Text Based Diagnosis of COVID-19 Using Data Mining Techniques: A Comparative Study

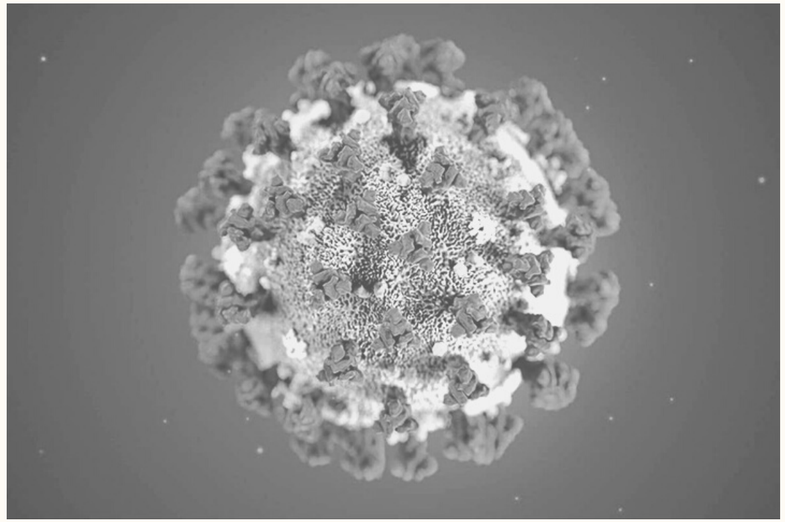
Aadarsh Gupta[1*] Aastha Valecha[2*] Sapna Mishra[3] Tapan Gandhi[4]

Indian Institute of Technology Delhi

[1]aadarsh.iitd@gmail.com [2]aasthavalecha9@gmail.com [3]eez208443@iitd.ac.in [4]tgandhi@ee.iitd.ac.in

Background

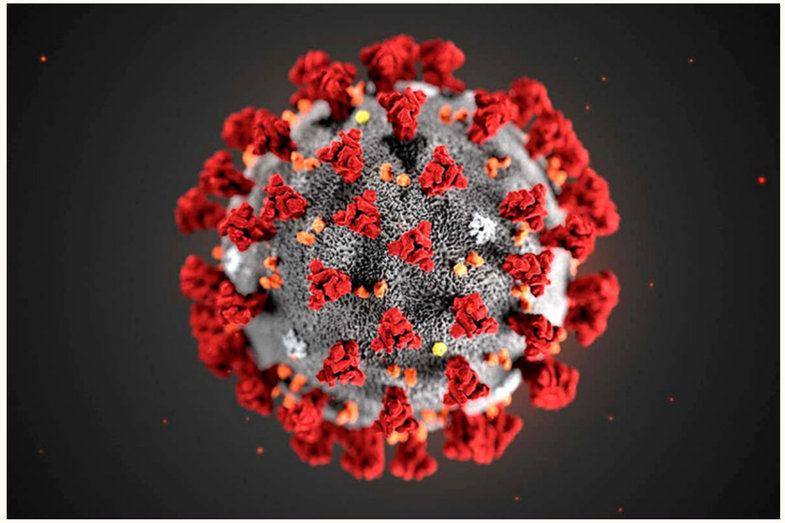
Wuhan City of China



COrona VIrus Disease of 2019 (COVID-19)

Background

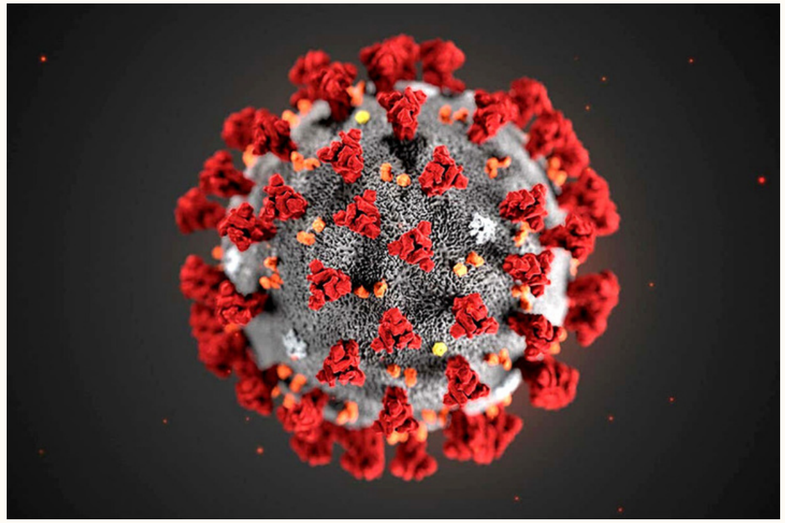
Wuhan City of China



COrona VIrus Disease of 2019 (COVID-19)

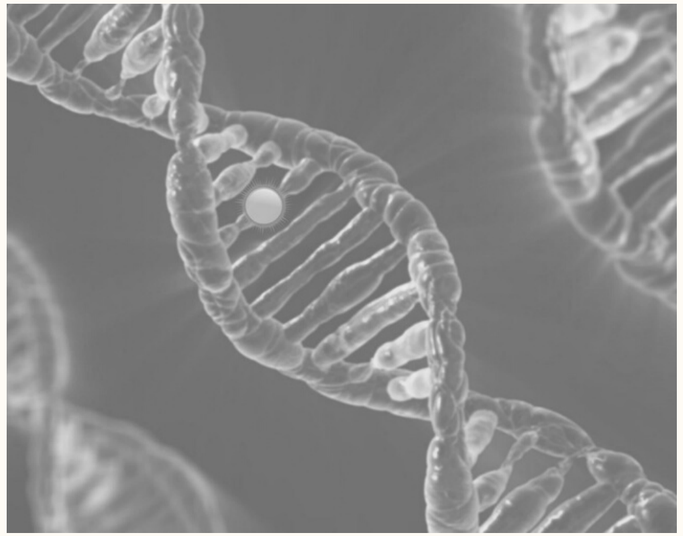
Background

Wuhan City of China



COrona VIrus Disease of 2019 (COVID-19)

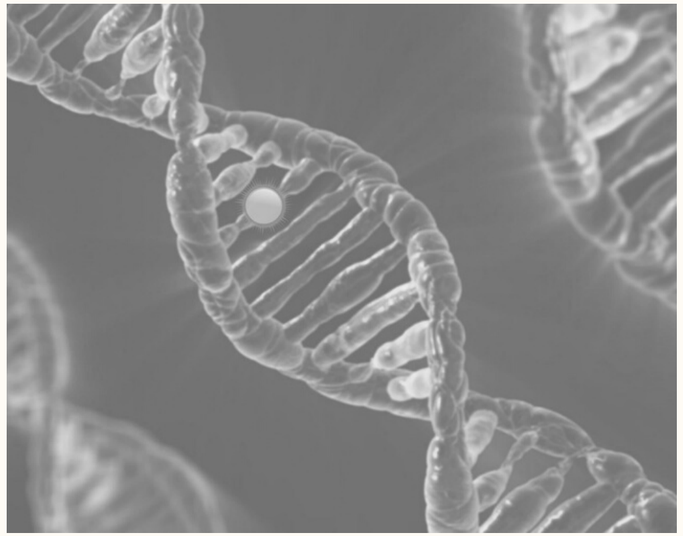
Background



Mutation in Severe Acute Respiratory Syndrome Coronavirus II (SARS-CoV-II)

Vaccines

Background



Mutation in Severe Acute Respiratory Syndrome Coronavirus II (SARS-CoV-II)

Vaccines

Background

Mutation in Severe Acute Respiratory Syndrome Coronavirus II (SARS-CoV-II)

Vaccines

Background

6.5 Million Deaths!

Mutation in Severe Acute Respiratory Syndrome Coronavirus II (SARS-CoV-II)

Vaccines

Challenges Posed by COVID19

- Gamification to prevent the Spread of Coronavirus
- Early Signs of Pending Pandemic
- New Technologies to Prevent the Transmission of Coronavirus
- Anti-COVID-19 Infection Protective Film
- Diminishing the Transmission of COVID-19
- Frontier Tech COVID Action: Ventilation Systems
- **Medicine/Vaccine**

Motivation

Symptomatic person

Treatment

Motivation

Motivation

Symptomatic person

Efficient and Faster diagnostic method

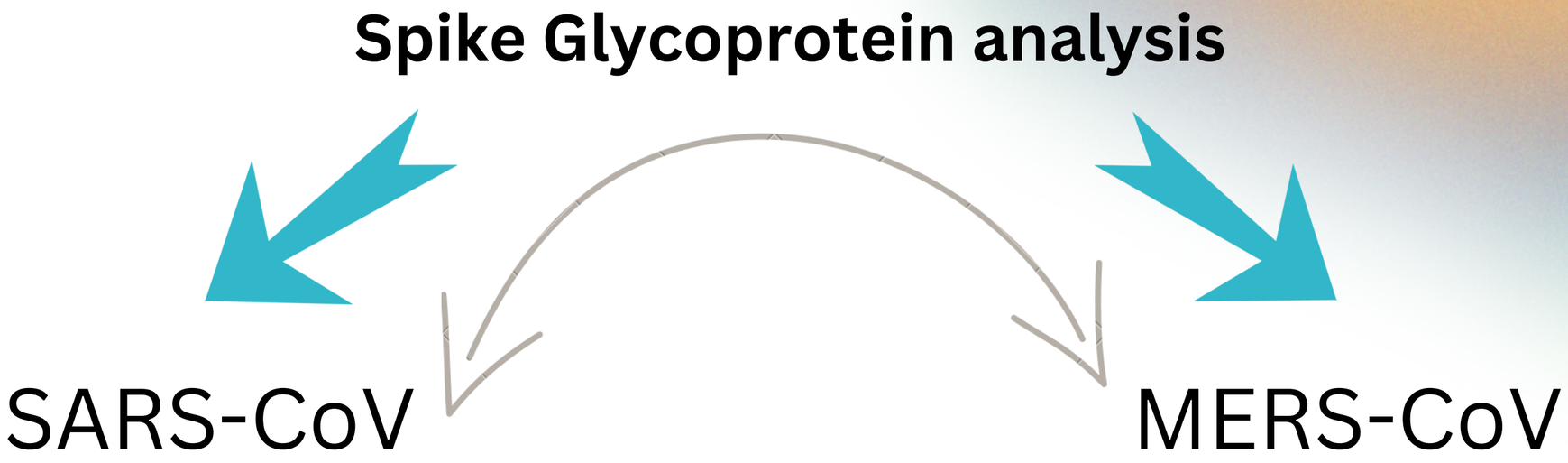
Treatment

Time Lag!

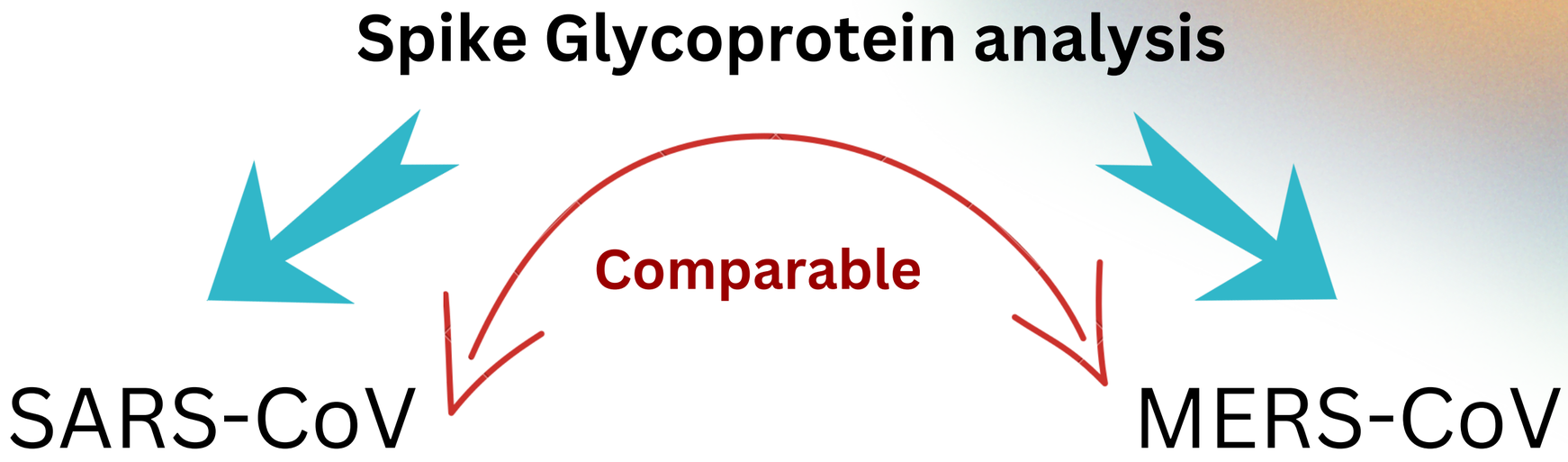
Correlation with Middle East Respiratory Syndrome-related CoronaVirus (MERS-CoV)

Spike Glycoprotein analysis

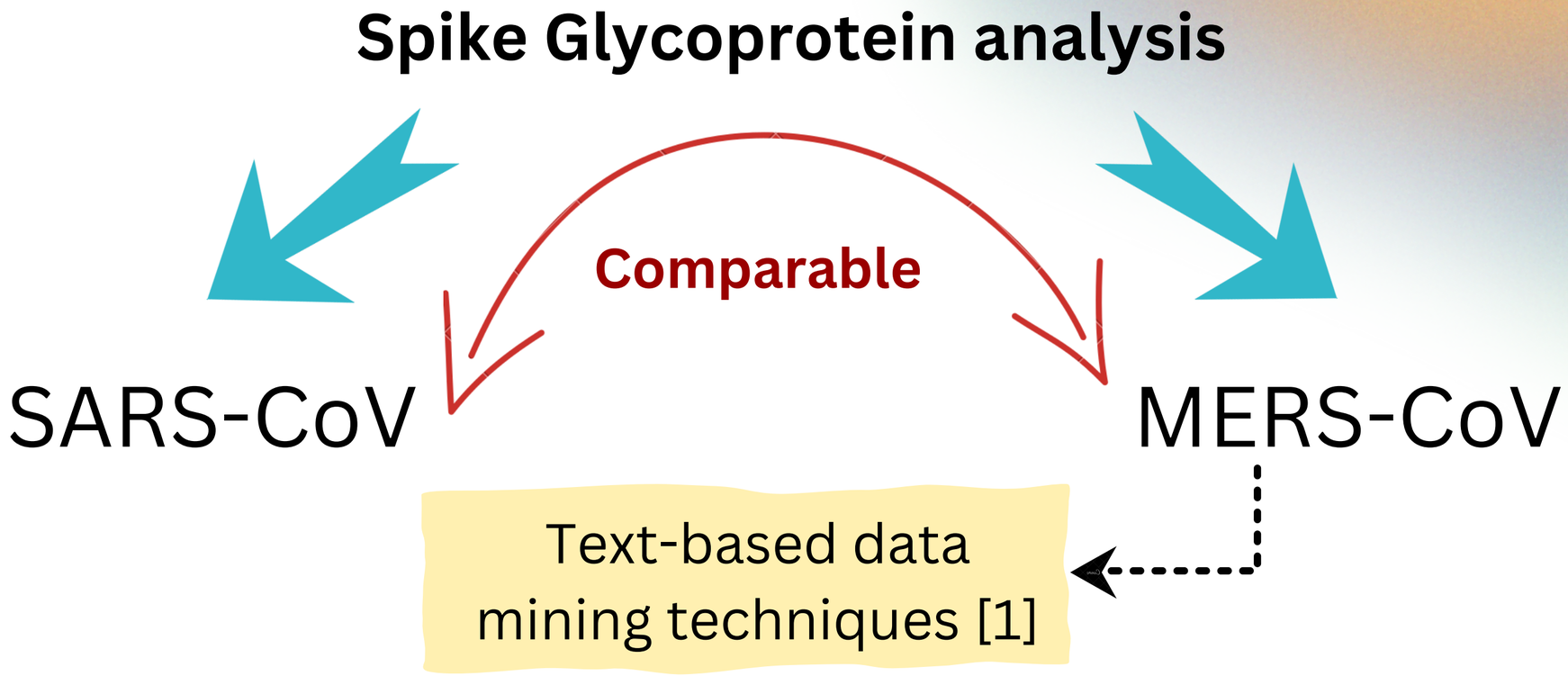
Correlation with Middle East Respiratory Syndrome-related CoronaVirus (MERS-CoV)



Correlation with Middle East Respiratory Syndrome-related CoronaVirus (MERS-CoV)

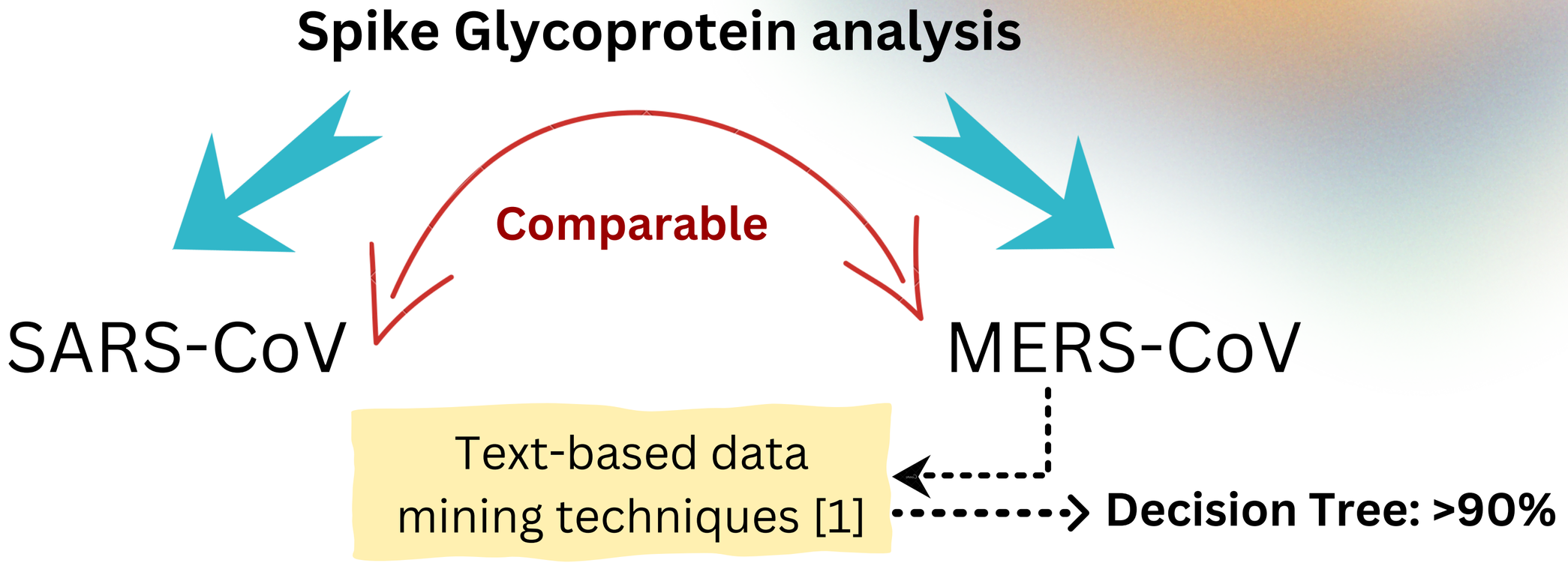


Correlation with Middle East Respiratory Syndrome-related CoronaVirus (MERS-CoV)



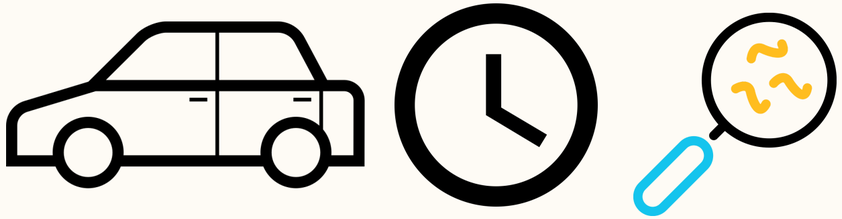
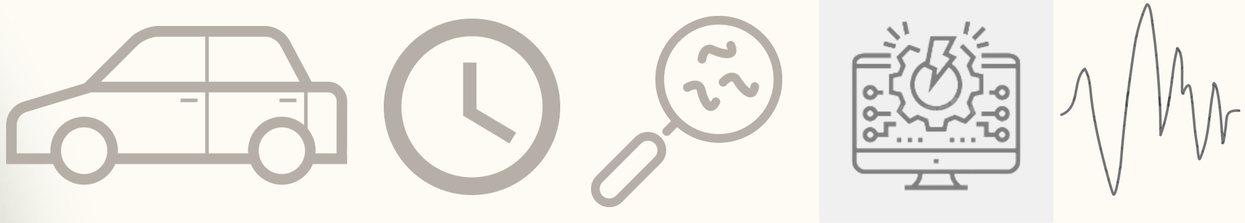
[1] H. Kurdia and N. AlMansour, “Identifying accurate classifier models for a text-based MERS-CoV dataset”

Correlation with Middle East Respiratory Syndrome-related CoronaVirus (MERS-CoV)



[1] H. Kurdia and N. AlMansour, “Identifying accurate classifier models for a text-based MERS-CoV dataset”

Examine existing diagnostics and our methods

Nucleic acid based		Travel & Cost ineffective High testing time Prone to contamination
X-Ray/Acoustic		
Text based (symptoms)		

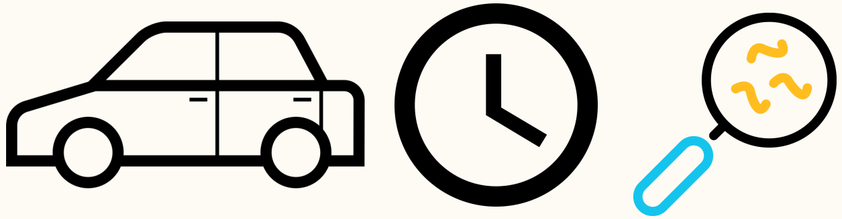
Examine existing diagnostics and our methods

Nucleic acid based		Travel & Cost ineffective High testing time Prone to contamination
X-Ray/Acoustic		Computationally expensive & noisy signal Erroneous devices
Text based (symptoms)		

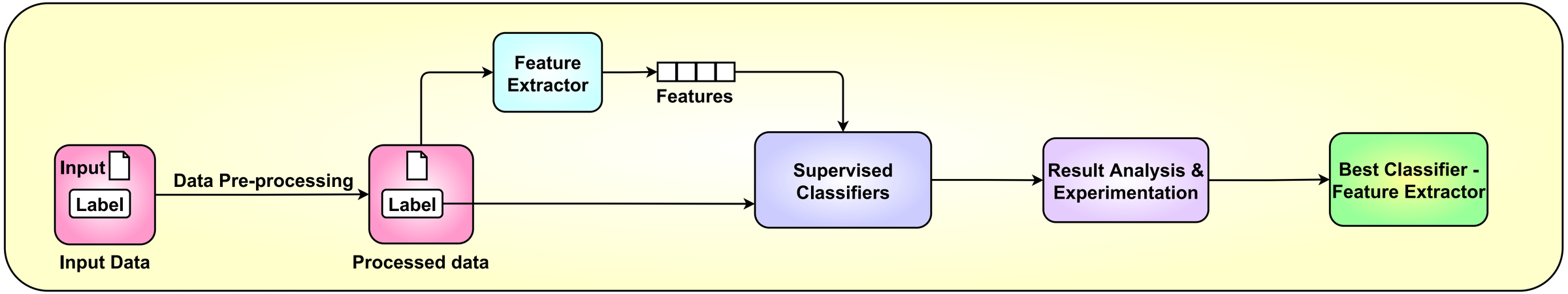
Examine existing diagnostics and our methods

Nucleic acid based		Travel & Cost ineffective High testing time Prone to contamination
X-Ray/Acoustic		Computationally expensive & noisy signal Erroneous devices
Text based (symptoms)		

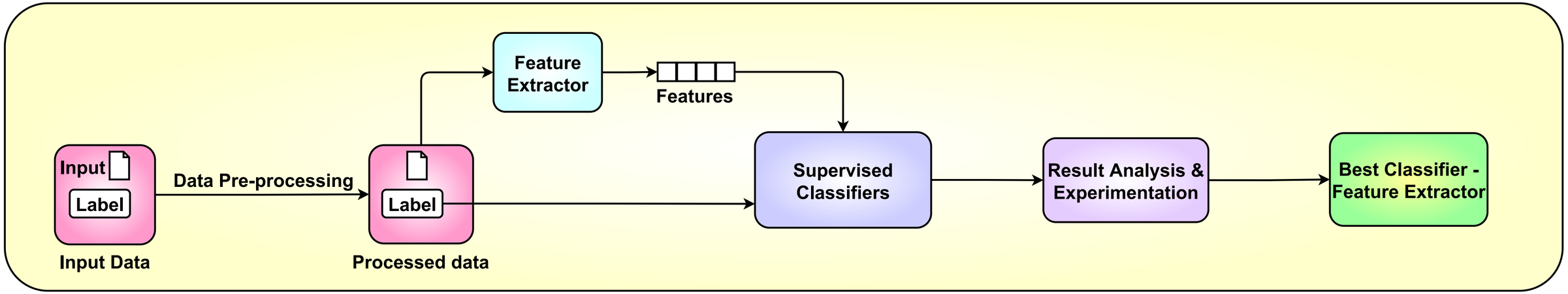
Examine existing diagnostics and our methods

Nucleic acid based		Travel & Cost ineffective High testing time Prone to contamination
X-Ray/Acoustic		Computationally expensive & noisy signal Erroneous devices
Text based (symptoms)		

Text based Diagnosis: Overview



Text based Diagnosis: Stages

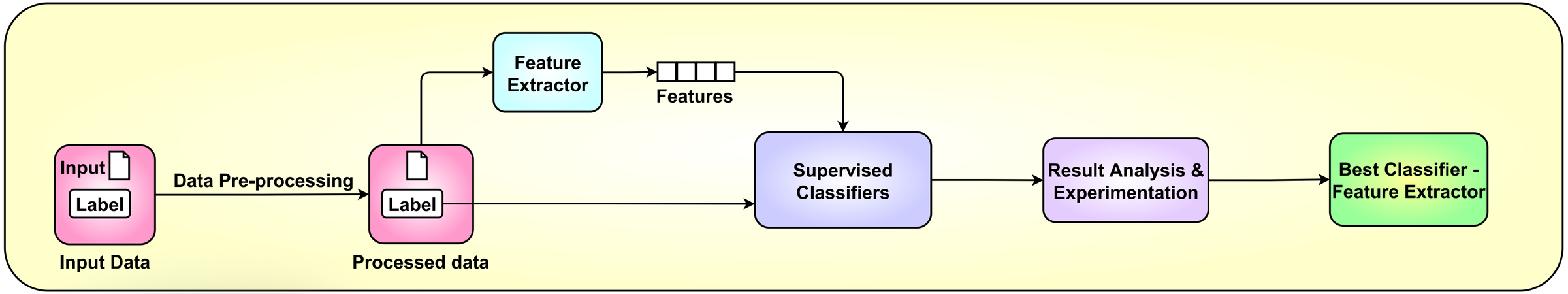


Coswara [2]

Symptoms and COVID Presence Data [3]

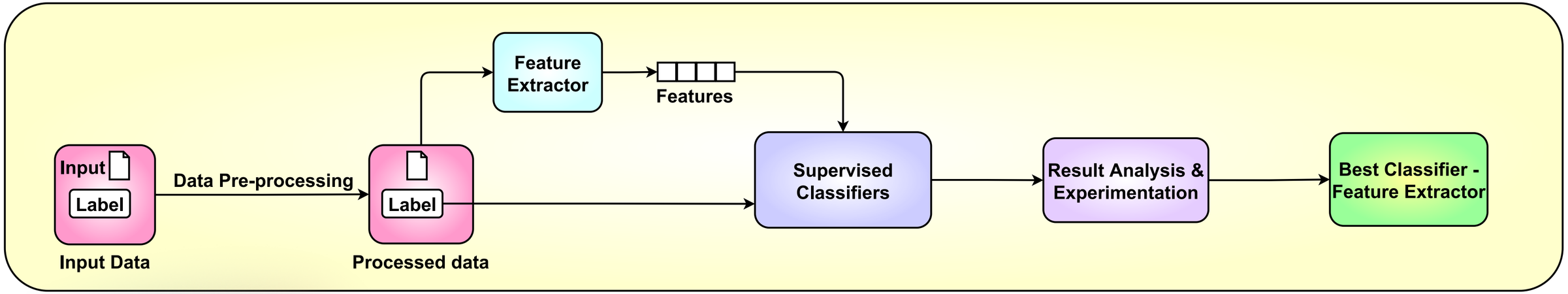
[2] N. Sharma et al., “Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis”
[3] “Symptoms and COVID Presence.” https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence

Text based Diagnosis: Stages



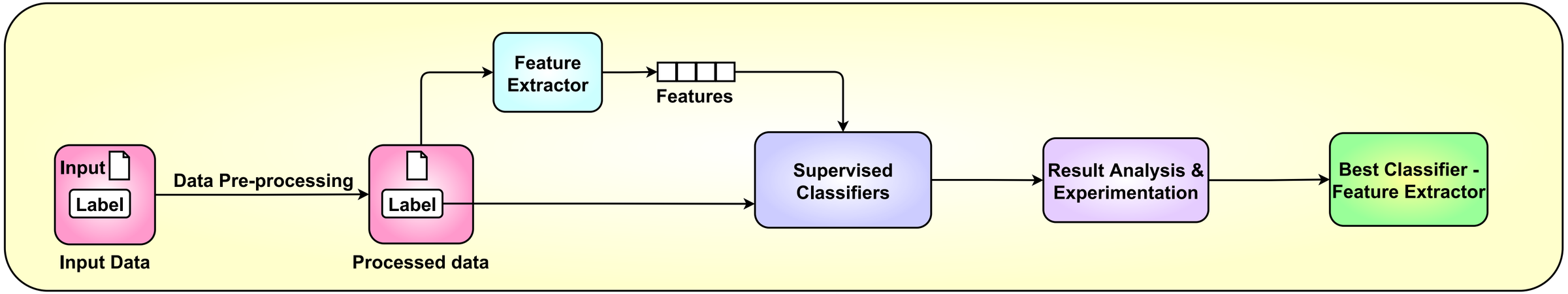
- Symptoms selection
- Label Encoding of text responses
- Ten-fold age categorization

Text based Diagnosis: Stages



- Principal Component Analysis (PCA)
- Manifold Approximation and Projection (UMAP)
- Isometric Mapping (ISOMAP)

Text based Diagnosis: Stages



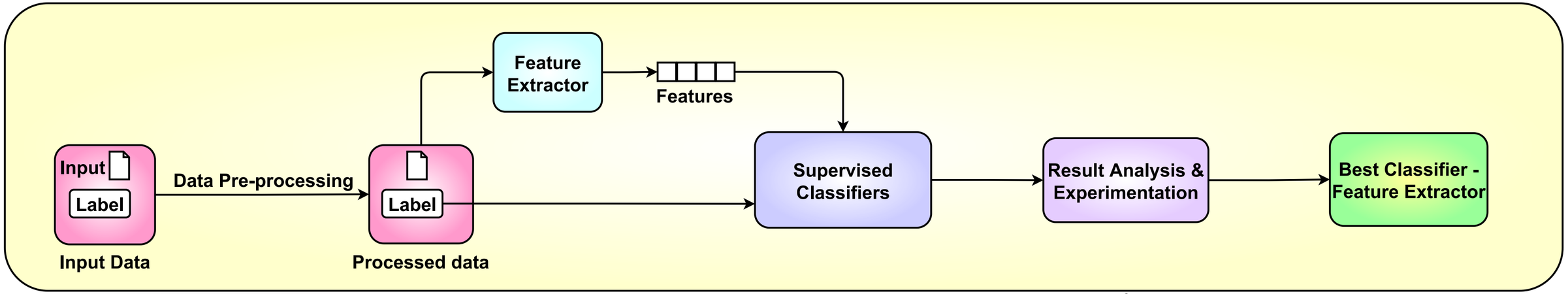
Support Vector Machines (SVM) ⟷ Logistic Regression (LR)

AdaBoost Classifier ⟷ K-Nearest Neighbor (KNN)

Gradient Boosting Classifier ⟷ Random Forest (RF)

Stacking Ensemble Classifier ⟷ Gaussian Naive Bayes (GNB)

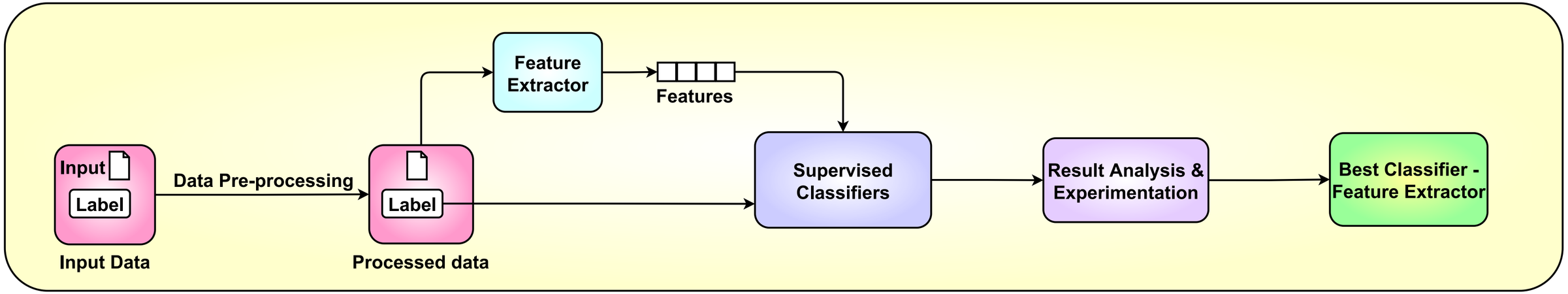
Text based Diagnosis: Stages



Accuracy

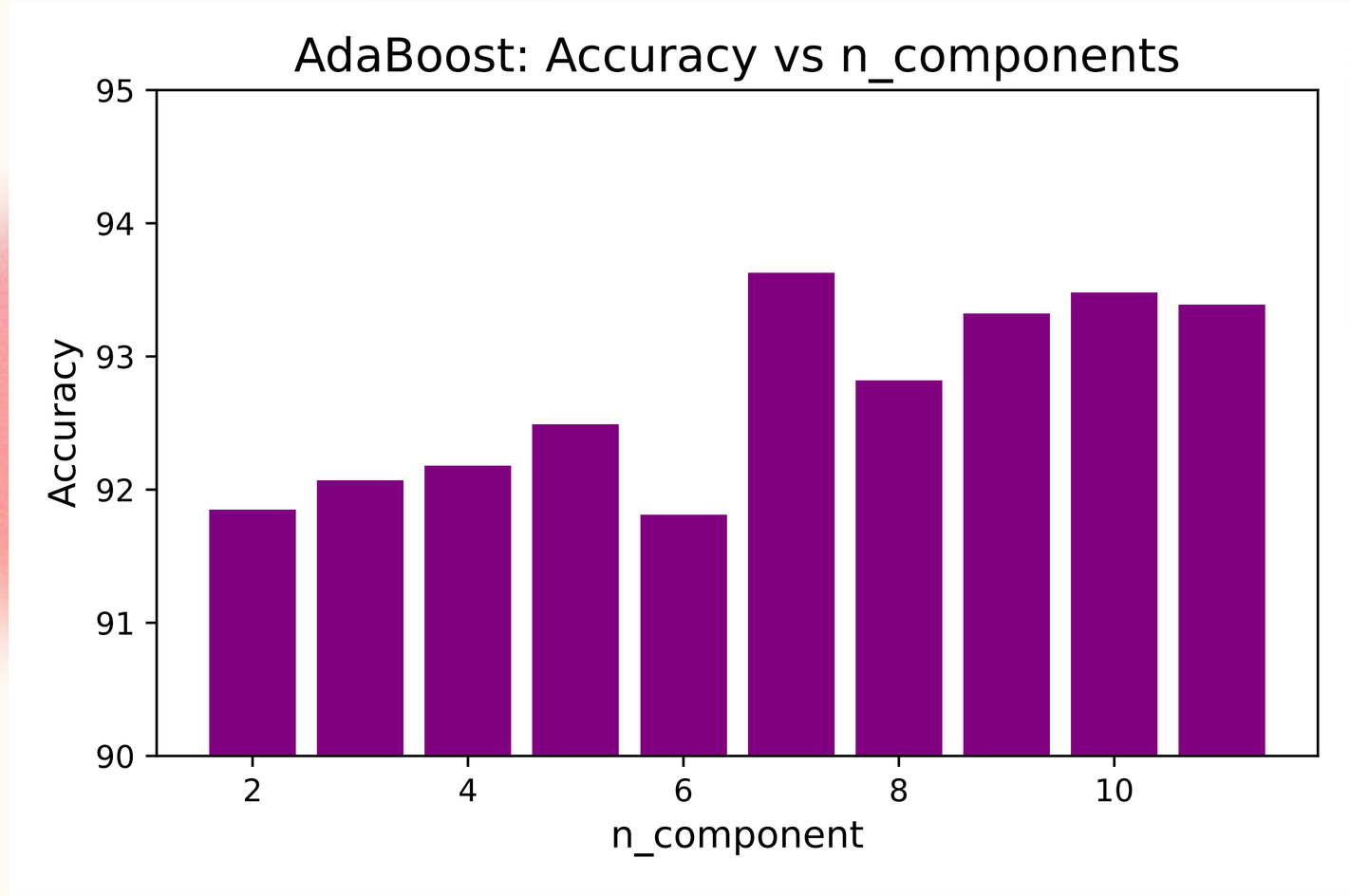
Key criterion due to low biasedness towards either class

Text based Diagnosis: Stages

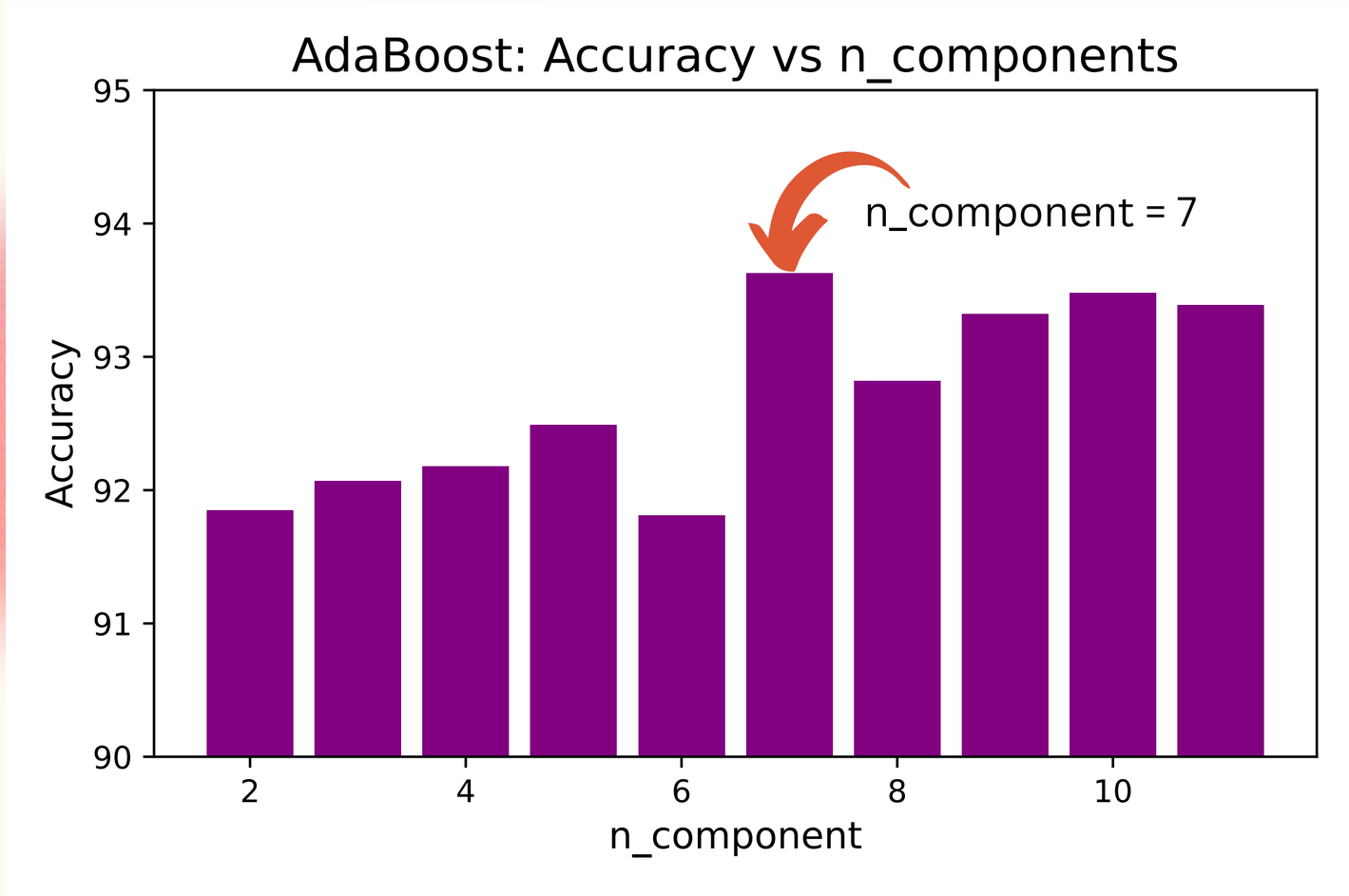


AdaBoost Classifier + PCA
Accuracy: 93.60

Optimal Results for PCA



Optimal Results for PCA

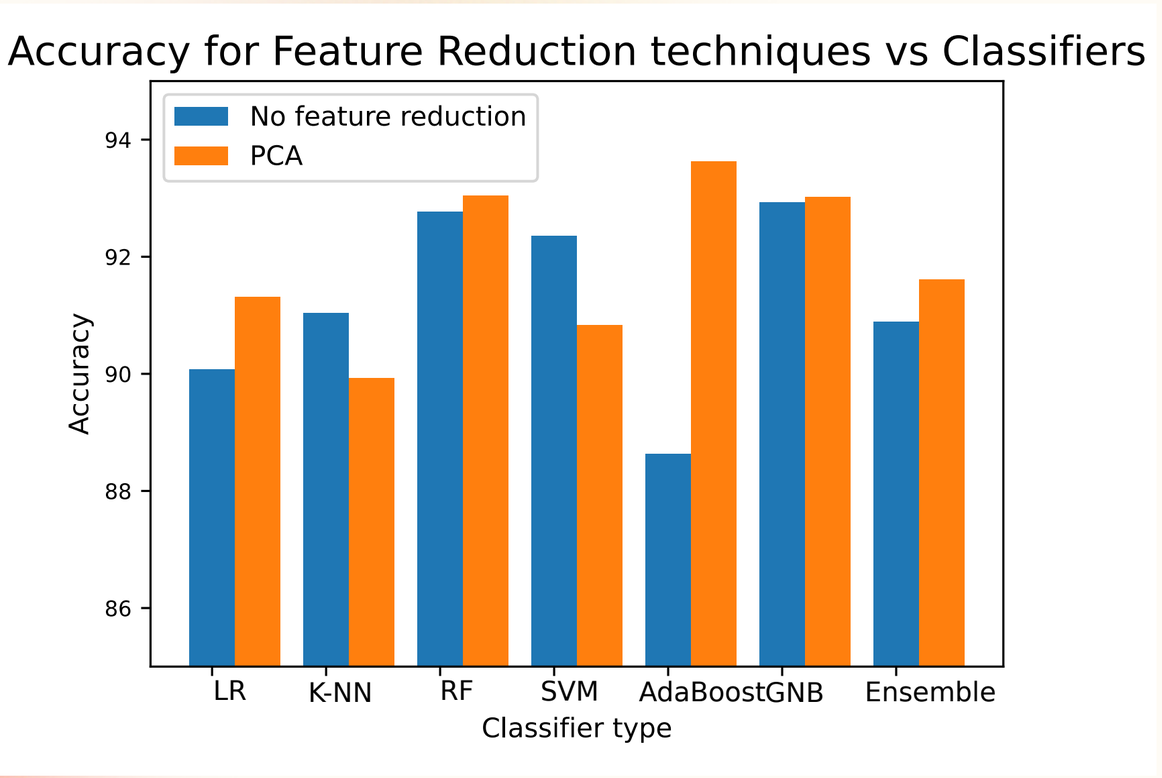


Quantitative Results

	Without FR	PCA	UMAP	ISOMAP
LR	90.080	91.310	80.180	89.930
K-NN	91.040	89.510	86.910	88.590
RF	**92.910**	92.640	**88.830**	91.150
GNB	65.400	92.400	81.970	90.650
SVM	92.360	90.830	83.270	90.030
AdaBoost	88.630	**93.630**	84.930	92.030
GBM	92.340	93.020	88.040	**92.510**
Ensemble	91.080	92.360	87.670	89.860

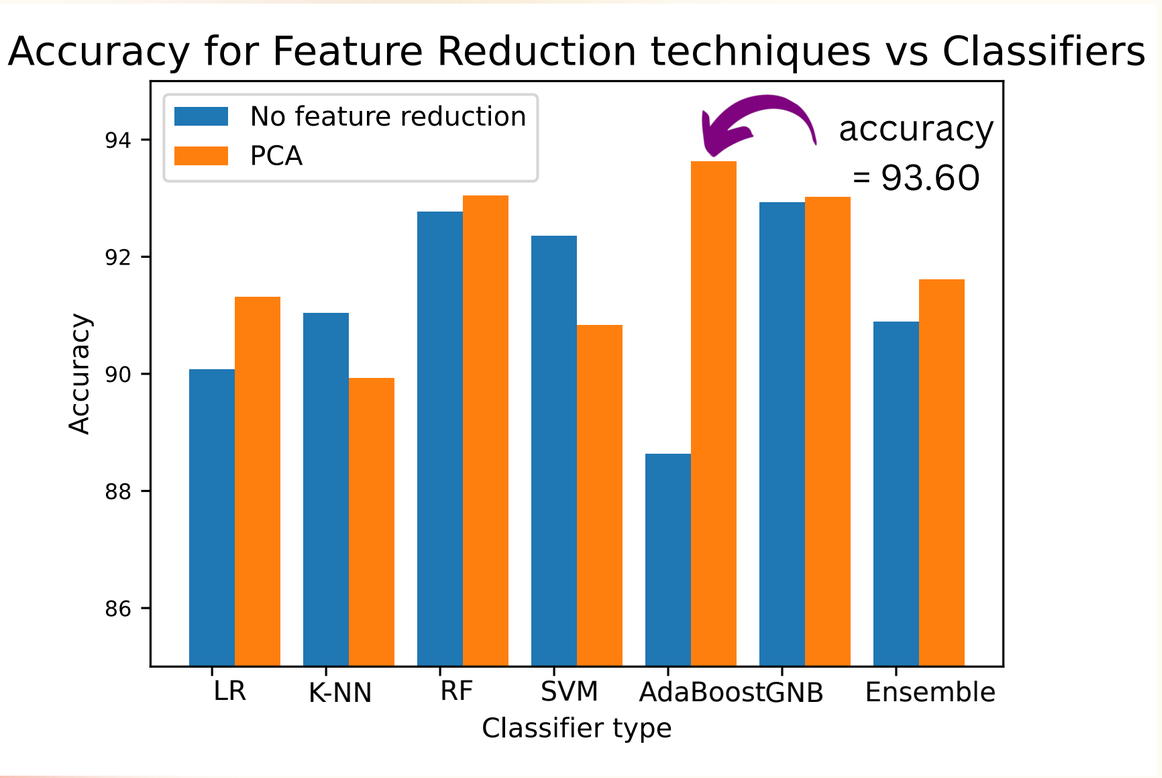
Quantitative Results

	Without FR	PCA	UMAP	ISOMAP
LR	90.080	91.310	80.180	89.930
K-NN	91.040	89.510	86.910	88.590
RF	**92.910**	92.640	**88.830**	91.150
GNB	65.400	92.400	81.970	90.650
SVM	92.360	90.830	83.270	90.030
AdaBoost	88.630	**93.630**	84.930	92.030
GBM	92.340	93.020	88.040	**92.510**
Ensemble	91.080	92.360	87.670	89.860

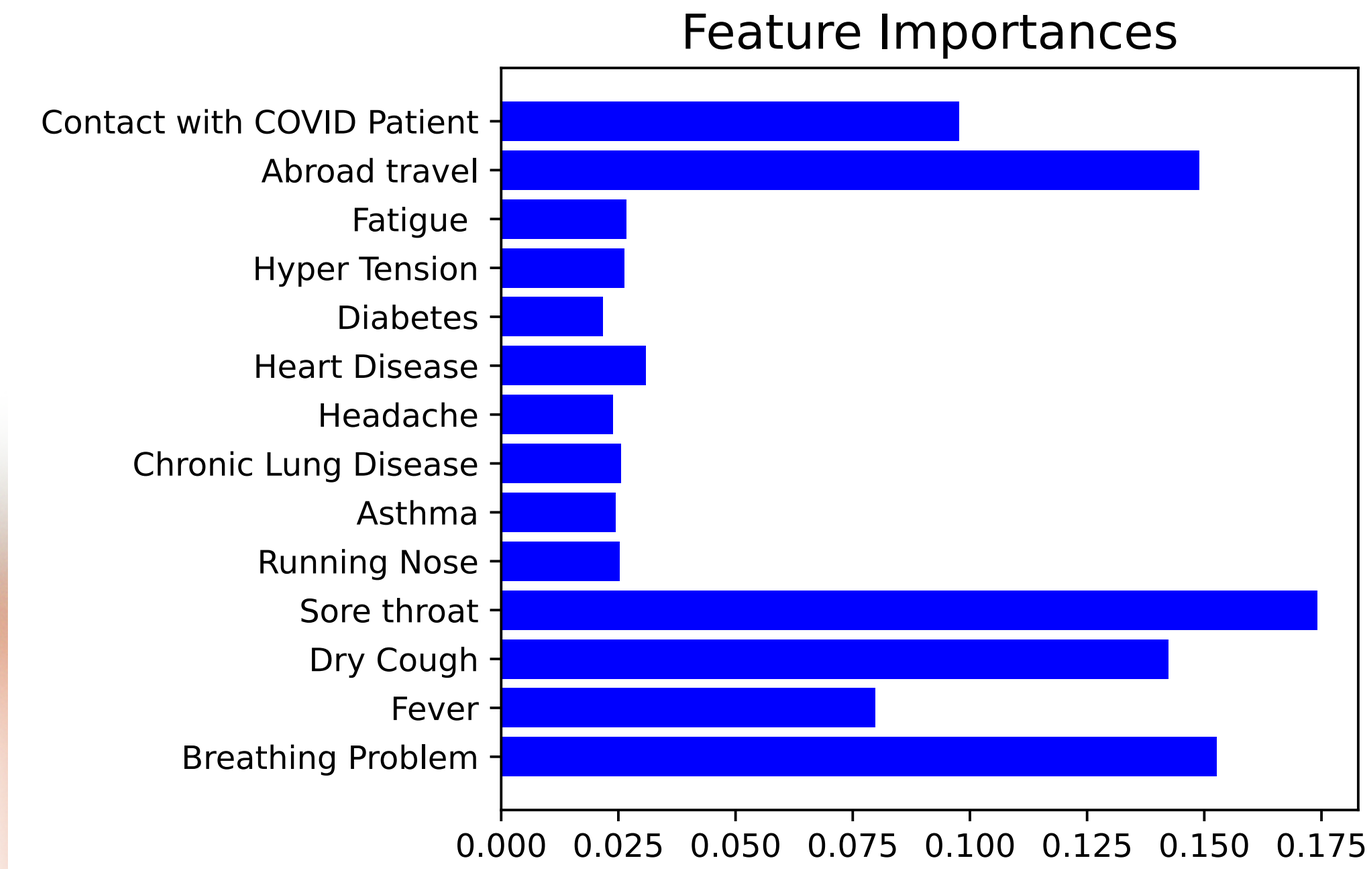


Quantitative Results

	Without FR	PCA	UMAP	ISOMAP
LR	90.080	91.310	80.180	89.930
K-NN	91.040	89.510	86.910	88.590
RF	**92.910**	92.640	**88.830**	91.150
GNB	65.400	92.400	81.970	90.650
SVM	92.360	90.830	83.270	90.030
AdaBoost	88.630	**93.630**	84.930	92.030
GBM	92.340	93.020	88.040	**92.510**
Ensemble	91.080	92.360	87.670	89.860

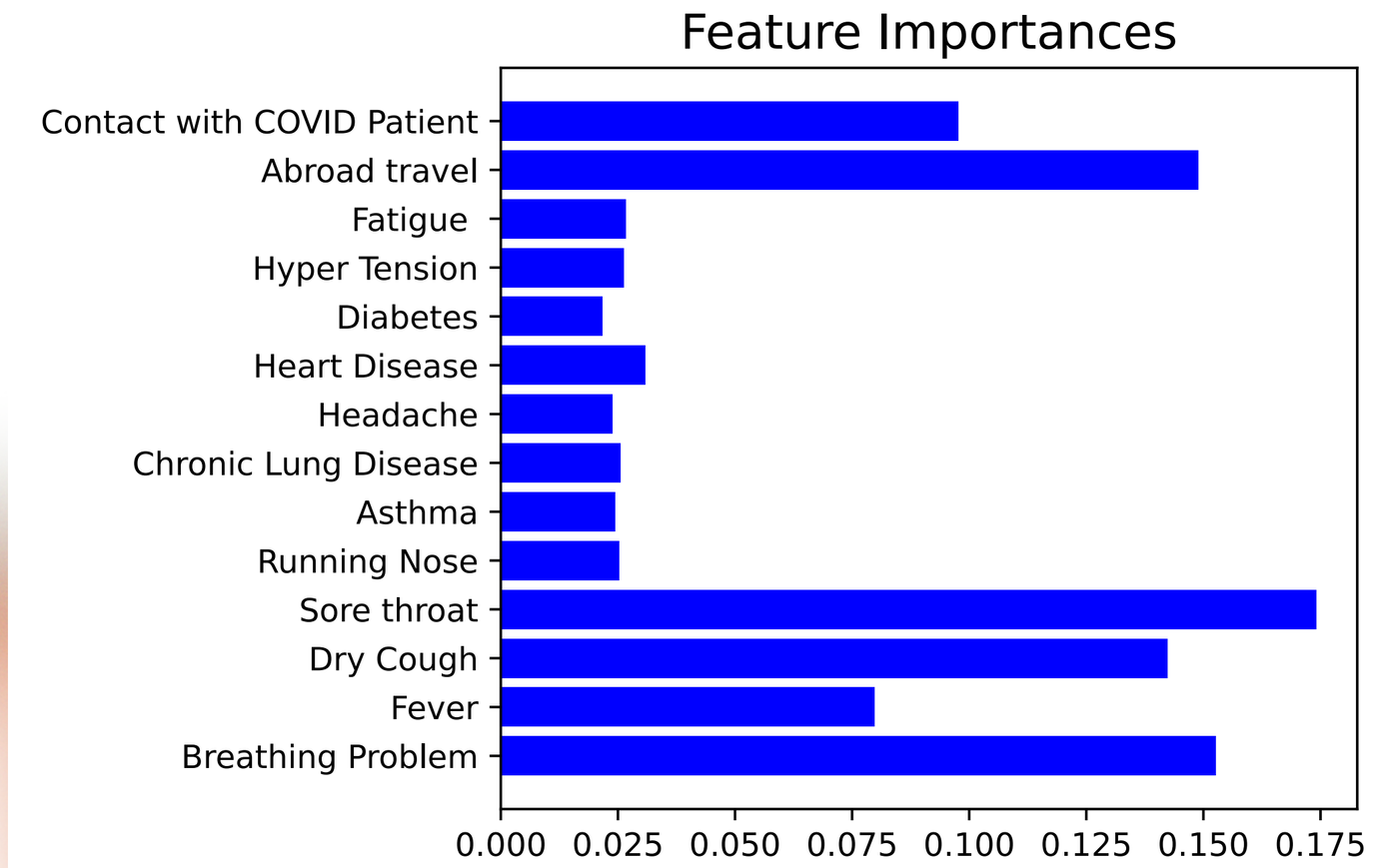


Further Analysis – Symptoms



Feature importance analysis using Random Forest (RF) classifier

Further Analysis – Symptoms



Important physiological symptoms

Feature importance analysis using Random Forest (RF) classifier

Future work

- On-line unsupervised methods to increase scalability of data collection
- Robust, reliable and large dataset collection: COUGHIUM [4] (in progress)
- Extending use of data mining techniques for diagnosis of other diseases

[4] "COUGHIUM: COVID-19 preliminary detection using Artificial Intelligence from Cough Recordings." https://web.iitd.ac.in/~eez208443/coughium/

Conclusion

- Text Based diagnostic methods enhance efficiency and portability
- Data mining techniques using supervised classification with feature extraction for optimal performance
- Interpretability of symptoms importance for physical diagnostics

Questions & Answers

Thank you for listening!