Ontario Tech University

NUCL 5050 – Applied Risk Analysis

Professor: Dr. Hossam Gaber

Name: Vallika Kasibhatla

ID: 100928820

Risk Project Report: Risk Analysis of AI in the medical field.

Table of Contents

1. Introduction 4
1.1 Problems in relying on AI 5
a. Cyber Attacks: 5
b. Systematic Bias 5
c. Lack of Emotional Intelligence 5
d. Causation vs Correlation 6
1.2 Systems under consideration: 6
1.3 Expected Hazards: 7
1.4 Target Risk Reduction Criteria 7
2. Literature Review 8
3. System Specification 10
Systems: 10
Entire System: 10
Development of an AI Model: 10
Image Classification Method: 11
4. Risk Analysis 12
4.1 Hazard Scenario: 12
4.2 Causes 12
4.3 Propagation and Escalation 13
4.4 Consequences 14
5. Risk estimation 17
5.1 Qualitative Analysis 17
5.2 Quantitative Analysis 18
5.3 Bayesian Belief Network 20
5.4 Human Error Analysis: 23
5.5 Risk Calculations 27
6. Defining Acceptable Risk and Reduction 28
6.1 Defining Acceptable risk. 28
6.2 Layers of Protection 29
6.3 Reduction of Failure probability 30
7. Reliability 32
7.1 Building an AI Model 32

7.2 AI-aging 32
8. Validation and Verification 35
9. Monitoring 37
10. Standards 37
11. Conclusion 38
12. References: 39
13. Appendix: 42

1. Introduction

Artificial Intelligence (AI) is a new field of study where a computer can mimic human behavior and intellect to solve problems like a human. The benefits of AI are limitless. Any machine or software can be trained to have the capacity to produce data from the environment, heuristics, and form decisions. AI's instantaneous, cognitive, adaptive nature makes it powerful in any industry, especially in the medical field. AI applications in medicine can be broadly classified into two categories: virtual and physical (Hamlet, 2017).

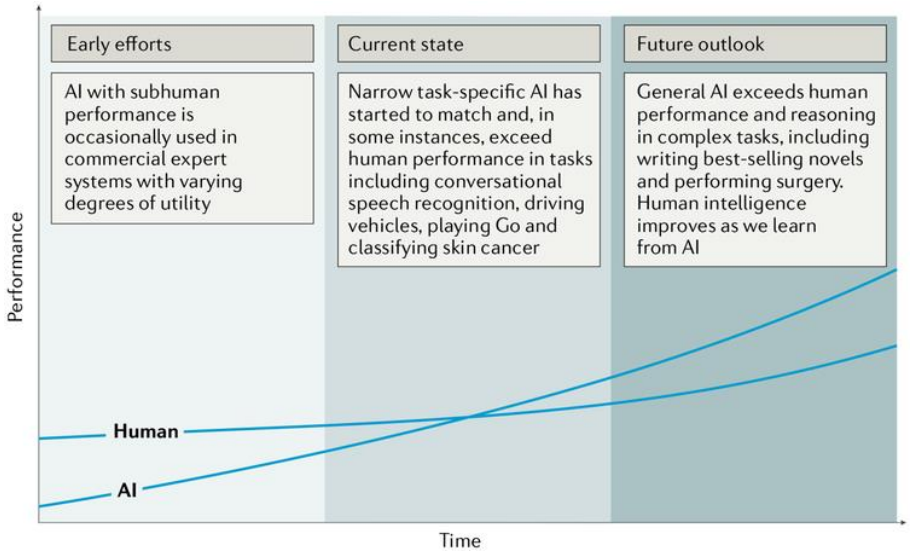


Figure 1: The change of AI and Human intelligence.

This chart shows how artificial intelligence (AI) and human intelligence have changed over the years, starting from early computers and going into the future. At first, AI wasn't as smart as humans, and it had mixed success. Now, we have AI that's good at specific tasks and can sometimes be even better than humans. Predictions say that soon, general AI could be better than humans in certain things. The way humans and AI work together could make humans even smarter—a sort of teamwork where both benefits.

A few common ways AI has been used were in diagnosis, treatment application, imaging, data storage, rehabilitative surgical practices, and virtual health assistants. This technology helps professionals find patterns among extensive data that would otherwise overlook. The following are unique medical inventions:

- Early detection of diabetic retinopathy through images based on deep learning (K., Kang, et al. 2021)
- Screening of chest X-rays using AI for early detection of Tuberculosis
- AI-powered Drug Discovery
- Wearable Devices and Sensors
- Surgical procedures robotics
- Virtual Reality simulations for surgeries and patients with Arteritis, etc.

However, this innovation comes with many speculations. The main reason for this doubt is that AI technologies perform as a black-box neural network (Kiener, 2021). This means that coders know the input and the output but may need to learn precisely how the system finds associations among

data. These internal works cannot be easily verified and visible. This can lead to more uncertainties and a lack of accountability for the customers and the doctors.

1.1 Problems in relying on AI

a. Cyber Attacks:

As any AI technologies are systems that are connected to computers, there will always be a need to be cautious against cyber-attacks. With academic research involvement, cybersecurity, even in medicine, is highly advised.

Issues (Kiener, 2021):

- One cyber-attack is input attacks (Kiener,2021), which means that the data in computer systems can be manipulated. Issues such as changes in dosage, medical histories, patient information, and schedules can be disrupted.
- Additionally, a more critical issue would be that features in medical machines, such as MRIs, can be altered, and a false diagnosis can be shown. These risks are hard to detect as they aren't within the system. Input attacks only need to manipulate the input data, and the AI system will assume that there are no bugs.
- Even robust AI systems are susceptible to hacking, potentially leading to a complete system shutdown and leaving critical medical emergencies unattended.

b. Systematic Bias

Gathering varied and thorough information on patients and test findings is one of the basic elements in making a medical forecast. Data augmentation, picture preparation, cleanup, and handling of missing values come after data preprocessing. The most effective and straightforward model is then found by evaluating the fine-tuning hyperparameters and trained again on the labeled data to create predictive analysis.

Issues:

- The main issue in finding comprehensive data (supervised data) is that it may not encompass every case, and if the system gets a new input, it may not predict it accurately. Also, as the data needs to be labeled, this may require the intervention of doctors and clinicians (unsupervised data); we would face a scenario in which there could be some misdiagnosis or a biased input (human error).
- This is a breach of privacy as the model trains on the patient records, and if there is a cyber-attack, it may not be safe.
- If data is missing commonly, the data is duplicated or estimated, which may reduce the model's accuracy. If the image isn't clear in any image processing method, it won't detect it correctly.
- Additionally, When the model is trained and fine-tuned, some distinctive trends might be suppressed due to their uniqueness (atypical symptoms). The predictions need to be more accurate. This is known as AI bias.

c. Lack of Emotional Intelligence

Sometimes to patients aren't aware of AI's cost, convenience, or informativeness. Instead, the resistance to embracing medical AI is rooted in the notion that AI overlooks individuals' unique characteristics and personal situations (Harvard Business Review, 2019). People see themselves as a unique being, which also extends to their health. So, when AI is applied to these people, they

believe that there is no care given. AI-driven medical care is seen as rigid and standardized, well-suited for addressing the needs of an "average" patient. Psychologically, the lack of emotion from these AI machines creates a notion that patients aren't taken care of.

d. Causation vs Correlation

The current systems aren't equipped to differentiate between causations and simple correlations, which raises worries that some treatment may not be customized to the needs of each patient individually.

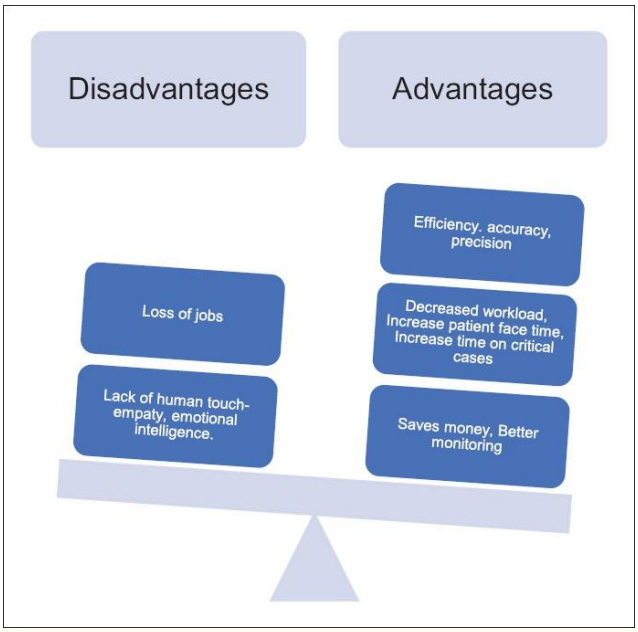


Figure 2: Advantages and disadvantages of artificial intelligence in medicine

1.2 Systems under consideration:

- Firstly, the collection of data is essential to develop models and facilitate predictions. This encompasses many documents, such as electronic health records, patient information, and online datasets.
- Additionally, some machines may possess pre-installed machine-learning models that aid in diagnosing medical conditions. These AI models play a significant role in providing valuable insights and make certain processes easier. Such as scanning ID cards for auto filling.
- Moreover, powerful computers systems are crucial for optimal performance. This includes high-performing computing resources like digital screens, GPUs (Graphics Processing Units), and imaging tools. Given the high demand for these systems, speed, accuracy and reliability is the bare minimum features to look for.
- Clinical experts are needed for training and evaluating any AI system. Their involvement is necessary for tasks such as data labeling, evaluations, and examinations so that the AI systems are trained properly for all medical cases.
- Promoting patient comfort with AI-driven tools is important to enhance accuracy and identify areas where improvements are possible. Patient engagement can help research to move forward by knowing that there are customers.

- Lastly, an essential aspect of effectively implementing AI into medical practice lies in training healthcare professionals and data scientists on methodologies related to artificial intelligence ethics. It is necessary to equip them with strategies to mitigate the potential consequences of utilizing this technology.

1.3 Expected Hazards:

- Problems with Data Accuracy: Using biased or missing data might result in incorrect diagnosis and treatment suggestions, which can be fatal to the patient.
- Errors in Algorithmic Procedures: Machine learning models may provide false-positive or false-negative results due to lack of evaluations and insufficient data. If a wrong prediction is made it can fatal.
- Security Vulnerabilities: Through any cyberattacks or inner attacks, it can lead to revealing sensitive data, security breaches, hampering of data.
- Difficulties in Medical Professionals and AI Interaction: Humans who largely rely on artificial intelligence, and not evaluate AI, can lead to miscommunications and serious mistakes being made during treatment operations.
- Ethical Considerations: The choices made by AI systems may lead to moral problems such instances of biased decision-making. The patients always need to be involved in any medical decision.

1.4 Target Risk Reduction Criteria

- There needs to be an eye on privacy rules, preventing data leaks, and any security breaches. A specific emphasis should be placed on digital hygiene (Argaw, 2019)—good digital security practices including selecting stringent privacy settings and robust password protection (Argaw, 2019).
- Accuracy and Precision should be tested. Any model used should have a high accuracy rate, especially in the medical field. Various evaluation matrices can be derived from the values found in a confusion matrix. These are used to measure if the model is reliable. They can be summarized as follows:

	Predicted class		
Actual class		Class 0	Class 1
	Class 0	*True Positive*	*False Negative*
	Class 1	*False Positive*	*True Negative*

Table 1: Confusion Matrix

Accuracy: Measures how well a model correctly classified a certain class.

$$Accuracy = \frac{TN+TP}{TN+FP+TP+FN}$$

Precision: Measures the proportion of true positive predictions out of all the positive predictions made by the model.

$$Precision = \frac{TP}{FP+TP}$$

Recall: Measures the proportion of true positive predictions out of all the actual positive instances in the dataset.

$$Recall = \frac{TP}{TP+FN}$$

F1-score: Measures how well a model can predict both positive and negative classes as a function of both precision and recall.

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall}$$

2. Literature Review

2.1 Statistics on Misdiagnosis:

a. Radiology Misdiagnosis

In the medical field, one of the main areas where artificial intelligence is used is to detect and diagnose accurately and effectively. The few advantages of AI for radiology are that:

- It is faster to detect abnormalities through common patterns and a trained model. It can also help to form early predictions of diseases.
- The inputs can be varied. The AI model can be trained on various data, such as images, databases, and doctors' information.
- It can provide quantitative information and an overall data analysis about the patient for a holistic report.

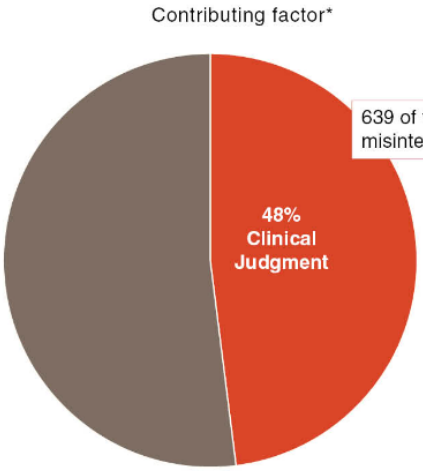
According to Siegal et al. (2017), a Meta study of 29,777 cases asserted between 2010 and 2014 were conducted to understand better the problems patients face in the healthcare delivery system. The interpretation of the studies indicated the following:

- Out of 1325 cases, under the analysis of Radiology, there have been 766 cases that were incorrectly diagnosed.

Radiology-related Allegations	# Cases	Total incurred
1. Diagnosis-related	766	$202,714,000
Cognitive/clinical judgment (interpretation)		
Communication (to and from, providers & patients)		
2. Medical treatment	287	$40,880,000
Improper performance of tx/procedure		
Improper management of treatment course		
Retained foreign body		
3. Equipment-related	90	$3,609,000
Improper inspection/maintenance		
Equipment malfunction/failure (inc. user error)		
4. Safety & Security	75	$2,844,000
Fail to ensure safety, falls		
Fail to ensure safety, other injury during care		

Figure 4: Number of cases for few Radiology related allegations.

- Even the slightest mistake in healthcare can have profound and far-reaching consequences for patients. This mistake resulted in 48% of the patients (639) missing their breast and lung cancer detection, among other types.
- In Figure 2, the breakdown of 48% of the cases illustrates the portion of individuals who could not complete their radiology procedures. This resulted in a $263M total incurred losses.



Misinterpretation of diagnostic studies (639 cases)	
Top procedures	% Cases
CT scan (abd, head, chest)	18%
Diagnostic radiology (e.g., CXR)	13%
Mammography	10%
MRI	10%
Diagnostic ultrasound	3%

Figure 5: Percentage of cases under few procedures

- Additionally (Siegal et al., 2017), apart from just the misdiagnosis, there were issues with communication between the medical professionals and the patients. There were around 23% of radiology cases where there were communication failures.

Also, failure to properly document any findings can lead to delayed or unsuccessful follow-up and result in unfavorable results.

b. Covid Misdiagnosis

- In a study published in "Nature Machine Intelligence," a group of University of Washington researchers examined various models for detecting COVID-19 using chest X-rays. The study found that these models didn't truly understand medical factors but took shortcuts, creating unreliable links between non-medical factors and the disease. Interestingly, the models often didn't consider important medical signs and instead used dataset-specific features, like text markers or patient positioning, to predict COVID-19 (Healthcare-in-Europe, 2021). This study highlights the difficulties and issues when using AI models in medical diagnosis.
- Shortcut learning is less reliable compared to a genuine understanding of medical pathology, and typically results in the model's inability to perform well in different settings beyond its original environment. DeGrave explained (Healthcare-in-Europe, 2021), "A model that depends on shortcuts often operates effectively only within the specific hospital where it was created. When you introduce the system to a new hospital, it becomes ineffective. This ineffectiveness can potentially mislead doctors by suggesting an incorrect diagnosis and inappropriate treatment."

3. System Specification

Systems:

Entire System:

The following block diagram illustrates the architecture of a typical system that uses images for classification. The focus of the system is the flow of data from the Database to the Deep Learning Models and vice versa. The input to this system would be an image (Based on the radiology type.) This image can be taken from machines such as MRI's, CT's, Radio graphics etc. The input will be fed through a web application or any interface. This image would also be input for the image classification model, which is retrieved from the server. Once the analysis is done, the results will be stored in the database where the model relearns. Finally, this result, which could be a simple string will be accessed by the doctor for further actions. The main systems are the server, database, interface, and the ML model.

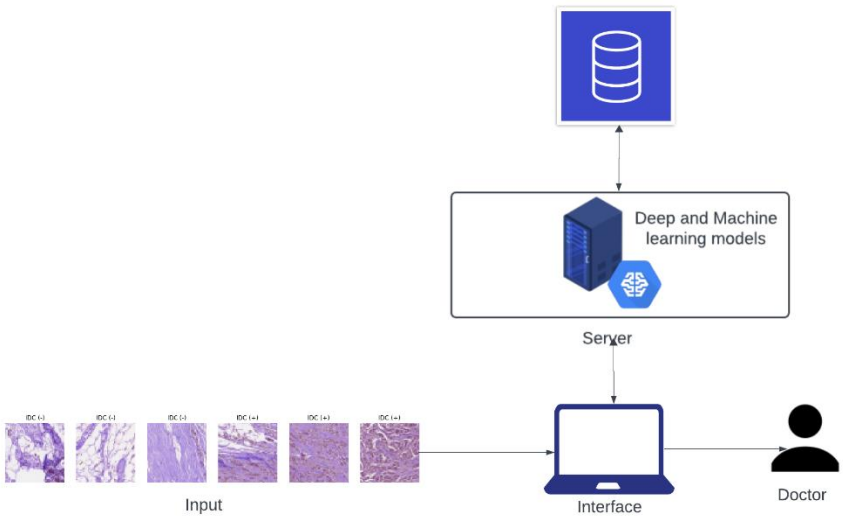


Figure 6: Entire system block diagram

Development of an AI Model:

a) Training:

Data Collection: To train a model, there is a need for relevant data along with accurate labels. Suppose we take the example of collecting image data that the AI model uses to make predictions. This data can come from various sources, including cameras, medical imaging devices, and databases. The labels are provided, maybe through expert judgment or historical medical records.

Preprocess Data: Processing the input is crucial to creating an accurate model. For feature extraction, specific steps need to be taken, such as image resizing and data augmentation of the information. In cases where missing data is identified, it is either eliminated from the dataset if possible or if removal is not feasible, synthetic data is generated as a replacement.

Feature Extraction: This vector would be created from the data collected. Depending on the AI model, there may be a feature extraction step. For example, in deep learning models, convolutional layers can automatically extract relevant features from images, such as tumor size or decolorization. This vector format will be stored in the database so that the model can be trained, and for future input, it will find the pattern based on the features extracted.
Feature Matrix: A matrix will be created by adding the labels and the feature extracted vector. This forms a new database that will be analyzed.
Machine Learning Algorithm: Image classification allows a model to predict the class of an object or scene in an image. This system is a classification problem as it has classes as the label. For image processing, CNN (Convolutional Neural Network) would be used.
Model: Once the Algorithm uses the Feature Matrix, a model will be saved in the database and used for any future input.

b) Predicting:
The main steps for predicting an image include feature extraction using the predictive model, which results in an output in the form of a string. Input is in photos, first analyzed by the preprocessing and feature extraction model (retrieved from the database). This will result in a vector that can be inputted into the model (retrieved from the database). As this is a classification problem, the output may be in terms of a string that gives a diagnosis.

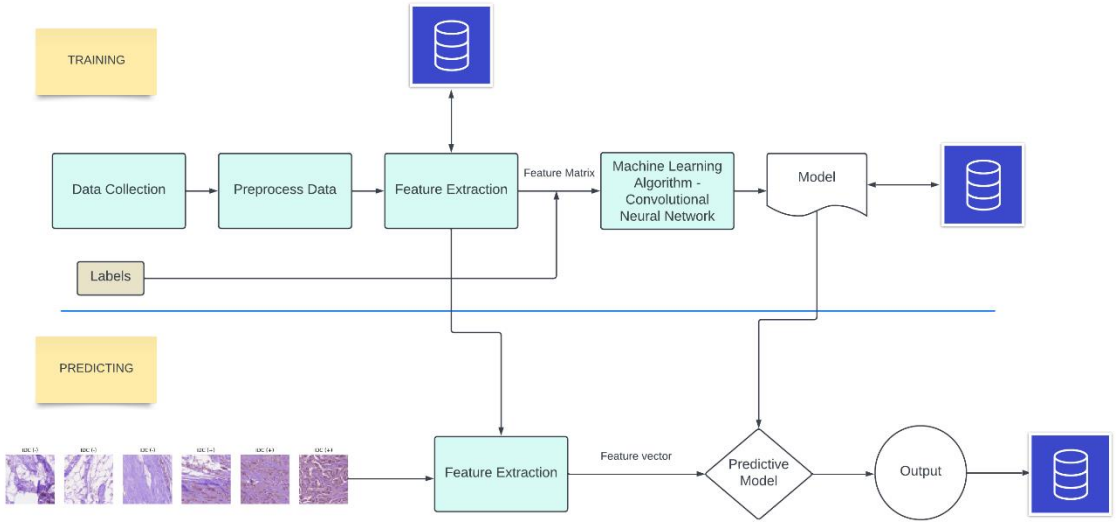


Figure 7: Flowchart for AI Model

Image Classification Method:
The high-level step involves filtering out specific image files and extracting features from the remaining images. The feature extraction process results in a condensed set of features that capture the visual content of an image.

The selected features, in this case, represent the dominant characteristics of the image. The reasonable number of features selected was 20, which effectively encapsulates the essential visual traits of the image while maintaining a manageable feature dimensionality. Ultimately, the function computes the average of these features over the spatial dimensions, resulting in a single feature vector. This feature vector can then serve as an input to a classifier.

The third step involves training the dataset using these image features and corresponding labels. These features and labels are organized and converted into a format suitable for training a neural network classifier, in this case, a Convolutional Neural Network (CNN). CNNs are chosen for their ability to achieve high classification accuracy in image-related tasks.

The CNN architecture consists of three main components: convolutional layers, pooling layers, and fully connected (FC) layers. The model operates on one-dimensional image data. After each convolutional layer, an activation function is applied to introduce non-linearity and dropout regularization is used to mitigate overfitting. Subsequently, a max pooling layer is applied to downsample the output from the initial convolutional layers. Finally, a dense layer processes the input data and categorizes it into one of three classes.

The dense layer's output is passed through a softmax activation function to determine the probability distribution of the classes. In this context, the model receives image data with a specific dimension and produces a probability distribution for the classification of the diagnosis.

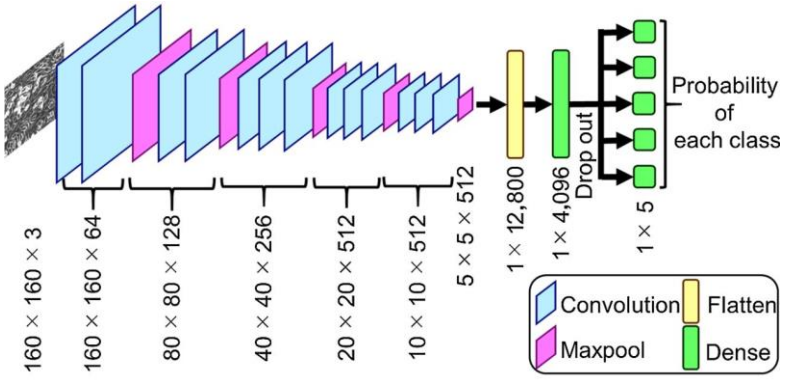


Figure 8: CNN model

4. Risk Analysis

4.1 Hazard Scenario:

A potential hazard scenario involves the risk of misdiagnosis deriving from either human error or the inadequacy of the employed model in analyzing exceptional cases through predictions.

4.2 Causes (According to ENISA, 2023):

a. Model Failure (MF)

1. Input Data

- Manipulating Medical Data: If the system gets hacked and patients' data is deleted or altered even in the slightest way, it can lead to incorrect inputs for an AI model and may result in a diagnosis that can be dreadful.
- Unfair Processing: Using biased or missing data might result in incorrect diagnosis and treatment suggestions, which can be fatal to the patient. We would face a scenario where there could be misdiagnosis or biased input (human error).

- Medical Record Mishaps: Doctors may need to properly store patient records, if not, potentially leading to issues in patient care and data security. Errors in record-keeping can lead to miscommunications, treatment delays, and compromise patients' safety.

2. Incorrect Weights: Normalization typically involves scaling the features of the input data so that they have similar ranges. Normalization in the context of AI models often refers to a preprocessing technique applied to input data to ensure that it has a consistent and standardized format.

b. External Failure

1. Overreliance on AI/Lack of Communication: Excessive reliance on AI systems can lead to diagnostic errors as the system may only consider some of the necessary inputs. So, there needs to be effective human-AI collaboration and communication in decision-making.
2. Poisoning (Label modifications): During the data collection and cleaning, the doctors could annotate the data and force certain variables in the model to result in a targeted prediction. Unauthorized label modifications on medications or substances pose a severe risk of poisoning.
3. Databases unreliability: Online databases, while often convenient and accessible, come with specific reliability challenges. The inconsistency in data quality is erratic, as the information within these databases can vary widely in accuracy, completeness, and relevancy.
4. Misinterpretation of Output: Misunderstanding or incorrect interpretation of the results or predictions provided by an artificial intelligence system designed to assist in medical diagnosis, treatment planning, or decision-making.

4.3 Propagation and Escalation

Causes	Escalation	Propagation
Misinterpretation of Diagnosis	If there are no skilled workers and the information isn't validated, then the problem would escalate.	There could be a community outbreak.
Incorrect Weights	If there is some issue with the dataset, and the information wasn't corrected it can be an issue.	There could be misdiagnosis and issues with the overreliance on AI.
Manipulating Medical Data	As the manipulation of medical data becomes more subtle and sophisticated, the risk of causing severe harm through altered patient records or inaccurate diagnostic information escalates.	Manipulated data can propagate through interconnected healthcare systems, affecting multiple healthcare providers and potentially leading to widespread patient safety issues.
Overreliance on AI/Lack of Communication	Blindly relying on AI without human oversight and not fostering effective communication channels can escalate the risk of overlooking critical errors or malicious manipulations in AI-generated outputs.	If overreliance on AI becomes a pervasive trend without effective communication about its limitations, the lack of critical oversight can propagate across healthcare institutions.

Poisoning (Label modifications)	Intentional poisoning of medical data labels can escalate as attackers devise more sophisticated methods to infiltrate and modify datasets without detection.	Poisoned data can propagate through the training sets of various AI models, impacting predictions across different medical applications and potentially compromising patient care.
Medical Record Mishaps	Mishandling of medical records can escalate if there are inadequacies in secure storage, retrieval, and sharing mechanisms, leading to more frequent and severe incidents.	Mishandled medical records can propagate through electronic health record systems, affecting patient care in different healthcare settings and potentially leading to widespread breaches of privacy and trust.
Biased/Missing Data while Training the Model	Introducing biased or missing data during model training can escalate as more AI systems are developed without sufficient attention to the representativeness of the training data.	Biased models can propagate disparities in healthcare outcomes, affecting diverse patient populations and potentially leading to systemic issues if not addressed in the development and deployment of AI systems.
Database Unreliability	The unreliability of databases can escalate if there's a lack of robust data governance, leading to more frequent data corruption, loss, or unauthorized access.	Unreliable databases can propagate errors across different healthcare applications and systems, affecting various aspects of patient care and public health.

Table 2: Escalations and Propagation for causes.

4.4 Consequences (According to ENISA, 2023)

Once these underlying causes happen and initiate the main event, that is the misdiagnosis, they can result in a range of outcomes, for instance:

1. Physical and Emotional Suffering: Patients who receive a misdiagnosis may suffer needless physical and emotional anguish and distress.
2. Lawsuits: Misdiagnosis often leads to lawsuits when patients or their families believe they have suffered harm or injury due to the misdiagnosis.
3. Reputation degradation: Can occur when instances of misdiagnosis become public, damaging the reputation of medical facilities and professionals.
4. Loss of trust: It follows misdiagnosis, as patients may lose confidence in the hospitals and their doctors.
5. Delayed Treatment: Misdiagnosis can lead to physical harm, potentially causing adverse health effects or complications. This can result in a worsening of the patient's health due to inappropriate care.
6. Increased Healthcare cost: Misdiagnosis can result in elevated healthcare costs, given the need to investigate why the diagnosis was wrong and to mitigate it in the future.

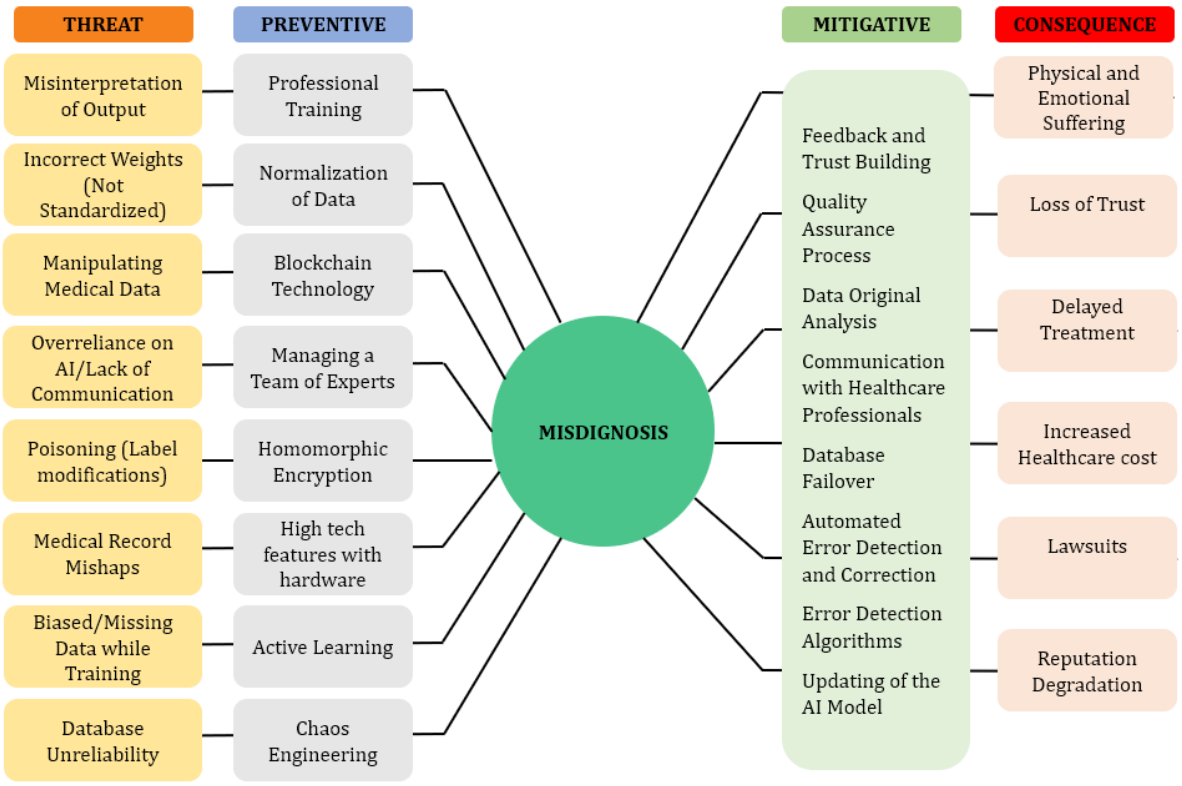


Figure 10: Bow Tie Diagram of causes, prevention, top event, mitigation and consequences.

Causes	Prevention		Recovery	Consequences
Misinterpretation of Output	Professional Training: • How to properly interact with the AI system and interpret its output. • To have an understanding on how the output should be and accordingly to make sure if the model is in the right path. • Provide comprehensive training for healthcare professionals on how to interpret and use AI-generated outputs effectively.		Feedback and Trust Building: • Have a meeting for decisioning on how and why this case has occurred. Gather insights from the professionals on how to prevent this misjudgment. • Rebuilding trust in AI applications in healthcare	Physical and Emotional Suffering Loss of Trust Delayed Treatment Increased Healthcare cost
Incorrect Weights (Not Normalized)	Normalization of Data: • Normalize data to a common scale to avoid biasing the model towards variables with higher magnitudes. This is		Quality Assurance Processes Enhancement: • Strengthen quality assurance processes for model development to	Physical and Emotional Suffering Delayed Treatment

	particularly important for algorithms sensitive to the scale of input features.	**HAZARD SCENARIO**	prevent similar incidents in the future. • Implement additional checks and validations related to variable weighting, feature scaling, and other critical aspects of model training.	
Manipulating Medical Data	Blockchain Technology: • Each transaction is recorded in a block, which is then added to a chain of blocks. • It provides a secure and transparent way to store and share data. In a blockchain-based decentralized ledger, data transactions are recorded in a distributed database that is maintained by a network of computers.		Data Original Analysis: • Leverage data provenance analysis tools to trace and identify manipulated data. • Retrieving previous versions of the data when misdiagnosis is detected.	Physical and Emotional Suffering Lawsuits Reputation Degradation
Overreliance on AI/Lack of Communication	Managing a Team of Experts: • Employ a team of experts on the field to provide a consensus of the diagnosis. • Find a confidence interval that would find doctors know when there is a need of a human intervention to verify the AI output.		Communication with Healthcare Professionals • Engage with healthcare professionals who use or are affected by the AI model. Provide clear communication about the incident, the corrective measures taken, and the steps being implemented to prevent similar incidents in the future.	Loss of Trust Reputation Degradation Physical and Emotional Suffering
Poisoning (Label modifications)	Homomorphic Encryption: • Apply homomorphic encryption to protect sensitive information during model training. • Robust testing using fake inputs and confirming results with experts.		Database Failover: • Retrain Model on new data and switch approaches to find other ways to avoid reliance on the AI model. • Retrieving previous versions of the data when misdiagnosis is detected.	Loss of Trust Delayed Treatment Increased Healthcare cost
Medical Record Mishaps	High tech features with hardware: • Real-time data entry through hardware		Automated Error Detection and Correction:	Physical and Emotional Suffering

	appliances such as heart monitors and dosages. • Training of clinical professionals • Open communication with patients to way the risks and surety		• Use automated systems to detect and correct errors in real-time. • Establish a process for manual verification and correction if needed	Delayed Treatment Increased Healthcare cost
Biased/Missing Data while Training the Model	Active Learning: • Involve Expert Judgement on certain cases to fill Missing data. • Create an Active learning system for the AI model to detect biases.		Error Detection Algorithms: • Retraining strategies to address biases in the deployed model.	Increased Healthcare cost Reputation Degradation
Database Unreliability	Chaos Engineering: • Simulate database failures using chaos engineering principles. • Implement automated failover mechanisms to ensure continuous service.		Updating of the AI Model: • Implement a feedback loop to update the model based on misdiagnosis. • Retrain the model to override certain recommendations	Delayed Treatment Increased Healthcare cost Lawsuits

Table 3: The prevention and recovery of the causes

5. Risk estimation

5.1 Qualitative Analysis

Risk Matrix:

They are mainly used to determine the size of a risk and whether or not the risk is sufficiently controlled. The combination of probability and severity will give any event a place on a risk matrix (Dean, 2023).

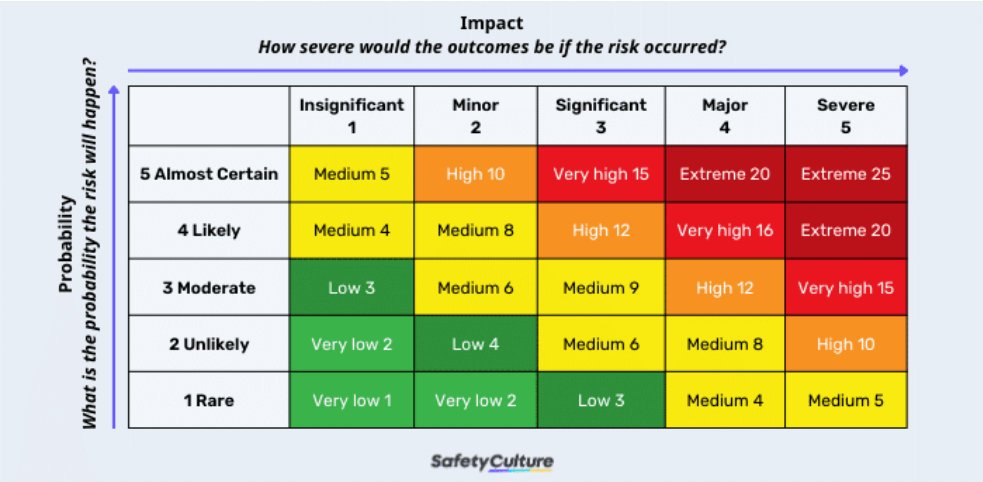


Figure 11: Risk Heat Matrix (Safety Culture, 2023)

Consequences	Likelihood	Impact on Healthcare	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 4: Risk Matrix Analysis

5.2 Quantitative Analysis

Fault Tree Analysis

Fault tree analysis (Rockwell Automation, 2022), sometimes called event tree analysis, is a method used to identify the potential reasons for a system failure. It uses a visual fault tree diagram to show the different possible causes of a failure. FTA helps you understand what factors lead to an event, often called a failure, and assess the probability of it happening (Rockwell Automation, 2022).

Top events (TEs): These events are found at the top of the fault tree, signaling the start of the investigation into the system's failure. They have one input and don't have any related outputs because they mark the beginning of the failure analysis.

Intermediate events (IE): These events typically result from one or more prior events and have both inputs and outputs. They often lead to additional explanations.

Basic events (BE): These events are usually the causes of the top event. They are at the lowest level.

OR gate: This type of gate may have one or more inputs, and an output event will occur if one or more of the input events happen.

AND gate: This type of gate may have one or more inputs, and an output event will occur if all the inputs happen.

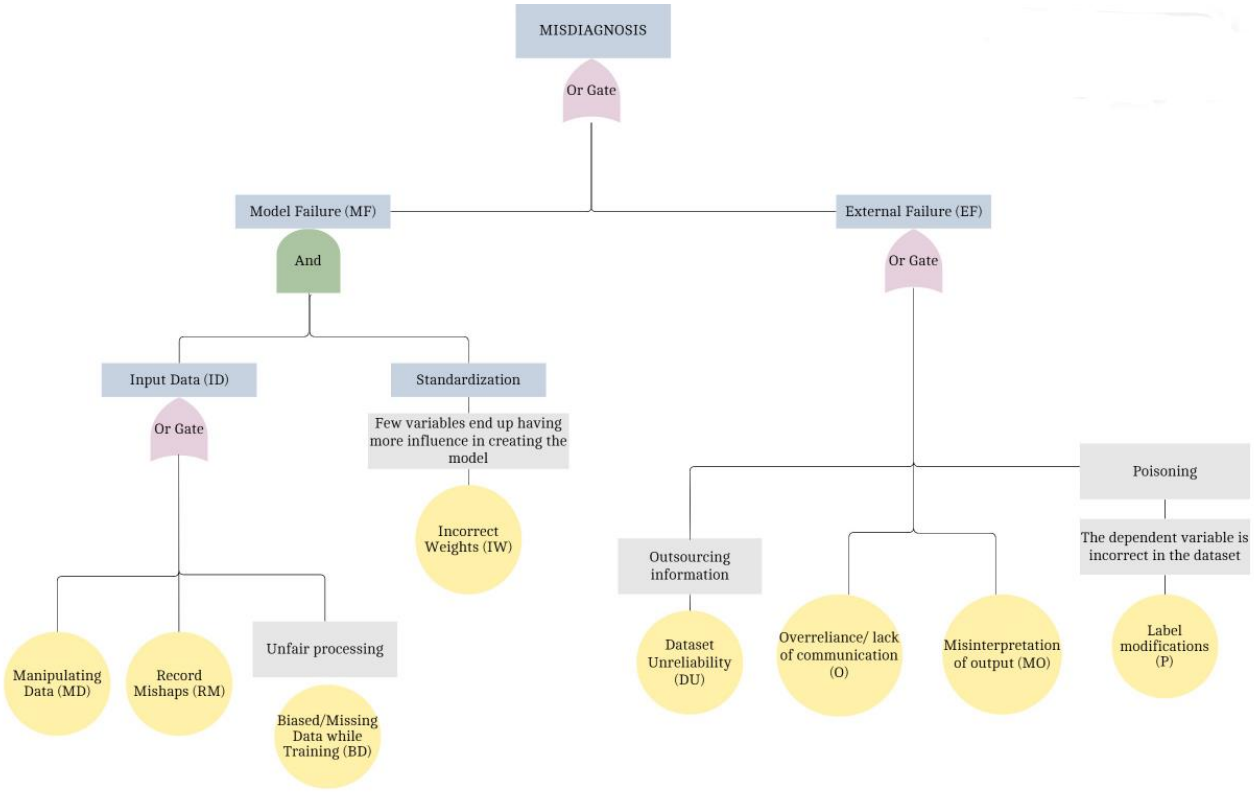


Figure 12: Fault Tree

Minimal Cut Set

Cut sets (Weibull, 1992) can also be used to discover single point failures (one independent element of a system which causes an immediate hazard to occur and/or causes the whole system to fail.)

Values:

Causes	**Probability**
Misdiagnosis [M]	= 0.11
Model Failure (MF) • Incorrect weight/Having wrong weightage [IW] • Input Data (ID) - Record Mishaps [RM] - Manipulating Medical Data [MD] - Biased/Missing Data while Training the Model [BD]	 = 0.026 = 0.023 = 0.027 = 0.016
External Failure (EF) • Datasets unreliability [DU] • Overreliance on AI/Lack of Communication [O] • Misinterpretation of Output [MO] • Poisoning (Label modifications) [P]	 = 0.044 = 0.03 = 0.017 = 0.01

Mocus Method (According to figure 4):

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

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

Formula:

$$P_0(t) = 1 - \prod_{j=1}^{k}[1 - \breve{P}_j(t)]$$

$$P(Input\ Data) = P(RM) + P(MD) + P(BD) = 0.066$$
$$P(Model\ Failure) = P(Input\ Data) * P(IW) = 0.066 * 0.026 = 0.0017$$
$$P(External\ Failure) = P(DU) + P(O) + P(MO) + P(P) = 0.103$$

$$So, P(Misdignosis) = 1 - [(1 - (0.066 * 0.026) * (1 - 0.103)] = \mathbf{0.105}$$

5.3 Bayesian Belief Network

Probabilistic models help create diagrams for variables and find probabilities about each event with their connections. The **Bayesian Belief Network** represents relations among variables in a graphical manner. The main assumption is that the variables are conditionally independent, meaning no dependency exists.

Mitchell (Machine Learning, Page 184, 1997) states, "*... it provides an intermediate approach that is less constraining than the global assumption of conditional independence made by the naive*

Bayes classifier, but more tractable than avoiding conditional independence assumptions altogether."

The Bayesian belief network consists of two main components (Jason Brownlee, 2019):

- **Nodes:** these are random variables or events. Each node corresponds to a specific variable or event you want to model.
- **Edges:** These represent probabilistic dependencies or causal relationships between nodes. An edge between two nodes implies that one node influences the other. The direction of the arrow on the edge represents the direction of the influence, and it's often interpreted as a causal relationship. If there is no edge, it has a conditional independence between the nodes.

Formula: Suppose that B is an event and A1…. A partition. Then:

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{\sum_{j=1}^{n} P(B|A_i) * P(A_i)} - - - -(1)$$

Conditional probability tables (CPTs) are used in Bayesian belief networks to quantify the interactions between nodes. CPTs provide information about the conditional probabilities of each node in respect to its parent nodes. It shows how each variable in the network depends on or is impacted. This is in the form of a truth table.

The following is the general form of Bayes' theorem:

$$\boldsymbol{P\big(\,Xi \mid Parents(Xi)\,\big) = \frac{P\big(\,Xi \mid Parents(Xi)\,\big) \cdot P\big(Parents(Xi)\big)}{P(Xi)}} - - - -(2)$$

Assumption: The parent's nodes are independent of each other. Two variables are independent if:

$$\boldsymbol{\forall xy = P(x, y) = P(x)P(y)} - - - -(3)$$

Equation 3 states that their joint distribution factors into a product of two simpler distributions.

A **joint probability distribution** is a statistical distribution that lists all potential outcomes of several random variables along with their respective probabilities. It offers a thorough overview of the probability of different combinations of these variables' values happening at the same time.

Assumption: There is conditional independence.

The joint probability distribution formula for multiple factors is:

$$\boldsymbol{P(x_1, x_2 \ldots x_n) = \prod_{i=1}^{n} P\big(x_i|parent(x_i)\big)} - - - -(4)$$

Probabilities for Causes: This is a converging connection where CC, HE, ML are the parents of M.

The following probabilities have been derived from historical data, expert opinions, or simply estimations.

M	
P(M)	0.011
P(~M)	0.989
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

M	
P(M)	0.011
P(~M)	0.989
IW	
P(IW|M)	0.015
P(~IW|M)	0.985
MD	
P(MD|M)	0.03
P(~MD|M)	0.97
O	
P(O|M)	0.029
P(~O|M)	0.971
P	
P(P|M)	0.03
P(~P|M)	0.97
RM	
P(RM|M)	0.08
P(~RM|M)	0.92
BD	
P(BD|M)	0.025
P(~BD|M)	0.975
DU	
P(DU|M)	0.09
P(~DU|M)	0.91
MO	
P(MO|M)	0.021
P(~MO|M)	0.979

Table 5: Likelihood of the causes and top event occurring given that event M has occurred.

How did we get these values?

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 the 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.

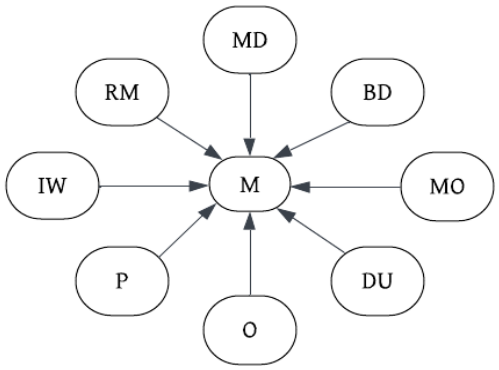


Figure 13: Bayesian Diagram

Conditional probability table for M
The conditional probabilities show the likelihood of different combinations of values for these variables.

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

Table 6: Conditional probabilities of M given the causes for few cases.

5.4 Human Error Analysis:
Human Reliability Analysis (HRA), also known as human reliability assessment, is a systematic approach employed to evaluate the influence of human factors on risk. The primary objectives of HRA are realized through its essential functions, which encompass the identification of potential errors (human error identification), the quantification of the likelihood of these errors (error

quantification), and ultimately, the identification of strategies to diminish the likelihood and consequences of errors (error reduction) (Kirwan, 2017).

Human Reliability (Rausand 2011): The probability that a person

- Correctly performs some system-required activity in a required time period (if time is a limiting factor) and
- Perform no extraneous activity that can degrade the system.

The main steps of a typical quantitative HRA are:

Task Analysis method: Identify critical operations where human errors could lead to accidents and operational problems. Human error identification method: Identify potential human error modes and error causes and performance influencing factors. Human error quantification method: Determine the human error probabilities for each error mode and for the complete task.

THERP

According to Sutton Technical Books, A method for assessing human reliability is an extension of probabilistic risk assessment (PRA), which considers that humans can make mistakes and errors like equipment failure. The THERP (Technique for Human Error Rate Prediction) is a technique developed in the 1950s to predict human error rates. According to Sutton, 2022, The Technique for Human Error Rate Prediction (THERP) is a systematic method used to estimate the probability of human errors in complex systems. There are three key stages: data input stage, AI processing stage for prediction, and after prediction diagnosis.

Before the incident: Data Input Stage	1. Incorrect Input to the model (II) This error involves inaccuracies or mistakes in the input data provided to the AI model. It could be from inaccurate patient information entry or issues with data integrity while building the AI model. The medical professionals are responsible for inputting the necessary details into the model. 2. Medical Record Mishaps (MM) This error refers to mistakes in the medical records, which could impact the quality and accuracy of the data input into the AI model. If the system gets hacked and patients' data is deleted or altered even in the slightest way, it can lead to incorrect inputs for an AI model and may result in a diagnosis that can be dreadful.	Probabilities: 1. 0.02 - Probability of incorrectly entering patient information as an input in the model. 2. 0.01 - Probability of incorrect information in medical records
During the incident: AI processing Stage for prediction	1. Overreliance on AI (O) Overreliance occurs when healthcare professionals trust the AI model's predictions without critically evaluating or validating the results against their clinical assessment. Medical professionals should always use their judgment, experience, and expertise with AI technology to provide the best patient care.	Probabilities: 1. 0.02 - Probability of overreliance on AI output. 2. 0.03 - Probability of misinterpreting AI-generated results.

	2. Misinterpretation of the result (MR) Healthcare professionals need to understand or interpret the predictions generated by the AI model, leading to potential diagnostic errors. Misinterpretation of the result could occur due to a lack of understanding of the AI model's limitations or incorrect interpretation of the results.	
After the incident: After Prediction Diagnosis	1. Lack of Communication (C) Lack of communication is an error that may occur after the prediction has been made by the AI model. It means that there is a lack of communication between the healthcare professionals and the patients. 2. Mistake in diagnosis (MD) Mistakes in diagnosis are potential errors that may occur after the prediction has been made by the AI model. Mistakes in diagnosis occur when the healthcare professionals misinterpret the results or rely too heavily on AI technology. Therefore, it is essential to validate the results with clinical assessments and seek a second opinion, if necessary.	Probabilities: 1. 0.01 - Probability of inadequate communication. 2. 0.02 - Probability of not considering AI output in the diagnosis and creating a mistake while diagnosis,

Probability:

For the overall probability of the failure:

$$F1 = 0.02$$

$$F2 = 0.98 * 0.01 = 0.001$$

$$F3 = 0.98 * 0.99 * 0.02 = 0.019$$

$$F4 = 0.98 * 0.99 * 0.98 * 0.03 = 0.029$$

$$F5 = 0.98 * 0.99 * 0.98 * 0.97 * 0.01 = 0.019$$

$$F6 = 0.98 * 0.99 * 0.98 * 0.97 * 0.99 * 0{,}02 = 0.018$$

$$Total\ Failure = F1 + F2 + F3 + F4 + F5 + F6 = 0.106 = 10.6\%$$

For the overall Success:

$$S = S1 * S2 * S3 * S4 * S5 * S6$$

$$Total\ Success = 0.98 * 0.99 * 0.98 * 0.97 * 0.99 * 0.98 = 0.894 = 89.4\%$$

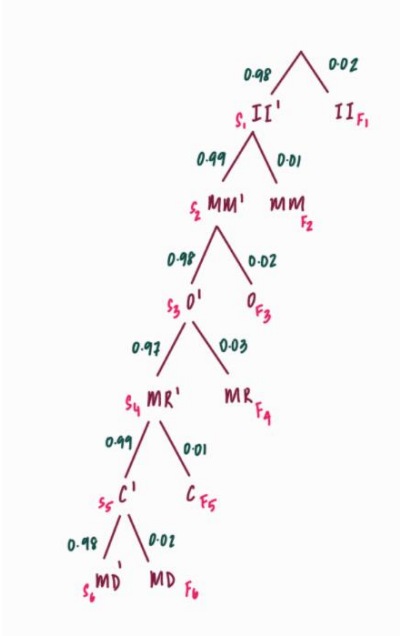


Figure 14: Human error event tree in the medical field using AI for diagnosis

Causes and Mitigation Strategies

		Causes	Mitigation
1. Data Input Stage	Incorrect Input to the Model	Human factors such as data entry errors, lack of standardized data entry procedures, or insufficient data validation checks.	Implement rigorous data validation checks to identify and correct input errors. Provide comprehensive training to healthcare professionals on accurate data entry procedures.
	Medical Record Mishaps	Inadequate updating of medical records, system glitches, or issues in the electronic health record (EHR) management.	Establish regular procedures for updating medical records promptly. Implement robust EHR management systems with built-in checks for accuracy and completeness.

2. AI Processing Stage for Prediction	Overreliance on AI	Lack of awareness of AI limitations, insufficient training on interpreting AI outputs, or an overly optimistic perception of AI capabilities.	Provide comprehensive training on understanding the strengths and limitations of the AI model. Encourage a collaborative approach, emphasizing that AI is a tool to aid clinical decision-making rather than a substitute for it.
	Misinterpretation of the Result	Complex AI outputs, lack of clarity in presentation, or inadequate training on interpreting AI-generated results.	Enhance the interpretability of AI results through clear visualization and explanations. Provide ongoing education to healthcare professionals on interpreting and validating AI outputs.
3. After Prediction Diagnosis	Lack of Communication	Ineffective communication protocols, lack of standardized reporting procedures, or breakdowns in information sharing.	Establish clear communication channels and protocols for disseminating AI-generated results. Encourage interdisciplinary communication and collaboration among healthcare professionals.
	Mistake in Diagnosis	Failure to consider AI-generated results, miscommunication, or errors in the diagnostic process.	Integrate AI-generated predictions into existing diagnostic workflows. Emphasize the importance of considering AI results as complementary information

5.5 Risk Calculations

The formula "Risk = Likelihood * Consequence" is a simplified representation of the risk assessment process.

1. **Risk:**
   In the context of this formula, "risk" refers to the uncertainty associated with the occurrence of an undesired event.
2. **Likelihood:**
   Likelihood, also known as probability, is the measure of how probable it is that a particular risky event will occur. It is often expressed as a percentage or a fraction.
3. **Consequence:**
   Consequence refers to the impact or severity of the risk event should it occur. It is an assessment of the extent of harm or damage that could result from realizing the risk.

Consequences cost:

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

How did I get the values?
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.

			With 10.5% Failure rate
No.	**Consequences**		**Cost in Dollars**
Risk #1	Physical and Emotional Suffering =	40,880,000	4292400
Risk #2	Loss of Trust =	2,844,000	298620
Risk #3	Delayed Treatment =	775,000,000	81375000
Risk #4	Reputation degradation =	1,435,000	150675
Risk #5	Lawsuits =	3,609,000	378945

6. Defining Acceptable Risk and Reduction

According to the failure rate through the fault tree analysis, which was 10.5%, there should be metrics to determine if the percentage is acceptable in the medical field. Various causes trigger the top event, which is misdiagnosis. There should be an analysis of minimizing this risk, as the consequences are life-threatening and severe.

6.1 Defining Acceptable risk.

It is essential to determine if the risk is acceptable to ensure that the system's preventive measures have been exhausted to get the minimal consequence. Risk and cost/benefit analysis are important tools in informing the public about the actual risk and cost as opposed to the perceived risk and cost involved in an activity (Manuele, 2010).

According to the Health and Safety Executive, the complete form of "ALARP" is "As low as reasonably practicable." It weighs the risk with cost, time, and effort to sense reasonableness and determine if corrective measures are essential. It is a figure that shows different regions of chance. This unacceptable region needs immediate action. The tolerable region would need some remedial action to be taken at an appropriate time. It doesn't have high priority but needs to be addressed, and to find a way to lower this risk. The low region shows that the system is in a safe region, and this level is maintained. According to a meta-study, an unacceptable region/ not tolerable is 6.4% above, while a Probability of failure between 3.6% and 6.4% would be considered a Tolerable risk. The safe region has a probability of less than

3.6%. According to the failure rate from the risk assessment, it is 10.5%, then it falls in the Not Tolerable region. This has a high urgency to minimize the probability of failure below 6.4%.

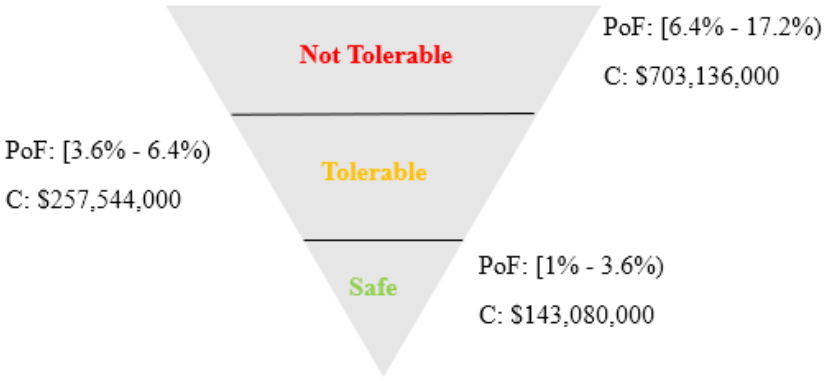


Figure 15: Acceptability Matrix

6.2 Layers of Protection

To minimize risks, we follow the onion model. The core involves training an AI model, followed by continuous monitoring and supervision. This is necessary as medical professionals would need to know when and how the model is degrading. If the detection of degradation is detected, then they can rebuild the model or find an alternative. The third layer is the manual intervention, where there could be a need to rebuild the model. If the reconstruction is unsuccessful, the model is discarded, and human expert judgment can be used for diagnosis. In the event of a misdiagnosis, the patient should be informed. Additionally, the medical professions should be included for further testing. If there were many cases of misdiagnosis, information should be shared with the community to increase community awareness.

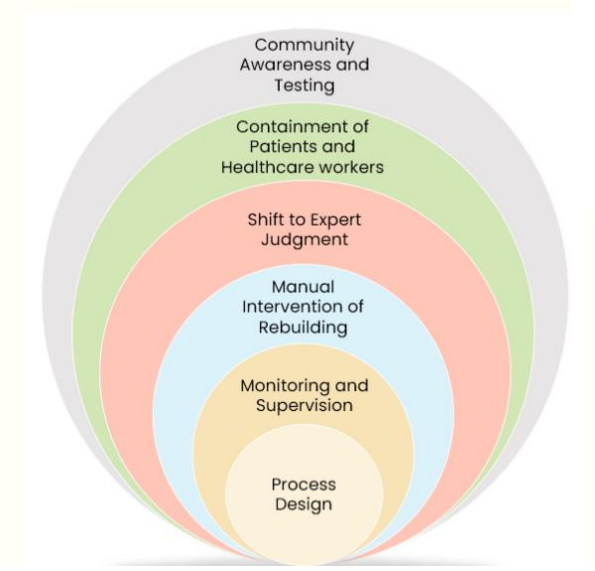


Figure 16: Onion Model

6.3 Reduction of Failure probability

The reduction is focused on the process design element, to lower the chances of failure, there are three areas where improvement is possible. These areas involve addressing the causes of manipulating medical data, overreliance on AI, and dealing with database unreliability. These specific issues were identified as having higher risks through qualitative analysis. The calculations were revisited, considering the preventive measures. These preventive measures are:
Blockchain Technology:

- Each transaction is recorded in a block, which is then added to a chain of blocks. It provides a secure and transparent way to store and share data. In a blockchain-based decentralized ledger, data transactions are recorded in a distributed database that is maintained by a network of computers.

Managing a Team of Experts:

- Employ a team of experts on the field to provide a consensus of the diagnosis. Find a confidence interval that would find doctors know when there is a need for a human intervention to verify the AI output.

Chaos Engineering:

- Simulate database failures using chaos engineering principles. Implement automated failover mechanisms to ensure continuous service.

The following figure shows the calculation that shows the reduction of the failure rate of the causes.

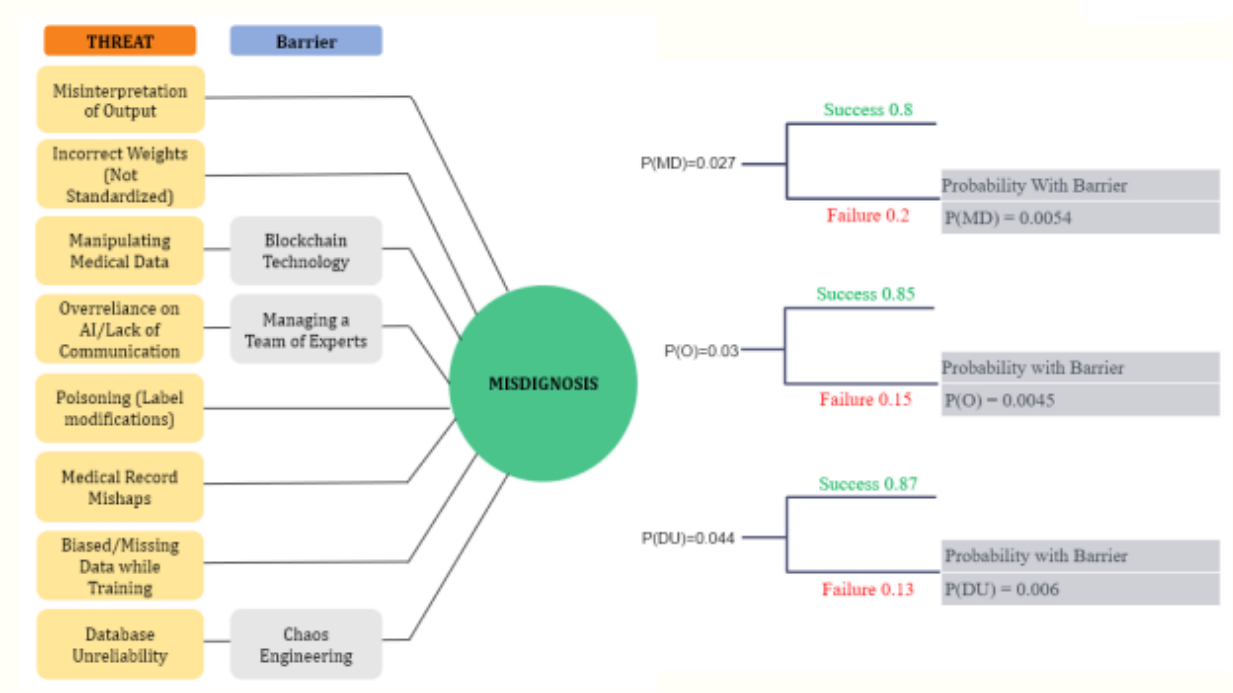


Figure 17: Reduction with Barriers

Probability Calculation:

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

Causes	Probability
Model Failure (MF) • Incorrect weight/Having wrong weightage [IW] • Input Data (ID) - Record Mishaps [RM] - Manipulating Medical Data [MD] - Biased/Missing Data while Training the Model [BD]	 = 0.026 = 0.023 = 0.0054 = 0.016
External Failure (EF) • Datasets unreliability [DU] • Overreliance on AI/Lack of Communication [O] • Misinterpretation of Output [MO] • Poisoning (Label modifications) [P]	 = 0.006 = 0.0045 = 0.017 = 0.01

$$P(Input\ Data) = P(RM) + P(MD) + P(BD) = 0.044$$
$$P(Model\ Failure) = P(Input\ Data) * P(IW) = 0.044 * 0.026 = 0.0012$$
$$P(External\ Failure) = P(DU) + P(O) + P(MO) + P(P) = 0.047$$

$$So, P(Misdignosis) = 1 - [(1 - (0.044 * 0.026) * (1 - 0.047)] = \mathbf{0.048} = 4.8\%$$

According to the acceptability risk matrix, the probability of failure reduced from 10.5% to 4.8%. This percentage is in the tolerable range.

Cost for Solutions:

Cost of Blockchain Implementation (approx.) = $50,000

Cost of Chaos Engineering (approx.) = $500,000

Risk Calculations:

The following are the risk calculations using the consequences. The costs were able to be reduced.

			With 4.8% Failure rate
No.	**Consequences**		**Cost in Dollars**
Risk #1	Physical and Emotional Suffering =	40,880,000	1962240
Risk #2	Loss of Trust =	2,844,000	136512
Risk #3	Delayed Treatment =	775,000,000	37200000
Risk #4	Reputation degradation =	1,435,000	68880
Risk #5	Lawsuits =	3,609,000	173232

So, it takes less amount of money to install these solutions but saves more in the consequences.

7. Reliability

7.1 Building an AI Model

To further understand how to evaluate an AI model. I had built a classification model using a dataset that has information and images of chest x-rays for detecting pneumonia. The code will be attached in the Appendix. For Reliability calculations and Validation/Verification the following dataset will be used.

	Image	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Emphysema	Fibrosis	Hernia	Infiltration	Mass	Nodule	PatientId
0	00000001_000.png	0	1	0	0	0	0	0	0	0	0	0	1
1	00000001_001.png	0	1	0	0	0	1	0	0	0	0	0	1
2	00000001_002.png	0	1	0	0	1	0	0	0	0	0	0	1
3	00000002_000.png	0	0	0	0	0	0	0	0	0	0	0	2
4	00000004_000.png	0	0	0	0	0	0	0	0	0	1	1	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...
99311	00030801_001.png	0	0	0	0	0	0	0	0	0	1	0	30801
99312	00030802_000.png	0	0	0	0	0	0	0	0	0	0	0	30802
99313	00030803_000.png	0	0	0	0	0	0	0	0	0	0	0	30803
99314	00030804_000.png	0	0	0	0	0	0	0	0	0	0	0	30804
99315	00030805_000.png	0	0	0	0	0	0	0	0	0	0	0	30805

99316 rows × 16 columns

Figure 18: Dataset from Kaggle.

7.2 AI-aging

They are concerned about the system. Machine Learning Models are trained on a database. This means the model may no longer perform well if the data changes or is outdated. So, as time passes, the AI model could be degraded. This is known as AI aging. A testing framework was developed to find to what extent the model degrades for identifying temporal model degradation (MIT, 2023). The researchers ran 20,000 experiments of this type for each dataset-model pair to report an aging model chart for each dataset-model team. The following chart represents the model's performance and how the model ages increase. They concluded that:

1. **The error increases over time:** the model becomes less and less performant as time passes. This may happen due to a drift in any of the model's features or concept drift.
2. **The error variability increases over time:** The gap between the best and worst-case scenarios increases as the model ages. When an ML model has high error variability, it performs well and sometimes severely. The model performance is not just degrading, but it has erratic behavior.

Additionally, through patterns, they observed that 91% of the AI model degraded over time. The variability of error changes depending on the dataset and field. There is a medium (50th percentile) AI model aging pattern for medical datasets. For example, in the following image, the black line represents an average degradation rate.

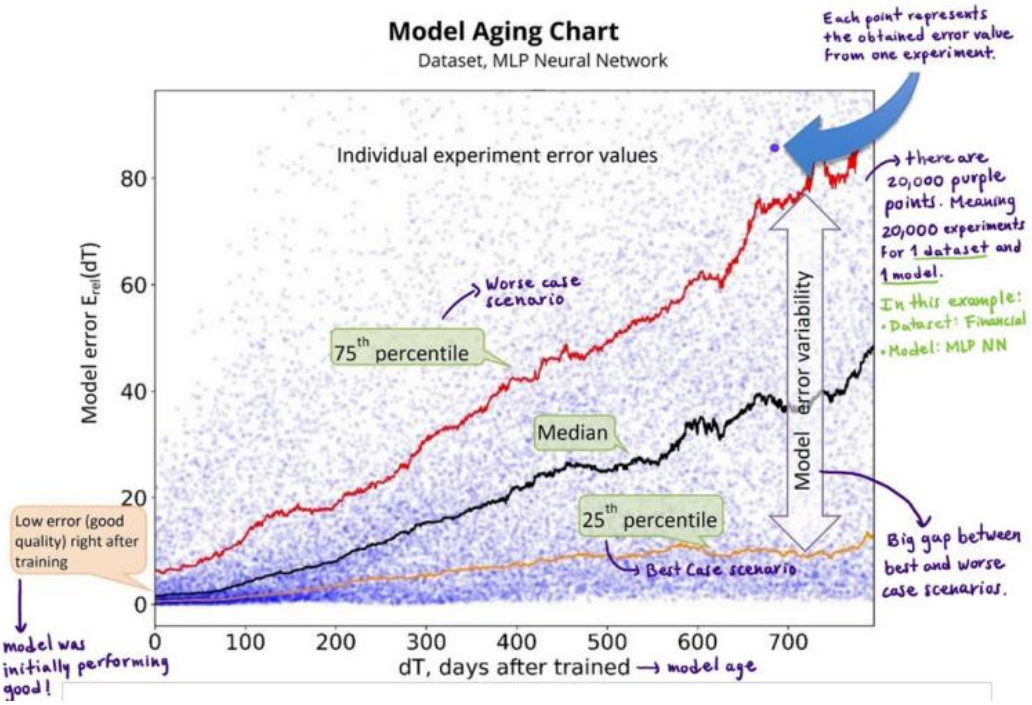


Figure 19: Model Aging example

Reliability (r(t)) is defined as the probability that a system, which is an AI model in our case, will work effectively for a given duration.

The **Failure rate (f(t))** can be defined as the number of times a system fails in each period; in our case, we would see the misclassification rate of an AI model.

The blue represents Early life, the green represents Useful life and then the red represents the Wear out stage.

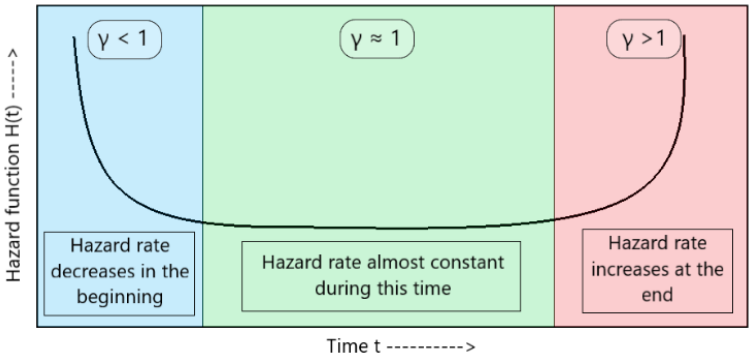


Figure 20: Structure of Hazard plot

The following values were calculated through creating an ML model and observing the accuracies over multiple repetitions.

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	1.02041E-05	1E-05
20	97319.7484	0.0201	0.9799	10	9.9316	1E-05	1.02051E-05	1E-05
30	97309.8168	0.0202	0.9798	10	69.5212	7E-05	7.14432E-05	7.1E-05
40	97240.2956	0.0209	0.9791	10	9.9316	1E-05	1.02135E-05	1E-05
50	97230.364	0.021	0.979	10	49.658	5E-05	5.10725E-05	5.1E-05
60	97180.706	0.0215	0.9785	10	49.658	5E-05	5.10986E-05	5.1E-05
70	97131.048	0.022	0.978	10	198.632	0.0002	0.000204499	0.0002
80	96932.416	0.024	0.976	10	595.896	0.0006	0.000614754	0.00061
90	96336.52	0.03	0.97	10	198.632	0.0002	0.000206186	0.00021
100	96137.888	0.032	0.968	50	1787.688	0.00036	0.000371901	0.00037
150	94350.2	0.05	0.95	50	1986.32	0.0004	0.000421053	0.00042
200	92363.88	0.07	0.93	50	4965.8	0.001	0.001075269	0.00108
250	87398.08	0.12	0.88	50	3476.06	0.0007	0.000795455	0.0008
300	83922.02	0.155	0.845	100	8441.86	0.00085	0.001005917	0.00101
400	75480.16	0.24	0.76	100	7945.28	0.0008	0.001052632	0.00105
500	67534.88	0.32	0.68	100	12911.08	0.0013	0.001911765	0.00191
600	54623.8	0.45	0.55	100	8938.44	0.0009	0.001636364	0.00164
700	45685.36	0.54	0.46	100	15890.56	0.0016	0.003478261	0.00348
800	29794.8	0.7	0.3	100	18870.04	0.0019	0.006333333	0.00633
900	10924.76	0.89	0.11	100	9931.6	0.001	0.009090909	0.00909
1000	993.16	0.99	0.01	10	993.16	0.001	0.1	0.1
1010	0	1	0	-	0	0	0	0

Table 6: AI Aging over time (Built using Kaggle Dataset and code)

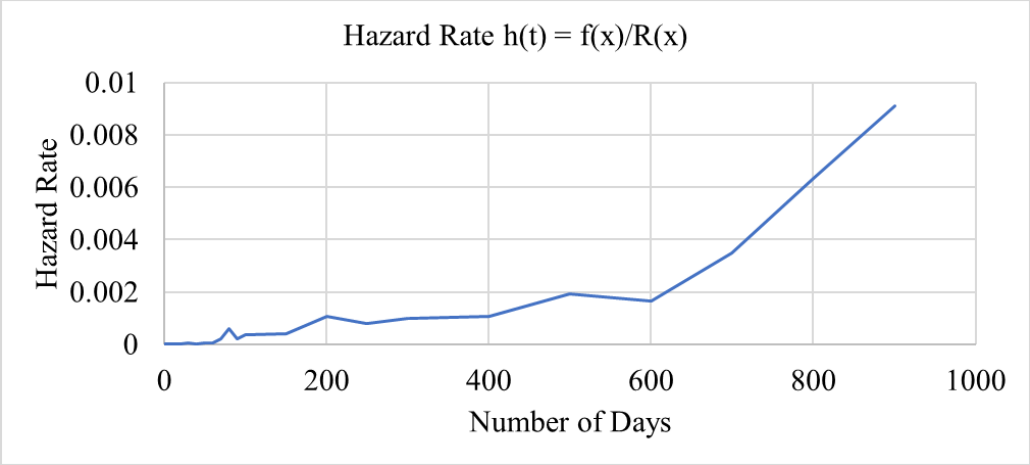


Figure 21: Hazard Rate

Hazard Rate (h(t)): It tells us about the measure of risk of failure. It gives us information on the probability of failure of the object in a study in (t+1) time, assuming it has survived till time t. So the higher the hazard value, the higher the risk of the object's failure in the study. The Weibull distribution also has a hazard function, h(t), which essentially tells us the prior information about an event yet to occur. The function is:

$$h(t) = (t / \alpha)^{\gamma}$$

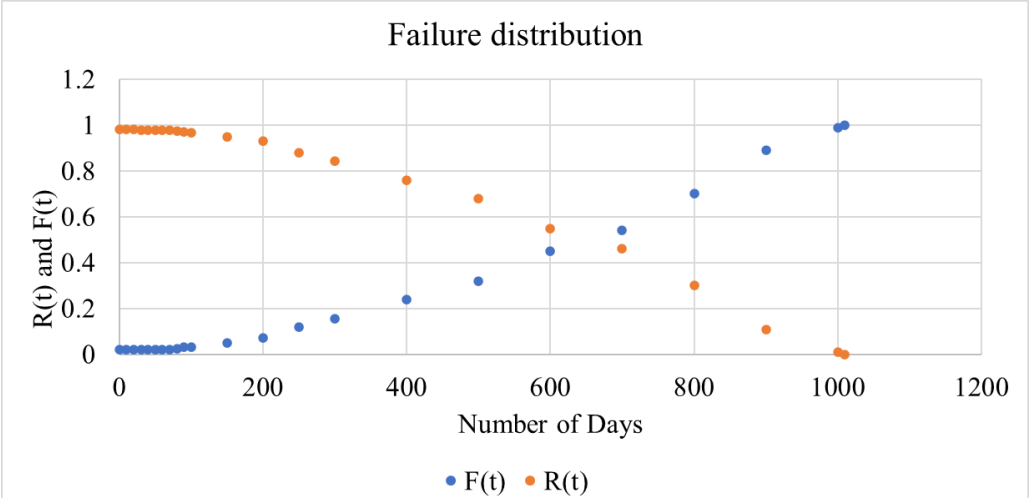


Figure 22: Failure rate and Reliability Rate

Mean Time to Failure (MTTF): It is the average time a non-repairable part or piece of equipment remains in operation until it needs to be replaced.

- Apply to non-repairable systems.
- Measures the amount of time between a device's first day of use and the last
- Represents the average lifespan,

$$MTTF = \frac{1}{Average\ of\ r(T)}$$

$$MTTF = \frac{1}{0.0058} \approx 172\ days$$

https://limblecmms.com/metrics/mean-time-to-failure/#:~:text=Mean%20Time%20To%20Failure%20(MTTF)%20is%20the%20average%20time%20a,components%20that%20cannot%20be%20repaired

8. Validation and Verification

Verification involves checking that the AI model is designed and implemented correctly according to its specifications. This process focuses on the correctness of the model's code, algorithms, and overall design. Validation focuses on assessing the model's performance against real-world data and ensuring that it meets the specified requirements and objectives. Using the testing data of the dataset and running the model to verify if the model performance is well is compulsory to make sure that the model performance is.

Evaluation Matrix:

A confusion matrix is a table used in machine learning to evaluate the performance of a classification model. It provides a summary of the model's predictions compared to the actual ground truth across different classes.

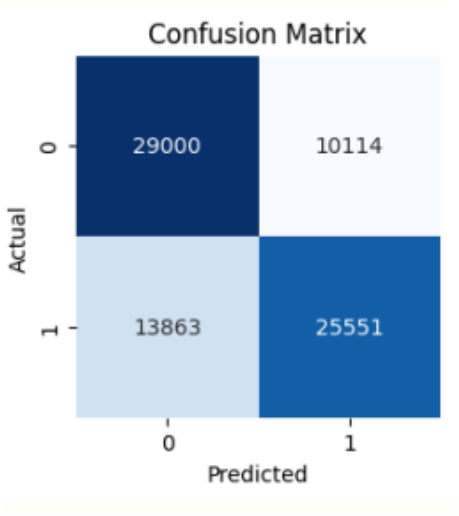


Figure23: Confusion Matrix

The False negative (13863 instances that are positive, but the model incorrectly predicted them as negative) and False positive (10114 instances that are negative, but the model incorrectly predicted them as positive) means that these are misclassified. These values allow you to calculate various performance metrics, such as accuracy, precision, recall, and F1 score, which provide insights into the model's performance on different aspects of classification. A classification report is a summary of the performance of a machine learning classification model. It provides a comprehensive overview of how well the model is doing in terms of various evaluation metrics. The accuracy is 69% which means that 69% of the time the model is right.

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

Figure 24: Classification Report

The following confidence interval shows the surety of the prediction. This also gives a validation of every diagnosis that can be used by the medical professionals to further evaluate and come to a proper diagnosis. The first bound is the prediction surety of class 0 (doesn’t have pneumonia) and the second bound is surety of class 1(have pneumonia).

```
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: 0
  Actual Class: 0
  Confidence Interval: [0.337, 0.663]
```

Figure 25: Confidence Interval

9. Monitoring

NannyML 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. Normally it is easy to determine the diagnosis is right when we already have the labels (the right class) but now as we don't have the labels, NannyML helps give an estimated guess to what the label would have been and compares it with the model's predictions.

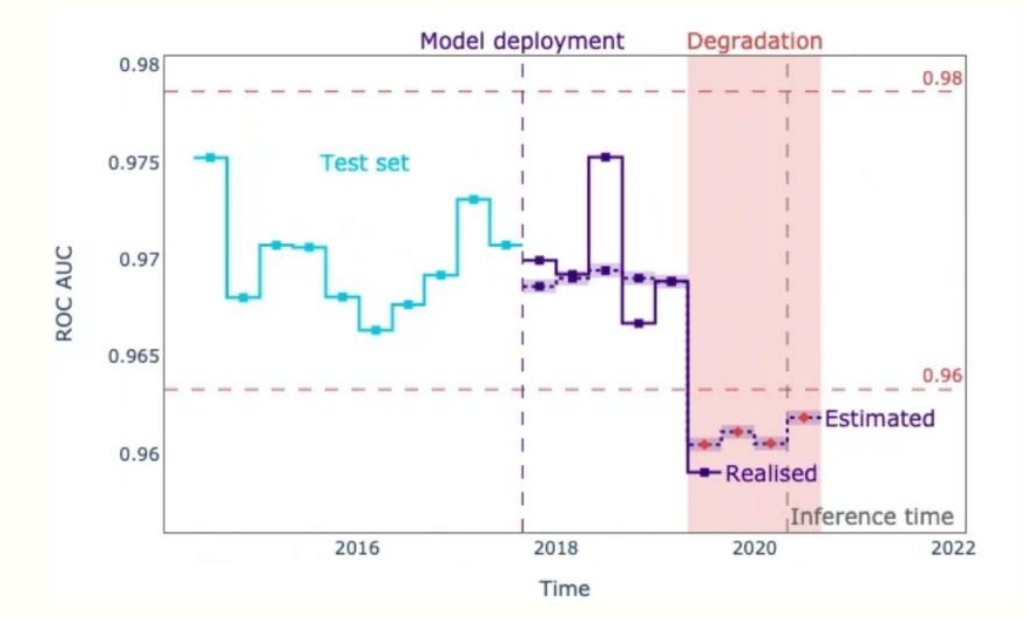


Figure 26: NannyML program

10. Standards

According to Health Canada, 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). There are 10 principles that are used for a good AI model:

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

Additionally, according to the International Electrotechnical Commission, there are two standards for medical equipment and machinery in AI.

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

11. Conclusion

- There could be many risks associated with using AI in the medical field.
- The risk of misdiagnosis with respect to Misinterpretation of Output, Incorrect weights, Manipulating Medical Data, Overreliance on AI/Lack of Communication, Poisoning (Label modifications), Medical Record Mishaps, Biased/Missing Data while Training the Model, database Unreliability is 10.5%. This is in the not tolerable range.
- The reduction can be done by using barriers such as Blockchain technology, Managing a team of experts, Chaos Engineering.
- The reduction rate of failure is 4.8%, which is in the safe zone.

12. References:

Kiener, M. *Artificial intelligence in medicine and the disclosure of risks.* AI & Soc 36, 705–713 (2021). https://doi.org/10.1007/s00146-020-01085-w

Ai can outperform doctors. so why don't patients trust it? Harvard Business Review. (2019, October 30). https://hbr.org/2019/10/ai-can-outperform-doctors-so-why-dont-patients-trust-it

Argaw, S.T., Bempong, NE., Eshaya-Chauvin, B. *et al. The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review*. *BMC Med Inform Decis Mak* **19**, 10 (2019). https://doi.org/10.1186/s12911-018-0724-5

Hamlet , P., & Tremblay , J. (n.d.). Artificial Intelligence in medicine. Metabolism: clinical and experimental. https://pubmed.ncbi.nlm.nih.gov/28126242/

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. Nature reviews. Cancer, 18(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

Amisha, Malik, P., Pathania, M., & Rathaur, V. K. (2019). Overview of artificial intelligence in medicine. Journal of family medicine and primary care, 8(7), 2328–2331. https://doi.org/10.4103/jfmpc.jfmpc_440_19

Siegal, D., Stratchko, L. & DeRoo, C. (2017). *The role of radiology in diagnostic error: A medical malpractice claims review*. Diagnosis (Berlin, Germany). https://pubmed.ncbi.nlm.nih.gov/29536933/

Cybersecurity and privacy in AI - Medical Imaging Diagnosis. ENISA. (2023, July 5). https://www.enisa.europa.eu/publications/cybersecurity-and-privacy-in-ai-medical-imaging-diagnosis

Ai shortcuts could lead to misdiagnosis of covid-19. in. (2021, June 3). https://healthcare-in-europe.com/en/news/ai-shortcuts-could-lead-to-misdiagnosis-of.html

Bow Tie Risk Management methodology. SKYbrary Aviation Safety. https://skybrary.aero/articles/bow-tie-risk-management-methodology#:~:text=Description,place%20to%20minimise%20the%20risk.v

What is Fault Tree Analysis (FTA)? Rockwell Automation. Fiix. (2022, November 18). https://fiixsoftware.com/glossary/fault-tree-analysis/

Weibull, R. *Minimal cut sets*. (1992) weibull.com -- Free Data Analysis and Modeling Resources for Reliability Engineering. https://www.weibull.com/hotwire/issue63/relbasics63.htm

Dean, P. (2023, October 13). *How to read a risk matrix used in a risk analysis*. Plant Assessor. https://www.assessor.com.au/resources/news-articles/how-to-read-a-risk-matrix

5x5 risk matrix: What is it & how to use it? Safety Culture. (2023, June 6). https://safetyculture.com/topics/risk-assessment/5x5-risk-matrix/

World Health Organization. (n.d.). *Patient safety*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/patient-safety#:~:text=Diagnostic%20errors.,in%20their%20lifetime%20(13).

Tiwary, A., Rimal, A., Paudyal, B., Sigdel, K. R., & Basnyat, B. (2019, January 22). *Poor communication by health care professionals may lead to life-threatening complications: Examples from two case reports*. Wellcome open research. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6694717/

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Brownlee, J. (2019, September 24). *A gentle introduction to Bayesian Belief Networks*. MachineLearningMastery.com. https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/

Robinson PJ, Wilson D, Coral A, Murphy A, Verow P. *Variation between experienced observers in the interpretation of Accident and emergency radiographs*. The British journal of radiology 1999. 72: 323–30. doi: 10.1259/bjr.72.856.10474490

WebMD. (2023, July 19). *Misdiagnosis seriously harms 795,000 people annually: Study*. WebMD. https://www.webmd.com/a-to-z-guides/news/20230719/misdiagnosis-seriously-harms-people-annually.

Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R., & Khan, R. A. (2020, May 13). *Healthcare data breaches: Insights and implications*. Healthcare (Basel, Switzerland). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349636/

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.

Biggest pharmaceutical lawsuits by settlement amount. Pharmaceutical Technology. (2023, August 9). https://www.pharmaceutical-technology.com/features/biggest-pharmaceutical-lawsuits/?cf-view

Nezamodini Z S, Khodamoradi F, Malekzadeh M, Vaziri H. (2016, March 29). Nursing errors in intensive care unit by human error identification in Systems Tool: A Case Study https://www.researchgate.net/profile/Zeynab-Nezamodini/publication/301673724_Nursing_Errors_in_Intensive_Care_Unit_by_Human_Error_Identification_in_Systems_Tool_A_Case_Study

BahooToroody, A., Mohammad , M., Gholamnia, R., Mohammad Bahoo Torody, M., & Hekmat Nejad, N. *Developing a Risk-Based Approach for Optimizing Human Reliability*

Assessment in an Offshore Operation. Article Citations - References - Scientific Research Publishing. https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55))/reference/ReferencesPapers.aspx

Srinivasan, R., Srinivasan, B., & Mohd , U. I. (2022, May 28). *Human factors in digitalized process operations*. Methods in Chemical Process Safety. https://www.sciencedirect.com/science/article/pii/S2468651422000071

Sutton I,. *THERP (technique for human error rate prediction): Sutton technical books*. THERP (Technique for Human Error Rate Prediction) | Sutton Technical Books. https://iansutton.com/topics/therp-technique-human-error-rate-prediction

Sutton, I. (2022, December 7). *OSHA's PSM update: What's not there - human error*. Net Zero by 2050. https://netzero2050.substack.com/p/oshas-psm-whats-not-there-part-1

NannyML. (n.d.). NannyML Quickstart. Retrieved from https://www.nannyml.com/blog/nannyml-quickstart

Fiddler AI. (n.d.). 91 Percent of ML Models Degrade Over Time. Retrieved from https://www.fiddler.ai/blog/91-percent-of-ml-models-degrade-over-time#:~:text=A%20recent%20study%20by%20Harvard,to%20advance%20real%2Dlife%20applications

Ribeiro, A., Singh, S., & Guestrin, C. (2022). Beyond model drift: concept drift and model decay in machine learning. Scientific Reports, 12(1), 2077. https://www.nature.com/articles/s41598-022-15245-z

Srinivas, A. (2020). Concept Drift and Model Decay in Machine Learning. Retrieved from https://towardsdatascience.com/concept-drift-and-model-decay-in-machine-learning-a98a809ea8d4

DeepChecks. (n.d.). ML Model Monitoring Checklist: Things You Should Look Out For. Retrieved from https://deepchecks.com/ml-model-monitoring-checklist-things-you-should-look-out-for/

International Electrotechnical Commission. (n.d.). Standards for medical devices. Retrieved from https://www.iec.ch/blog/standards-medical-devices

International Electrotechnical Commission. (n.d.). Publication 2612. Retrieved from https://webstore.iec.ch/publication/2612

Health Canada. (n.d.). Good machine learning practice for medical device development. Retrieved from https://www.canada.ca/en/health-canada/services/drugs-health-products/medical-devices/good-machine-learning-practice-medical-device-development.html

13. Appendix:

```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from tensorflow import keras
import librosa
import librosa.display

import tensorflow as tf
from matplotlib.pyplot import specgram
import keras
df_main = pd.read_csv('/content/sample_data/train-all.csv',encoding="ISO-8859-1")
from sklearn.utils import resample

# Assuming df is your DataFrame
# Separate majority and minority classes
df_majority = df_main[df_main['Pneumonia'] == 0]
df_minority = df_main[df_main['Pneumonia'] == 1]

# Upsample minority class
df_minority_upsampled = resample(df_minority, replace=True, n_samples=len(df_majority),
random_state=42)

# Combine majority class with upsampled minority class
df_balanced = pd.concat([df_majority, df_minority_upsampled])

# Shuffle the dataset
df_balanced = df_balanced.sample(frac=1, random_state=42)
print(df.head())

import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
y= df['Pneumonia']
X= df.drop('Pneumonia', axis=1)
X=X.drop('PatientId',axis=1)
X=X.drop('Image',axis=1)
# using the train test split function
X_train, X_test, y_train, y_test = train_test_split(X,y ,
                                    random_state=104,
                                    test_size=0.4,
                                    shuffle=True)
```

```
from lightgbm import LGBMClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_curve, auc, roc_auc_score
from sklearn.datasets import make_classification

# Create and train a classification model (Logistic Regression in this case)
model = LogisticRegression(random_state=42)
model.fit(X_train, y_train)

# Predict probabilities for both training and testing sets
y_train_prob = model.predict_proba(X_train)[:, 1]
y_test_prob = model.predict_proba(X_test)[:, 1]
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Make predictions on the test set
y_test_pred = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_test_pred)
print(f"Accuracy: {accuracy:.2f}")

# Generate and print the classification report
class_report = classification_report(y_test, y_test_pred)
print("Classification Report:\n", class_report)

# Generate and print the confusion matrix
conf_matrix = confusion_matrix(y_test, y_test_pred)
print("Confusion Matrix:\n", conf_matrix)
import numpy as np
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Make predictions on the test set
y_test_pred = model.predict(X_test)

# Assuming your model supports predict_proba
if hasattr(model, "predict_proba"):
    y_test_proba = model.predict_proba(X_test)

    # Calculate confidence intervals for each prediction
```

```
    confidence_intervals = np.zeros_like(y_test_proba)
    for i in range(len(confidence_intervals)):
        confidence_intervals[i] = np.percentile(y_test_proba[i], [2.5, 97.5])

    # Display the confidence intervals, predicted class, and actual class for the first few
predictions
    print("Predictions with Confidence Intervals:")
    for i in range(5):  # Display for the first 5 predictions
        predicted_class = y_test_pred[i]
        actual_class = y_test.iloc[i] if hasattr(y_test, 'iloc') else y_test[i]
        lower_bound, upper_bound = confidence_intervals[i]

        print(f"Sample {i+1}:")
        print(f"  Predicted Class: {predicted_class}")
        print(f"  Actual Class: {actual_class}")
        print(f"  Confidence Interval: [{lower_bound:.3f}, {upper_bound:.3f}]")
        print("---")

    # You can use the confidence intervals as needed for your analysis
else:
    print("Model does not support predict_proba.")
```