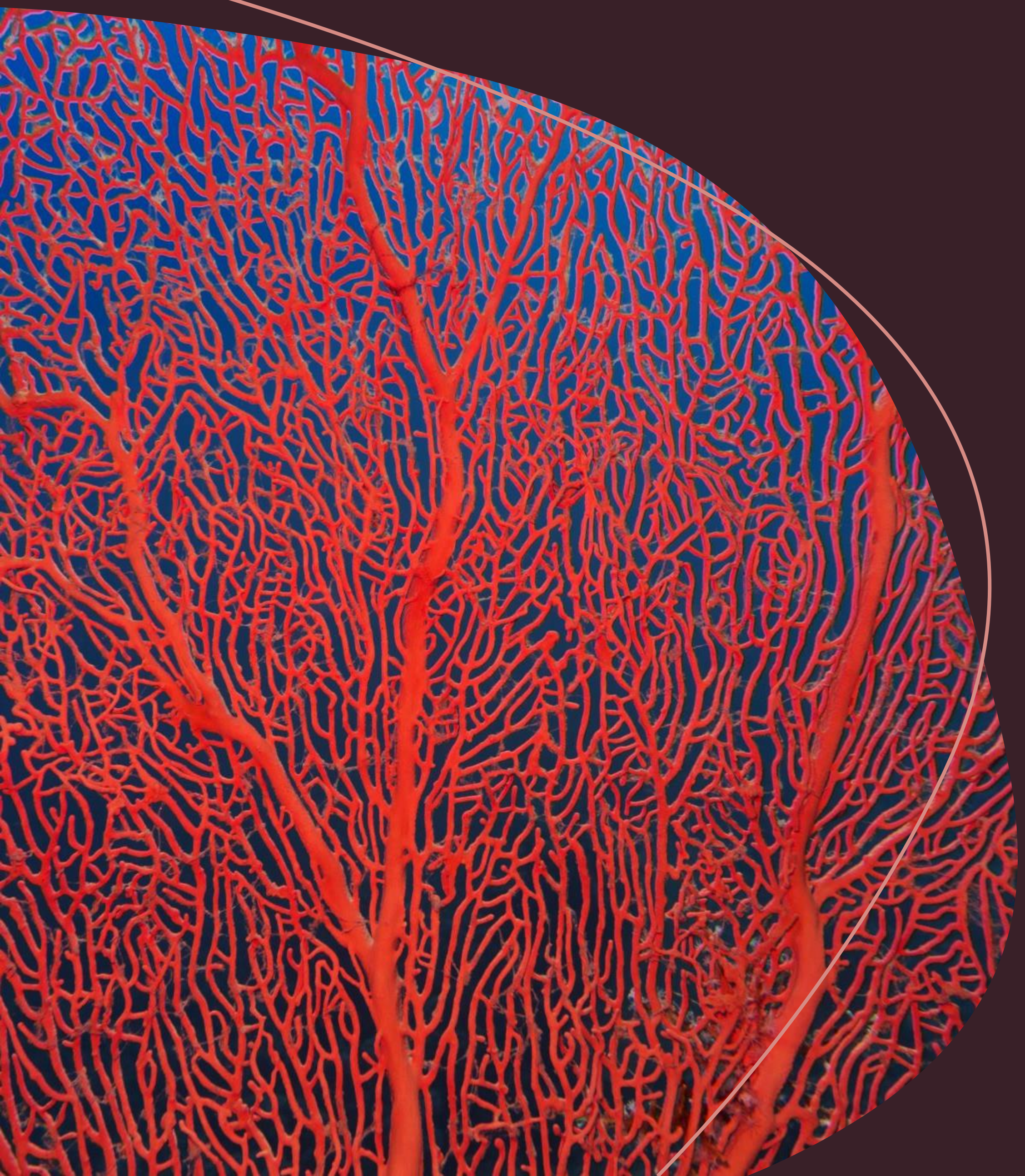


Data Mining in Bioinformatics

Fundamentals of Data Science

Project Presentation

(order of presenting)

Hosna Eltarras, Menna Mohammed,

Aisha Hagar, Yomna Jehad

What is Bioinformatics ?

- Understand Biological Data
- Biological data: Genes, RNA, Protein, EHRs and DNA Sequences
- DNA Language

ACGT

DNA Alphabet

AGGGAGAATGTTGAAACACAAGC

DNA Sentence

- English Human Language

A B C D E F ..

English Alphabet

I Love Bioinformatics

English Sentence

Reference: [1]

Explosion of Biological Data

Reference: [https://en.wikipedia.org/wiki/Genome]

Main WorkFlow

- Data Preprocessing
- Clustering
- Classification
- Deep learning

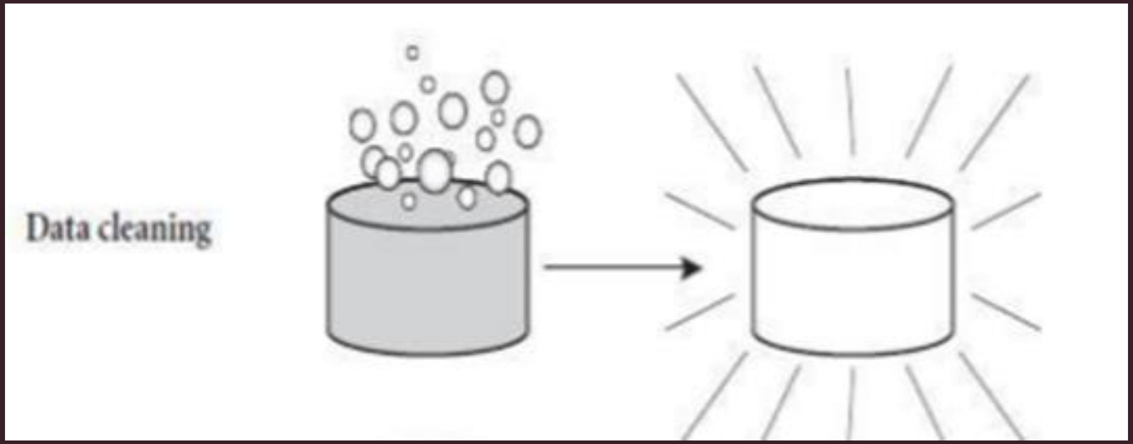
Pre-Processing

- Data cleaning
- Data transformation
- Data reduction

References: [2,3]

Data cleaning

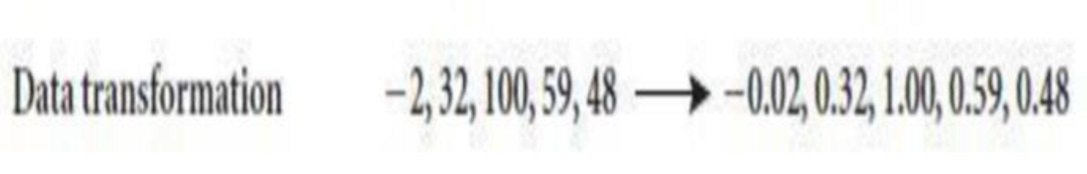
- Missing data
  - Ignore the tuples
  - Fill the missing values
- Noisy data
  - Binning
  - Regression
  - clustering



References: [2,3]

Data transformation

- Normalization
- Attribute selection
- Generalization

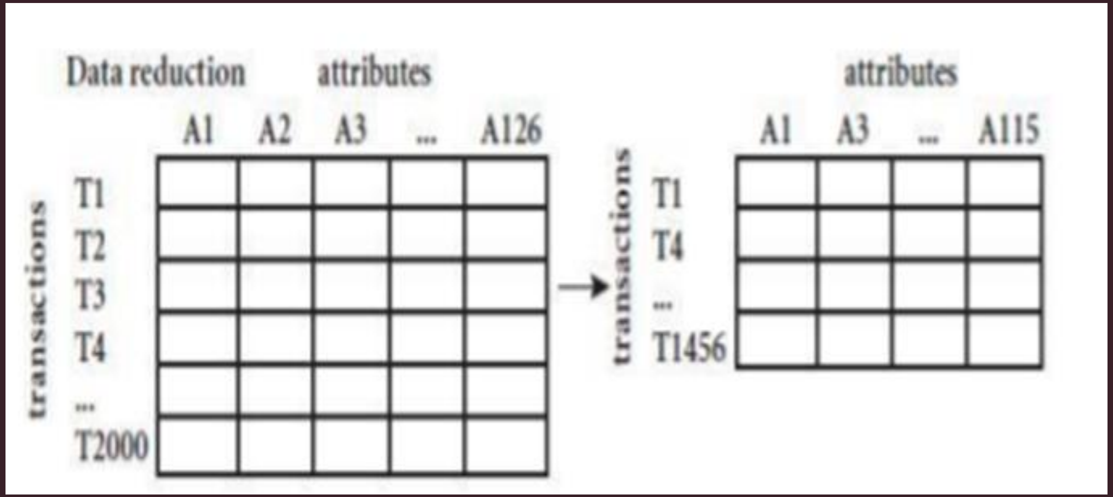


Data transformation −2, 32, 100, 59, 48 ⟶ −0.02, 0.32, 1.00, 0.59, 0.48

References: [2,3]

Data Reduction

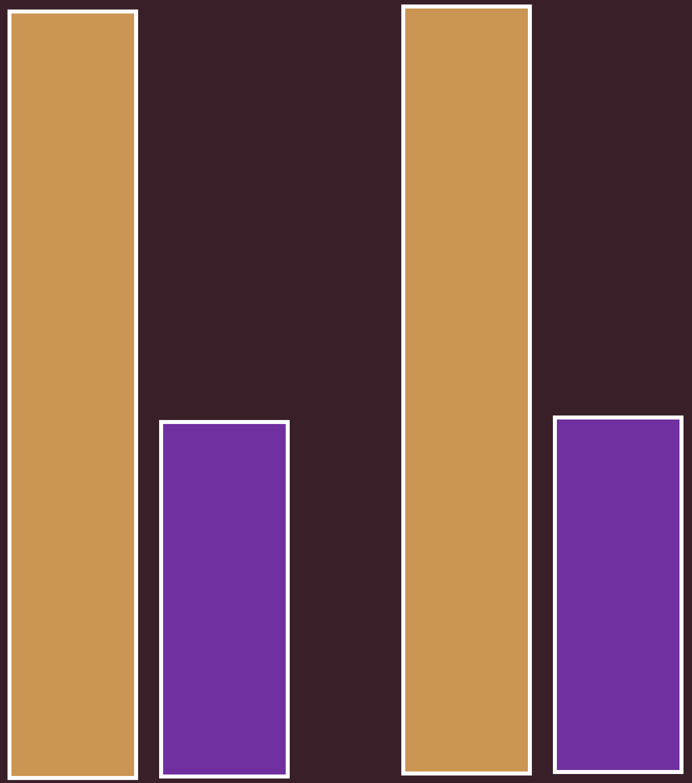
- Data cube aggregation
- Dimensionality reduction
- Numerosity reduction



References: [2,3]

Imbalanced Data

- Asymmetric Distribution
  - Hospitals
  - General Population
  - Rare Diseases

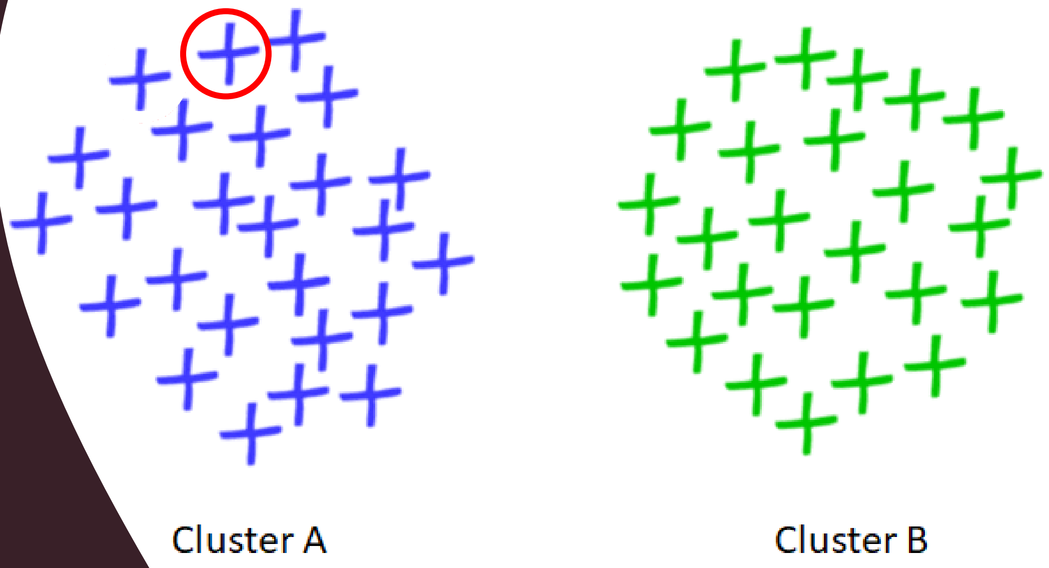


References: [4]

Clustering

Exploring hidden structure

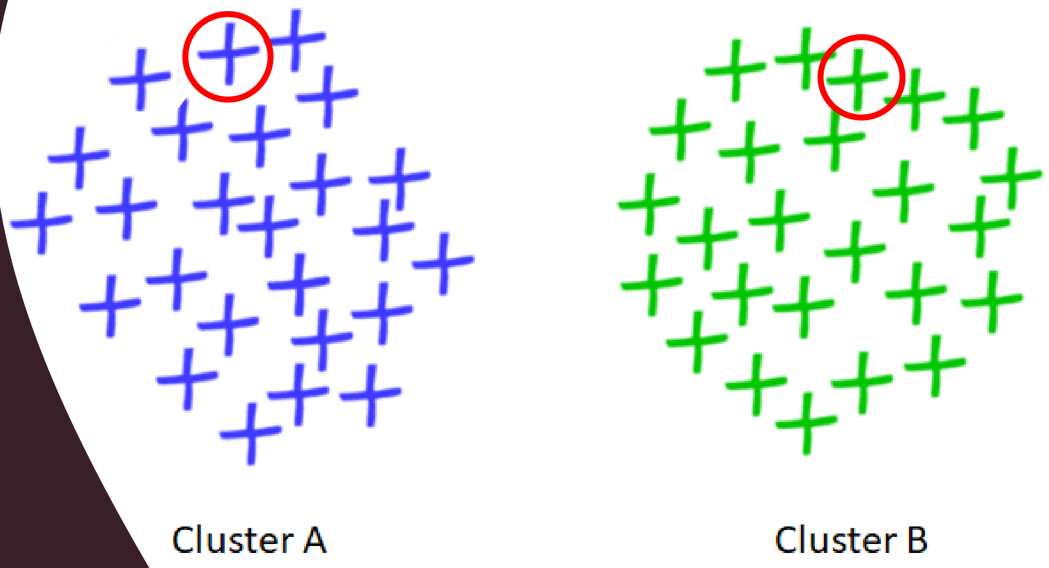
- What is the function of the Red Gene ?



References: [4,5]

Fuzzy Clustering

One gene in **Two** clusters !

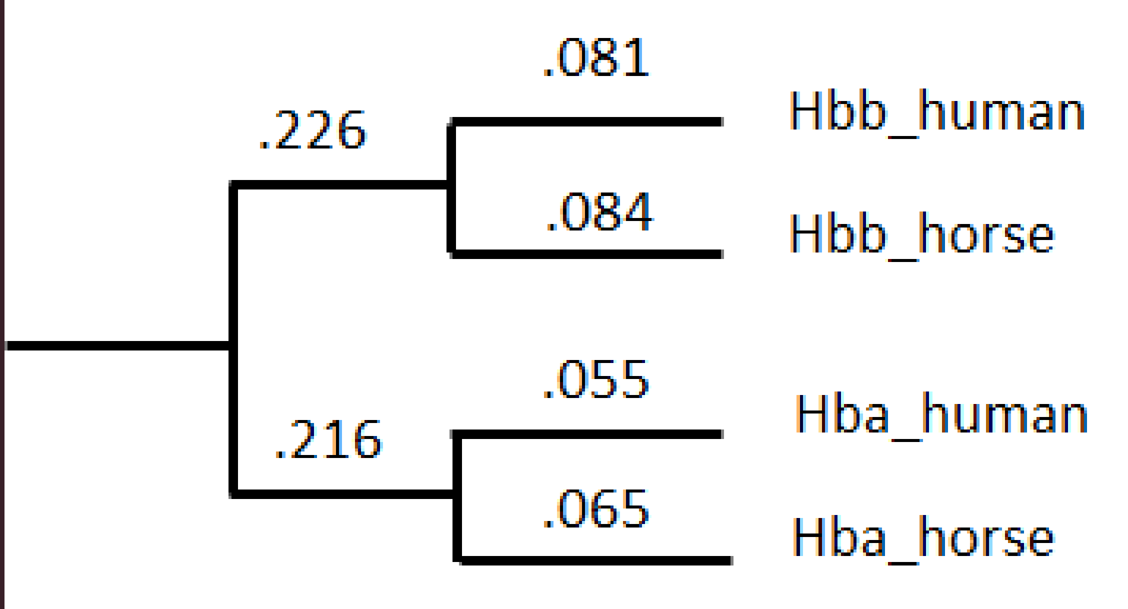


References: [4,5]

Hierarchical clustering:
phylogenetic trees

- Cure
- Clinical Trials

- Covid 19 and Bats,
  How did we know?



References: [4,5,6]

Classification in Bioinformatics

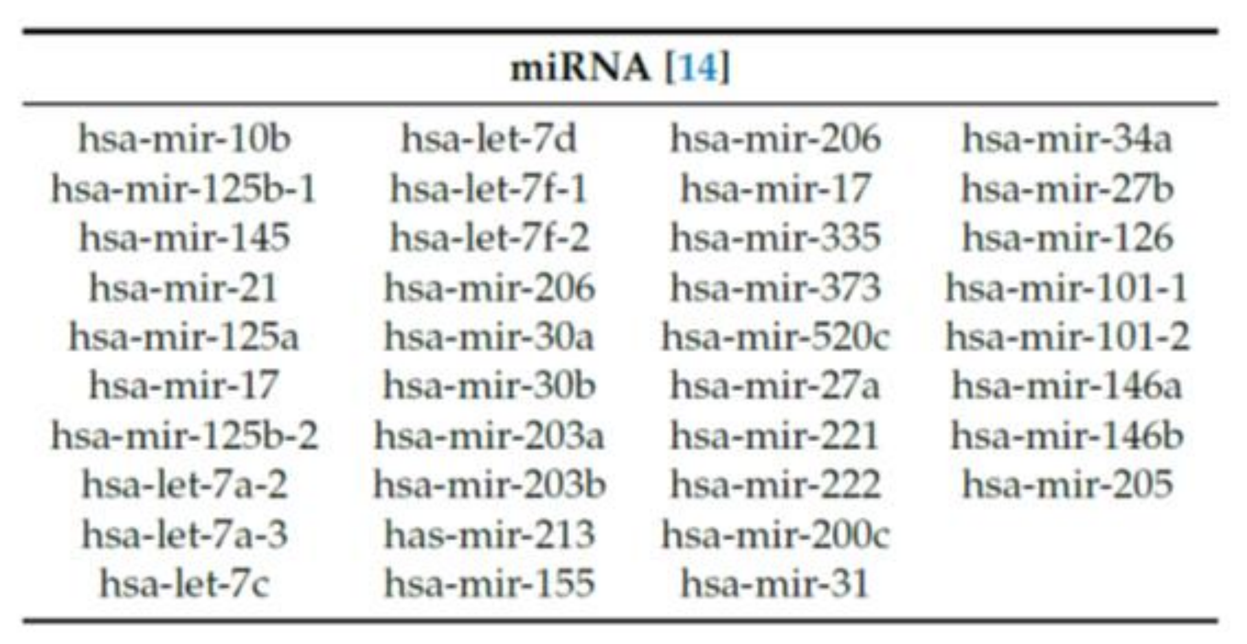
- Applications

Ex: Identify the gene signature of a disease, Classify patients' data for medical diagnosis, Evaluation of disease severity,… etc.

References: [7]

Breast Cancer Detection

- Dataset Features: MicroRNAs as biomarkers.



miRNA [14]			
hsa-mir-10b	hsa-let-7d	hsa-mir-206	hsa-mir-34a
hsa-mir-125b-1	hsa-let-7f-1	hsa-mir-17	hsa-mir-27b
hsa-mir-145	hsa-let-7f-2	hsa-mir-335	hsa-mir-126
hsa-mir-21	hsa-mir-206	hsa-mir-373	hsa-mir-101-1
hsa-mir-125a	hsa-mir-30a	hsa-mir-520c	hsa-mir-101-2
hsa-mir-17	hsa-mir-30b	hsa-mir-27a	hsa-mir-146a
hsa-mir-125b-2	hsa-mir-203a	hsa-mir-221	hsa-mir-146b
hsa-let-7a-2	hsa-mir-203b	hsa-mir-222	hsa-mir-205
hsa-let-7a-3	has-mir-213	hsa-mir-200c	
hsa-let-7c	hsa-mir-155	hsa-mir-31	

Fig: Clinically Verified MiRNAs

References: [7]

Flow of the Experiment

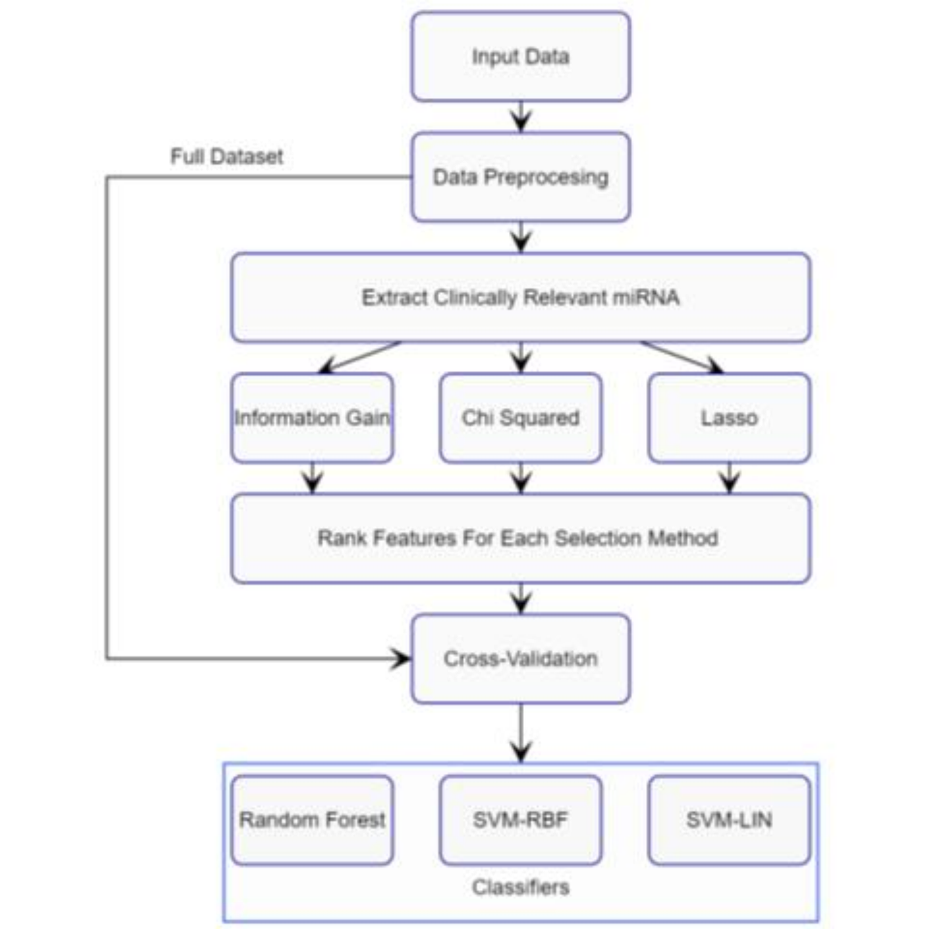
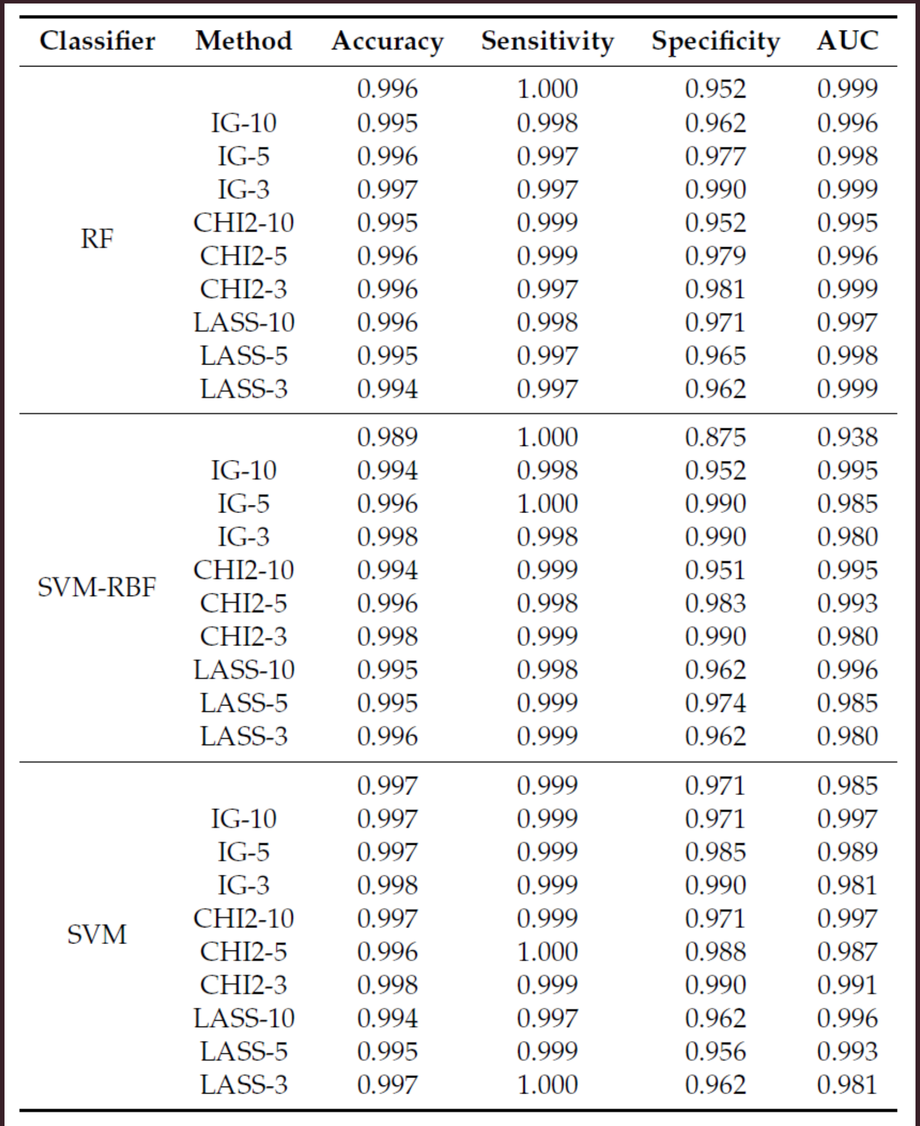


Fig: Schematics for Cancer Detection with Machine Learning

References: [7]

Evaluation

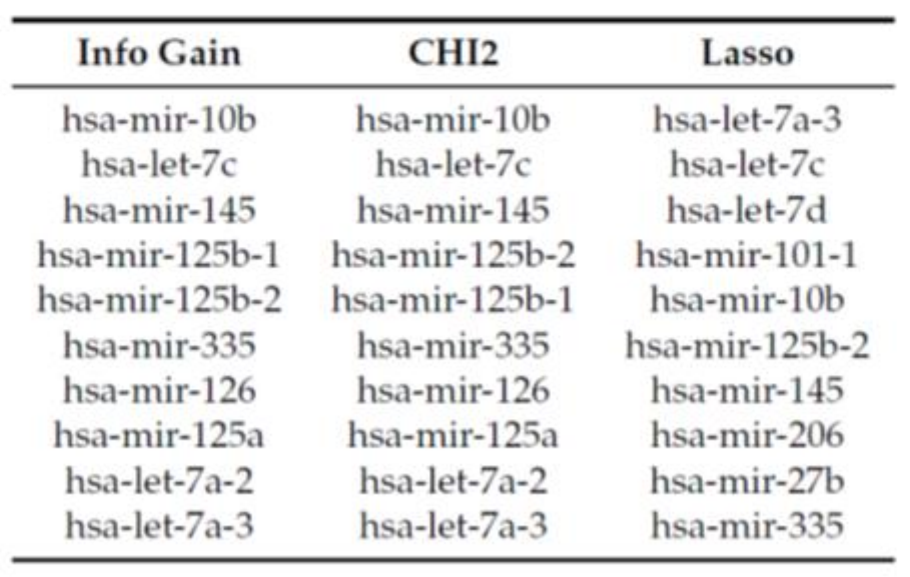


Classifier	Method	Accuracy	Sensitivity	Specificity	AUC
RF		0.996	1.000	0.952	0.999
	IG-10	0.995	0.998	0.962	0.996
	IG-5	0.996	0.997	0.977	0.998
	IG-3	0.997	0.997	0.990	0.999
	CHI2-10	0.995	0.999	0.952	0.995
	CHI2-5	0.996	0.999	0.979	0.996
	CHI2-3	0.996	0.997	0.981	0.999
	LASS-10	0.996	0.998	0.971	0.997
	LASS-5	0.995	0.997	0.965	0.998
	LASS-3	0.994	0.997	0.962	0.999
SVM-RBF		0.989	1.000	0.875	0.938
	IG-10	0.994	0.998	0.952	0.995
	IG-5	0.996	1.000	0.990	0.985
	IG-3	0.998	0.998	0.990	0.980
	CHI2-10	0.994	0.999	0.951	0.995
	CHI2-5	0.996	0.998	0.983	0.993
	CHI2-3	0.998	0.999	0.990	0.980
	LASS-10	0.995	0.998	0.962	0.996
	LASS-5	0.995	0.999	0.974	0.985
	LASS-3	0.996	0.999	0.962	0.980
SVM		0.997	0.999	0.971	0.985
	IG-10	0.997	0.999	0.971	0.997
	IG-5	0.997	0.999	0.985	0.989
	IG-3	0.998	0.999	0.990	0.981
	CHI2-10	0.997	0.999	0.971	0.997
	CHI2-5	0.996	1.000	0.988	0.987
	CHI2-3	0.998	0.999	0.990	0.991
	LASS-10	0.994	0.997	0.962	0.996
	LASS-5	0.995	0.999	0.956	0.993
	LASS-3	0.997	1.000	0.962	0.981

Fig: Performance Metrics of Classifiers with Different Feature Selection Methods Over MiRNAs Subsets(3, 5, 10)

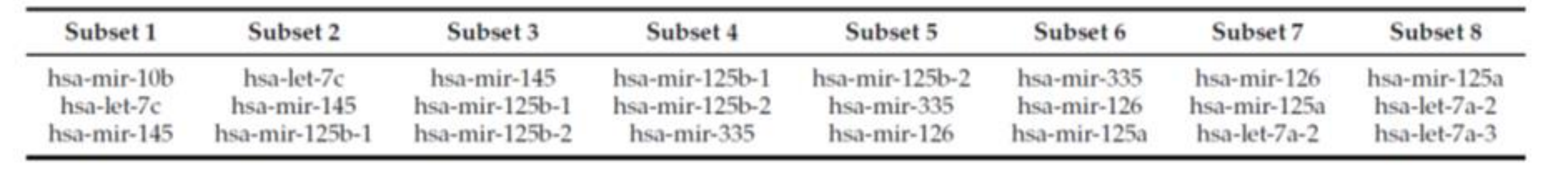
References: [7]

Selecting Fewer Features for Classification



Info Gain	CHI2	Lasso
hsa-mir-10b	hsa-mir-10b	hsa-let-7a-3
hsa-let-7c	hsa-let-7c	hsa-let-7c
hsa-mir-145	hsa-mir-145	hsa-let-7d
hsa-mir-125b-1	hsa-mir-125b-2	hsa-mir-101-1
hsa-mir-125b-2	hsa-mir-125b-1	hsa-mir-10b
hsa-mir-335	hsa-mir-335	hsa-mir-125b-2
hsa-mir-126	hsa-mir-126	hsa-mir-145
hsa-mir-125a	hsa-mir-125a	hsa-mir-206
hsa-let-7a-2	hsa-let-7a-2	hsa-mir-27b
hsa-let-7a-3	hsa-let-7a-3	hsa-mir-335

Fig: Top Ranked Features Under Different Feature Selection Techniques



Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8
hsa-mir-10b	hsa-let-7c	hsa-mir-145	hsa-mir-125b-1	hsa-mir-125b-2	hsa-mir-335	hsa-mir-126	hsa-mir-125a
hsa-let-7c	hsa-mir-145	hsa-mir-125b-1	hsa-mir-125b-2	hsa-mir-335	hsa-mir-126	hsa-mir-125a	hsa-let-7a-2
hsa-mir-145	hsa-mir-125b-1	hsa-mir-125b-2	hsa-mir-335	hsa-mir-126	hsa-mir-125a	hsa-let-7a-2	hsa-let-7a-3

Fig: Subsets of Ranked miRNAs

Performance Evaluation Over Different Subsets

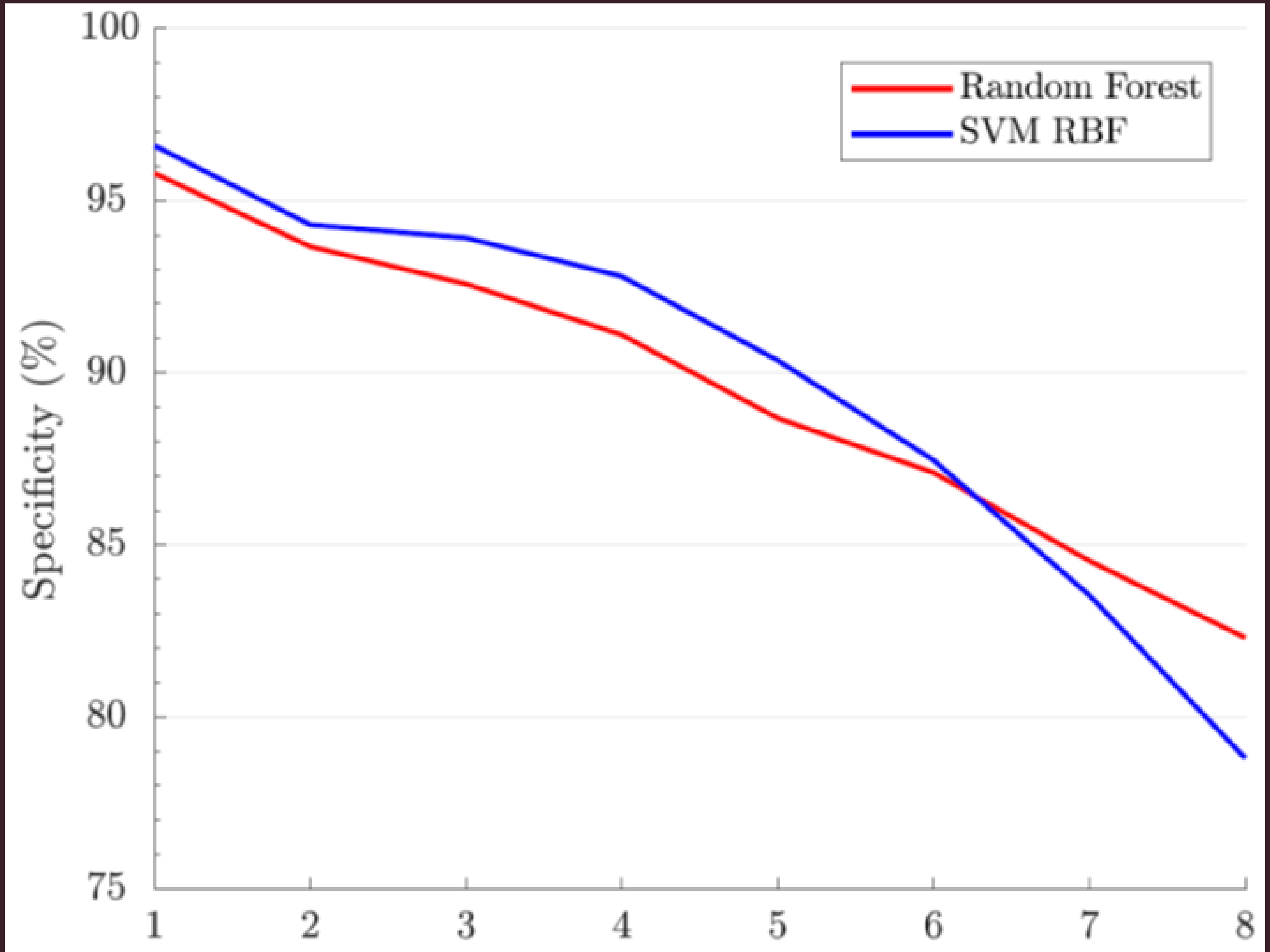
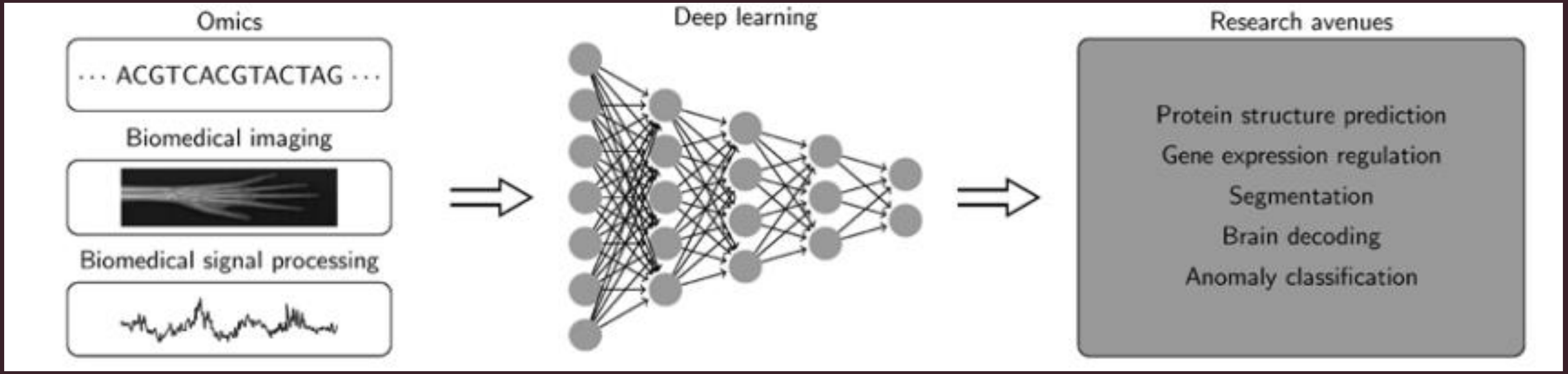


Fig: Specificity Across Different Clinical miRNA Subsets

References: [7]

Deep learning in bioinformatics

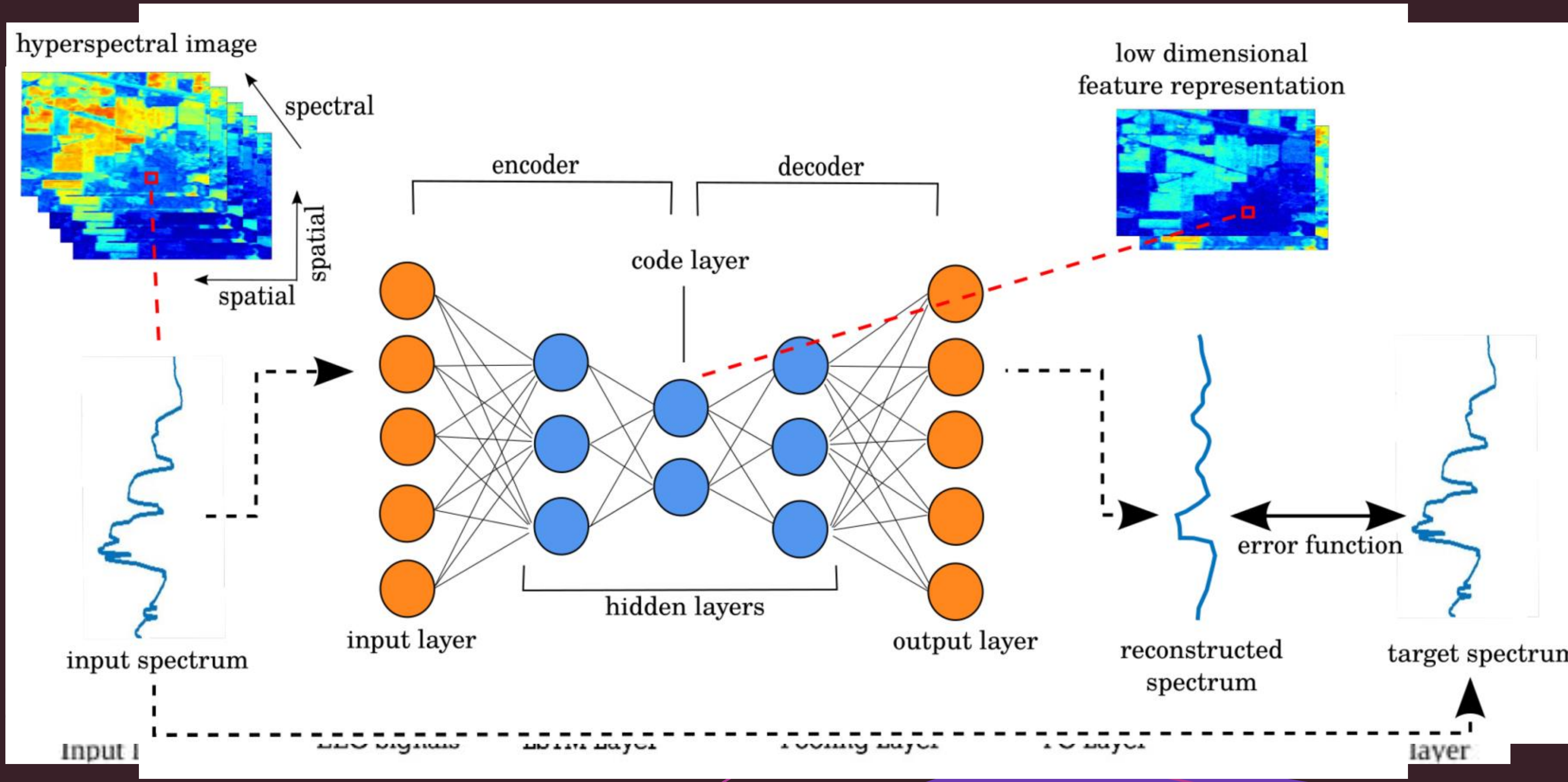


- Increasing computational capacity and the improved algorithms.
- Massive amount of data.
- Automatic feature extraction.

References: [8,9]

Deep learning Architectures & Applications

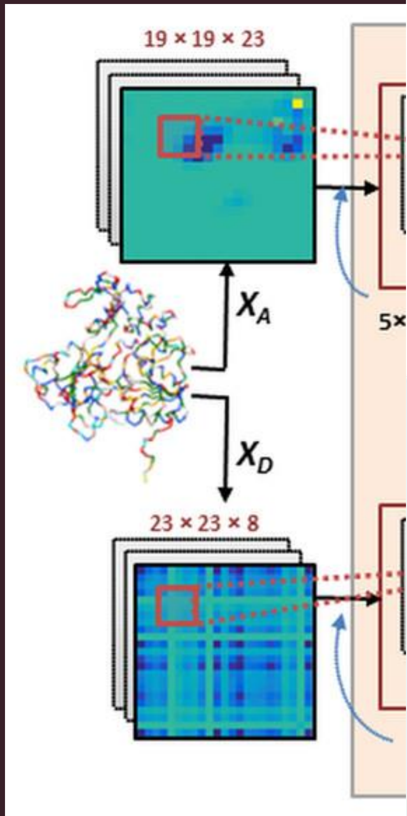
- Convolutional neural networks.
- Recurrent neural networks.
- Autoencoder.

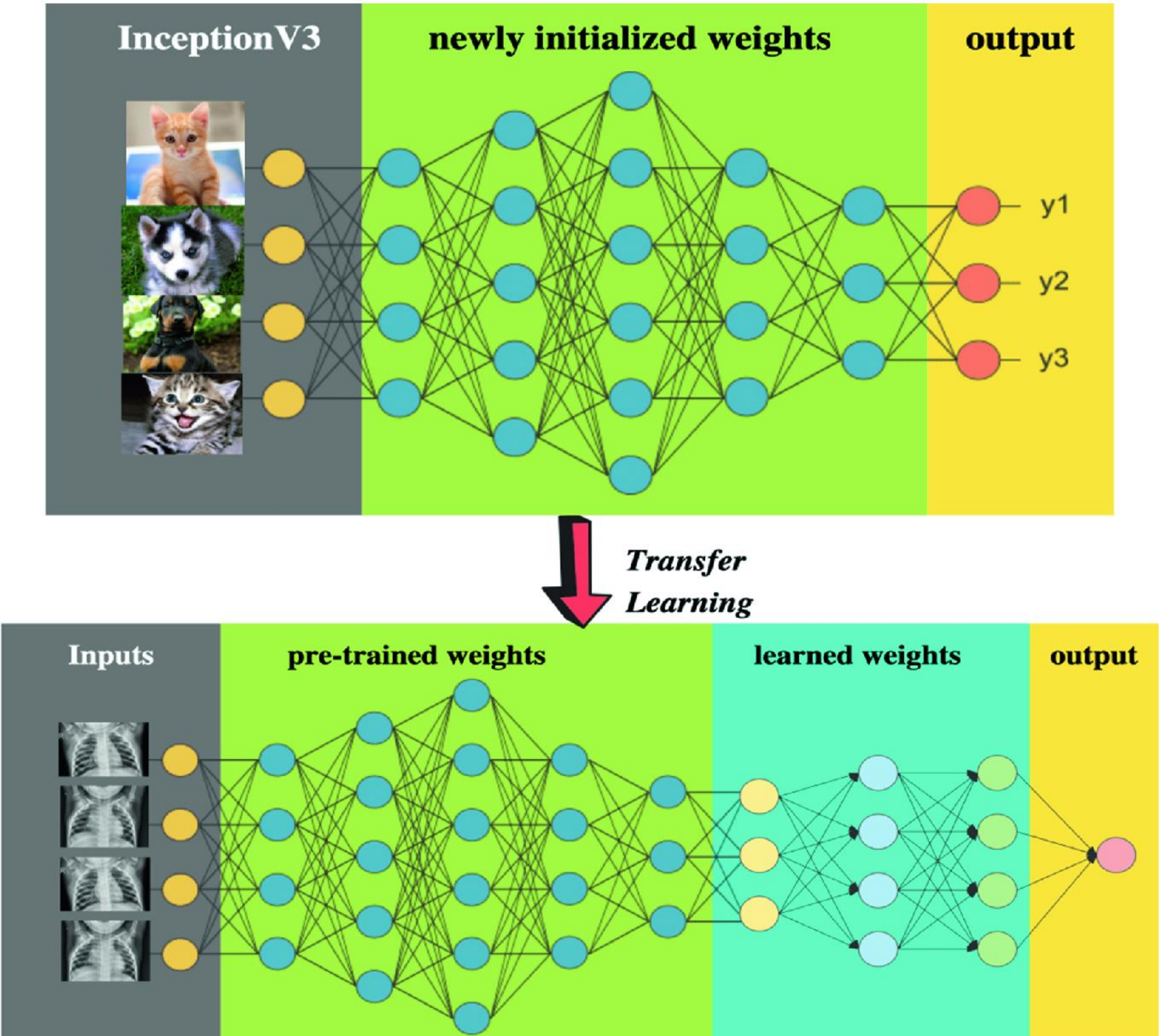


References: [8,9]

Deep learning Architectures & Applications

- Ensemble deep learning in bioinformatics.
- Transfer Learning Deep Learning.





References: [8,9]

Challenges

- Heterogeneous data
- No standard schema
- Imbalanced data
- Interpretability of models
- Computational challenges

CURSE of Dimentionality

References: [4,10]

References

1. Zhiqiang Zeng, Hua Shi, Yun Wu,et al. (2016) Survey of Natural Language Processing Techniques in Bioinformatics, College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

2. A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining 1A. Sivakumar and 2R.Gunasundari 1Department of Computer Science, Karpagam University, Coimbatore. sivamgp@gmail.com 2Department of Information Technology, Karpagam University, Coimbatore.

3. Jamshed, H., Ali Khan, M. S., Khurram, M., Inayatullah, S. y Athar, S. (2019). Data Preprocessing: A preliminary step for web data mining, Pakistan.

4. Lan, K., Wang, Dt., Fong, S. *et al.* A Survey of Data Mining and Deep Learning in Bioinformatics. *J Med Syst* **42,** 139 (2018). https://doi.org/10.1007/s10916-018-1003-9

5. Vincent Limo, (2019) A REVIEW OF DATA MINING IN BIOINFORMATICS, CENTRIA UNIVERSITY OF APPLIED SCIENCES Information Technology

6. Marco Cascella; Michael Rajnik; Abdul Aleem, et al. (2021) Features, Evaluation, and Treatment of Coronavirus (COVID-19)

7. Rehman O, Zhuang H, Muhamed Ali A, Ibrahim A, Li Z. Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach. *Cancers*. 2019; 11(3):431. https://doi.org/10.3390/cancers11030431

8. Min, S., Lee, B., & Yoon, S. (2016). Deep learning in bioinformatics. *Briefings in Bioinformatics*, bbw068. https://doi.org/10.1093/bib/bbw068

9. Tang, B., Pan, Z., Yin, K., & Khateeb, A. (2019). Recent Advances of Deep Learning in Bioinformatics and Computational Biology. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00214

10. Manisha Mathur (2018) Biomedical Challenges: A review, Post Graduate Institute of Veterinary Education & Research, Rajasthan University of Veterinary and Animal Science, Jaipur, Rajasthan, India