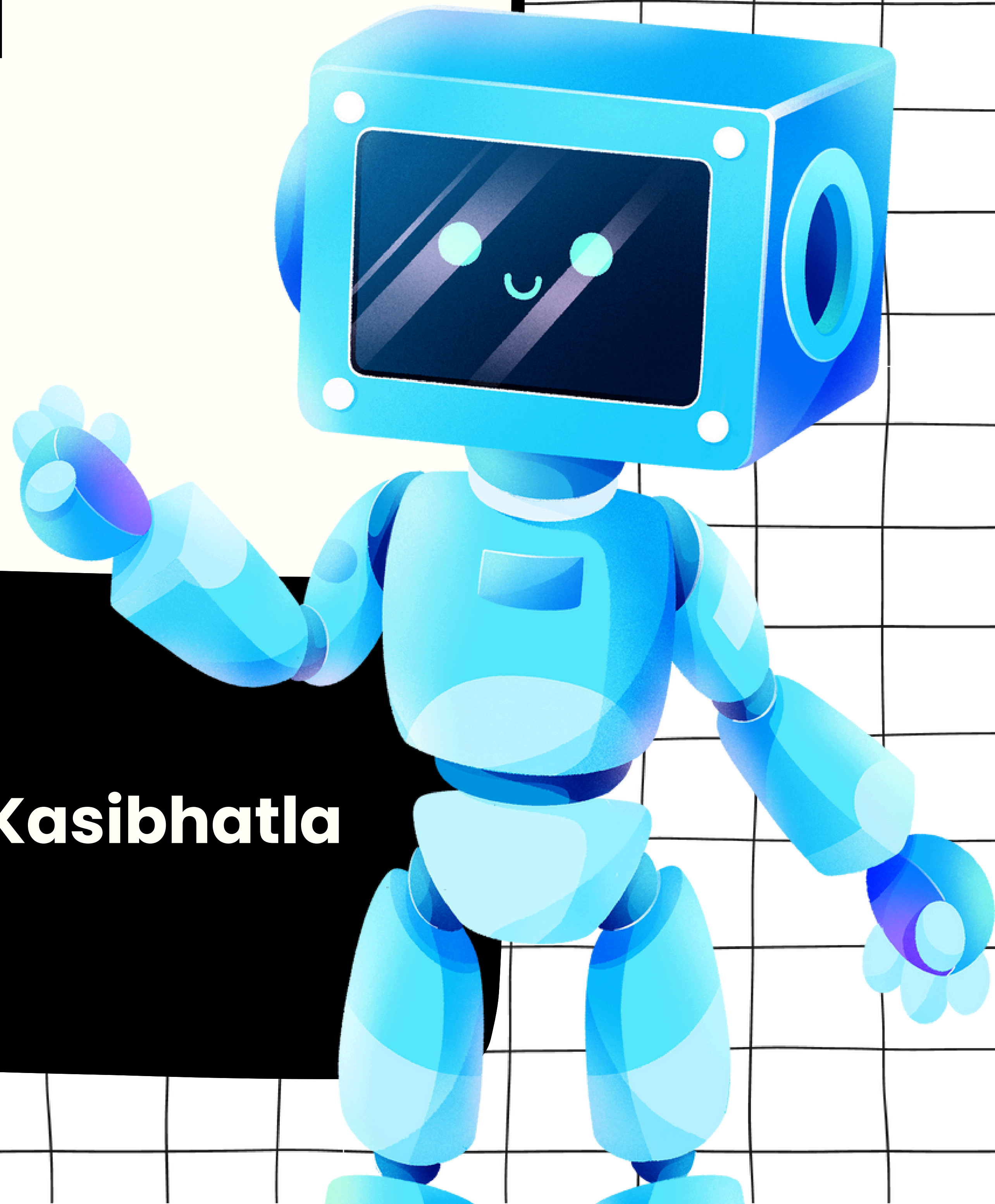
Risk Analysis for an AI Model in the Medical Field

Professor: Dr. Hossam Gaber
NUCL 5050 - Applied Risk Analysis
Ontario Tech University
Date: 24th Nov 2023

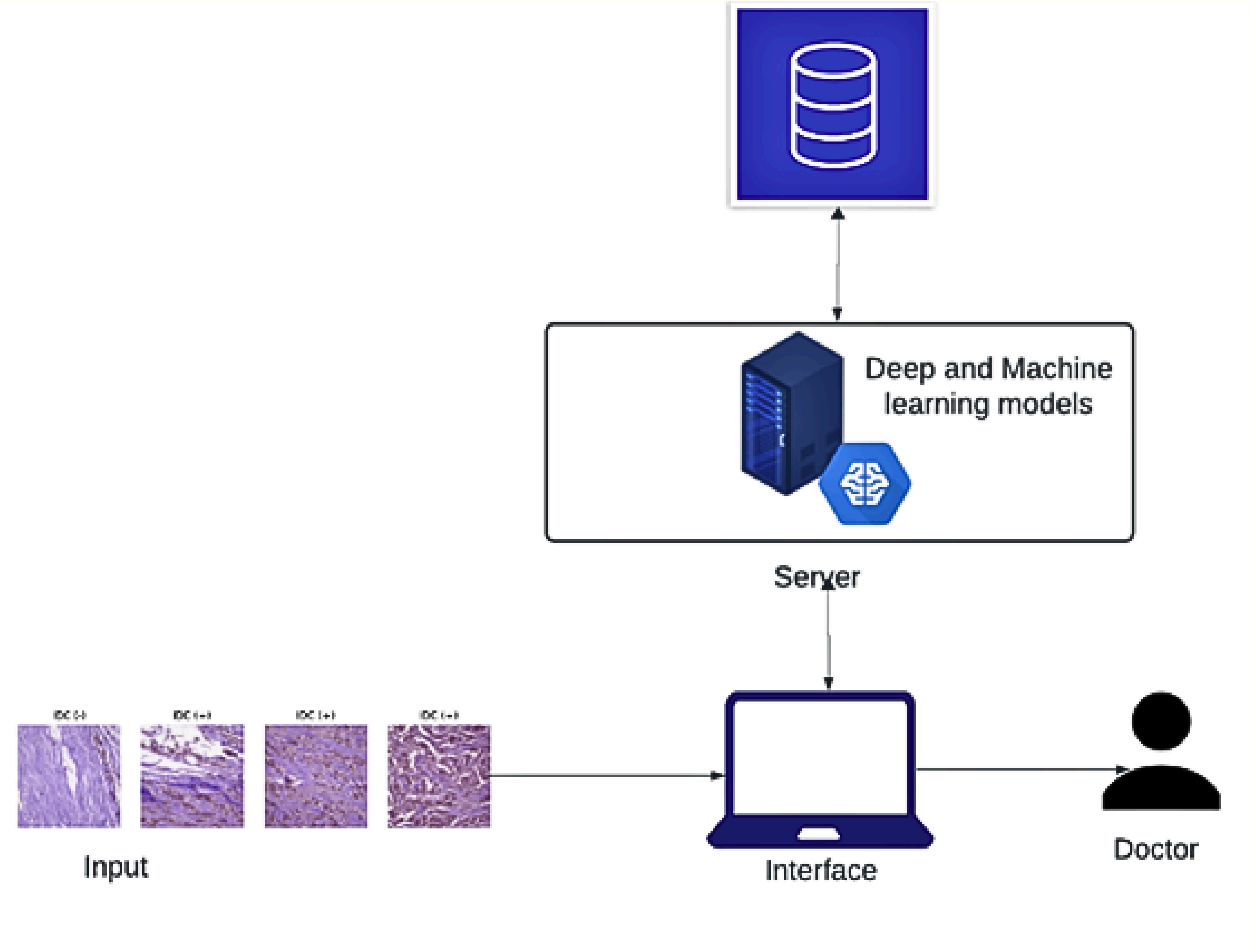
Name: Vallika Kasibhatla
Group 12



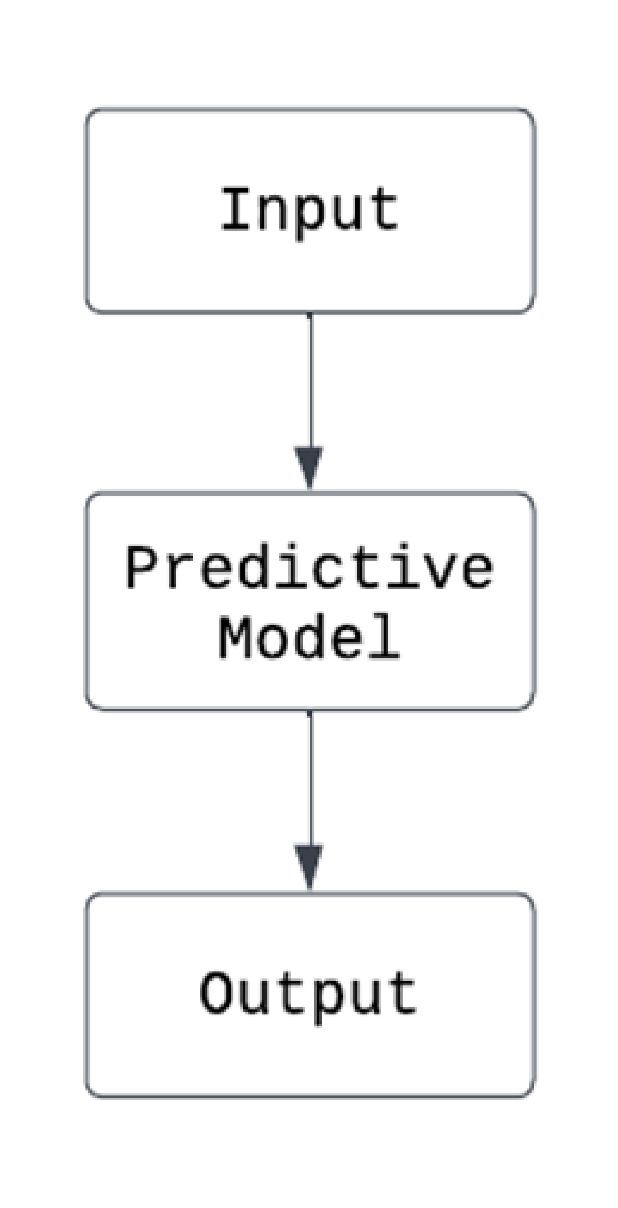
Agenda

1. Entire System
2. Main Scope
3. Bow Tie
4. Qualitative Analysis
5. Quantitative Analysis
6. Human Error Analysis
7. Reduction
8. Robustness of the System
9. Validation and Verification
10. Monitoring
11. Standards

Systems



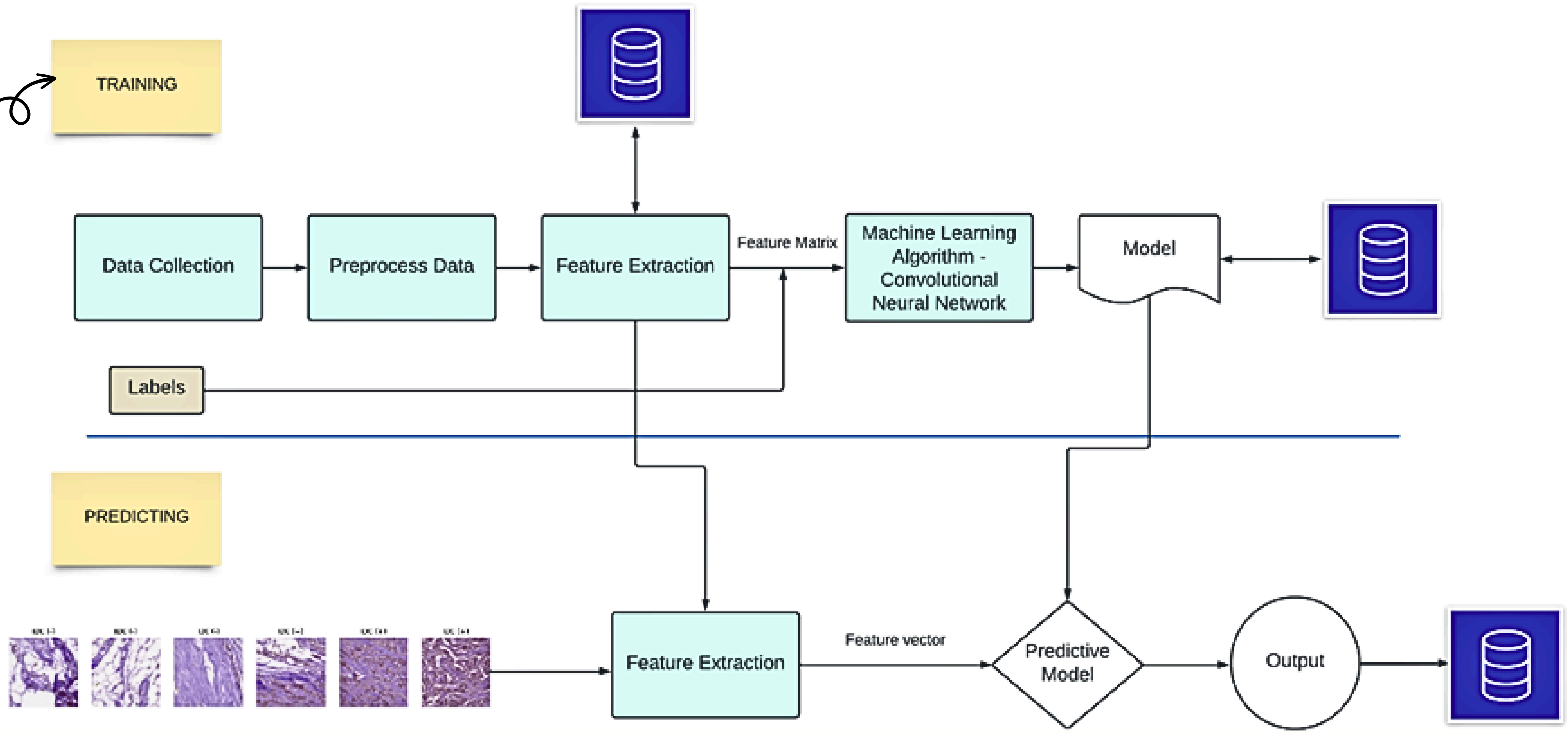
Entire System



Flow Chart

Systems

Entire System



Process Design

Causes

Probability

Model Failure (MF)

- Incorrect weight/Having wrong weightage [IW] = 0.026
- Input Data (ID)
  - Record Mishaps [RM] = 0.023
  - Manipulating Medical Data [MD] = 0.027
  - Biased/Missing Data while Training the Model [BD] = 0.016

External Failure (EF)

- Datasets unreliability [DU] = 0.044
- Overreliance on AI/Lack of Communication [O] = 0.03
- Misinterpretation of Output [MO] = 0.017
- Poisoning (Label modifications) [P] = 0.012

# bp	# su	rbc	classification
50 – 180	0 – 5	normal 50% [null] 38% Other (47) 12%	ckd 62% notckd 38% Other (2) 1%
80.0	0.0		ckd
50.0	0.0		ckd
80.0	3.0	normal	ckd
70.0	0.0	normal	ckd
80.0	0.0	normal	ckd
90.0	0.0		ckd
70.0	0.0		ckd
	4.0	normal	ckd
100.0	0.0	normal	ckd
90.0	0.0	abnormal	ckd
60.0	4.0		ckd
70.0	0.0	abnormal	ckd
70.0	1.0		ckd

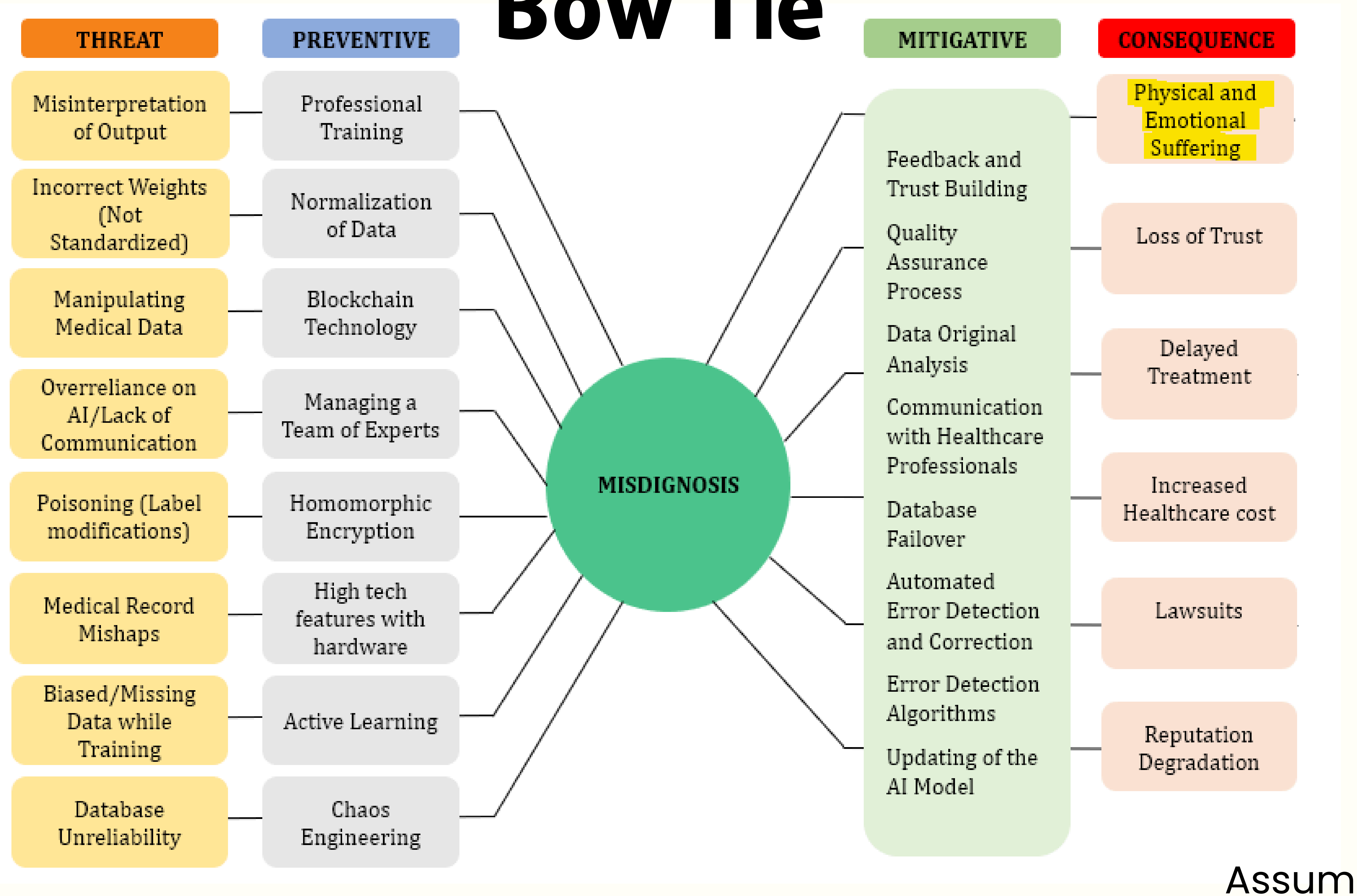
Dataset

Top Event: Model isnt accurate -> Misdiagnosis

- chest_xray
  - test
    - NORMAL
    - PNEUMONIA
  - train

Labels

Bow Tie



Assumption*

Qualititative Analysis

[1,2] Physical and Emotional Suffering $40,880,000
Loss of Trust $2,844,000
Delayed Treatment $20,714,000
Lawsuits $775,000,000
Reputation degradation $1,435,000
Increased Healthcare cost $3,609,000

Impact
How severe would the outcomes be if the risk occurred?

Probability
What is the probability the risk will happen?

	Insignificant 1	Minor 2	Significant 3	Major 4	Severe 5
5 Almost Certain	Medium 5	High 10	Very high 15	Extreme 20	Extreme 25
4 Likely	Medium 4	Medium 8	High 12	Very high 16	Extreme 20
3 Moderate	Low 3	Medium 6	Medium 9	High 12	Very high 15
2 Unlikely	Very low 2	Low 4	Medium 6	Medium 8	High 10
1 Rare	Very low 1	Very low 2	Low 3	Medium 4	Medium 5

SafetyCulture

Causes	Likelihood	Impact	LxC	Risk Rating	Response (Action)
Misinterpretation of Output	1	4	4	Low	Training to make sure the mistake won't repeat. And monitor the patients who could have been affected.
Incorrect weights	3	3	9	Medium	Rebuild the model with better preprocessing steps with quality checks.
Manipulating Medical Data	3	4	12	High	Restoring the data from backups, and investigate on who/ how this data was altered
Overreliance on AI/Lack of Communication	4	3	12	High	Rebuild trust with the patients by open communication. The final decision should always be analyzed with human intervention
Poisoning (Label modifications)	1	4	4	Low	Make the findings aware to the public to not compromise safety. Find the vendors/ professionals to identity the incident
Medical Record Mishaps	2	4	12	Medium	Improve the data entry procedures with root cause analysis and add data integrity safeguards.
Biased/Missing Data while Training the Model	1	4	4	Low	Transparency and compliance with the affected individuals. Start an audit of data handling practices
Database Unreliability	5	4	20	Extreme	Re-analyze and consult an expert judgment. Find the vendors/ professionals to identity the incident

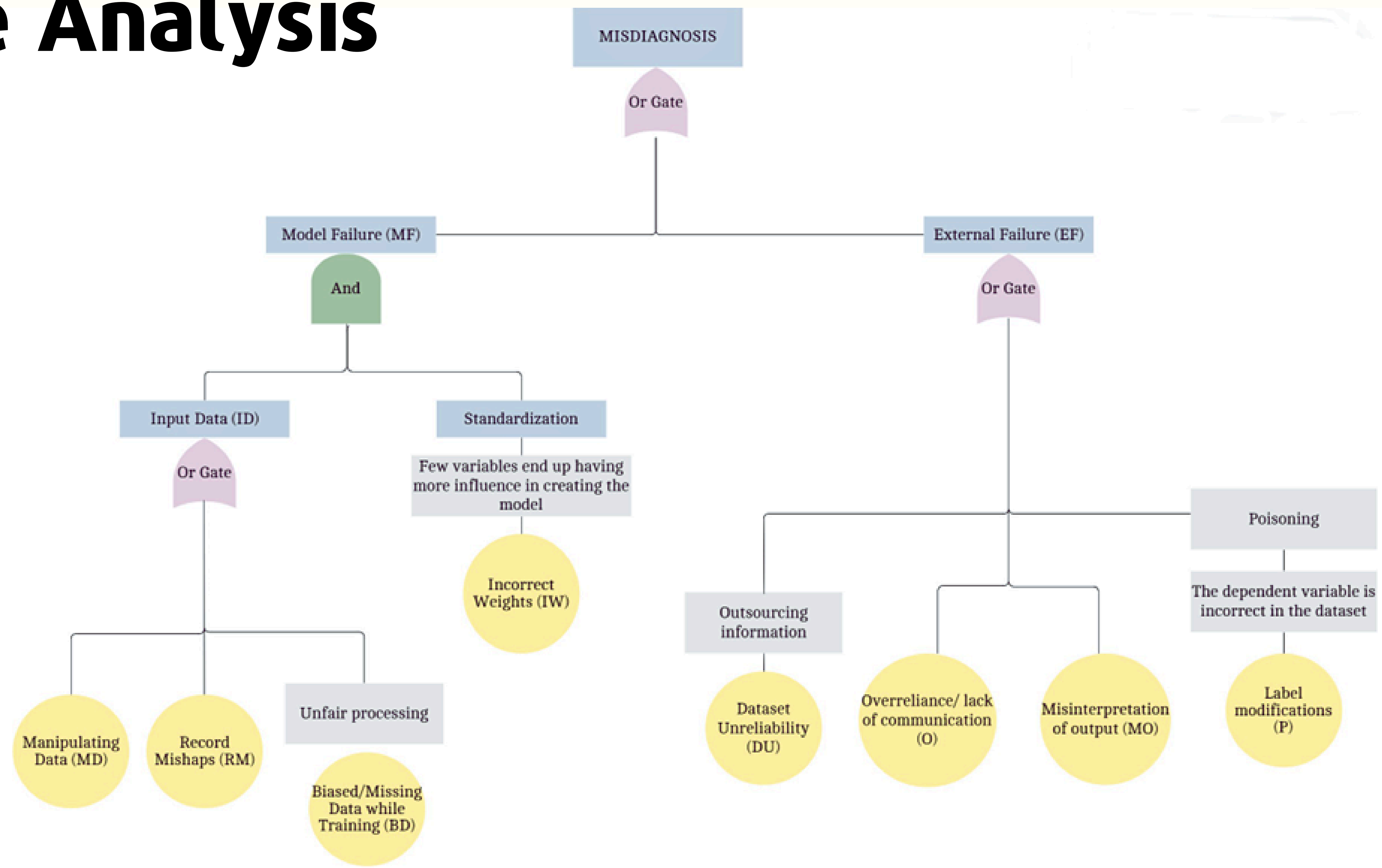
Table: Risk Matrix Analysis

Quantitative Analysis

Fault Tree

$$\frac{\frac{\frac{Misdiagnosis}{Model\ Failure \quad External\ Failure}}{ID, IW \quad External\ Failure}}{\{RM, IW\} \quad \{MD, IW\} \quad \{BD, IW\} \quad External\ Failure}$$

$$\frac{External\ Failure}{\{RM, IW\} \quad \{MD, IW\} \quad \{BD, IW\} \quad \{DU\} \quad \{O\} \quad \{MO\} \quad \{P\}}$$

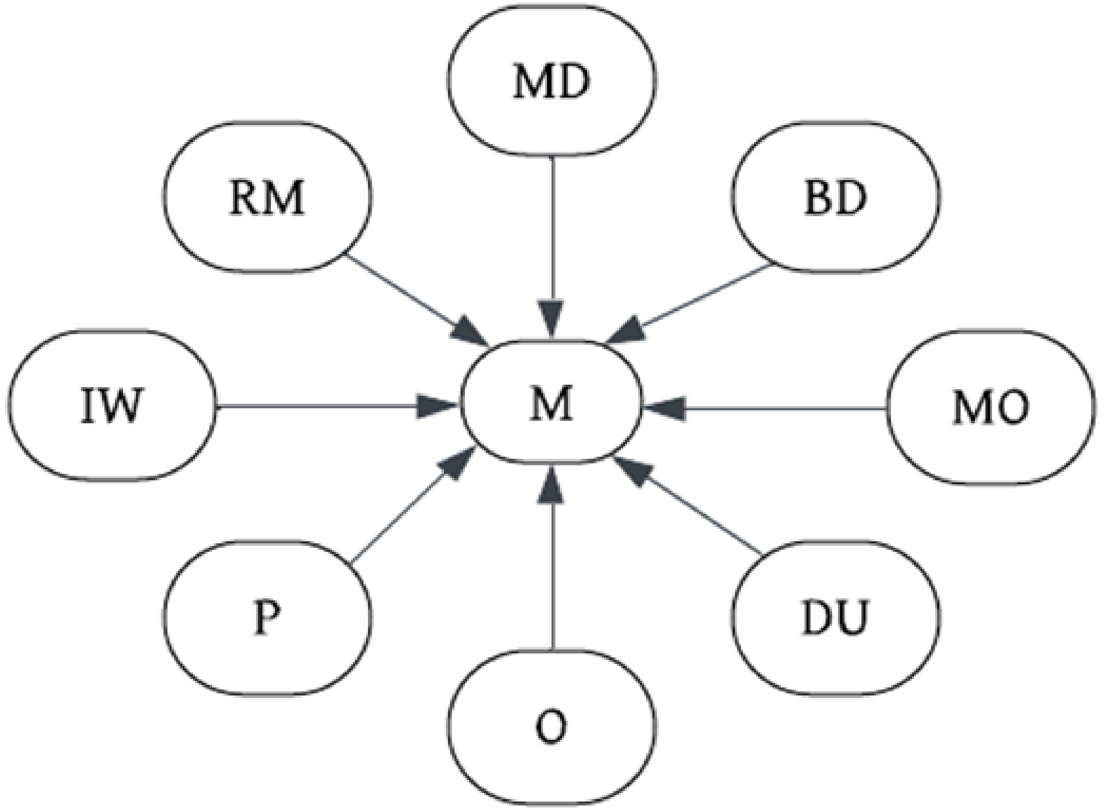


Mocus Method and Minimal Cut Set

$$P(M) = P(RM) * P(IW) + P(MD) * P(IW) + P(BD) * P(IW) + P(DU) + P(O) + P(MO) + P(P)$$

Probability of Failure = 10.5%

Bayesian Belief Network



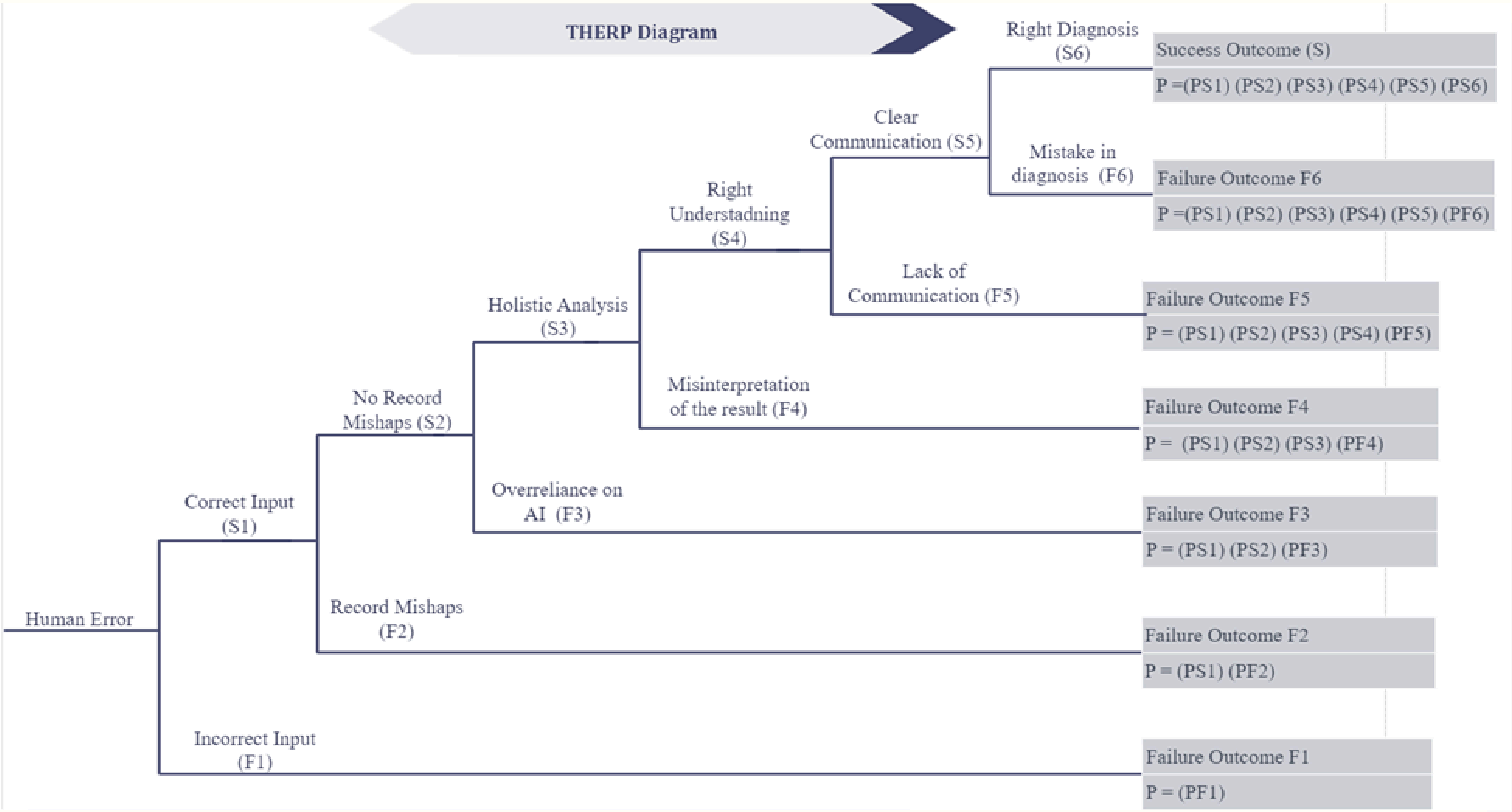
M	
P(M)	0.011
P(~M)	0.89
IW	
P(IW)	0.026
P(~IW)	0.974
MD	
P(MD)	0.027
P(~MD)	0.973
O	
P(O)	0.03
P(~O)	0.97
P	
P(P)	0.012
P(~P)	0.988
RM	
P(RM)	0.023
P(~RM)	0.977
BD	
P(BD)	0.016
P(~BD)	0.984
DU	
P(DU)	0.044
P(~DU)	0.956
MO	
P(MO)	0.017
P(~MO)	0.983

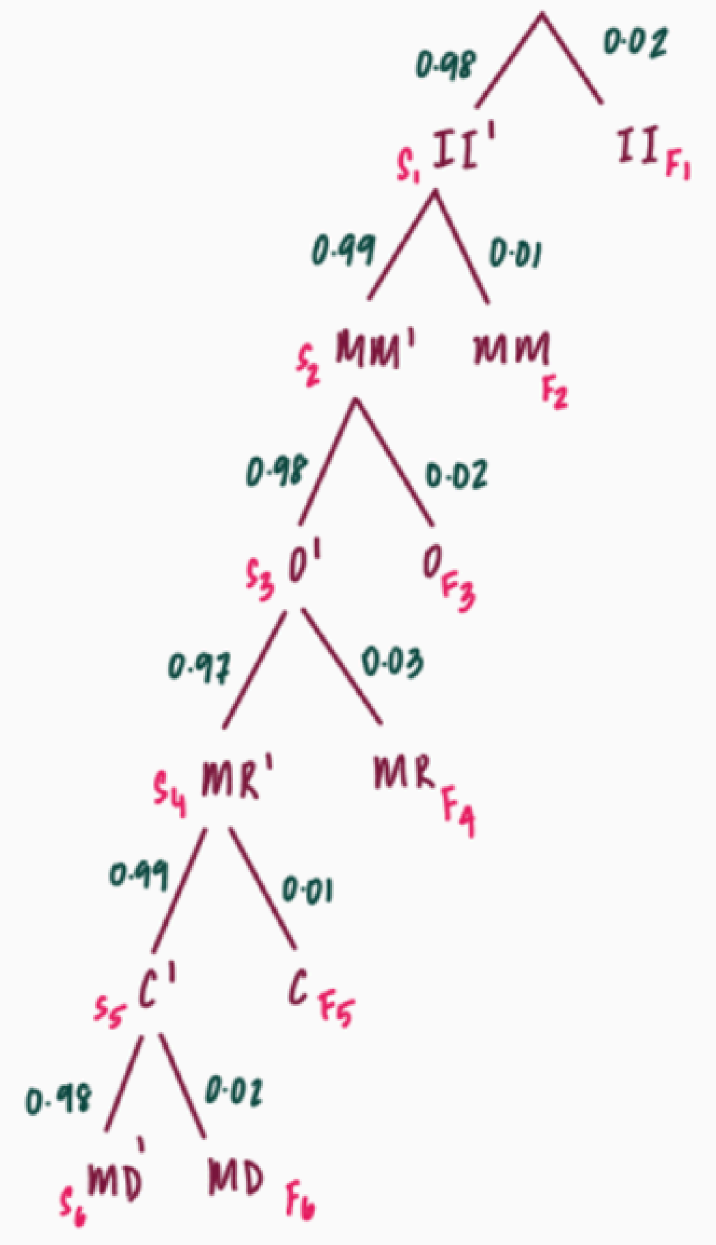
Formula Example:

$$p(M|x,y,x) = \frac{p(M,x,y,z)}{p(x,y,z)} = \frac{p(M) * p(x|M) * p(y|M) * p(z|M)}{p(x) * p(y) * p(z)}$$

RM	IW	MD	BD	DU	O	MO	P	M	M=T	M=F
1	1	0	0	0	0	0	0	1	0.02	0.98
0	1	1	0	0	0	0	0	1	0.01	0.99
1	0	1	1	0	0	0	0	0	0.06	0.94
1	1	1	1	0	0	0	0	1	0.04	0.96
0	0	0	0	1	0	0	0	1	0.02	0.98
1	0	0	0	1	1	1	1	1	0.18	0.82
1	1	1	1	1	1	1	1	1	0.19	0.81
0	0	0	0	1	0	0	1	0	0.05	0.95
0	0	0	0	0	1	0	0	1	0.01	0.99
0	1	0	0	0	1	0	0	1	0.01	0.99
0	0	0	0	0	0	0	0	0	0.005	0.995
1	0	0	1	1	1	1	1	1	0.29	0.71

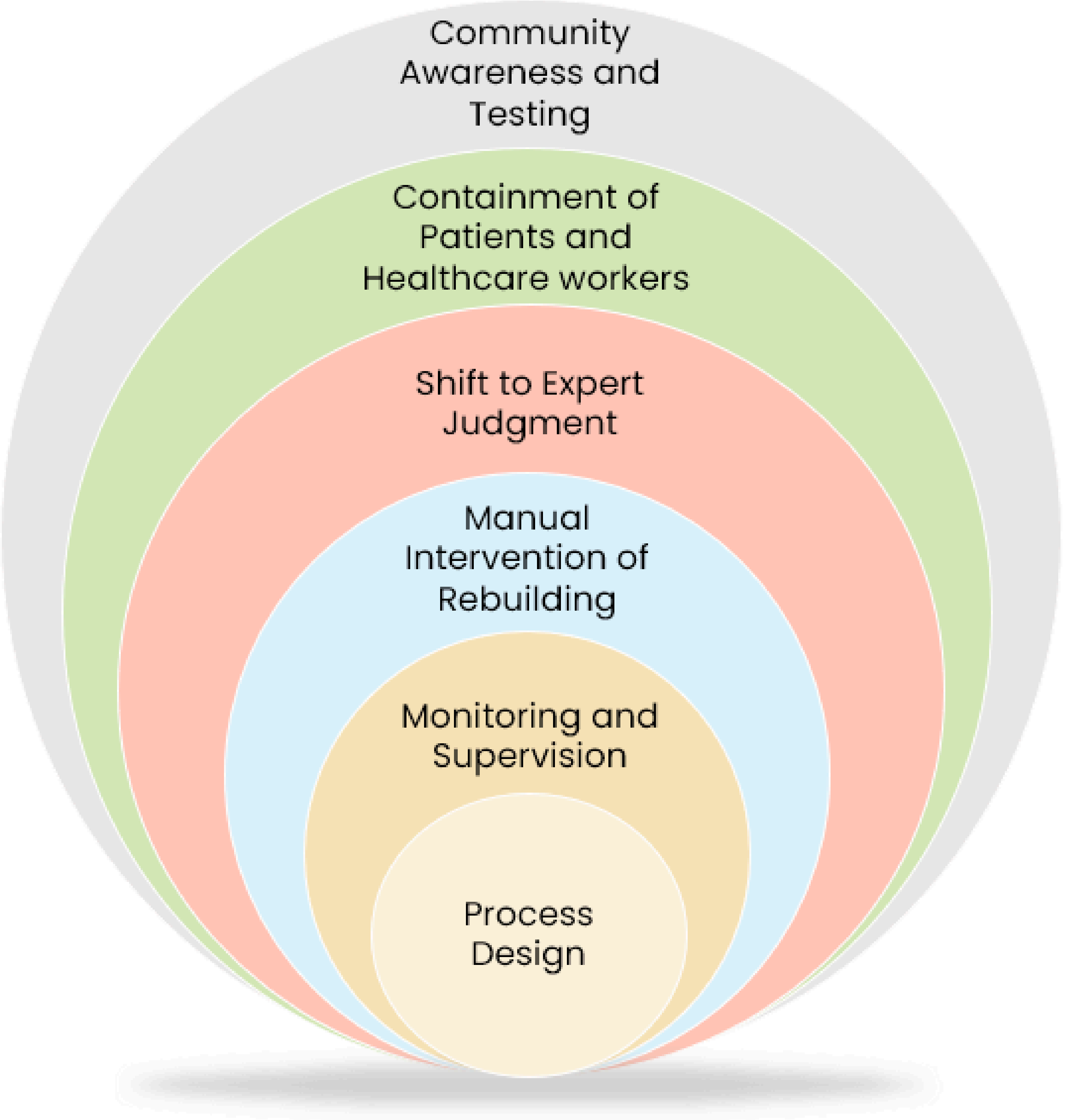
Human Error Analysis

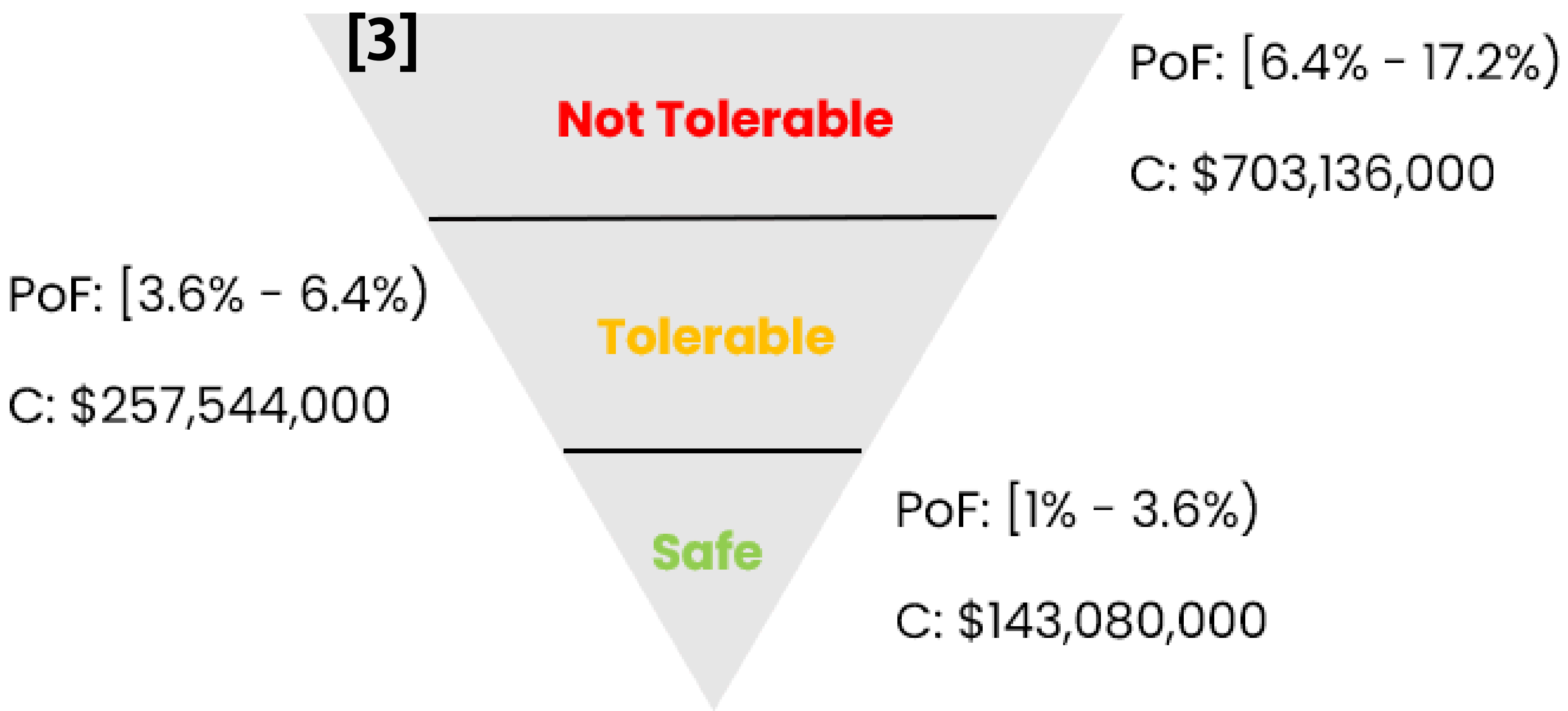




Total Failure= F1+F2+F3+F4+F5+F6=0.106=10.6%
Total Success= 0.98*0.99*0.98*0.97*0.99*0.98=0.894=89.4%

Reduction





Probability of Failure = 10.5%

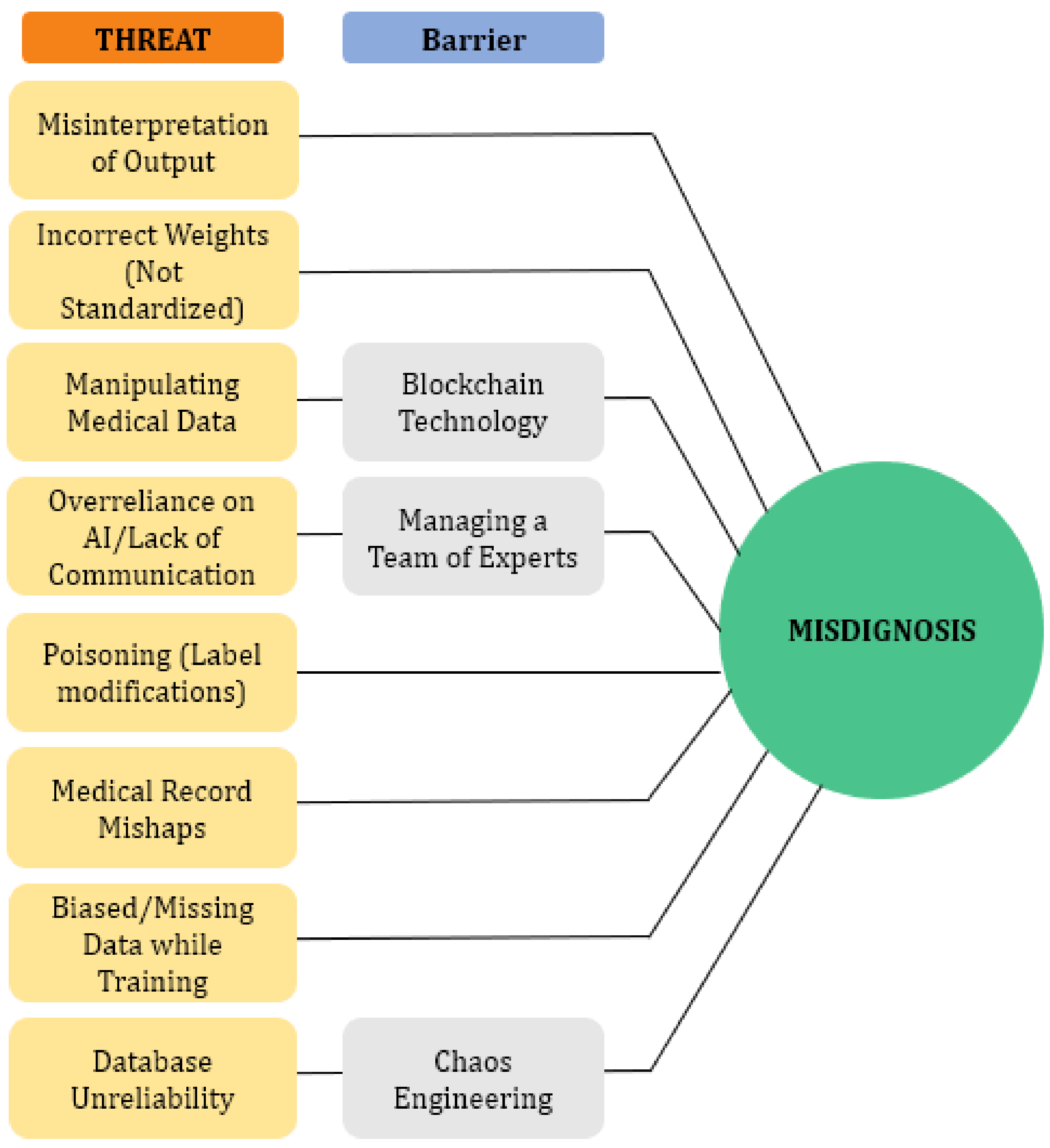
Consequences	**Likelihood**	**Impact**	**LxC**	**Risk Rating**
Misinterpretation of Output	1	4	4	Low
Incorrect weights	3	3	9	Medium
Manipulating Medical Data	3	4	12	High
Overreliance on AI/Lack of Communication	4	3	12	High
Poisoning (Label modifications)	1	4	4	Low
Medical Record Mishaps	2	4	12	Medium
Biased/Missing Data while Training the Model	1	4	4	Low
Database Unreliability	5	4	20	Extreme

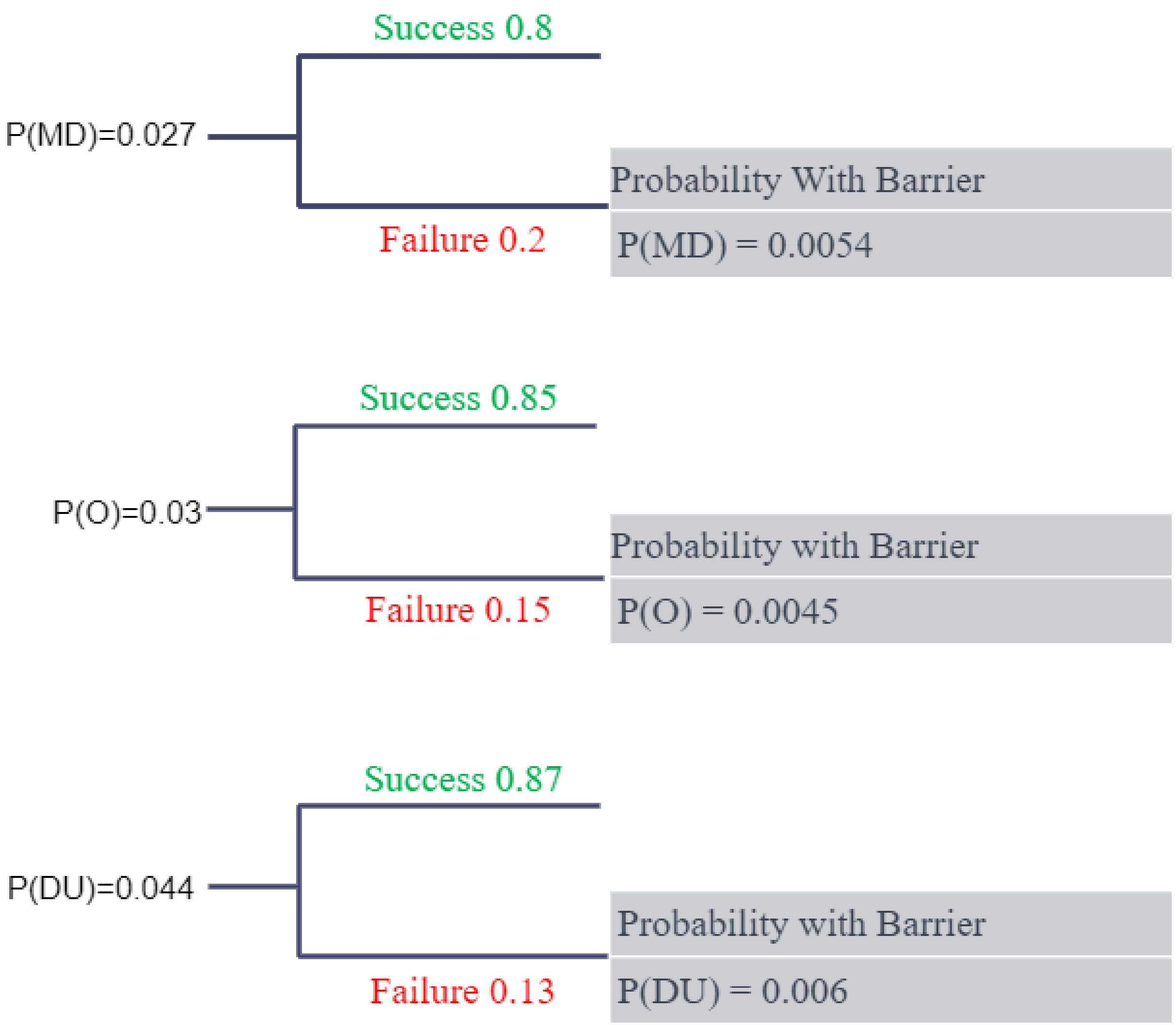
Reduction

Blockchain Technology:
Each transaction is recorded in a block, which is then added to a chain of blocks.

Managing a Team of Experts:
Employ a team of experts on the field to provide a consensus on the diagnosis.

Chaos Engineering:
Simulate database failures using chaos engineering principles.





Before

Probability of Failure = 10.5%
Consequences = $429,240,000

After

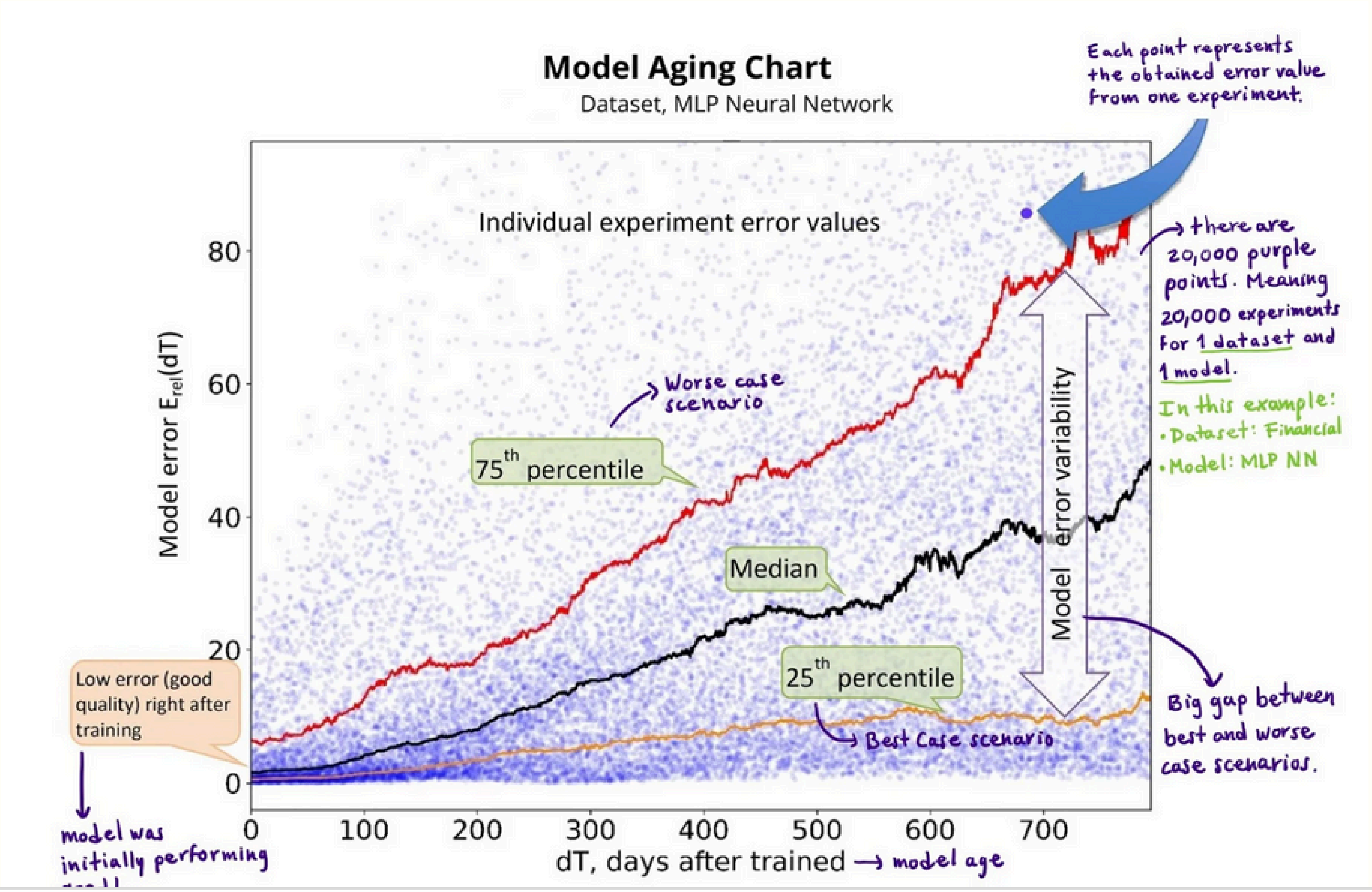
Probability of Failure = 4.8%
Consequences = $196,224,000

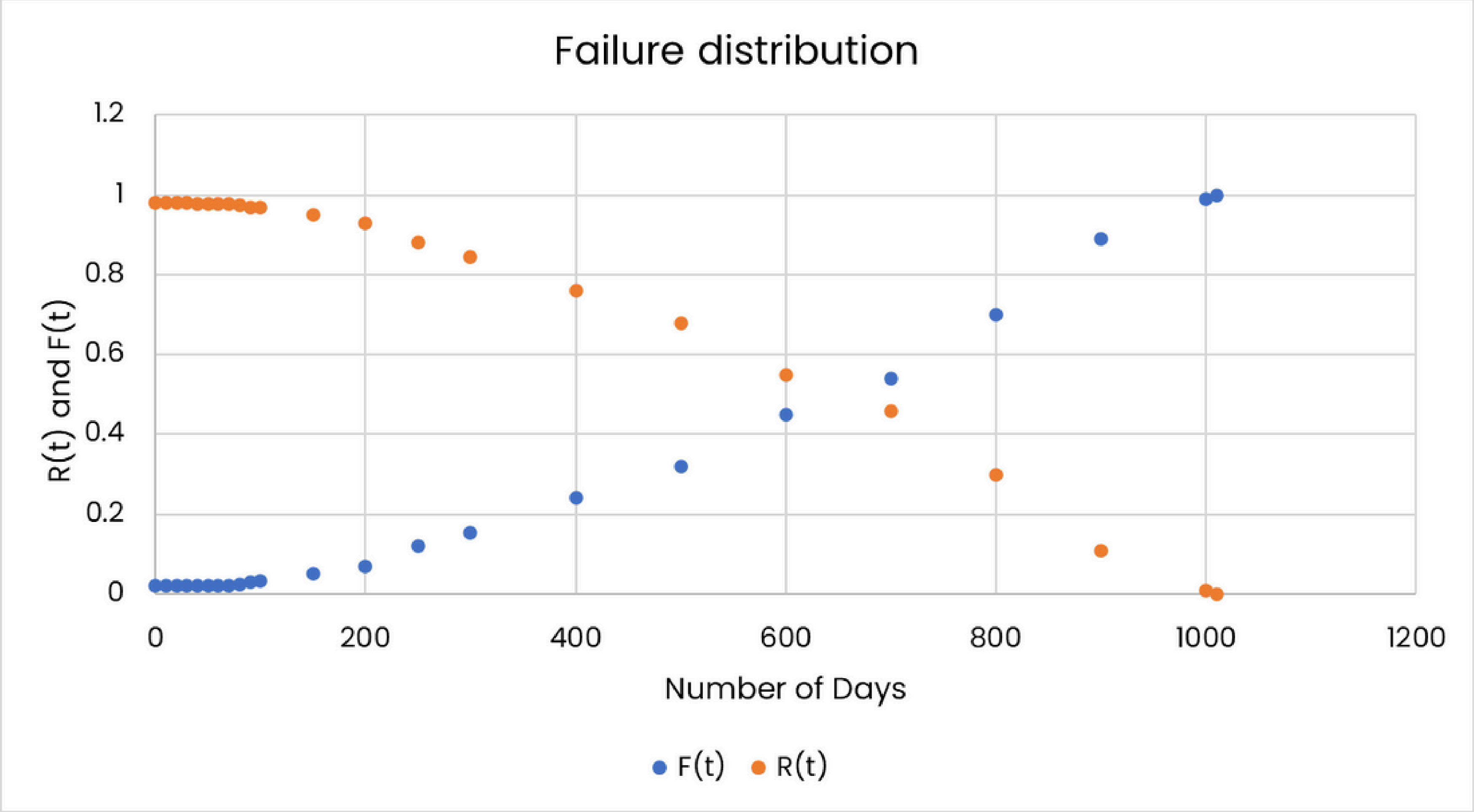
Robustness [4]

Days	# of cases	F(t)	R(t)	Delta	n(t+delta) -n (t)	f(t)	r(t)	h(t)
0	97329.68	0.02	0.98	10	0	0	0	0
10	97329.68	0.02	0.98	10	9.9316	1E-05	1.0204E-05	1E-05
20	97319.7484	0.02	0.98	10	9.9316	1E-05	1.0205E-05	1E-05
30	97309.8168	0.02	0.98	10	69.5212	7E-05	7.1443E-05	7E-05
40	97240.2956	0.021	0.979	10	9.9316	1E-05	1.0213E-05	1E-05
50	97230.364	0.021	0.979	10	49.658	5E-05	5.1073E-05	5E-05
60	97180.706	0.022	0.979	10	49.658	5E-05	5.1099E-05	5E-05
70	97131.048	0.022	0.978	10	198.632	0.0002	0.0002045	2E-04
80	96932.416	0.024	0.976	10	595.896	0.0006	0.00061475	6E-04
90	96336.52	0.03	0.97	10	198.632	0.0002	0.00020619	2E-04
100	96137.888	0.032	0.968	50	1787.688	0.00036	0.0003719	4E-04
150	94350.2	0.05	0.95	50	1986.32	0.0004	0.00042105	4E-04
200	92363.88	0.07	0.93	50	4965.8	0.001	0.00107527	0.001
250	87398.08	0.12	0.88	50	3476.06	0.0007	0.00079545	8E-04
300	83922.02	0.155	0.845	100	8441.86	0.00085	0.00100592	0.001
400	75480.16	0.24	0.76	100	7945.28	0.0008	0.00105263	0.001
500	67534.88	0.32	0.68	100	12911.08	0.0013	0.00191176	0.002
600	54623.8	0.45	0.55	100	8938.44	0.0009	0.00163636	0.002
700	45685.36	0.54	0.46	100	15890.56	0.0016	0.00347826	0.003
800	29794.8	0.7	0.3	100	18870.04	0.0019	0.00633333	0.006
900	10924.76	0.89	0.11	100	9931.6	0.001	0.00909091	0.009
1000	993.16	0.99	0.01	10	993.16	0.001	0.1	0.1

Average r(t)= 0.0058

MTTF = 171.3 Days





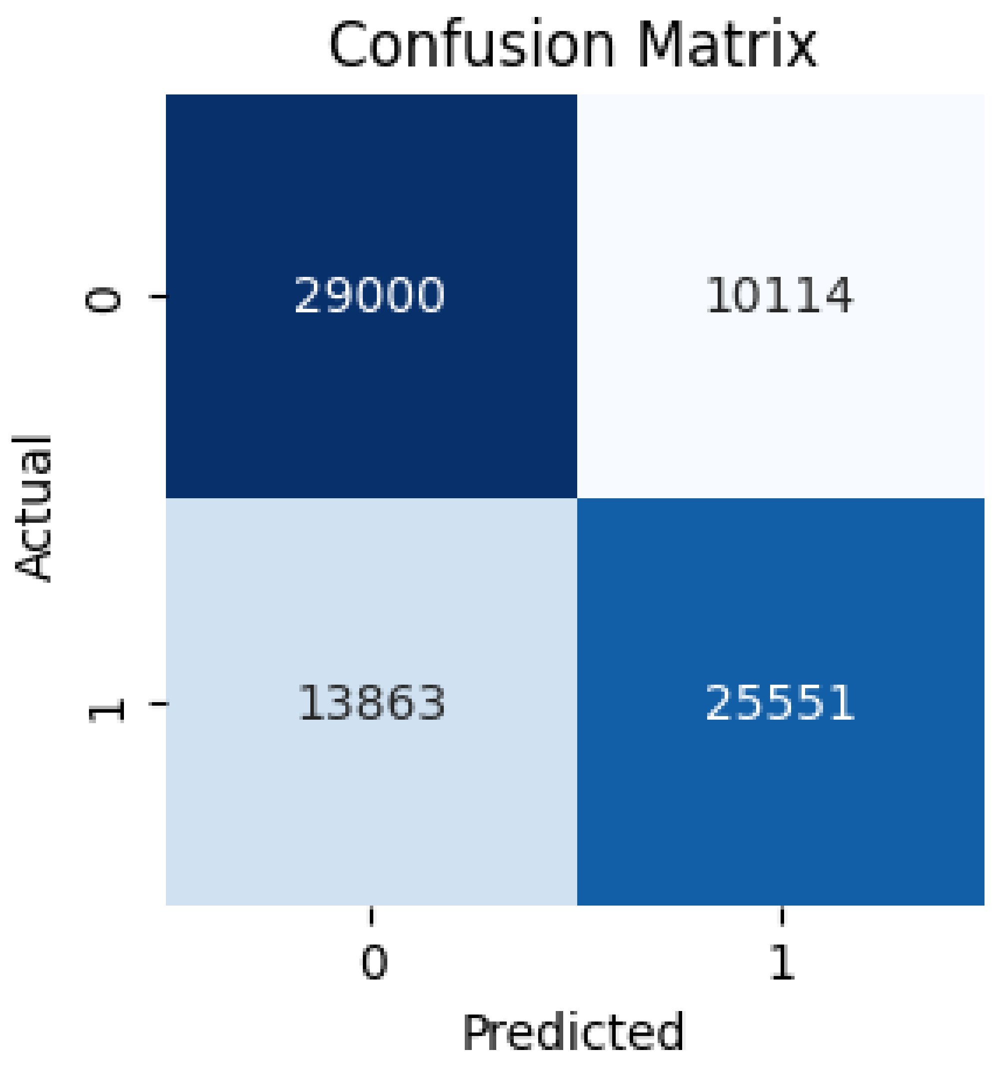
Validation and Verification

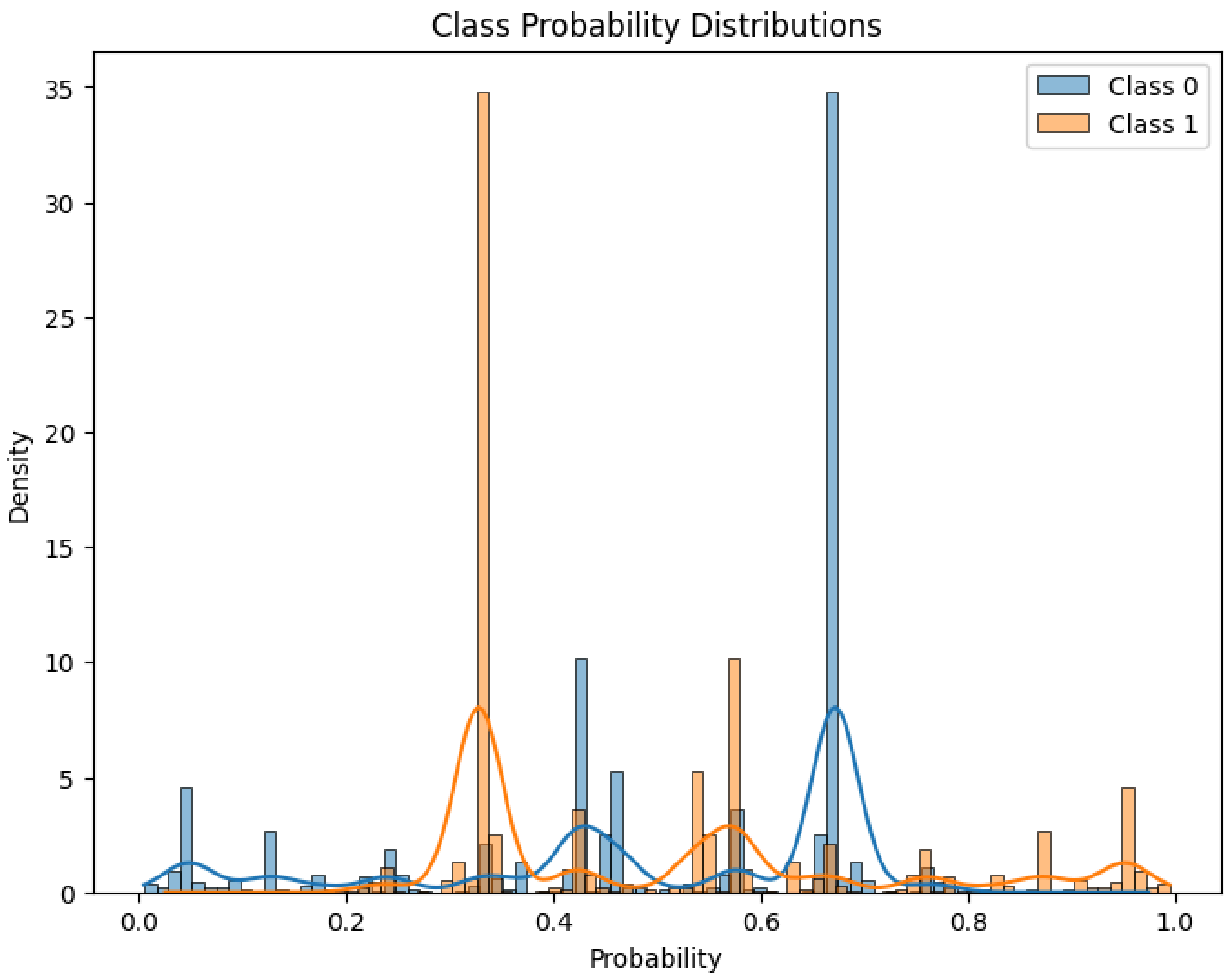
- Precision
- Accuracy
- Confidence Interval
- ROC, AUC

```
Accuracy: 0.69
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.74      0.71     39114
           1       0.72      0.65      0.68     39414

    accuracy                           0.69     78528
   macro avg       0.70      0.69      0.69     78528
weighted avg       0.70      0.69      0.69     78528
```

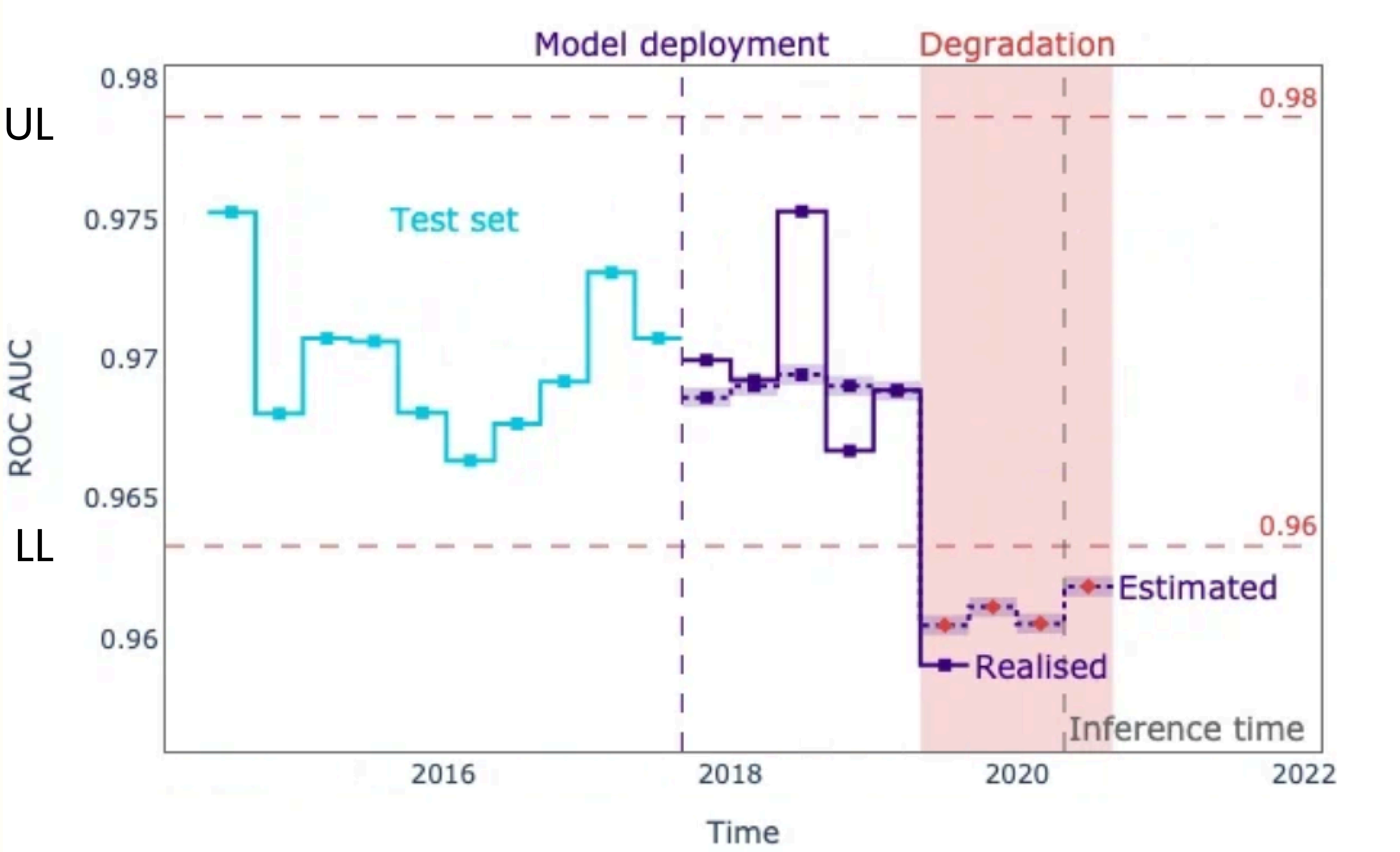





```
Predictions with Confidence Intervals:
Sample 1:
  Predicted Class: 1
  Actual Class: 1
  Confidence Interval: [0.073, 0.927]
---
Sample 2:
  Predicted Class: 1
  Actual Class: 1
  Confidence Interval: [0.251, 0.749]
---
Sample 3:
  Predicted Class: 1
  Actual Class: 1
  Confidence Interval: [0.435, 0.565]
---
Sample 4:
  Predicted Class: 1
  Actual Class: 1
  Confidence Interval: [0.427, 0.573]
---
Sample 5:
  Predicted Class: 1
  Actual Class: 0
  Confidence Interval: [0.337, 0.663]
---
```

Monitoring

NannyML [5] is an open-source Python library that allows you to estimate post-deployment model performance, detect data drift, and intelligently link data drift alerts back to changes in the model performance.



Standards

According to the International Electrotechnical Commission [6,7]

- **TC 62/SC 62A -** Common aspects of medical equipment, software, and systems which include AI.
- **11.040.01 -** Medical equipment in general

According to Health Canada [8] , The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of **Good Machine Learning Practice (GMLP).** The 10 guiding principles identify areas where the International Medical Device Regulators Forum (IMDRF), international standards organizations and other collaborative bodies could work to advance GMLP.
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle

1. **Good Software** Engineering and Security Practices Are Implemented
2. Training **Data Sets** Are Independent of Test Sets
3. Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population
4. Selected Reference Datasets Are Based Upon Best Available Methods
5. Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
6. Focus Is Placed on the **Performance of the Human-AI Team**
7. Testing Demonstrates Device Performance During Clinically Relevant Conditions
8. **Users** Are Provided Clear, Essential Information
9. Deployed Models Are **Monitored for Performance** and Re-training Risks Are Managed
10. Multi-Disciplinary Expertise Is Leveraged Throughout the Total **Product Life Cycle**

References

Consequence Calculations:
[1] Biggest pharmaceutical lawsuits by settlement amount. Pharmaceutical Technology. (2023, August 9). https://www.pharmaceutical-technology.com/features/biggest-pharmaceutical-lawsuits/?cf-view
[2] Janagama, S. R., Strehlow, M., Gimkala, A., Rao, G. V. R., Matheson, L., Mahadevan, S., & Newberry, J. A. (2020, February 27). Critical communication: A cross-sectional study of Signout at the prehospital and hospital interface. Cureus.

Risk Acceptability:
[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8832260/#:~:text=Results,CI%206.4%25%2D17.2%25).

Robustness:
https://www.youtube.com/watch?v=OA949HXzZMI -> Celeste Fralick, Chief Data Scientist at McAfee
[4] https://www.nannyml.com/blog/91-of-ml-perfomance-degrade-in-time

Monitoring:
[5] https://www.nannyml.com/blog/nannyml-quickstart

Standards:
[6] https://www.iec.ch/blog/standards-medical-devices
[7] https://webstore.iec.ch/publication/2612
[8] https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/good-machine-learning-practice-medical-device-development.html

Statistics

According to Robinson (1999), prior studies have shown that in general radiographic examinations, the human error rate falls between 3% and 6%. Surprisingly, during external validation, the same AI algorithm, which didn't meet expectations in our case, displayed a higher error rate of 13%. Additionally, according to O'Mary (2023) "…on average, the researchers estimated that 11% of medical problems result in a misdiagnosis." While the rest of values were generated based on expert judgment. According to World Health Organization, "These occur in 5–20% of physician–patient encounters. According to doctor reviews, harmful diagnostic errors were found in a minimum of 0.7% of adult admissions. Most people will suffer a diagnostic error in their lifetime." According to Tiwary et al, in a 2008 study conducted by Bartlett et al., it was found that issues in communication with patients contribute to a rise in preventable adverse effects, primarily those related to drugs. The study estimated that a significant portion, specifically 27%, of medical malpractice cases can be attributed to failures in communication. Enhanced communication has the potential to mitigate medical errors and reduce patient injuries. On the other hand, inadequate communication can lead to adverse consequences, including diminished adherence to treatment, dissatisfaction among patients, and inefficient utilization of resources. failures.

According to an IBM report, the average cost of a data breach in 2019 was $3.92 million, while a healthcare industry breach typically costs $6.45 million And, inefficiencies in communication contribute to an estimated annual loss of $12 billion in the United States alone according to Janagama et al, in National Center for Biotechnology Information. Bayer and J&J jointly resolved approximately 25,000 claims filed in the US federal and state courts against their anticoagulant drug, Xarelto, in 2019. The patients filed complaints stating that Xarelto's use led to internal bleeding, stroke, and even death. The lawsuit was for $755 million dollars. And by taking information from Figure 4.

Thank you