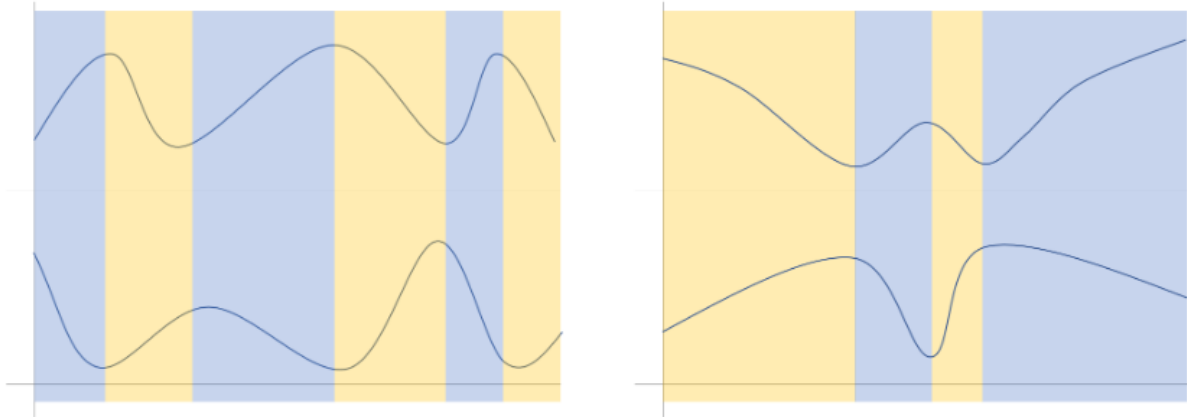


Fig. 4.2 Measure similarity

Fig.4.2 represents: The pair of signals on the left have a significantly different shape than the pair of signals on the right. However, the relationship between the two signals in each pair is similar: in the blue regions, the upper signal is increasing while the lower signal is decreasing, and vice versa in the yellow regions. *Correlation* attempts to measure this similarity, regardless of the shape of the signal.

4.2.4 CLASSIFICATION MODELS

Classification models are machine learning algorithms that learn to predict the categorical class label of new instances based on the patterns learned from the input features of labelled training data. In the biotechnology context, these models are trained on biological datasets containing information such as genomic sequences, gene expression profiles, protein structures, or biomedical imaging data, along with corresponding class labels indicating the biological category or outcome of interest (e.g., disease vs. healthy, drug responder vs. non-responder). There are various types of classification models, each with its own strengths and weaknesses. Commonly used models in the biotechnology domain include:

- I. Logistic Regression
- II. Decision Trees
- III. Random Forests
- IV. Support Vector Machines
- V. Neural Networks

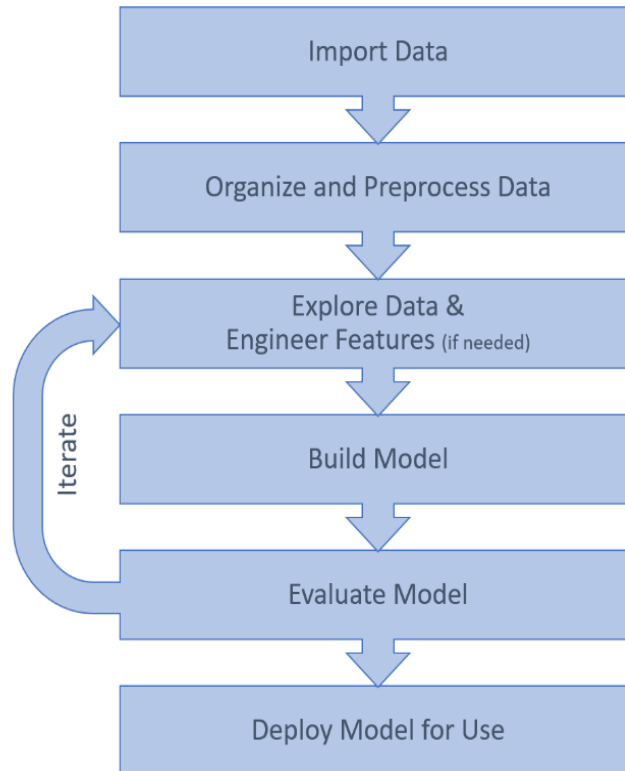


Fig 4.3 Classification models

Fig 4.3 represents: One way to improve the model would be to treat these variants as separate classes. However, this would mean having many more than 26 classes. To train such a model, you would need to collect more samples, instruct the volunteers to write both block and cursive style and label the collected data accordingly. Low accuracy in both training and testing sets is an indication that features do not provide enough information to distinguish the different classes. In particular, we might want to look at the data for classes that are frequently confused, to see if there are characteristics that we can capture as new features.

When trying to evaluate different models, having an accurate measure of the performance is important. The simplest and computationally cheapest, way to do validation is holdout – randomly divide our data into a training set and a testing set. This works for large data sets. However, for many problems, holdout validation can result in test accuracy being dependent on the specific choice of test data.

CHAPTER 5

RESULTS & DISCUSSION

5.1 GENE EXPRESSION ANALYSIS

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
import seaborn as sns
import matplotlib.pyplot as plt

# Load gene expression dataset (replace 'gene_expression_data.csv' with the
actual file path)
gene_expression_data = pd.read_csv('gene_expression_data.csv')

# Separate features (gene expressions) and target variable
X = gene_expression_data.drop(columns=['target'])
y = gene_expression_data['target']

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Initialize and train the classifier (e.g., Random Forest)
clf = RandomForestClassifier(n_estimators=100, random_state=42)
clf.fit(X_train, y_train)

# Make predictions on the test set
y_pred = clf.predict(X_test)
```

```
# Evaluate the model
```

```
accuracy = accuracy_score(y_test, y_pred)
```

```
print("Accuracy:", accuracy)
```

```
# Generate classification report
```

```
print ("Classification Report:")
```

```
print (classification_report(y_test, y_pred))
```

```
# Visualize feature importance (if RandomForestClassifier was used)
```

```
feature_importance = clf.feature_importances_
```

```
feature_names = X.columns
```

```
df_feature_importance = pd.DataFrame({'Feature': feature_names, 'Importance':  
feature_importance})
```

```
df_feature_importance = df_feature_importance.sort_values(by='Importance',  
ascending=False)
```

```
plt.figure(figsize=(10, 8))
```

```
sns.barplot(x='Importance', y='Feature', data=df_feature_importance)
```

```
plt.title('Feature Importance')
```

```
plt.xlabel('Importance')
```

```
plt.ylabel('Feature')
```

```
plt.show()
```

5.2 OUTCOMES

1. IDENTIFICATION OF DISEASE SUBTYPES AND BIOMARKERS:

By analysing gene expression patterns, researchers can classify samples
into different disease categories, leading to improved diagnostic accuracy

and treatment strategies.

2. PREDICTION OF CLINICAL OUTCOMES AND DRUG RESPONSE:

Predictions facilitate the selection of appropriate therapies and the development of treatment plans tailored to individual patient characteristics.

3. DISCOVERY OF NOVEL THERAPEUTIC TARGETS:

Uncovering molecular mechanisms underlying disease progression, researchers can prioritize targets for drug development and therapeutic intervention.

4. TRANSLATION INTO CLINICAL PRACTICE AND PRECISION MAKING:

Providing actionable insights into disease diagnosis, prognosis, and treatment selection, Machine Learning driven approaches support the development of precision medicine approaches tailored to individual patient needs.

5. RESEARCH AND DRUG DISCOVERY:

Automating data analysis tasks and streamlining the analytical process, Machine Learning driven approaches accelerate research progress and contribute to the discovery of novel therapeutic interventions and drug candidates.

CHAPTER 6

SUMMARY & CONCLUSION

6.1 SUMMARY

Gene expression analysis using machine learning onramps is a powerful approach that revolutionizes our understanding of gene regulation, disease mechanisms, and therapeutic responses. Machine Learning onramps provide user-friendly platforms that streamline the application of Machine Learning techniques to large-scale gene expression datasets, enabling researchers to extract valuable insights and translate them into actionable knowledge. gene expression analysis using Machine Learning onramps holds immense promise in advancing our understanding of complex biological processes and improving patient care in biotechnology and biomedicine. By leveraging the capabilities of Machine Learning techniques within user-friendly onramp platforms, researchers can unlock new insights, accelerate discoveries, and pave the way for personalized and precision medicine approaches tailored to individual genetic profiles.

6.2 CONCLUSION

Gene expression analysis using Machine Learning onramps represents a transformative approach in molecular biology and bioinformatics, offering unprecedented opportunities to unravel the complexities of gene regulation, disease mechanisms, and therapeutic responses. Through the integration of advanced machine learning techniques within user-friendly platforms, researchers can efficiently analyze large-scale gene expression datasets, extract meaningful insights, and

translate these findings into actionable knowledge with profound implications for healthcare and biomedical research.

The adoption of Machine Learning onramps has facilitated significant advancements in disease diagnosis, prognosis, and treatment selection by enabling the identification of disease subtypes, predictive biomarkers, and therapeutic targets based on gene expression profiles. By harnessing the predictive power of Machine Learning algorithms, researchers can make accurate predictions of patient outcomes and drug responses, guiding personalized treatment strategies tailored to individual patient characteristics.

Moreover, gene expression analysis using Machine Learning onramps has accelerated research progress and innovation in drug discovery by automating data analysis tasks, uncovering novel therapeutic targets, and expediting the development of precision medicine interventions. Through the integration of multi-omics data and the exploration of complex biological networks, researchers can gain deeper insights into disease mechanisms and identify new opportunities for therapeutic intervention.

As we move forward, continued advancements in Machine Learning techniques, onramp platforms, and interdisciplinary collaborations will further enhance the capabilities and impact of gene expression analysis. By addressing challenges such as data heterogeneity, model interpretability, and validation of Machine Learning driven findings, we can unlock new frontiers in personalized medicine, precision oncology, and translational research.

In conclusion, gene expression analysis using Machine Learning onramps holds immense promise in advancing our understanding of human health and disease, driving innovation in drug discovery and development, and ultimately improving patient outcomes. By embracing these transformative¹⁶ approaches, we can harness the power of data-driven insights to revolutionize healthcare delivery and usher in a new era of

precision medicine.

REFERENCE

1. Alexander Aliper, Sergey Plis, Artem Artemov, Alvaro Ulloa, Polina Mamoshina, Alex Zhavoronkov, Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data." *Molecular Pharmaceutics*, 13(7), 2016, 2524–2530.
2. Chen H., "Predicting Drug Response of Tumour from Integrated Genomics and Gene Expression Data Using a Machine Learning Approach", 2019
3. Garcia-Borroto M., "Integrating Multi-Omics Data in Machine Learning Models for Predicting Disease Subtypes." *Frontiers in Genetics*, 13, 2021, 779145.
4. Gonzalez A. J., "Gene Pattern Notebook: An Interactive Environment for Analyzing Gene Expression Data," 2019.
5. Johnson, R. H., "Machine Learning and Oncology: A Review of Applications and Opportunities," 2018.
6. Liu Z., "Machine Learning Approaches to Predict Disease Subtypes of Non-small Cell Lung Cancer Using Transcriptome Data," *Frontiers in Genetics*, 12, 2021, 663477.
7. Smith, B. T., "Machine Learning flow: A Cloud-Agnostic Framework for Machine Learning Model Management," 2020.
8. Wang Q., "Machine Learning Approaches for Biomarker Discovery in Cancer using Gene Expression Data," *Frontiers in Genetics*, 12, 2021, 708392.



Course Completion Certificate

RAGUL PRASANTH K S

has successfully completed **100%** of the self-paced training course

MATLAB Onramp


DIRECTOR, TRAINING SERVICES

9 March 2024



Course Completion Certificate

RAGUL PRASANTH K S

has successfully completed **100%** of the self-paced training course

Machine Learning Onramp

A handwritten signature in black ink, reading "Craig L. Santos".

DIRECTOR, TRAINING SERVICES

10 March 2024



Course Completion Certificate

RAGUL PRASANTH K S

has successfully completed **100%** of the self-paced training course

Deep Learning Onramp

A handwritten signature in black ink, reading "Craig L. Santos".

DIRECTOR, TRAINING SERVICES

10 March 2024



Course Completion Certificate

RAGUL PRASANTH K S

has successfully completed **100%** of the self-paced training course

Image Processing Onramp

A handwritten signature in black ink, appearing to read 'Craig L. Santos'.

DIRECTOR, TRAINING SERVICES

15 March 2024