LEISHMANIA DISEASE DETECTION

A PROJECT REPORT

Submitted by

BL.EN.U4AIE21015	B Sai Abhishek
BL.EN.U4AIE21017	B Ruchith balaji
BL.EN.U4AIE21038	Chillakuru Hari

BACHELOR OF TECHNOLOGY

IN

ARTIFICIAL INTELLIGENCE

for the course

21BIO211-Intelligence in Biological Systems-4

Guided and evaluated by

Dr. Siva Kumar

Dept. of Chemisty

AMRITA SCHOOL OF COMPUTING, BANGALORE

AMRITA VISHWA VIDYAPEETHAM

BANGALORE 560 035

June – 2023

TABLE OF CONTENTS

Acknowledgement:

We offer our sincere pranams at the lotus feet of Universal guru, MATA AMRITANANDA MAYI DEVI who showered her blessings throughout this project. We are very thankful to Br. Viswamrita Chaitanya Swamiji, Director, Amrita School of Engineering, Bangalore, for his valuable support. We are indebted to thank Dr. Sriram Devanathan, Principal, Amrita Vishwa Vidyapeetham, Bangalore for engraving a path for us to utilize the available resources to the fullest and there by widen our perspective of education and growth through it.

. It is our privilege to express our sincerest regards to our project guide, Dr. Siva Kumar, Lecturer, Department of Chemistry, for the valuable inputs, guidance, encouragement, wholehearted cooperation and constructive criticism throughout the duration of our project.

Abstract:

Leishmaniasis is a parasitic disease caused by the Leishmania parasite, which is transmitted through the bite of infected sandflies. Early and accurate detection of Leishmaniasis is crucial for effective treatment and control of the disease. In this study, we propose a method for Leishmaniasis detection using DNA sequences.

The approach involves the analysis of DNA sequences to identify the presence of Leishmania parasites. We start by collecting DNA sequences from Leishmania-positive and Leishmania-negative samples. The sequences are processed and pre-processed to extract relevant features. These features may include nucleotide composition, GC content, motifs, or other sequence characteristics.

Next, we employ machine learning techniques such as K-Nearest Neighbor (KNN) , Support Vector Machines (SVM) to train models using the extracted features. The trained models are then used to predict the presence of Leishmaniasis in unseen DNA sequences. To evaluate the performance of the proposed method, we employ metrics such as accuracy, precision, recall, and F1-score. Additionally, we utilize techniques like cross-validation and confusion matrix analysis to assess the robustness and reliability of the models.The results of our study demonstrate the effectiveness of the proposed approach in detecting Leishmaniasis using DNA sequences. By accurately identifying infected samples, this method can aid in early diagnosis, prompt treatment, and effective control of Leishmaniasis.In conclusion, our study presents a novel approach for Leishmaniasis detection through the analysis of DNA sequences. The method combines bioinformatics techniques, machine learning algorithms, to provide accurate and efficient identification of the disease. This approach has the potential to contribute significantly to the field of Leishmaniasis diagnosis and improve public health outcomes.

CHAPTER - 1

INTRODUCTION

Introduction to Leishmanisis:

Leishmaniasis is a vector-borne parasitic disease caused by the Leishmania parasite. It is transmitted to humans through the bites of infected female sandflies, which are small insects found in tropical and subtropical regions. Leishmaniasis is considered a neglected tropical disease and affects millions of people worldwide, particularly in regions with poor socio-economic conditions.

The disease has a wide spectrum of clinical manifestations, ranging from mild skin lesions to severe systemic infections. There are three main forms of leishmaniasis: cutaneous, mucocutaneous, and visceral. Each form is caused by different species of Leishmania parasites and presents with distinct symptoms.

Cutaneous Leishmaniasis: This is the most common form of the disease. It is characterized by skin lesions at the site of the sandfly bite. The lesions may appear as painless, raised bumps or ulcers with a central crater. They can be single or multiple and may heal on their own over several months, leaving scars.

Mucocutaneous Leishmaniasis: This form affects the mucous membranes of the nose, mouth, and throat in addition to the skin. It typically occurs as a complication of untreated cutaneous leishmaniasis. Mucocutaneous leishmaniasis can lead to severe damage and disfigurement of the affected areas, including the destruction of nasal and oral tissues.

Visceral Leishmaniasis (also known as Kala-azar): This is the most severe form of leishmaniasis. It affects multiple organs, primarily the spleen, liver, and bone marrow. Visceral leishmaniasis can cause prolonged fever, weight loss, fatigue, anemia, and enlargement of the spleen and liver. If left untreated, it can be fatal.

Diagnosis of leishmaniasis involves clinical evaluation, microscopic examination of tissue samples or fluid aspirates, and laboratory tests such as serological assays or molecular methods. Treatment options include antiparasitic drugs, with the choice of medication depending on the form and severity of the disease.Prevention and control strategies for leishmaniasis focus on vector control measures, such as insecticide-treated nets, environmental management, and personal protective measures. Additionally, research efforts are underway to develop vaccines against the disease.

Motivation:

The motivation for Leishmaniasis disease detection stems from the urgent need to diagnose and treat the infection promptly. The Motivations are

Early Diagnosis: Early detection of Leishmaniasis is crucial for initiating timely treatment. Detecting the infection at an early stage can help prevent the progression of the disease, reduce the severity of symptoms, and minimize complications. Timely detection also enables the implementation of appropriate control measures to prevent the spread of the disease.

Improved Patient Outcomes: Accurate and timely detection of Leishmaniasis can significantly improve patient outcomes. It allows healthcare providers to prescribe the appropriate treatment regimen tailored to the specific form and severity of the disease. Early intervention can help prevent long-term complications and reduce morbidity and mortality associated with Leishmaniasis.

Public Health Impact: Effective detection of Leishmaniasis plays a crucial role in public health interventions. It allows for the identification of high-risk areas and populations, facilitating targeted control measures. Disease surveillance and monitoring through detection methods help in tracking the prevalence and incidence of Leishmaniasis, which is essential for planning and implementing effective prevention and control strategies.

CHAPTER – 2
LITERATURE SURVEY

Paper - [1]: A two-step approach-machine learning, variational autoencoder, and weighted gene co-expression network analysis identify key signature genes and pathways implicated in active visceral leishmaniasis.

In this study, they aimed to detect key signature genes and pathways implicated in active visceral leishmaniasis. We have fivegenes IFNG, SC5D, LSM1, CMC2, and SAR1B, with higher interamodular connectivity, as the key signature genes. These genes were validated using a machine learning algorithm by creating new gene features by applying a variational autoencoder. These genes are involved in processes like cytokine-cytokine receptor interaction, IL-17 signaling pathway, T-cell receptor signaling pathway, and Th1 and Th2 cell differentiation. Besides, hsa-miR-340-5p, hsa-miR-325-3p, hsa-miR-182-5p, hsa-miR-1271-5p/hsamiR-96-5p miRNAs were found to target key signature genes.

Now we are going to Detect the Leishmania species in DNA through a different approach. We are going to use Bio-informatics analysis and Machine learning techniques to detect the Leishmaniasis in the given DNA sequence accurately and Correctly.

Paper - [2]: Molecular identification of Leishmania RNA virus in cutaneous leishmaniasis patients.

This research is the first molecular and phylogenetic study based on the LRV2 strains from rodent reservoirs and CL patients in Isfahan province. This study is the first comprehensive epidemiological study in the Isfahan province region regarding LRV2 in Leishmania species. The findings of this study were consistent with the results of previous studies and Leishmania isolates in the present study were from CL patients and rodent reservoirs in Isfahan province so could not be generalized to other geographical zones of Iran and the around countries.

Now we are going to Detect the Leishmania species in DNA through a different approach. We are going to use Bio-informatics analysis and Machine learning techniques to detect the Leishmaniasis in the given DNA sequence accurately and Correctly.

CHAPTER – 3
SYSTEM DESIGN & IMPLEMENTATION

Brief:

Leishmaniasis:

Leishmaniasis disease detection using DNA/RNA analysis involves the identification and characterization of the Leishmania parasite's genetic material in patient samples. DNA/RNA-based techniques provide a sensitive and specific approach to diagnose Leishmaniasis, particularly in cases where other diagnostic methods may yield inconclusive results. Here is an overview of some commonly used DNA/RNA-based methods for Leishmaniasis detection:

1) **Polymerase Chain Reaction (PCR):** PCR is a widely used technique that amplifies specific DNA/RNA sequences of the Leishmania parasite in patient samples. It involves the use of specific primers that target conserved regions of the Leishmania genome. By amplifying the target DNA/RNA, PCR enables the detection and quantification of the parasite. Different PCR variants, such as conventional PCR, real-time PCR, and nested PCR, can be employed depending on the specific requirements of the diagnostic assay.

2) **-Mediated Isothermal Amplification (LAMP):** LAMP is an isothermal amplification method that can rapidly and specifically amplify DNA/RNA sequences. It offers advantages over PCR, such as simplicity, rapidity, and robustness, as it does not require sophisticated thermal cycling equipment. LAMP has been successfully utilized for the detection of Leishmania DNA/RNA in patient samples.

3) **Bio-Informatics Analysis**

4) **K – Nearest Neighbour (Machine Learning)**

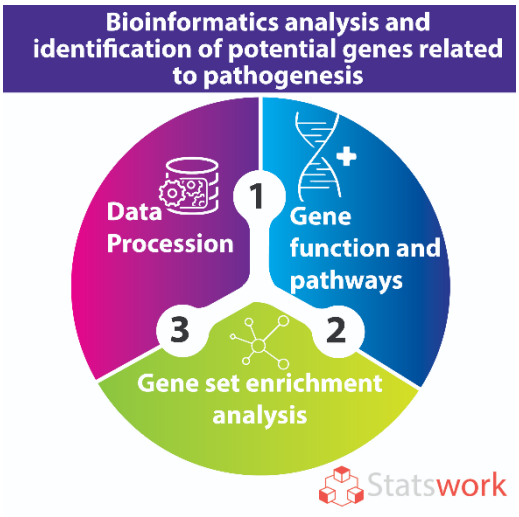
5) **Support Vector Machines(Machine Learning)**

In the above 5 techniques we are going to use Bio-Informatics Analysis , K – Nearest Neighbour (Machine Learning) , Support Vector Machines(Machine Learning).

1) Bio-Informatics Analysis:

Detecting Leishmania using bioinformatics analysis involves several steps, including data retrieval, sequence alignment, and identification. Here's an outline of the process:

1) **Data Retrieval:** Obtain the genetic sequence data for Leishmania from public databases like GenBank or other relevant sources. You can search for specific Leishmania species or use broader queries to retrieve a diverse set of sequences.
2) **Sequence Alignment:** Align the obtained Leishmania sequences with reference sequences to identify similarities and variations. Multiple sequence alignment tools like ClustalW, MAFFT, or Muscle can be used for this step. The alignment helps identify conserved regions across different Leishmania strains.
3) **Feature Extraction**: Extract relevant features from the aligned sequences, such as conserved motifs, protein domains, or regions of interest. This information can be useful for designing specific primers or probes for targeted Leishmania detection.
4) **Data Analysis and Interpretation**: Analyze the experimental results and interpret them in the context of Leishmania detection. Compare the obtained sequences with reference sequences, evaluate the amplification products, and assess the sensitivity and specificity of the assay.



Leishmania DNA taken at the start of the Analysis it is of the length 2400 Nucleotides and Human sequence of 2500 length is taken and both are compared with Bio informatics analysis. If the Leishmania matches with Human DNA it return True that person have Leishmania.

2) K-Nearest Neighbor:

- From the given Data set find dependent and independent Variables.
- Divide the variables to training and testing data, it means Data pre-processing.
- Fit the train and test data.
- Using the KNN Classifier taking n neighbors find the predicted for test data
- From predicted data find Accuracy, Precision, Recall, F1-Score etc.
- Specifically, four different distance functions, which are Euclidean distance, cosine similarity measure, correlation, and Chi square, are used in the k-NN classifier respectively

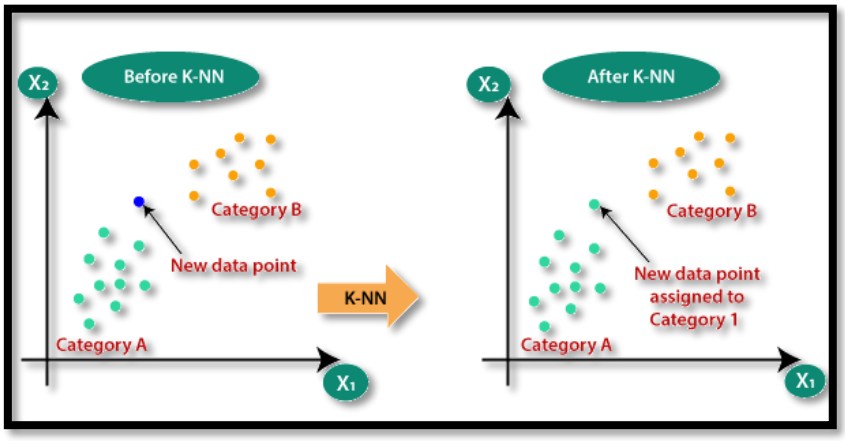
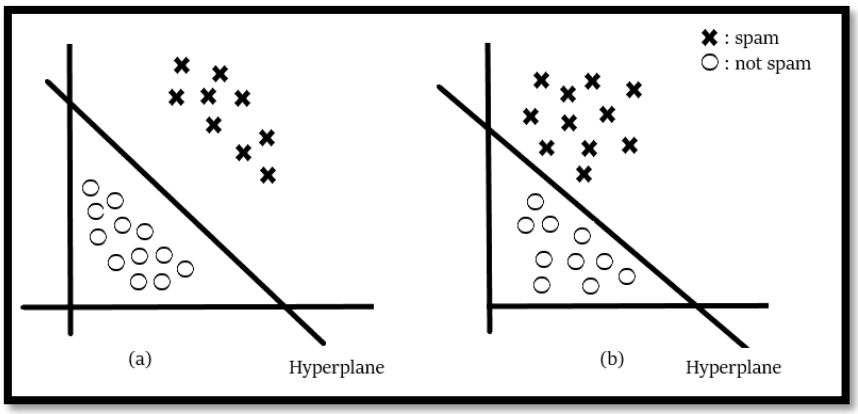


Fig-1

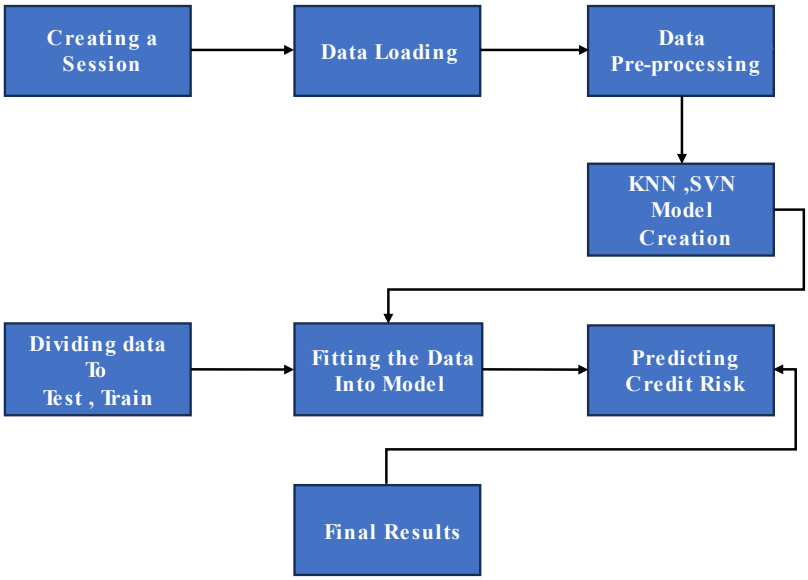
3) Support Vector Machines:

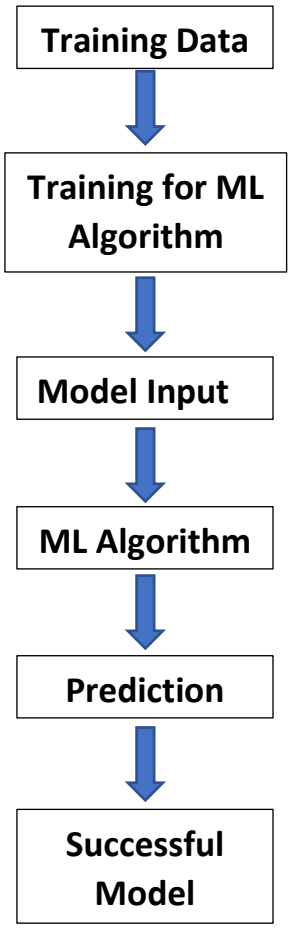
- From the given Data set find dependent and independent Variables.
- Divide the variables to training and testing data, it means Data pre-processing.
- Fit the train and test data.
- Using the SVC Classifier taking path as linear, curve find the predicted for test data
- From predicted data find Accuracy, Precision, Recall, F1-Score etc.

Three different types of SVM-Kernels are displayed below. The polynomial and RBF are especially useful when the data-points are not linearly separable



Design:


Creating a Session
Data Loading
Data Pre-processing
KNN ,SVN Model Creation
Dividing data To Test , Train
Fitting the Data Into Model
Predicting Credit Risk
Final Results


Training Data
Training for ML Algorithm
Model Input
ML Algorithm
Prediction
Successful Model

Implementation

1) Data Set Loading:

a) For KNN & SVM:

```
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   age                       299 non-null    float64
 1   fever                     299 non-null    int64
 2   creatinine_phosphokinase  299 non-null    int64
 3   weight_loss               299 non-null    int64
 4   ejection_fraction         299 non-null    int64
 5   Liver_swelling            299 non-null    int64
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64
 9   Fatigue                   299 non-null    int64
 10  Abnormal_blood_tests      299 non-null    int64
 11  time                      299 non-null    int64
 12  Leishmania_predict        299 non-null    int64
```

b) For Bio-Informatics Analysis:

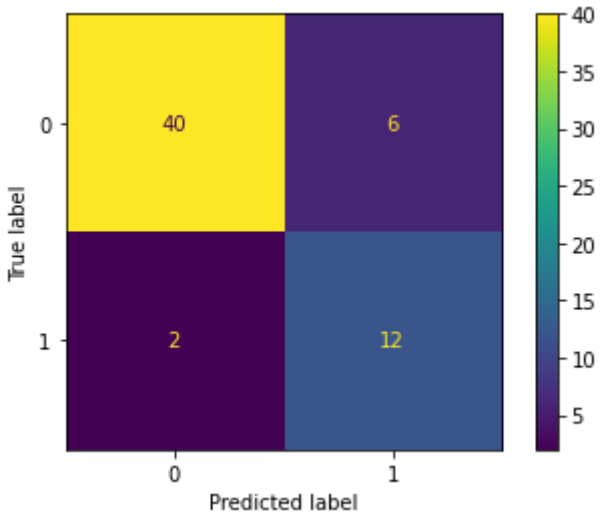
Lesishmania.fasta file:

GATTTAAGTGAATAGCTTGGCTATCTCACTTCCCCTCGTTCTCTTGCAGAACTTTGATTTTAACGAACTT
AAATAAAAGCCCTGTTGTTTAGCGTATCGTTGCACTTGTCTGGTGGGATTGTGGCATTAATTTGCCTGCT
CATCTAGGCAGTGGACATATGCTCAACACTGGGTATAATTCTAATTGAATACTATTTTTCAGTTAGAGCG
TCGTGTCTCTTGTACGTCTCGGTCACAATACACGGTTTCGTCCGGTGCGTGGCAATTCGGGGCACATCAT
GTCTTTCGTGGCTGGTGTGACCGCGCAAGGTGCGCGCGGTACGTATCGAGCAGCGCTCAACTCTGAAAAA
CATCAAGACCATGTGTCTCTAACTGTGCCACTCTGTGGTTCAGGAAACCTGGTTGAAAAACTTTCACCAT
GGTTCATGGATGGCGAAAATGCCTATGAAGTGGTGAAGGCCATGTTACTTAAAAAGGAGCCACTTCTCTA
TGTGCCCATCCGGCTGGCTGGACACACTAGACACCTCCCAGGTCCTCGTGTGTACCTGGTTGAGAGGCTC
ATTGCTTGTGAAAATCCATTCATGGTTAACCAATTGGCTTATAGCTCTAGTGCAAATGGCAGCCTGGTTG
GCACAACTTTGCAGGGCAAGCCTATTGGTATGTTCTTCCCTTATGACATCGAACTTGTCACAGGAAAGCA
AAATATTCTCCTGCGCAAGTATGGCCGTGGTGGTTATCACTACACCCCATTCCACTATGAGCGAGACAAC
ACCTCTTGCCCTGAGTGGATGGACGATTTTGAGGCGGATCCTAAAGGCAAATATGCCCAGAATCTGCTTA
AGAAGTTGATTGGCGGTGATGTCACTCCAGTTGACCAATACATGTGTGGCGTTGATGGAAAACCCATTAG
TGCCTACGCATTTTTAATGGCCAAGGATGGAATAACCAAACTGGCTGATGTTGAAGCGGACGTCGCAGCA
CGTGCTGATGACGAAGGCTTCATCACATTAAAGAACAATCTATATAGATTGGTTTGGCATGTTGAGCGTA
AAGACGTTCCATATCCTAAGCAATCTATTTTTACTATTAATAGTGTGGTCCAAAAGGATGGTGTTGAAAA
CACTCCTCCTCACTATTTTACTCTTGGATGCAAAATTTTAACGCTCACCCCACGCAACAAGTGGAGTGGC
GTTTCTGACTTGTCCCTCAAACAAAAACTCCTTTACACCTTCTATGGTAAGGAGTCACTTGAGAACCCAA
CCTACATTTACCACTCCGCATTCATTGAGTGTGGAAGTTGTGGTAATGATTCCTGGCTTACAGGGAATGC
TATCCAAGGGTTTGCCTGTGGATGTGGGGCATCATATACAGCTAATGATGTCGAAGTCCAATCATCTGGC
ATGATTAAGCCAAATGCTCTTCTTTGTGCTACTTGCCCCTTTGCTAAGGGTGATAGCTGTTCTTCTAATT
GCAAACATTCAGTTGCTCAGTTGGTTAGTTACCTTTCTGAACGCTGTAATGTTATTGCTGATTCTAAGTC
CTTCACACTTATCTTTGGTGGCGTAGCTTACGCCTACTTTGGATGTGAGGAAGGTACTATGTACTTTGTG

2) Preprocessing:

age	fever	creatinine_phosphokinase	weight_loss	ejection_fraction	Liver_swelling	platelets	serum_creatinine	serum_sodium	Fatigue	Abnormal_bloo
000000	0.088006	-0.081584	-0.101012	0.060098	0.093289	-0.052354	0.159187	-0.045966	0.065430	0.
088006	1.000000	-0.190741	-0.012729	0.031557	0.038182	-0.043786	0.052174	0.041882	-0.094769	-0.
081584	-0.190741	1.000000	-0.009639	-0.044080	-0.070590	0.024463	-0.016408	0.059550	0.079791	0.
101012	-0.012729	-0.009639	1.000000	-0.004850	-0.012732	0.092193	-0.046975	-0.089551	-0.157730	-0.
060098	0.031557	-0.044080	-0.004850	1.000000	0.024445	0.072177	-0.011302	0.175902	-0.148386	-0.
093289	0.038182	-0.070590	-0.012732	0.024445	1.000000	0.049963	-0.004935	0.037109	-0.104615	-0.
052354	-0.043786	0.024463	0.092193	0.072177	0.049963	1.000000	-0.041198	0.062125	-0.125120	0.
159187	0.052174	-0.016408	-0.046975	-0.011302	-0.004935	-0.041198	1.000000	-0.189095	0.006970	-0.
045966	0.041882	0.059550	-0.089551	0.175902	0.037109	0.062125	-0.189095	1.000000	-0.027566	0.
065430	-0.094769	0.079791	-0.157730	-0.148386	-0.104615	-0.125120	0.006970	-0.027566	1.000000	0.
018668	-0.107290	0.002421	-0.147173	-0.067315	-0.055711	0.028234	-0.027414	0.004813	0.445892	1.
224068	-0.141414	-0.009346	0.033726	0.041729	-0.196439	0.010514	-0.149315	0.087640	-0.015608	-0.
253729	0.066270	0.062728	-0.001943	-0.268603	0.079351	-0.049139	0.294278	-0.195204	-0.004316	-0.

3) Correlation Matrix:



CHAPTER – 4

RESULTS AND ANALYSIS

Result:

a) Bioinformatics:

```
No evidence of Leishmania disease.
No evidence of Leishmania disease.
No evidence of Leishmania disease.
No evidence of Leishmania disease.
No evidence of Leishmania disease.
No evidence of Leishmania disease.
Leishmania disease detected!
No evidence of Leishmania disease.
No evidence of Leishmania disease.
No evidence of Leishmania disease.
```

b) KNN & SVM:

```
Accuracy : 85.0
Recall: 72.82608695652173
Precision_Score: 82.02614379084967
F1_Score: 75.79560735096369
```

```
Accuracy : 86.66666666666667
Recall: 86.33540372670807
Precision_Score: 80.95238095238095
F1_Score: 82.95454545454545
```

Analysis:

In the abpve 2 techniques SVM is showing accuracy more and it has the more predicting power.

CHAPTER – 5

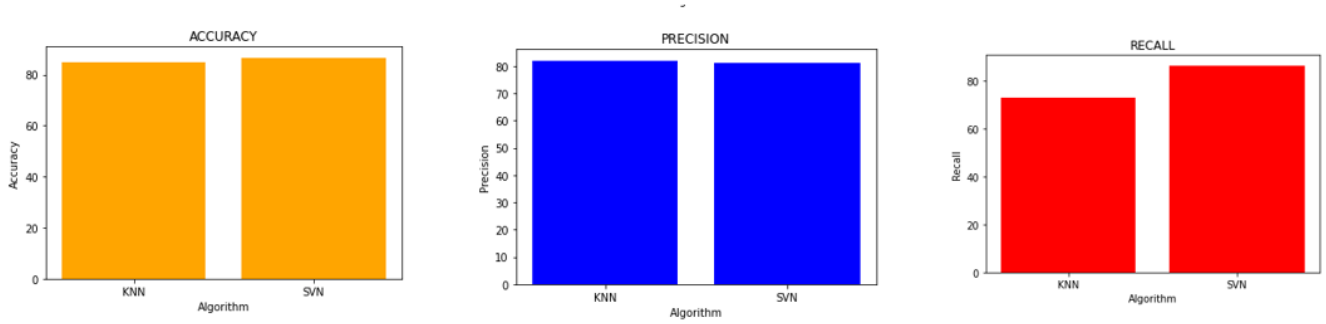
CONCLUSION

Conclusion:

In core research paper they have used genes for predicting the Leishmania using Auto Encoders and weighted gene co-expression network analysis. But we have used only Symptoms of Leishmaniasis and with those symptoms we Combined with K-Nearest Neighbour (KNN) & Support vector Machines (SVM) ML Algorithms and we got an efficient Output.

We got an accuracy, precision & Recall as follows:

```
Accuracy : 86.66666666666667
Recall: 86.33540372670807
Precision_Score: 80.95238095238095
F1_Score: 82.95454545454545
```



And also we have taken 2 sequences (one with Leishmaniasis sequence & other is Human DNA sequence) Both sequences are analysed with the Bio Informatics Analysis, we have taken a sample of 2400 length fasta file of leishmaniasis and 2556 length of human DNA sequence and both are compared. The result is the algorithm calculates the GC content in the to be tested Sequence and also calculates the score to compare with Leishmania sequence.

We have taken 2 human sequences and one got having leishmania and other don’t have leishmaniasis.

CHAPTER – 5

REFERENCES

1.WHO: Leishmaniasis Fact Sheet - Google Scholar https://www.who.int/news-room/fact-sheets/detail/leishmaniasis

2. Croft, S. L., Sundar, S., & Fairlamb, A. H. (2006). Drug Resistance in Leishmaniasis. Clinical Microbiology Reviews, 19(1), 111–

126. https://doi.org/10.1128/CMR.19.1.111-126.2006

3. Ma, H., & Zhao, H. (2013). Drug target inference through pathway analysis of genomics data. Advanced Drug Delivery

Reviews, 65(7), 966–972. https://doi.org/10.1016/j.addr.2012.12.004

4. Rabinowitz, J. D., Purdy, J. G., Vastag, L., Shenk, T., & Koyuncu, E. (2011). Metabolomics in Drug Target Discovery. Cold Spring

Harbor Symposia on Quantitative Biology, 76, 235–246. https://doi.org/10.1101/sqb.2011.76.010694

5. Zinzalla, G., & Thurston, D. E. (2009). Targeting protein–protein interactions for therapeutic intervention: A challenge for the

future. Future Medicinal Chemistry, 1(1), 65–93. https://doi.org/10.4155/fmc.09.12