DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to Visvesvaraya Technological University (VTU), Belagavi, Approved by AICTE and UGC, Accredited by NAAC with 'A' grade & ISO 9001 – 2015 Certified Institution)

Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Alternate Assessment Tool Report submitted for the subject

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING – IPCC22IS52

Submitted by

Aditya Singh (1DS22IS008)
Harsh Joshi (1DS22IS057)

Under the Guidance of
Dr. Annapurna P Patil
Dean Academics, Professor and HOD
Department of Information Science and Engineering.
DSCE, Bengaluru

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANASANGAMA, BELAGAVI-590018,
KARNATKA 2024-25

DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to Visvesvaraya Technological University (VTU), Belagavi, Approved by AICTE and UGC, Accredited by NAAC with 'A' grade & ISO 9001 – 2015 Certified Institution)
Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

(Accredited by NBA Tier 1: 2022-2025)

CERTIFICATE

This is to certify that the Alternate Assessment Tool (AAT) entitled "Prediction of Students Grades using KNN" as part of ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (IPCC22IS52) is a bonafide work carried out by Aditya Singh (1DS22IS008), Harsh Joshi (1DS22IS057) in partial fulfilment for the 5th semester of Bachelor of Engineering in Information Science & Engineering of the Visvesvaraya Technological University, Belgaum during the year 2024-2025. The AAT report has been approved as it satisfies the academic requirements prescribed for the Bachelor of Engineering degree.

DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to Visvesvaraya Technological University (VTU), Belagavi, Approved by AICTE and UGC, Accredited by NAAC with 'A' grade & ISO 9001 – 2015 Certified Institution)
Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-560078

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

(Accredited by NBA Tier 1: 2022-2025)

DECLARATION

We declare that we abide by the ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice. The work submitted in this report of ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (IPCC22IS52), Vth Semester BE, ISE has been compiled by referring to the relevant online and offline resources to the best of our understanding and in partial fulfilment of the requirement for the award of the degree of Bachelor of Engineering in Information Science & Engineering, at Dayananda Sagar College of Engineering, an autonomous institution affiliated to VTU, Belagavi during the academic year 2024-2025.

We hereby declare that the same has not been submitted in part or full for other academic purposes.

(1DS22IS017 ADITYA SINGH)
(1DS22IS019 HARSH JOSHI)

Place:

Date:

ABSTRACT

The work presents a report on the study, design, implementation, and evaluation of various Machine Learning Algorithms, emphasizing an understanding of their performance. As part of the Alternate Assessment Tool (AAT) for the subject ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING (IPCC22IS52), a team of five students completed and submitted one assignment. This assignment covered the Basics of Artificial Intelligence in Part A.

In Part B, the team analysed the working principles of Harnessing AI for Breakthroughs in Computational Biology, observed its implementation, and evaluated its performance. The relevance of the implemented algorithm to real-world applications was also understood, and the team successfully presented their findings.

Keywords:

1. **Logistic**
2. **Regression**
3. **Prediction**
4. **Diagnosis**
5. **Features**
6. **Accuracy**
7. **Metrics**
8. **Precision**
9. **Recall**
10. **Confusion**
11. **Malignant**
12. **Benign**

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and support of a large number of individuals who have been responsible for the successful completion of this project.

We take this opportunity to express our sincere gratitude to Dayananda Sagar College of Engineering for having provided us with a great opportunity to pursue our Bachelor Degree in this institution.

In particular we would like to thank Dr. B G Prasad, Principal, Dayananda Sagar College of Engineering for his constant encouragement and advice.

Special thanks to Dr. Annapurna P Patil, Professor and HoD, Department of Information Science Engineering, Dayananda Sagar College of Engineering for her motivation and invaluable support well through the development of this project.

We are highly indebted to our faculties Dr. Annapurna P Patil, Professor and HoD & Mrs. Shalini K B, Assistant Professor, Department of Information Science & Engineering, Dayananda Sagar College of Engineering for their constant support and guidance. Our guide has been a great source of encouragement throughout the course of this project.

CONTENTS

PART A

1. Assignment 1 1

PART B

1. Introduction 14
2. Objectives 15
3. System Architecture 16
4. Applications 17
5. Key Techniques 19
6. Results 24
5. Future Prospects 25
6. Conclusion 27

References 28

PART A

Assignment - 1
Week - 1

Submitted by
Aditya Singh

Date: 21-09-2024

1 Explain the steps involved in Resolution Refutation Proof. Discuss an algorithm to convert predicate formula to casual form.

Steps involved

1. Convert the given axiom to casual form.
2. Apply and prove the given Goal using negation rule.
3. Use the literals which all needed to prove.
4. Iteratively and solution to set and add resolvent to the set.
5. Continue until no further resolution can be added or a null clause can be obtained.

Algorithm to convert to clausal form:

1) Eliminate implications : $p \to q$ replace by $\neg p \vee q$
2) Eliminate biconditional : $p \Leftrightarrow q$ replace $(p \to q) \wedge (q \to p)$
3) Take negation inwards
4) Apply Distributive laws:
   $(p \vee q) \wedge (r \vee s)$

2. What are some common risks and benefits of AI?

Risk

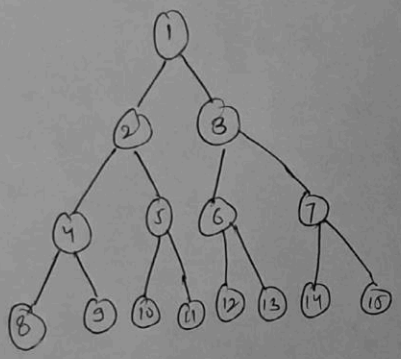
1) Reducing creativity individuals.
2) Creating unemployment
3) Reducing the knowledge in Subject.

Benefits:

1) Reducing human efforts
2) Do dull monotonous jobs easily
3) Performs the same task any number of times without getting bored.
4) Can perform life threatning jobs too
5) Can work of any number of hours. Do not have physical constraints unlike humans

3. Question 3:

a. State Space



BFS (order):

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15

Depth limit Search (3): 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, (11) → found

Iterative Deepening Search: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11.

Bidirectional Search: This would not work well.

Branching factor is 2. No. of nodes to be explored in both the directions grows exponent-ially.

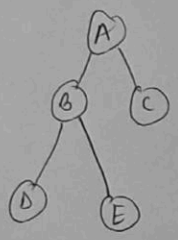
4. Prove : or Give Counter Example:

1 BFS is a special case of Uniform cost search

* In BFS all edges have equal cost.
* UCS Explores all the nodes based to their cummulative cost g(n).
* Since in BFS all nodes based on their cummulative ~~cost~~ g(n) is equal to depth of node
* UCS explores nodes in increasing order of depth, which was same as BFS.

Proved.

Ex. 2. DFS is a special case of BFS :- (Best fit search)



BFS : A B D E C

To construct the BFS we require heuristic functions. Based on heuristic function, the BFS varies

3. Uniform cost search is special case of A* algorithm.

Proof:

1. A* uses, $f(n) = g(n) + h(n)$. g(n) → cummulative cost
   h(n) → heuristic function

2. If $h(n) = 0$, A* reduces to UCS

∴ UCS is a special case of A*

5. Question 5.

P → Anyone passing his history exams
q → win lottery
r → Being happy

1. $\forall x$ (Pass History $(x) \wedge$ winLottery $(x) \rightarrow$ Happy (x))
2. $\forall x$ (Study $(x) \vee$ Lucky $(x) \rightarrow \forall y$ (pass them $[x, y]$)
3. $\sim$ Study (John) $\wedge$ Lucky (John)
4. $\forall x$ Lucky $(x) \rightarrow$ winLottery (x).

C ~ F :- C1. $\sim$ passHistory $(x) \vee \sim$ winLottery $(x) \vee$ Happy (x)

C2. $\sim$ Study $(x) \wedge \sim$ Lucky $(x) \vee$ pass them (x, y)

C3. $\sim$ Study (John) C4. Lucky (John)

C5. $\sim$ Lucky (John) $\vee$ winLottery (x)

∴ John is happy.

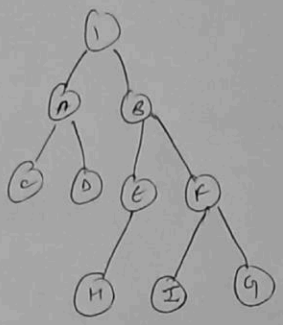
Lucky (John) $\quad \sim$ Lucky (John) $\quad \vee$ winLottery (x)

winLottery $(x) \quad \sim p$ History $(x) \vee \sim w$ Lottery $(x) \vee$ Happy (x)

$\sim p$ History $(x) \vee$ Happy (x)

$\sim$ Happy (John)

6. Trace the operation of A* search on any given Case Study and comment on performance of the same.



Node	h(n)
S	13
A	12
B	4
C	7
D	0
E	8
F	2
G	4
H	9
I	10

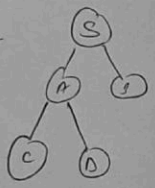
Step 1. f(n)

S→A = 14

S→B = 15

S→A

Step 2: f(n) =



S→A→C →17

S→A→D = 21

S→A→D

Time complexity : $O(b^d)$

Space complexity : $O(b^d)$

7. Question 7

1. True : Because DFS explores without knowing goal whereas A* goes with heuristic fn as a result explores lesser nodes than DFS.
2. True : h(n) = 0 is admissible heuristic for 8 puzzle because it never overestimate the cost to reach the goal.
3. False : A* is useful in robotics
4. False : If 0 cost all allowed BFS explore nodes indefinitely
5. True : Because heuristic ignores underestimates real cost as it goes potential obstacles.

PART B

Harnessing AI for Breakthroughs in Computational Biology

1. Introduction

Computational biology, an interdisciplinary field that combines biology, computer science, and mathematics, is pivotal in analyzing and understanding biological systems. It has revolutionized how researchers process complex biological data, such as DNA sequences, protein structures, and cellular pathways. However, the exponential growth of biological data due to advanced technologies like high-throughput sequencing has created challenges in data analysis and interpretation.

Artificial Intelligence (AI), with its ability to model complex patterns and make predictions from large datasets, has emerged as a transformative tool in computational biology. By leveraging techniques like machine learning and deep learning, AI enables breakthroughs in genomics, proteomics, drug discovery, and personalized medicine. AI-powered tools are not only accelerating research but also offering unprecedented insights into biological processes that were once beyond human comprehension.

This report explores the integration of AI into computational biology, highlighting its applications, challenges, and potential to reshape the future of biological research and healthcare.

2. Objectives

1. To investigate the transformative role of Artificial Intelligence (AI) in advancing computational biology and its implications for modern biological research.

2. To comprehensively analyze the key applications of AI across diverse domains, including genomics, proteomics, drug discovery, and precision medicine.

3. To evaluate the methodologies, algorithms, and computational techniques employed in AI to process, interpret, and derive insights from complex biological datasets.

4. To review and synthesize real-world case studies that illustrate the efficacy of AI in addressing critical challenges in biology and healthcare.

5. To critically assess the technical, ethical, and practical limitations of integrating AI within the framework of computational biology.

6. To explore future directions and emerging innovations that underscore the evolving synergy between AI and computational biology, fostering breakthroughs in science and medicine.

3. System Architecture

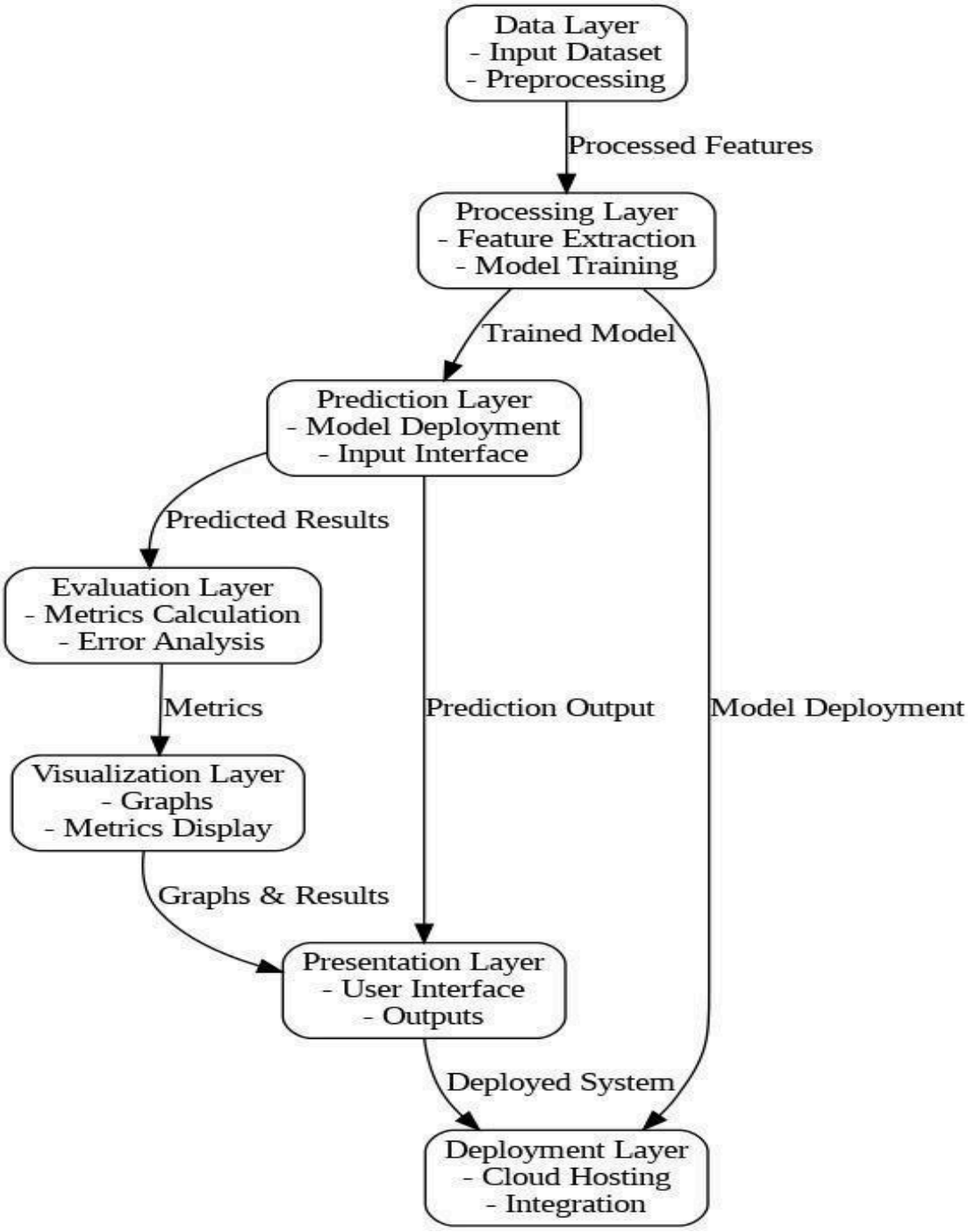


Figure 1: System Architecture

4. Applications

Artificial Intelligence (AI) has become a cornerstone in computational biology, providing sophisticated tools to analyze vast amounts of biological data. Below are some of the key applications where AI has demonstrated significant impact:

1. Genomics

AI has revolutionized genomics by improving the accuracy and efficiency of DNA sequence analysis.

- **Genome Annotation**: AI algorithms are used to identify genes and functional elements within genomic sequences.
- **Variant Detection**: Machine learning models detect genetic variations associated with diseases.

2. Proteomics

Understanding protein structures and functions is critical in biology, and AI has made remarkable strides in this area.

- **Protein Structure Prediction**: Deep learning models, such as AlphaFold, predict protein 3D structures with near-experimental accuracy.
- **Protein-Protein Interactions**: AI models analyze interactions to understand cellular mechanisms and identify therapeutic targets.

3. Drug Discovery and Development

AI accelerates the drug discovery process, reducing costs and development timelines.

- **Virtual Screening**: AI models predict the binding affinity of drug molecules to target proteins.
- **De Novo Drug Design**: Generative AI creates novel molecular structures with desired properties.

4. Disease Diagnosis and Precision Medicine

AI-powered tools enable earlier and more accurate diagnosis of diseases and the development of personalized treatment plans.

- **Disease Prediction**: AI models analyze genetic and clinical data to predict disease risks.
- **Biomarker Discovery**: Machine learning identifies biomarkers for targeted therapies.
- **Customized Therapies**: AI tailors treatment plans based on individual genetic and lifestyle profiles.

5. Epidemiology and Public Health

AI contributes to understanding and controlling disease outbreaks.

- **Disease Surveillance**: AI models predict disease outbreaks based on environmental and social factors.
- **Public Health Planning**: AI assists in resource allocation and intervention strategies.

5. Key Techniques in AI for computational biology

1. Machine Learning (ML)

Machine Learning is a foundational AI approach that enables systems to learn patterns from data and make predictions or decisions without explicit programming.

- **Supervised Learning**: Models are trained on labeled datasets to predict outcomes, such as identifying genetic mutations linked to diseases.
- **Unsupervised Learning**: Clustering and dimensionality reduction techniques (e.g., PCA, t-SNE) uncover hidden structures in biological data, such as grouping similar cell types.

2. Deep Learning

Deep Learning, a subset of ML, utilizes neural networks with multiple layers to model complex relationships in data.

- **Convolutional Neural Networks (CNNs)**: Used in image analysis, such as identifying cancerous cells in histopathological images.
- **Recurrent Neural Networks (RNNs) and Transformers**: Analyze sequential data, such as DNA or RNA sequences, and predict their functions.

3. Generative Models

Generative AI creates new data instances that resemble the original data.

- **Generative Adversarial Networks (GANs)**: Generate synthetic datasets for training, such as simulated gene expression profiles.

4. Bayesian Methods

Bayesian approaches incorporate prior knowledge into probabilistic models, making them suitable for uncertain or noisy biological data.

- **Bayesian Networks**: Model dependencies among variables, such as gene regulatory networks.
- **Bayesian Inference**: Used in phylogenetics to estimate evolutionary relationships.

5. Natural Language Processing

NLP techniques allow AI to process and interpret textual data, making them invaluable in biomedical research.

- **Text Mining**: Extracting biological information from research papers and clinical notes.
- **Semantic Analysis**: Understanding relationships between biological entities in databases and literature.

Cancer Prediction :-

Cancer Prediction

Dataset Information:

Target Variable (y):

Diagnosis (M = malignant, B = benign) Ten features (X) are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (coastline approximation - 1)

For each characteristic three measures are given:

a. Mean b. Standard error c. Largest/ Worst

```
# Step 1 : import library
import pandas as pd


# Step 2 : import data
cancer = pd.read_csv('https://github.com/YBIFoundation/Dataset/raw/main/Cancer.csv')
```

```
cancer.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.

5 rows × 33 columns

```
cancer.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      569 non-null    int64
 1   diagnosis               569 non-null    object
 2   radius_mean             569 non-null    float64
 3   texture_mean            569 non-null    float64
 4   perimeter_mean          569 non-null    float64
 5   area_mean               569 non-null    float64
 6   smoothness_mean         569 non-null    float64
 7   compactness_mean        569 non-null    float64
 8   concavity_mean          569 non-null    float64
 9   concave points_mean     569 non-null    float64
 10  symmetry_mean           569 non-null    float64
```

```
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

```
cancer.describe()
```

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_me
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.0000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.0887
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.0797
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.0000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.0295
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.0615
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.1307
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.4268

8 rows × 32 columns

```
# Step 3 : define target (y) and features (X)
cancer.columns
```

```
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],
      dtype='object')
```

```
y = cancer['diagnosis']
X = cancer.drop(['id','diagnosis','Unnamed: 32'],axis=1)
```

```
# Step 4 : train test split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size=0.7, random_state=2529)
```

```
# check shape of train and test sample
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((398, 30), (171, 30), (398,), (171,))
```

```
# Step 5 : select model
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(max_iter=5000)
```

```
# Step 6 : train or fit model
model.fit(X_train,y_train)
```

```
LogisticRegression
```

```
model.intercept_
```

```
array([-30.20269391])
```

```
model.coef_
```

```
array([[-0.8644508 , -0.1823121 ,  0.26510852, -0.02688942,  0.13284582,
         0.19445151,  0.40918278,  0.20206338,  0.17199488,  0.03798515,
         0.0192444 , -1.13284188, -0.13597054,  0.11911954,  0.02266663,
        -0.03006638,  0.04691738,  0.02805721,  0.03329433, -0.00980702,
        -0.27140621,  0.44034405,  0.16566196,  0.01286379,  0.2719812 ,
         0.59704539,  1.06177846,  0.40903862,  0.51193487,  0.08436947]])
```

```
# Step 7 : predict model
y_pred = model.predict(X_test)
y_pred
```

```
array(['B', 'M', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'B', 'M', 'B',
       'M', 'B', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'B', 'B', 'M',
       'B', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'M', 'B', 'B', 'B',
       'M', 'M', 'M', 'M', 'M', 'B', 'B', 'M', 'M', 'M', 'B', 'B', 'B',
       'B', 'B', 'B', 'B', 'B', 'M', 'M', 'M', 'B', 'M', 'B', 'M', 'M',
       'M', 'M', 'B', 'M', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'B',
       'M', 'M', 'M', 'B', 'B', 'M', 'M', 'M', 'B', 'B', 'B', 'B', 'M',
       'B', 'B', 'B', 'M', 'B', 'M', 'B', 'B', 'M', 'B', 'M', 'B', 'B',
       'B', 'M', 'B', 'B', 'M', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B',
       'M', 'B', 'B', 'M', 'B', 'M', 'B', 'M', 'M', 'B', 'B', 'B', 'M',
       'M', 'B', 'M', 'M', 'M', 'B', 'B', 'M', 'B', 'M', 'B', 'M', 'B',
       'M', 'B', 'M', 'B', 'B', 'M', 'B', 'M', 'M', 'B', 'B', 'B', 'B',
       'B', 'M', 'M', 'M', 'M', 'B', 'B', 'B', 'M', 'B', 'M', 'B', 'B',
       'B', 'B'], dtype=object)
```

```
# Step 8 : model accuracy
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
confusion_matrix(y_test,y_pred)
```

```
array([[97,  5],
       [ 2, 67]])
```

```
accuracy_score(y_test,y_pred)
```

```
0.9590643274853801
```

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

```
              precision    recall  f1-score   support

           B       0.98      0.95      0.97       102
           M       0.93      0.97      0.95        69

    accuracy                           0.96       171
   macro avg       0.96      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171
```

6 Results

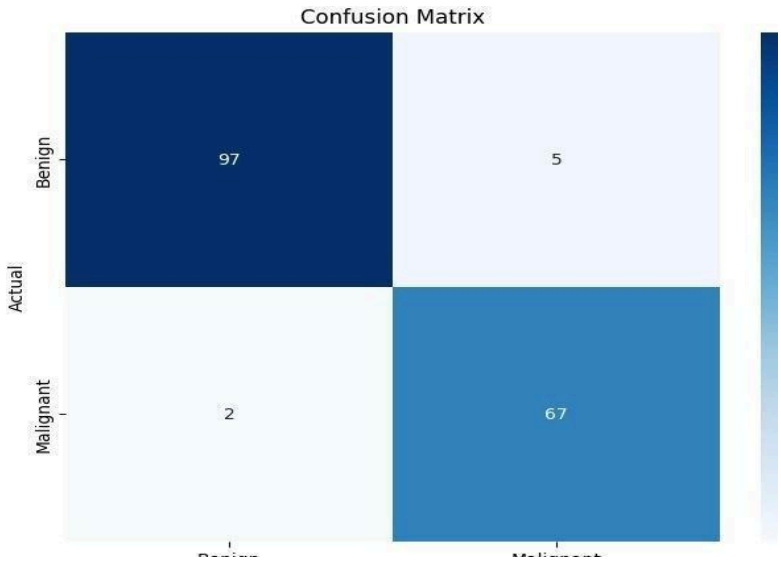


Fig 2: Confusion Matrix

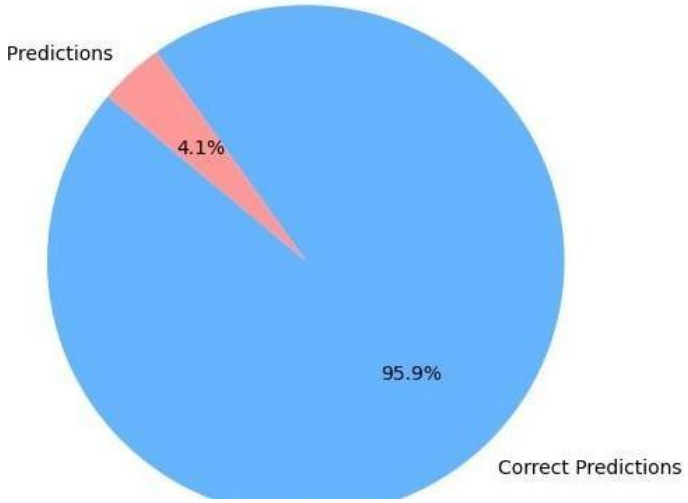


Fig 3: Model Accuracy

7. Future Prospects

The integration of Artificial Intelligence (AI) in computational biology is poised to revolutionize the field further, enabling breakthroughs in research, healthcare, and biotechnology. The future holds immense potential for AI to address existing challenges and unlock new possibilities.

1. Advancements in Multi-Omics Integration

AI will play a pivotal role in integrating data from genomics, transcriptomics, proteomics, metabolomics, and epigenomics.

- Holistic Understanding: Combining multi-omics data will enable comprehensive modeling of biological systems and diseases.
- Precision Medicine: AI-driven multi-omics integration will refine personalized treatment strategies for complex diseases like cancer and diabetes.

2. Revolutionizing Drug Discovery

AI will continue to accelerate the drug discovery process with greater precision and efficiency.

- AI-Driven Drug Design: Generative AI models will design highly specific drug molecules tailored to individual biological targets.
- Predictive Toxicology: AI systems will predict potential side effects and toxicity of drugs earlier in the development process.
- AI in Clinical Trials: Improved patient recruitment and stratification for clinical trials will enhance trial success rates and reduce costs.

3. Enhanced Protein Structure and Function Prediction

Building on successes like AlphaFold, future AI models will expand capabilities in structural biology.

- **Protein Dynamics**: AI will model dynamic protein behavior, providing deeper insights into enzymatic functions and interactions.
- **Post-Translational Modifications**: AI will predict modifications that impact protein function, aiding in drug targeting.

4. Synthetic Biology and Bioengineering

AI will transform synthetic biology by enabling the design and optimization of biological systems.

- **Custom Organism Design**: AI will guide the creation of organisms with specific traits for industrial, agricultural, or medical applications.
- **Bioinformatics Workflows**: AI will automate and optimize synthetic biology experiments, reducing trial-and-error processes.

5. AI-Powered Disease Prediction and Prevention

The next generation of AI tools will focus on proactive healthcare.

- **Real-Time Monitoring**: AI will integrate data from wearable devices and medical records to predict disease risks in real-time.
- **Global Health Surveillance**: AI models will predict and track disease outbreaks, aiding in rapid response and containment.

8. Conclusion

The integration of Artificial Intelligence (AI) into computational biology marks a transformative era in biological research and healthcare. AI's ability to process and analyze complex biological data has led to significant advancements in areas such as genomics, proteomics, drug discovery, and personalized medicine. By addressing long-standing challenges, such as data complexity and scalability, AI has opened new avenues for understanding biological systems and improving human health.

Despite its tremendous potential, the field faces challenges, including data quality issues, model interpretability, and ethical concerns. Overcoming these limitations requires continued innovation, interdisciplinary collaboration, and the development of transparent, scalable, and ethical AI solutions.

Looking ahead, the synergy between AI and computational biology promises groundbreaking discoveries, from designing life-saving drugs to engineering synthetic organisms and predicting disease outbreaks. As technology evolves, AI will play an increasingly central role in shaping the future of biological research, fostering a deeper understanding of life and transforming the way we approach global healthcare challenges.Highlighting the immense opportunities and responsibilities that come with leveraging AI in computational biology, emphasizing the need for a balanced, ethical, and forward-thinking approach to ensure its benefits are realized for science and society.

REFERENCES

[1] Zou, J., Huss, M., Abid, A., et al. (2019). "A primer on deep learning in genomics."*Nature Genetics*, 51(1), 12–18.

[2] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018)."Opportunities and obstacles for deep learning in biology and medicine."*Journal of The Royal Society Interface*, 15(141).

[3] Esteva, A., et al. (2017)."Dermatologist-level classification of skin cancer with deep neural networks."*Nature*, 542, 115–118.

[4] Senior, A. W., et al. (2020)."Improved protein structure prediction using potentials from deep learning."*Nature*, 577, 706–710.

[5] Zhao H., Caflisch A. Molecular dynamics in drug design. Eur. J. Med. Chem. 2015;91:4–14. doi: 10.1016/j.ejmech.2014.08.004.

[6] DiMasi J.A., Grabowski H.G., Hansen R.W. Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ. 2016;47:20–33. doi: 10.1016/j.jhealeco.2016.01.012.

[7] Kore P.P., Mutha M.M., Antre R.V., Oswal R.J., Kshirsagar S.S. Computer-aided drug design: An innovative tool for modeling. Open J. Med. Chem. 2012;2:139–148. doi: 10.4236/ojmc.2012.24017

PRESENTATION SLIDES

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Harnessing AI for Breakthroughs in Computational Biology

Presented by:

Aditya Singh (1DS22IS008)
Harsh Joshi (1DS22IS057)

Under the guidance of:

Dr. Annapurna P Patil
Dean Academics, Professor and HOD
Department of Information Science and Engineering.
DSCE, Bengaluru

Key Applications

- **Genomics:**
  AI helps analyze DNA sequences to identify mutations.

Example: AlphaFold predicts protein structures.

- **Drug Discovery:**
  Accelerates drug design and testing.

Example: AI-assisted screening for COVID-19 treatments.

- **Medical Diagnostics:**
  Detects diseases from medical images.

Example: AI for cancer detection.

- **Systems Biology:**
  Models complex biological networks like gene interactions.

Introduction

Computational Biology:

- The use of computers, mathematics, and data analysis to understand biological systems like DNA, proteins, and cells.
- It helps analyze complex genetic data, simulate biological processes, and model systems such as organs.
- This speeds up discoveries and improves the accuracy of research.
- It is essential for advancements in healthcare, drug development, and disease understanding.

Example: Analyzing DNA sequences to find genetic mutations linked to diseases.

Techniques and Tools

- **Machine Learning (ML):**
  Used for predictive modeling and pattern recognition.
- **Deep Learning (DL):**
  A subset of ML for analyzing large datasets (e.g., images, genomics data).
- **Natural Language Processing (NLP):**
  Helps analyze and interpret biological texts (e.g., scientific papers).
- **Tools:**
- TensorFlow
- PyTorch
- Scikit-learn

Cancer Prediction Classification

```
cancer.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   id                      569 non-null    int64
 1   diagnosis               569 non-null    object
 2   radius_mean             569 non-null    float64
 3   texture_mean            569 non-null    float64
 4   perimeter_mean          569 non-null    float64
 5   area_mean               569 non-null    float64
 6   smoothness_mean         569 non-null    float64
 7   compactness_mean        569 non-null    float64
 8   concavity_mean          569 non-null    float64
 9   concave points_mean     569 non-null    float64
 10  symmetry_mean           569 non-null    float64
```

```
cancer.describe()
```

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_me
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.0000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.0887
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.0797
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.0000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.0295
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.0615
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.1307
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.4268

8 rows × 32 columns

Cancer Prediction Classification

Cancer Prediction

Dataset Information:

Target Variable (y):

Diagnosis (M = malignant, B = benign) Ten features (X) are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension (coastline approximation - 1)

For each characteristic three measures are given:

a. Mean b. Standard error c. Largest/ Worst

```
# Step 1 : import library
import pandas as pd

# Step 2 : import data
cancer = pd.read_csv('https://github.com/YBIFoundation/Dataset/raw/main/Cancer.csv')
```

```
cancer.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.

5 rows × 33 columns

Cancer Prediction Classification

```
# Step 3 : define target (y) and features (X)
cancer.columns
```

```
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],
      dtype='object')
```

```
y = cancer['diagnosis']
X = cancer.drop(['id','diagnosis','Unnamed: 32'],axis=1)
```

```
# Step 4 : train test split
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y, train_size=0.7, random_state=2529)
```

```
# check shape of train and test sample
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```


Cancer Prediction Classification

```
((398, 30), (171, 30), (398,), (171,))
```

```
# Step 5 : select model
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(max_iter=5000)
```

```
# Step 6 : train or fit model
model.fit(X_train,y_train)
```

```
LogisticRegression
```

```
model.intercept_
```

```
array([-30.20269391])
```

```
model.coef_
```

```
array([[-0.8644508 , -0.1823121 ,  0.26510852, -0.02688942,  0.13284582,
         0.19445151,  0.40918278,  0.20206338,  0.17199488,  0.03798515,
         0.0192444 , -1.13284188, -0.13597054,  0.11911954,  0.02266663,
        -0.03006638,  0.04691738,  0.02805721,  0.03329433, -0.00980702,
        -0.27148621,  0.44834405,  0.16566196,  0.01286379,  0.2719812 ,
         0.59704539,  1.06177846,  0.40903862,  0.51193487,  0.08436947]])
```

```
# Step 7 : predict model
y_pred = model.predict(X_test)
y_pred
```

```
array(['B', 'M', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'B', 'M', 'B',
       'M', 'B', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'B', 'M',
       'B', 'B', 'M', 'B', 'M', 'B', 'B', 'B', 'B', 'M', 'B', 'B', 'B',
       'M', 'M', 'M', 'M', 'M', 'B', 'B', 'M', 'M', 'M', 'B', 'B', 'B',
       'B', 'B', 'B', 'B', 'B', 'M', 'M', 'M', 'B', 'M', 'B', 'M', 'B',
       'M', 'M', 'B', 'M', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'B',
       'M', 'M', 'M', 'B', 'B', 'M', 'M', 'M', 'B', 'B', 'B', 'B', 'M',
       'B', 'B', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'B', 'M', 'M',
       'B', 'M', 'B', 'B', 'M', 'B', 'B', 'B', 'M', 'B', 'B', 'B', 'B',
       'M', 'B', 'B', 'M', 'B', 'M', 'B', 'M', 'M', 'B', 'B', 'B', 'M',
       'M', 'B', 'M', 'M', 'M', 'B', 'B', 'M', 'B', 'M', 'M', 'B', 'M',
       'M', 'B', 'M', 'B', 'B', 'M', 'B', 'M', 'M', 'B', 'B', 'B', 'B',
       'B', 'M', 'M', 'M', 'M', 'B', 'B', 'B', 'M', 'B', 'M', 'B', 'B',
       'B', 'B'], dtype=object)
```

Cancer Prediction Classification

```
# Step 8 : model accuracy
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report
confusion_matrix(y_test,y_pred)
```

```
array([[97,  5],
       [ 2, 67]])
```

```
accuracy_score(y_test,y_pred)
```

```
0.9590643274853801
```

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

```
              precision    recall  f1-score   support

           B       0.98      0.95      0.97       102
           M       0.93      0.97      0.95        69

    accuracy                           0.96       171
   macro avg       0.96      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171
```


Challenges and Ethics

- **Data Privacy:**

  Protecting genetic and health data using encryption, anonymization, and secure storage methods. Policies like GDPR and HIPAA ensure that data is not accessed or misused without proper consent.

- **Bias in AI Models:**

  Bias arises when training datasets lack diversity. Addressed by using balanced and representative data and techniques like fairness-aware machine learning to reduce inequities in AI predictions.

- **Understanding AI Decisions:**

  AI models, especially deep learning systems, are often complex and lack transparency. Explainable AI (XAI) techniques, such as feature importance mapping, are used to make decisions interpretable.

Conclusion

AI, computational mathematics, and biology are being combined to transform scientific research and healthcare.

- AI processes large biological datasets, helping to uncover patterns and predict outcomes.
- Computational models simulate biological processes like protein folding and gene interactions.
- This collaboration speeds up drug discovery, disease diagnosis, and personalized medicine.
- Ethical concerns, like data privacy and AI bias, need to be addressed to ensure these technologies benefit everyone.

References

- Zou, J., Huss, M., Abid, A., et al. (2019). "A primer on deep learning in genomics."*Nature Genetics*, 51(1), 12–18.
- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018)."Opportunities and obstacles for deep learning in biology and medicine."*Journal of The Royal Society Interface*, 15(141).
- Esteva, A., et al. (2017)."Dermatologist-level classification of skin cancer with deep neural networks."*Nature*, 542, 115–118.
- Senior, A. W., et al. (2020)."Improved protein structure prediction using potentials from deep learning."*Nature*, 577, 706–710.
- Zhao H., Caflisch A. Molecular dynamics in drug design. Eur. J. Med. Chem. 2015;91:4–14. doi: 10.1016/j.ejmech.2014.08.004.
- DiMasi J.A., Grabowski H.G., Hansen R.W. Innovation in the pharmaceutical industry: New estimates of R&D costs. J. Health Econ. 2016;47:20–33. doi: 10.1016/j.jhealeco.2016.01.012.
- Kore P.P., Mutha M.M., Antre R.V., Oswal R.J., Kshirsagar S.S. Computer-aided drug design: An innovative tool for modeling. Open J. Med. Chem. 2012;2:139–148. doi: 10.4236/ojmc.2012.24017