Introduction to Statistics and Machine Learning for Bioinformatics

Workshop conducted from 6th – 7th October, 2023

Organized by:

SGPC's
Guru Nanak Khalsa College
of Arts, Science & Commerce (Autonomous)

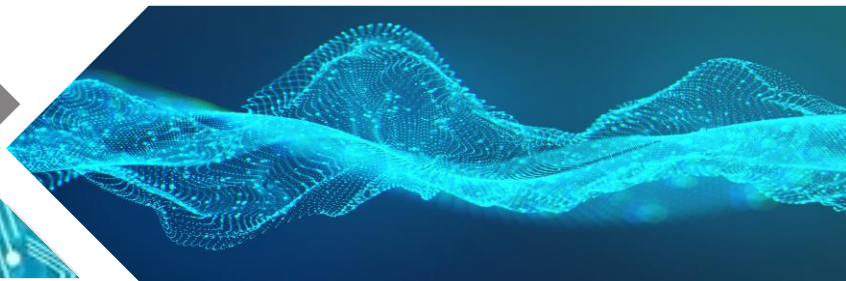
Shiromani Gurudwara Prabandhak Committee's Guru Nanak Khalsa College of Arts, Science and Commerce

In Association with:

Braask Education Private Limited's TechMedBuddy

Report Submitted by:

Ms. Prarthi Hrishit Kothari
Roll No. – 115
M. Sc. Bioinformatics – Part I



ACKNOWLEDGEMENT

Firstly, I would like to express my heartfelt gratitude to God for showering his blessings throughout the workshop.

I would like to express my sincere gratitude to Dr. H. S. Kalsi, The Principal and Dr. Gursimran Kaur Uppal, Head of Department, Bioinformatics and our esteemed professors Mrs. Aparna Patil Kose and Mrs. Sermarani Nadar at Guru Nanak Khalsa College of Arts, Science & Commerce (Autonomous) for organizing the informative workshop "Introduction to Statistics and Machine Learning in Bioinformatics" that provided us with such valuable insights based on the essential principles and fundamentals of statistics, artificial intelligence (AI) and machine learning, applicable in the field of bioinformatics. They have been extremely supportive and have worked actively to provide me with the required academic guidance.

Secondly, I sincerely appreciate the efforts taken by the esteemed speakers and learned scholars whose expertise, experiences, competence and dynamism has inspired me deeply –

1. Dr. Neel Das [Senior Data Scientist at Roche Healthcare]
2. Alok Anand [Founder of Braask Education Pvt. Ltd.]
3. Manas Pratiti [PhD Scholar at IIITD]
4. Dr. Tahseen Abbas [Associate Scientist at Excelra]
5. Dr. Hara Prasad Mishra [Doctor of Medicine (MD) at University of Delhi].

I offer my sincere appreciation to all the people for providing me with such a wonderful learning opportunity of utmost significance that has made a profound impact on my professional journey.

SHIROMANI GURUDWARA PARBHANDHAK COMMITTEE'S

GURU NANAK KHALSA COLLEGE OF ARTS, SCIENCE & COMMERCE

(Autonomous)

Accredited by NAAC With 'A' Grade With a CGPA of 3.54

GNIRD

Department Of Bioinformatics

TechMedBuddy

This certificate declares that

Prarthi Hrishit Kothari

has completed two days workshop on **"Introduction To Statistics and Machine Learning in Bioinformatics"** organized by Department of Bioinformatics at G.N. Khalsa college held on 6th and 7th October 2023.

Dr. Gursimran Kaur Uppal
Head, Dept of Bioinformatics
G.N. Khalsa College

Mr. Alok Anand
Founder, Braask Ed. Pvt. Ltd
IIT Delhi

Dr. H.S. Kalsi
I/C Principal
G.N. Khalsa College

INDEX

1. TABLE OF CONTENTS

2. LIST OF FIGURES

3. LIST OF TABLES

ABSTRACT

The primary objective of this academic report is to present a perspective on the two - day workshop titled "Introduction to Statistics and Machine Learning for Bioinformatics", conducted on the 6th and the 7th of October, 2023, by the Department of Bioinformatics, Guru Nanak Khalsa College of Arts, Science & Commerce (Autonomous).

During the workshop, various eminent and intellectual speakers from an array of scientific disciplines highlighted the necessity of Artificial Intelligence (AI), Machine Learning (ML) and fundamental statistics and biostatistics in the domain of bioinformatics to significantly enhance the industry of healthcare and medical biosciences in addition to refining their techniques and decision – making process with greater efficiency and accuracy.

This workshop created a space to learn, discuss, identify the problems in the healthcare industry, brainstorm ideas, and apply the knowledge gained to try to develop solutions based on scientific data obtained using various online biological databases and tools.

INTRODUCTION

Bioinformatics is defined as interdisciplinary field of science that combines the applications of computer science and biostatistics along with various sectors of biological and biomedical sciences majorly comprising genetics, cell biology, biochemistry, life sciences and bioengineering, to facilitate analysis, interpretation, management and storage of biological data obtained from experimental, observational and *in silico* studies.

The two - day workshop titled "Introduction to Statistics and Machine Learning for Bioinformatics" was oriented towards providing valuable insights into the essential principles and fundamentals of statistics Artificial Intelligence (AI) and Machine Learning (ML), applicable in the field of bioinformatics, majorly in the study of various omics including - genomics, transcriptomics, proteomics, metabolomics and phenomics.

The employment of Artificial Intelligence (AI) and Machine Learning (ML) in the scientific domain would be advantageous for the healthcare industry by providing improved diagnostic tools to detect diseases and life – threatening conditions at an early stage and a more unified and predictive approach to precision medicine and personalized care for better patient and treatment outcomes.

The workshop was conducted in multiple sessions, each session dedicated to an esteemed speaker that not only delivered information related to application of machine learning in enhancing the techniques and decision - making process of the healthcare and medical sciences industry, but also educated us regarding various skill sets necessary for pursuing a career in this domain.

CHAPTER 1

Timeline of the Sessions Conducted (Day – wise Schedule)

Table 1 – *Day – wise Session Timeline*

Date	Title of the Session	Name of Esteemed Speaker	Designation
Day 1 – 6th October, 2023	Application of AI (Artificial Intelligence) in Healthcare Industry	Dr. Neel Das	Senior Data Scientist at Roche Healthcare
	Introduction to Artificial Intelligence (AI) and Machine Learning (ML) and its Applications in Bioinformatics	Alok Anand	Founder of Braask Education Pvt. Ltd
	Fundamentals of Python Programming	Manas Pratiti	PhD Scholar at IIITD
Day 2 – 7th October, 2023	Introduction to the Fundamentals of Statistics and Biostatistics	Alok Anand	Founder of Braask Education Pvt. Ltd.
	Hands – On Session : Differential Expression and Enrichment Analysis	Manas Pratiti	PhD Scholar at IIITD
	Transitioning from Academia to Industry	Dr. Tahseen Abbas	Associate Scientist at Excelra
	AI – Enabled Clinical Decision Making and Support Systems (CDSS)	Dr. Hara Prasad Mishra	Doctor of Medicine (MD) at University of Delhi

CHAPTER 2

SESSIONS CONDUCTED ON DAY 1 – 6th OCTOBER, 2023

2.1 Applications of AI (Artificial Intelligence) in Healthcare Industry

Speaker – Dr. Neel Das

This session commenced with an informative presentation on the historical origin and the current application of AI (Artificial Intelligence) in the domain of healthcare and biomedical data science.

The AI - based tools AlphaFold, developed by DeepMind and ELIZA, an alternative to ChatGPT were introduced during the talk. A deep learning – based algorithm known as **AlphaFold,** is used for accurate protein structure prediction by incorporating multiple sequence alignments and understanding of the physical and biological attributes of the protein entity. AlphaFold has been widely used in the sectors of drug discovery and vaccine development. **ELIZA** chatbot engages participants in text – based conversations and functions by detecting and analyzing keywords in the input statements provided by the user and reflects them back in the form of brief remarks or inquiries.

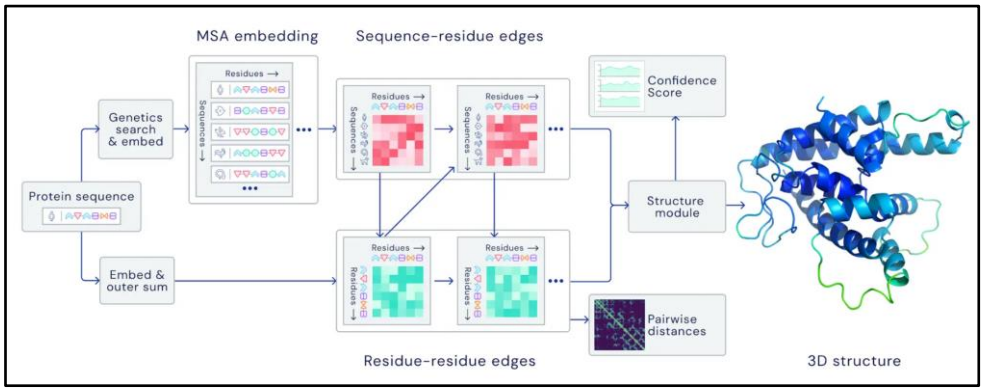


Figure 1 – Mechanism of AlphaFold
[Image Credits – DeepMind. doi: https://doi.org/10.1038/s41586-021-03819-2]

Emphasis was laid on the historical perspective that ranges from the origins and the birth of medical informatics in the 1950s to the era of deep learning in the current times.

The origins commenced with the involvement of mathematical, psychological and engineering – based perspectives with the sole purpose of testing all the possibilities for the development of an artificial brain. The earliest origins are represented by the Turing Test, a technique developed by Alan Turing that determines the capability of a computer to think like a human being. The birth of Artificial Intelligence (AI) materialized when Dr. John McCarthy coined the term "AI". The origins concluded with the birth of Medical Informatics represented by the research paper titled "Reasoning Foundations of Medical Diagnosis", authored and published by Robert S. Ledley and Lee B. Lusted. They proposed the possibility of connecting statistical fundamentals to medical diagnosis by employing multiple logics and theories, one of them being Boolean operations.

Symbolic AI is a term used to define a collection of techniques and strategies that are built on the basis of high – level symbolic human readable representations of problem and logic. This era also highlighted the DENDRAL Project developed by Stanford University, which was one of the most influential expert systems. One of the specific objectives of this project assist organic chemists in the construction of a hypotheses and identification of unknown molecules by automating the analysis and interpretation of the mass spectra data obtained by applying the chemistry expertise.

Inspired by the DENDRAL Project, various clinical and medical expert systems were developed including Interest – I, Mycin, PIP (Present Illness Program), CASNET (Casual Association Network) and EKLAVYA (developed by IIT Madras).

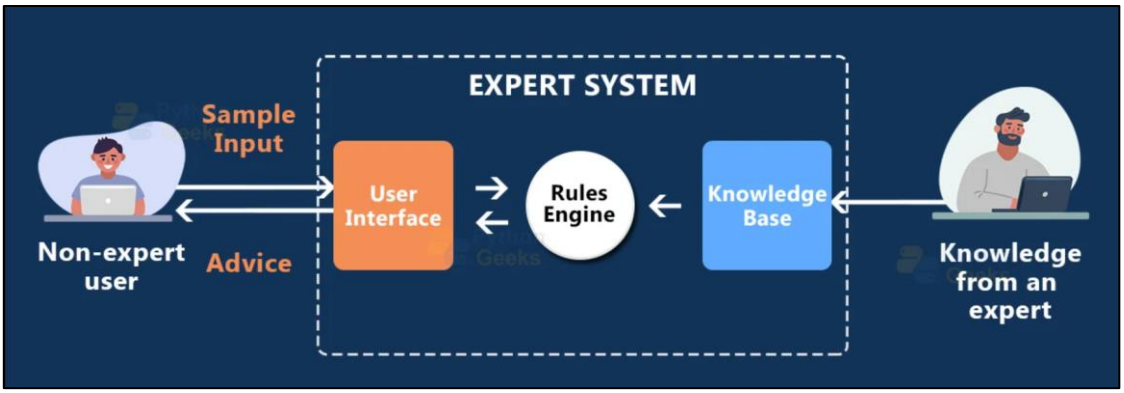


Figure 2 – Expert System in Artificial Intelligence (AI) developed during the era of Symbolic AI
[Image Source – Team, P. (2022, November 14). *Expert Systems in AI*. Python Geeks. https://pythongeeks.org/expert-systems-in-ai/]

AI Winters displayed a major setback due to reduced funding and reduced public interest. The major cause of this setback was the failure to provide any practical results due to the following reasons –

1. Brittle logic – Unforeseen circumstances may not be managed well by the expert systems.
2. Time consuming expertise
3. Severe computational limitations – Excessive time used for processing the data due to lack of computational resources.

The Resurgence marked the comeback of AI where it had become crucial to design novel approaches. Major factors supporting this comeback include – the growth of massive genomic datasets due to the improvement in the potential of computational resources to manage and process large amount of data and proliferation of Electronic Health Records (EHRs) in hospitals and other healthcare institutes.

Textural analysis of medical images was made possible by the massive development of AI where a region of interest is obtained from sources such as MRI scans, multiple different extraction features are applied. For instance, the features intensity, shape, texture of the tumor

may be applied to an MRI scan for tumor segmentation. Further characteristics and properties may be analyzed via machine learning algorithms. The only drawback of this application is the requirement of domain expertise.

During the **era of deep learning,** Deep Neural Networks were introduced which eliminated the need for domain expertise due to their in – built potential to determine and analyze which extraction features should be applied in order to obtain the desired results with accuracy. The foundational research paper that indicated the implementation of mathematical and biological perspective in the establishment of Deep Neural Networks was titled "Backpropagation for Neural Networks", authored by Rumelhart, Hinton and Williams in 1986. Although beneficial, Deep Neural Networks had their own limitations of requiring massive amounts of data and extensive computational techniques in addition to increased human error rates.

Yet, deep learning alongside **Natural Language Processing (NLP) algorithms** has been able to extend its applications in almost all biomedical sciences and plays an essential role in medical imaging, data analysis and study of brain and body machine interface.

The session concluded with the discussion of the current technical challenges that are existent till date including explainability, robustness that add noise to the original data, regulations around data and algorithmic bias.

Figure 3 – Session on Applications of Artificial Intelligence (AI) in Healthcare Industry with Dr. Neel Das

2.2 Introduction to Artificial Intelligence (AI) and Machine Learning (ML) and its Applications in Bioinformatics

Speaker – Alok Anand

Alok Anand sir commenced his educational speech by defining Machine Learning (ML) in terms of finding patterns from data, via different strategies, rules, criteria, parameters, and further using those patterns for analysis or predictions. The fundamental domains of Artificial Intelligence (AI) and Machine Learning (ML) were precisely stated including Predictive Modelling, Pattern Recognition, Natural Language Processing (NLP) and Multimodality. Diffusion models and ODE Methods were developed to retrieve the desired output from a given set of biological data via "prompts".

The significance of deciphering "omics" data (genomics, transcriptomics, proteomics, metabolomics and phenomics) plays an essential role in Precision Medicine. Precision Medicine is an amalgamation of Personalized Medicine, Data Science and Artificial Intelligence (AI) that designs personalized healthcare treatment course and individually tailored medicine doses taking into consideration an individual's biomedical history and lifestyle along with continuous monitoring of specific therapy targeted biomarkers.

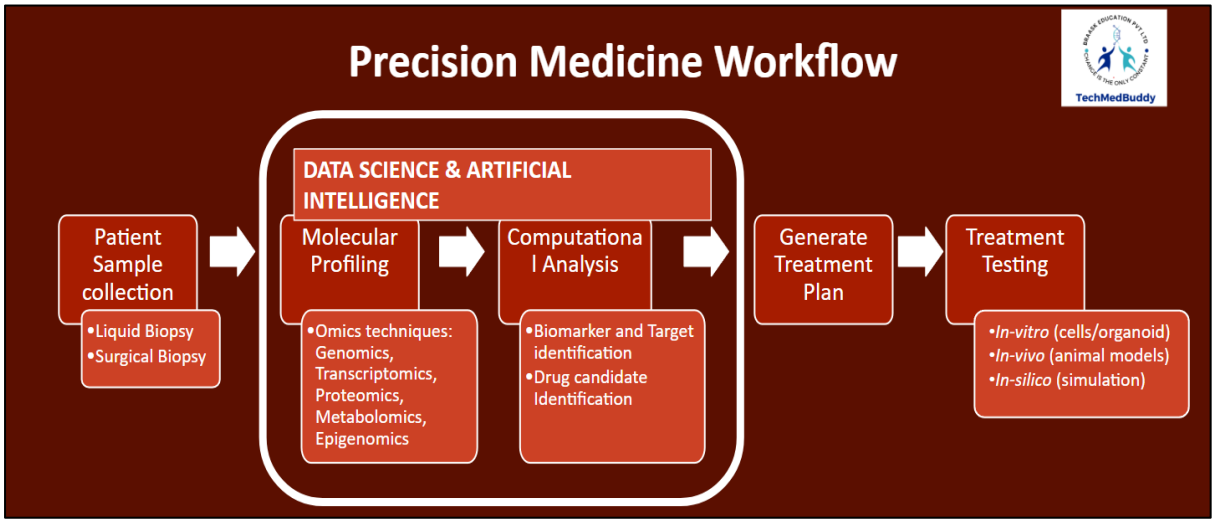


Figure 4 – Workflow of Precision Medicine

The concept of Data Visualization was simplified and explained using an instance of pre – attentive processing. The sole purpose of data visualization, which involves learning through hypothesis, goals, parameters, dynamic and static data, is –

1. Time Series – Series of events depending on time
2. Part to Whole
3. Ranking – represents level of intensity
4. Distribution
5. Correlation – represents degree of similarity
6. Deviation
7. Nominal comparison – represents variable names

In addition to explaining about the extensive range of biological databases available for retrieving scientific data such as –

1. **KEGG (Kyoto Encyclopedia of Genes and Genomes) –** A secondary database used for understanding and simulating higher – order functional behaviours of the cell / organism from its genome information.
2. **cBioPortal –** An open – access, open resource for interactive exploration of multidimensional cancer genomic data sets, that functions via data integration, data visualization and data correlation.
3. **TCGA (The Cancer Genome Atlas) –** A tool used to study genetic alterations and gene expression profiles in cancer along with clinical information of cancer patients.
4. **NCBI GEO (Gene Expression Omnibus) –** A gene repository that stores and shares a wide array of high – throughput genomics data.
5. **NCBI SRA (Sequence Read Archive) –** A public database that stores raw sequencing data and alignment information generated by various high – throughput sequencing technologies.

Figure 5 – Biological Databases

Emphasis was laid on the involvement of AI and ML tools in analyzing the results obtained from laborious and time - consuming experimental approaches such as Analysis of DNA Microarray and RNA - Seq Data. Sir's session significantly transformed our perspective as to how AI and ML may efficiently contribute to the research in the computational biology sector.

2.3 Fundamentals of Python Programming

Speaker – Manas Pratiti

Through this session on Python Programming covered by Manas Pratiti ma'am, it was extremely beneficial for us to have practical experience of understanding the programming language using Google's Colab platform and Jupyter Notebooks.

Google Colab is a hosted Jupyter Notebook service well – suited to machine learning and data science fundamentals and requires no setup to use. It provides free access to computing resources including GPUs and TPUs. The tool provides a programming interface that enables the user to write and execute a block of code, view the output through a terminal and create various data visualizations.

Figure 6 – Google Colab and Jupyter Notebooks

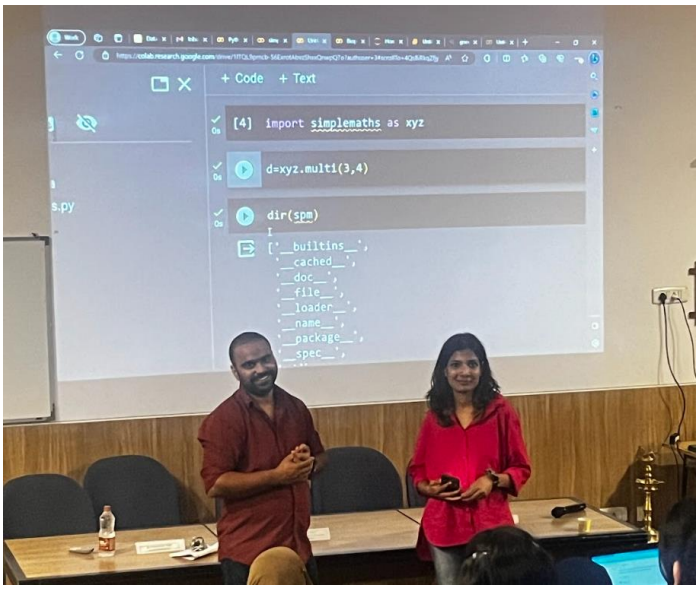


Figure 7 – Python Programming Session with Manas Pratiti ma'am

Before providing a detailed explanation of how AI can be used in Python programming to manipulate data to achieve a particular result, ma'am also explained the basic concepts of Python required to build a structure or a block of code. These basics included learning about variables, datatypes, conditional and iterative statements, nested statements, functions and various in - built libraries such as NumPy and math, which can be imported in the program for advanced purposes. Her session concluded with displaying the applications of Python programming to manage biological data sets.

CHAPTER 3

SESSIONS CONDUCTED ON DAY 2 – 7th OCTOBER, 2023

3.1 Introduction to the Fundamentals of Statistics and Biostatistics

Speaker – Alok Anand

Alok Anand sir's explanation of the fundamentals of statistics and biostatistics essentially taught us about a variety of terminologies that form the basis of statistical methodologies in bioinformatics.

These terminologies include samples, that represent a subset of a population which is further defined as the largest collection of entities according to one's interest at a specific moment of time. Following are the three primary types of variables that act as crucial parameters during a scientific study – Random variables (those variables whose value cannot be predicted in advance at a specific point of time), Quantitative Variables (variables that can be measured in numbers) and Qualitative Variables (variables that represent quality of data in terms of categorizing the data).

For a particular study, data may be derived from various sources that can be categorized into Primary Data and Secondary Data. Primary Data has been retrieved from actual experimental techniques and investigations, developed and conducted respectively by the researcher himself whereas Secondary Data is the data reliant on another source such as research papers and publications.

An integral component of this session comprised the Data Visualization Plots without which analyzing and interpreting complex biological data sets is challenging. Few of such essential plots include – Volcano plot, HeatMap, BoxPlot, and t - SNE plot.

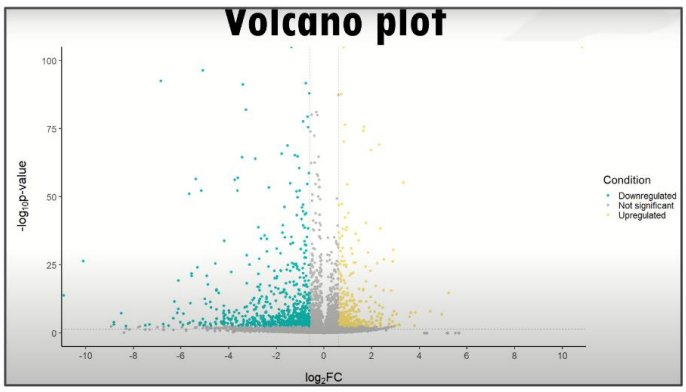


Figure 8 – Volcano Data Visualization Plots

The BoxPlot, also known as a Box – and – Whisker Plot, is considered one of the best plots comprising the potential to interpret the central tendency, spread and skewness of the data and potential outliers. HeatMaps enable the visual representation of patterns or relationships in a matrix of data, where the intensity of the colour represents the expression level of the intersecting gene and sample.

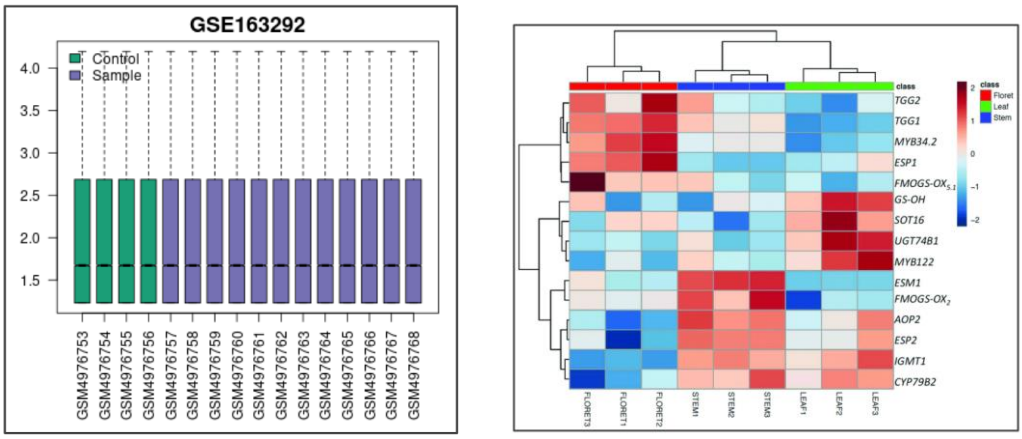


Figure 9 – BoxPlot and HeatMap Data Visualization Plots

Similarly, a t – SNE plot, that stands for t – Distributed Stochastic Neighbour Embedding, identifies clusters or patterns in complex datasets and represents each data point as a point in the reduced space. For instance, it transforms the visualization of a high – dimensional data into a lower – dimensional space.

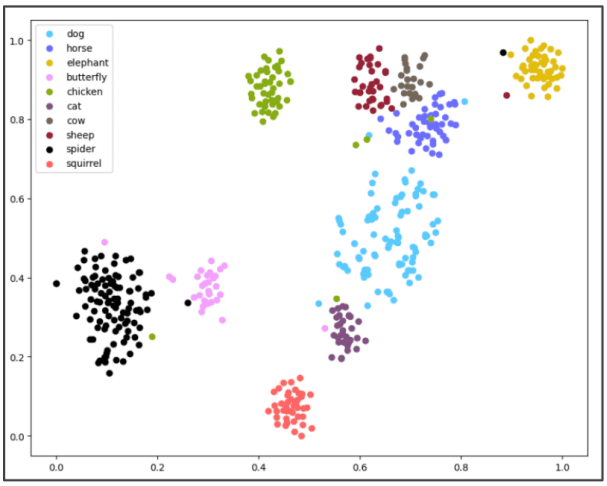


Figure 10 – t – SNE Data Visualization Plot

In addition to explaining various measurement scales such as Nominal scale, Ordinal scale, Interval scale, Ratio scale and Confidence Intervals, Alok sir also guided us in understanding the Confusion Matrix, that is represented by the following 4 categories –

Table 2 – *Categories of the Confusion Matrix*

Actual Labels	1	0	0	1
Predicted Labels	1	0	1	0
Interpretation	True Positive (TP)	True Negative (TN)	False Positive (FP) (Type I error)	False Negative (FN) (Type II error)
Representation	-	-	α	β

This categorization can further be used in computing and estimating the Precision, Specificity and Sensitivity of a data set. Moreover, Alok sir also discussed each step of the journey of scientific research minutely, covering each and every detail commencing from the idea, hypothesis formation and data collection to the statistical analysis via mathematical modelling and interpretation of the results obtained.

The three essential statistical analysis tests that play a critical role in inference and decision – making process, were addressed in the session's conclusion.

1. **T test –** The T test is applicable when comparing the mean of two groups under the assumption that data are continuous and randomly sampled from a population with a homogeneity of variance.

2. **ANOVA test –** Statistically significant difference can be estimated between two or more categorical groups using the ANOVA test, which can be sub – categorised into one – way ANOVA test and two - way ANOVA test with the number of independent variables acting as the differentiating factor between the two classifications.

3. **Kruskal Wallis Test –** A non – parametric statistical test by nature, the Kruskal Wallis Test acts as a medium for comparison of three or more independent groups with respect to their median values, by evaluating the ranks of all the observations across all the groups and statistically determining the competence of the set hypothesis on the basis of the summation of these rankings.

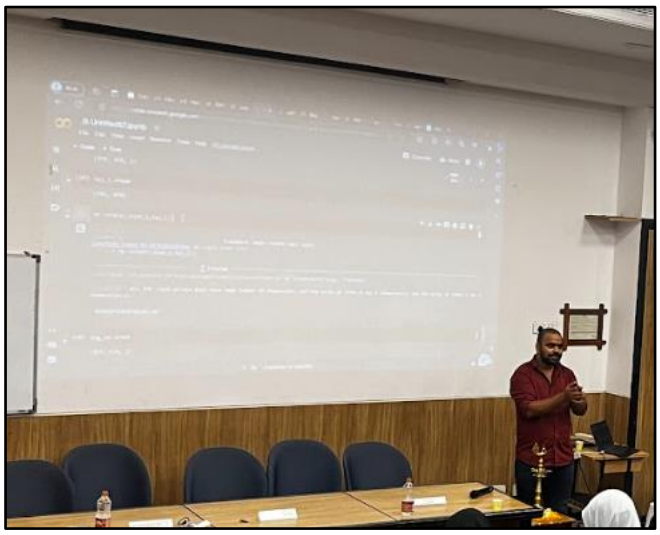


Figure 11 – Session on Introduction to the Fundamentals of Statistics with Alok Anand sir

3.2 Hands – On Session : Differential Expression and Enrichment Analysis

Speaker – Manas Pratiti

We had the opportunity to study the significance of differential expression and enrichment analysis through a practical hands-on session guided by Manas Pratiti ma'am, which further introduced us to the GEO2R (Gene Expression Omnibus) platform, an interactive bioinformatics tool used for identification and comparison of differentially expressed gene data sets.

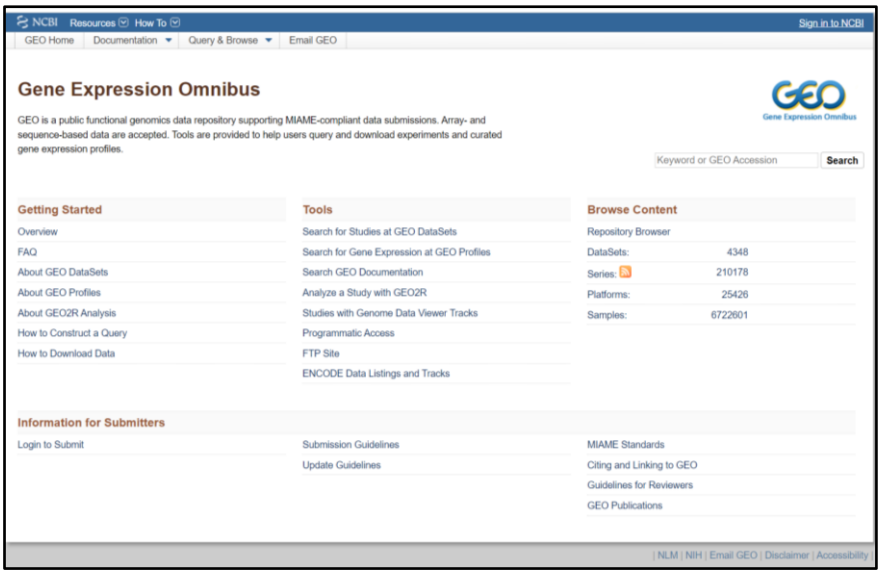


Figure 12 – Homepage of GEO2R Database

The GEO2R Platform employs unique identifiers for segregating curated gene expression data sets appropriately within the repository for locating the records of interest with a greater efficiency and accuracy and specific to the objective of the scientific domain under study. These essential unique identifiers include –

1. GEO Dataset (GDS) – It comprises a curated data set that has been derived from the existing data submitted by the investigator.

2. Series (GSE) – It displays a list of expression profiles conducted for the experiment.

3. Samples (GSM) – It consists of information specific to the biological samples used in the experiments, including extraction procedures.

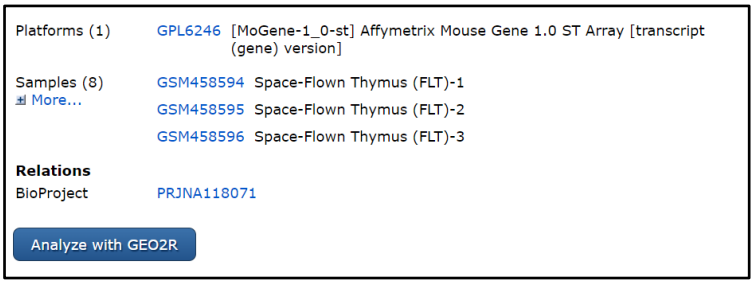
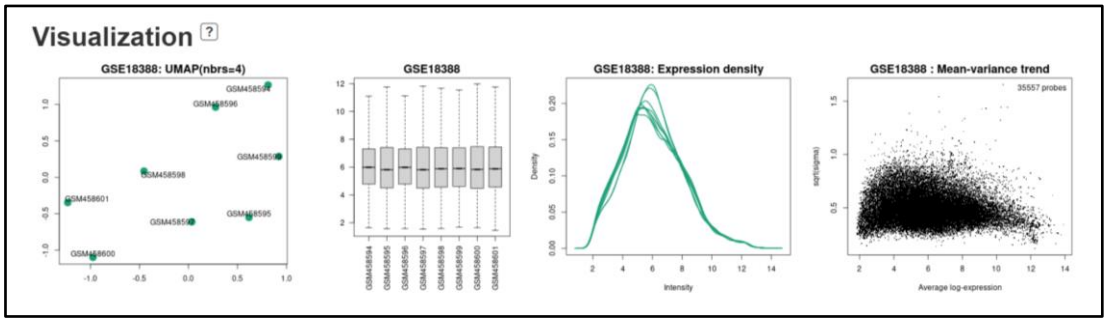


Figure 13 – Unique Identifiers utilized by the GEO2R Platform

This hands-on experience allowed us to apply the newly acquired understanding of the analytical tool in predicting and deriving valuable outputs from the large real – world biological datasets and interpreting the results through Data Visualization plots and parameters such as – P. Value, adj. P. Val, t and B parameters.

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
10358454	0.00509	1.96e-07	13.48	4.3773	1.384	Rbm3	RNA binding motif protein 3
10603469	0.00509	2.86e-07	12.92	4.24228	1.304	Rbm3	RNA binding motif protein 3
10556113	0.00546	4.61e-07	12.24	4.06259	1.408	Rbm3	RNA binding motif protein 3
10535904	0.00844	9.50e-07	-11.28	3.76621	-1.894	Hsph1	heat shock 105kDa/110kDa p...
10490946	0.03472	5.44e-06	-9.21	2.93118	-1.02	Hsp90aa1	heat shock protein 90, alpha (...
10514713	0.03472	5.86e-06	-9.13	2.89184	-0.948	Wdr78	WD repeat domain 78
10531415	0.03973	7.82e-06	-8.83	2.73562	-1.499	Cxcl10	chemokine (C-X-C motif) liga...
10391146	0.04368	9.83e-06	-8.59	2.60896	-1.232	Acly	ATP citrate lyase
10497079	0.0521	1.32e-05	-8.29	2.44138	-1.251	Ptger3	prostaglandin E receptor 3 (s...
10378848	0.05835	1.64e-05	-8.08	2.31365	-0.882	Hsp90aa1	heat shock protein 90, alpha (...

Figure 14 – Results obtained for the searched query "gse18388"



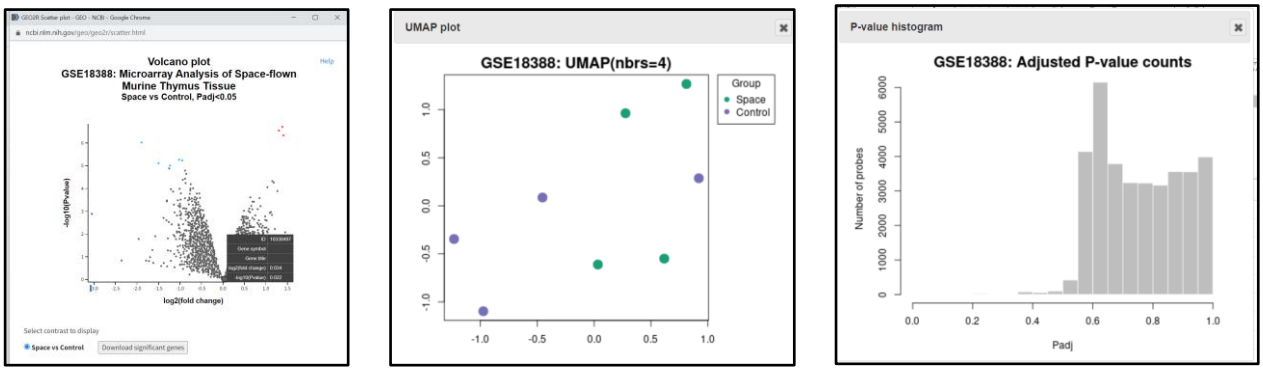


Figure 15 – Data Visualization Plots obtained for the searched query "gse18388"

3.3 Transitioning from Academia to Industry

Speaker – Dr. Tahseen Abbas

While sharing her own invaluable and motivational personal experiences, Dr. Tahseen Abbas ma'am, her talk offered an inspiration to further venture in the corporate world to gain knowledge of both the industrial and the scientific community. Emphasis was laid on developing the highly essential skill sets required to be proficient in the industry, some of which include programming and IT skills and pipeline developments, for instance, NGS Pipelines, apart from the soft skills such as communication, teamwork, and problem-solving required in this dynamic sector.

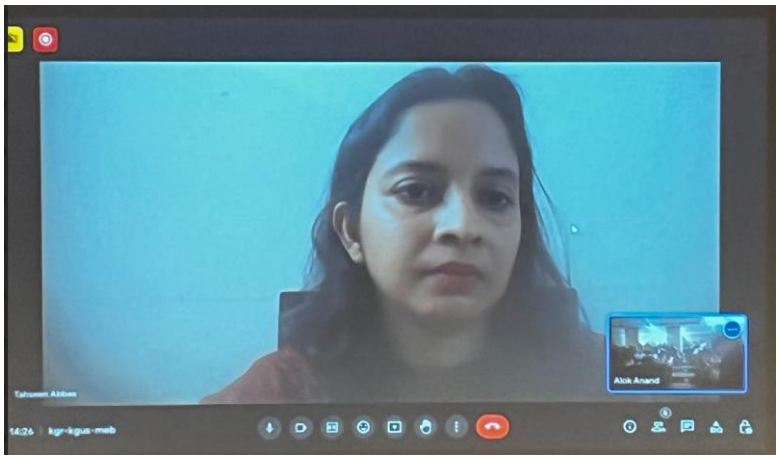


Figure 16 – Session for the Preparation for an Industrial Job with Dr. Tahseen Abbas ma'am

Proficiency in communication enables bioinformaticians to convey intricate scientific concepts to both technical and non – technical audiences with ease. Strong collaboration abilities enable bioinformaticians to work well with other scientists and professionals to accomplish shared objectives. Challenges faced during the scientific research can be identified and troubleshooted by bioinformaticians with good problem – solving abilities. Competence in the fundamental skill sets outlined by Dr. Tahseen Abbas ma'am makes bioinformaticians highly sought for and poised for substantial industry contributions.

3.4 AI – Enabled Clinical Decision Making and Support Systems (CDSS)

Speaker – Dr. Hara Prasad Mishra

During this informative talk delivered by Dr. Hara Prasad Mishra sir, the potential of AI - enabled Clinical Decision Making and Support Systems (CDSS) to revolutionize the healthcare sector was discussed. Any computer program created to assist medical personnel in making clinical choices is known as a clinical decision – support system.

To give patients individualized care, the healthcare sector is rapidly transforming with the advent of artificial intelligence (AI) models and tools. Large clinical datasets, which include details on a patient's symptoms, diagnosis, course of therapy, and results, can be managed by AI systems. Personalized care plans that are tailored to the individual requirements and characteristics of each patient can then be created using this Electronic Health Record (EHR).

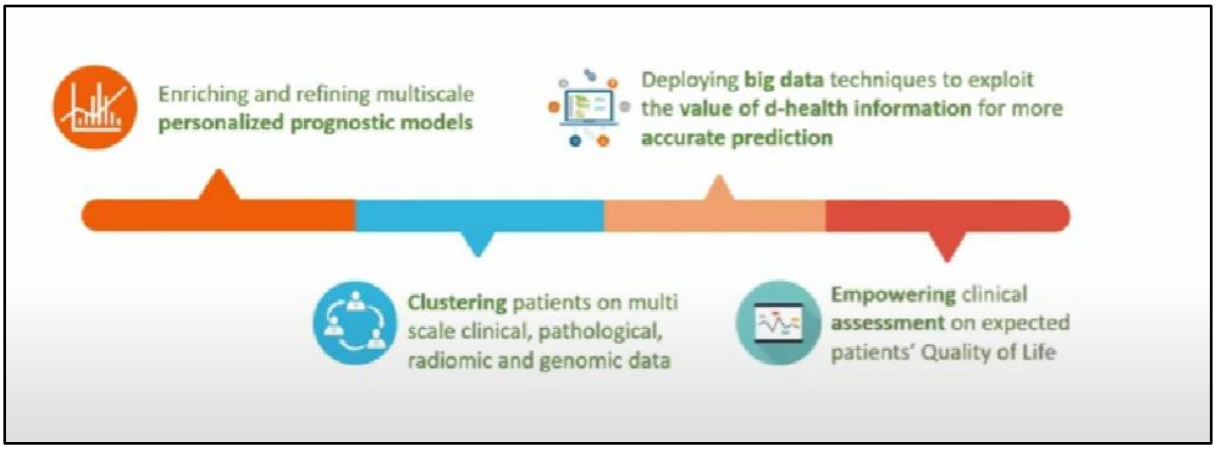


Figure 17 – Timeline of a CDSS System employed for Enhancing prognosis

Not only the patients, but also the healthcare professionals and practitioners are benefitting from the immense potential of AI Models and Tools in the sector of CDSS, which has significantly reduced their workload and enabled them to make more informed decisions regarding the diagnosis and treatment of a wide array of diseases. Through such a system, AI algorithms can be utilized for –

1. Identifying patterns in the records of a patient that may be challenging for humans to detect
2. Risk prediction and predicting patient outcomes
3. Recommend medicine doses and course of treatment that is based on real – world evidence generated, taking into consideration the needs, lifestyle and preferences of an individual.
4. Drug discovery
5. Real – time monitoring and predictive analysis

Although beneficial, CDSS has its own drawbacks. Multiple factors to be taken into consideration include – Data quality and Interoperability, ethical and legal considerations, trust and acceptance by healthcare professionals, integration of such as system into existing clinical workflows and algorithmic biasness and explainability.

Figure 18 – Session on Artificial Intelligence (AI) – Enabled Clinical Decision Making and Support Systems (CDSS) with Dr. Hara Prasad Mishra sir

KEY TAKEAWAYS FROM THE WORKSHOP

1. **Comprehensibility and Practical Awareness –** Having gained a concrete understanding of the techniques and strategies underlying a scientific research, practical exposure has been acquired to manage real – world dynamics in the bioinformatics domain.
2. **Competency in the fundamentals of statistics and biostatistics –** The workshop has been instrumental in providing a foundation in the domain of statistics. Statistics are the need of the hour in order to comprehend core concepts such as hypothesis formation and data visualization that are further utilized for the interpretation and insightful analysis, hence making informed decisions.
3. **Recognizing the Importance of Machine Learning in Healthcare domain –** AI models and tools have been developed to manage large clinical datasets, integrate them into personalized care of the patients and aid the professionals and practitioners of the healthcare industry in significantly reducing their workload and making more informed decisions regarding the diagnosis and treatment of a wide array of diseases.
4. **Creating a job – embedded professional development –** The workshop essentially imparted encouragement and a stimulus to advance deeper into corporate world to learn and gain insights about both the scientific and industrial communities.

CONCLUSION

This interactive and informative workshop based on the "Introduction to Statistics and Machine Learning in Bioinformatics" has played a significant role in broadening our vision to the immense potential of Artificial Intelligence (AI) and Machine Learning (ML) in not only the industrial IT sector but also in bioinformatics and various other biomedical fields. Their applications have essentially displayed their relevance in enhancing the healthcare industry with respect to both the esteemed professionals and the patients in diagnosing and providing them with personalized course of treatment. Through the knowledge acquired from this workshop, we have essentially gained a strong foundation in understanding various concepts and terminologies of machine learning, fundamentals of statistics, which can be practically used further to enhance and refine the interpretation of the results obtained for a particular biological data set. This workshop has been extremely beneficial as it can serve as the base knowledge in further exploration of the scientific domains.

REFERENCES

1. Moore J. H. (2007). Bioinformatics. *Journal of cellular physiology*, *213*(2), 365–369. https://doi.org/10.1002/jcp.21218
2. Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002 Apr 27;324(7344):1018-22. doi: 10.1136/bmj.324.7344.1018. PMID: 11976246; PMCID: PMC1122955.
3. Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2
4. Musen, M.A., Shahar, Y., Shortliffe, E.H. (2006). Clinical Decision-Support Systems. In: Shortliffe, E.H., Cimino, J.J. (eds) Biomedical Informatics. Health Informatics. Springer, New York, NY. https://doi.org/10.1007/0-387-36278-9_20