DISSERTATION REPORT

At

CARPL

A Report

By

Mr. Swastik Mitra

PG/23/120

Under the guidance of: Dr. Preetha GS

POST GRADUATE DIPLOMA IN HOSPITAL AND HEALTH MANAGEMENT 2023-2025

IIHMR DELHI

International Institute of Health Management Research New Delhi

(Completion of Dissertation from respective organization)

The certificate is awarded to

SWASTIK MITRA

in recognition of having successfully completed his internship in the department of

CUSTOMER SUCCESS AND ACCOUNT MANAGEMENT

and has successfully completed his Project on

AI in Neuroimaging: Validating a Diagnostic Tool for Responsible and Beneficial Clinical Use

(Feb 17 to May 17, 2025)

Organisation: CARPL.ai

He comes across as a committed, sincere & diligent person who has a strong drive & zeal for learning.

We wish him all the best for future endeavors.

Training & Development

Vaishali

Zonal Head-Human Resource

CARPL AI PVT. LTD. New Delhi

TO WHOMSOEVER IT MAY CONCERN

This is to certify that SWASTIK MITRA, a student of PGDM (Hospital & Health Management) from International Institute of Health Management Research, New Delhi has undergone internship training at CARPL.ai from 17th FEB 2025 to 17th May 2025.

The Candidate has successfully carried out the study designated to him during internship training and his approach to the study has been sincere, scientific and analytical.

The Internship is in fulfillment of the course requirements.

I wish him success in all his future endeavors.

Dr. Sumesh Kumar
Associate Dean, Academic and Student Affairs
IIHMR, New Delhi

Mentor

IIHMR, New Delh

INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT
RESEARCH,
NEW DELHI

CERTIFICATE BY SCHOLAR

This is to certify that the dissertation titled **AI in Neuroimaging: Validating a Diagnostic Tool for Responsible and Beneficial Clinical Use** and submitted by

SWASTIK MITRA

Enrollment No. **PG/23/120** under the supervision of **DR PREETHA GS** for award of PGDM (Hospital & Health Management) of the Institute carried out during the period

from 17TH FEB 2025 to 17TH MAY 2025 embodies my original work and has not formed the basis for the award of any degree, diploma associate ship, fellowship, titles in this or any other Institute or other similar institution of higher learning.

Signature

Certificate from Dissertation Advisory Committee

This is to certify that **SWASTIK MITRA,** a graduate student of the **PGDM (Hospital & Health Management)** has worked under our guidance and supervision. He is submitting this dissertation titled "**AI in Neuroimaging: Validating a Diagnostic Tool for Responsible and Beneficial Clinical Use**" at "CARPL.ai" in partial fulfillment of the requirements for the award of the **PGDM (Hospital & Health Management).**

This dissertation has the requisite standard and to the best of our knowledge no part of it has been reproduced from any other dissertation, monograph, report or book.

Dr Preetha G R
Professor & Dean (Research)

IIHMR Delhi

Amit Sahni
Senior Account Manager-Enterprise AI

CARPL.ai

CARPL.AI PVT. LTD. New Delhi

FEEDBACK FORM

Name of the Student: Swastik Mitra

Name of the Organisation in Which Dissertation Has Been Completed: CARPL AI

Area of Dissertation: Customer Success and Account Management

Attendance: 100%

Objectives achieved: - Validated an AI tool for brain hemorrhage detection through real world clinical data & dissertation report
- Identify gaps in model and proposed improvements
- Supported AI deployments in public health domain across India & ME

Deliverables:
- AI performance report using diff metrics
- Account management of the projects from Indian & ME
- Stakeholder communication & progress reporting

Strengths:
- Understanding of AI in health
- Cross functional work management (Clinical & Tech)
- Skilled in Communication, requirement gathering & engagements.

Suggestions for Improvement:
- Contribute more actively to academic publication & researches

Suggestions for Institute (course curriculum, industry interaction, placement, alumni):
- Add module related to AI in radiological workflows which will help students to explore the adoption & benefits of AI within this niche.

Signature of the Officer-in-Charge/ Organisation Mentor (Dissertation)

CARPL.AI PVT. LTD. New Delhi

Certificate of Approval

The following dissertation titled **AI in Neuroimaging: validating a diagnostic tool for responsible and beneficial clinical use** is hereby approved as a certified study in management carried out and presented in a manner satisfactorily to warrant its acceptance as a prerequisite for the award of **PGDM (Hospital & Health Management)** for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the dissertation only for the purpose it is submitted.

Dissertation Examination Committee for evaluation of dissertation.

Name	Signature
Dr. Sukesh Bhardwaj	Sukesh
Praveen Kumar	

IIHMR DELHI

INTERNATIONAL INSTITUTE OF HEALTH MANAGEMENT RESEARCH (IIHMR)
Plot No. 3, Sector 18A, Phase- II, Dwarka, New Delhi- 110075
Ph. +91-11-30418900, www.iihmrdelhi.edu.in

Name of Student (in block letter) : Mr. SWASTIK MITRA

Enrollment/Roll No.	PGDM/23/120	**Batch Year**	2023-25
Course Specialization (Choose one)	Health IT	2023-25	Healthcare IT
Name of Guide/Supervisor	Dr. Preetha GS		
Title of the Dissertation/Summer Assignment	**AI in Neuroimaging: Validating a Diagnostic Tool for Responsible and Beneficial Clinical Use**		
Plagiarism detect software used	"TURNITIN"		
Similar contents acceptable (%)	Up to **15** Percent as per policy		
Total words and % of similar contents Identified	10%		
Date of validation (DD/MM/YYYY)	20/07/2025		

CERTIFICATE ON PLAGIARISM CHECK

Guide/Supervisor
Name: Dr. Preetha GS
Signature:

Student
Name: Swastik Mitra
Signature:

Report checked by

Institute Librarian
Signature:
Date:
Library Seal

Dean (Academics and Student Affairs)
Signature:
Date:
(Seal)

ACKNOWLEDGEMENTS

I would like to express our sincere gratitude to the following individuals and organizations for their valuable contributions to this report:

I am extremely grateful to Dr. Vidhur Mahajan, CEO of CARPL AI, for giving me this dissertation opportunity under his esteemed guidance. His mentorship, invaluable insights, and constant encouragement have greatly enriched my learning experience.

Mr. Amit Sahni, for his insight and assistance during the entire research journey. Dr. Easwar, for making available relevant materials and his assistance in analyzing and interpreting them.

My sincere gratitude goes to Dr. Preetha GS, whom I consider my mentor from the International Institute of Health Management and Research. She provided unwavering encouragement, direction, and constructive input which has greatly influenced my academic and professional development. Through her understanding and dedication, I have been able to afford a solid foundation for my future career.

The faculty and staff of the International Institute of Health Management and Research also deserve my profound appreciation. Their enduring encouragement, coupled with a favorable paradigm for learning, has tremendously catalyzed my development and achievements.

I thank everyone who participated in this project for their support and motivation.

TABLE OF CONTENTS

List of Figures:

List of Tables:

ABBREVATIONS

AI	Artificial Intelligence
CARPL	CARING Analytics Platform
HCP	Healthcare Provider
PACS	Picture Archival Communication System
API	Application Programming Interface
PHI	Personal Health Information
AUC	Area Under Curve
ROC	Receiver Operating Curve
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
ICH	Intracranial Hemorrhage
IVH	Intraventricular Hemorrhage
SDH	Subdural Hematoma
SAH	Subarachnoid Hemorrhage
IPH	Intraparenchymal Hemorrhage
EDH	Epidural Hematoma
GT	Ground Truth

ABOUT THE ORGANIZATION:

1.1 ABOUT CARPL:

CARPL.ai is the first platform as a software company which provides the functionality to build, test, monitor and deploy medical imaging AI applications within the radiological workflows for optimization in context of Clinical, Financial and Operational domains. Ultimately the CARPL platform connects healthcare providers with radiological AI applications, helping improve access, affordability, and quality of medical care.

It is used by some of the world's leading healthcare providers, AI researchers, industry teams and startups. It is built with the single goal of making it easy for clinicians to get access to advanced analytics tools.

1.2 VISION

CARPL's vision is to be the back-end platform behind all medical imaging AI deployment globally by being the main point of contact for AI deployment at healthcare providers, and the preferred strategy for AI developers.

1.3 FEATURES

- Easy to Access: There is a fragmented ecosystem out there if any HCP wanted to deploy any radio diagnostic AI solution how they can contact AI developers providing similar solutions individualistically, but CARPL platforms have 100+ AI solutions of different AI developers on a simple single user- friendly platform
- Easy to Assess: CARPL is the first company to provide the feature of Validation and Testing & Monitoring of the AI solutions for the HCPs so they can analyze the efficiency of the AI model over their own Patient data.

- Easy to Integrate: If the HCPs start deploying AI solutions related to radiodiagnosis by multiple AI developers for different use cases they would require different RIS-PACS integrations but for accessing multiple AI solutions via CARPL requires single RIS-PACS setup.
- Reliability: All the AI solutions onboarded on CARPL's platform are regulatory compliant including FDA(US), CE(Europe), HSA(Singapore), ANVISA(Brazil).

PROJECT REPORT

ABSTRACT

AI-powered platforms are transforming diagnostic radiology by enhancing image quality, accelerating interpretation, and optimizing reporting workflows. From acquisition improving contrast and reducing noise across CT, MRI, ultrasound, X-ray, and mammography to real-time triage, AI embedded in PACS/RIS systems helps detect abnormalities, prioritize urgent cases, and prepare findings for radiologist review.

CARPL.ai exemplifies this shift as a vendor-neutral, enterprise-grade orchestration platform that unifies over 110–175 AI applications from 50–75+ global vendors via a single user interface, data channel, and procurement pipeline.
It streamlines pre-deployment validation and post-deployment monitoring, allows deployment on-premises or cloud, and boasts FDA 510(k) clearance for safe integration into clinical workflows
carpl.ai.
In urgent scenarios, such as non-contrast brain CT for intracranial hemorrhages epidural, subdural, subarachnoid, intraparenchymal, and intraventricular CARPL.ai supports rapid anomaly detection and case prioritization, aiding clinical decision making while reducing radiologist burden and turnaround times.

1. INTRODUCTION

The adoption of artificial intelligence (AI) in medical images became an important advance in diagnostic radiology. With an increasing global demand for high -speed interpretation and high precision of medical scans, AI technologies are increasingly positioned to support clinical decision making, particularly in environments where expert radiologists are overloaded or scarce. The radiological workflow enabled for AI is something that helps health systems to achieve clinical, operational and financial results, and that is why it leads to the growing adoption of the AI.
The integration and adoption of AI in the radiological workflow improves multiple stages of images and reports:

- Image acquisition and processing: IA is integrated into images diagnosis (CT, MRI, ultrasound, X -rays and mammography) to support technicians during the exploration acquisition. It helps to manage contrast levels, reduce noise and improve image clarity. For example, the subtle medical AI improves the quality of the image of MR and PET while reducing scanning times.

- Interpretation and triage: AI systems integrated into PACS (image storage software) and RIS can analyze real -time scans to identify abnormalities, prioritize cases based on urgency and prepare the findings for radiologist review. CARPL, for example, allows radiology AI models to connect to PACS/RIS systems used by medical care providers, which helps in decision support.

    - Report Generation and Workflow Management: AI -operated tools help to create reports, flag the missed findings, and standardize the diagnostic language. Additionally, on the operational side, AI supports administrative workflow-patrol registration and timetable billing and electronic health records (EHR) integration (eg, XSolis for AI-operated case management).

CARPL is the world's first test and deployment platform designed especially for medical imaging AI applications. It combines healthcare providers with third-party AI solutions, enabling improved access, affordable and quality medical care.
Working as a bridge between healthcare professionals and AI developers, CARPL A.I. The tools act as a central center for testing, valid and integrated in the workflow of radiology.
Reliable by leading hospitals, research institutes, startups and industry teams, CARPL's One mission is designed: to facilitate and accelerate the clinician in cutting-aging imaging analytics. CARPL uses a platform -based approach to integrate with hospitals' radiological workflow. The main product is a platform that helps users to cease to accuse multiple radiological AI solutions developed by third party partners around the world and meet their clinical use.

In the interpretation of non-disadvantaged computed tomography (CT) scans of the brain for the discovery of acute and chronic intracranial abnormalities, especially AI. It is a very crucial domain that has a significant promise.

Epidural (EDH), Subdural (SDH), Subarachnoid (SAH), intraparenchymal (IPH), and intraventricular (IVH) hemorrhages are the most urgent conclusions.

1.1 RATIONALE FOR THE STUDY

The current study is situated within this context of responsible AI deployment. Our team is in the process of onboarding a brain CT AI solution into our platform, with plans to deploy it across multiple healthcare sites. However, this deployment is contingent upon the tool's proven diagnostic performance when tested on real-world clinical data. Thus, a systematic validation study was undertaken using a dataset of 1261 CT brain scans obtained from a healthcare institute in Brazil. Each case included a radiologist-generated report, which served as the basis for establishing the ground truth.

This AI solution is designed to automatically detect several types of hemorrhagic abnormalities on brain CT, including:

- Intracranial Hemorrhage: A broad term that refers to any bleeding within the skull. This includes bleeding in brain tissue or surrounding areas such as the meninges. It is often caused by trauma, stroke, or aneurysm rupture and can lead to increased intracranial pressure, brain damage, or death if untreated.

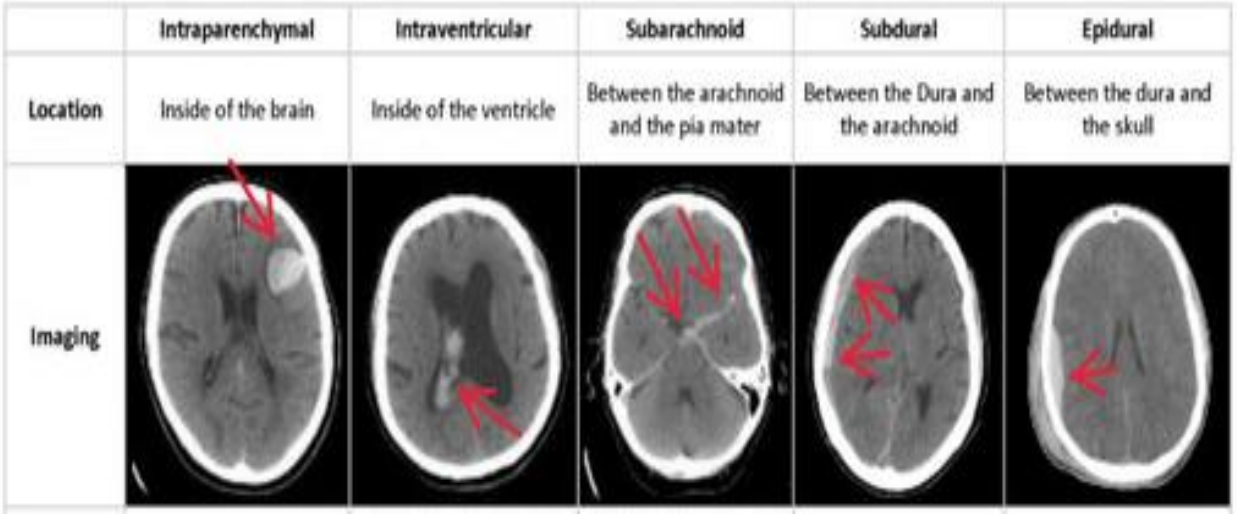
- Epidural Hemorrhage: Bleeding between the inner surface of the skull and the dura mater (the outermost protective membrane of the brain). It is usually caused by head trauma leading to arterial injury (commonly the middle meningeal artery).

- Subdural Hemorrhage: Bleeding between the dura mater and the arachnoid membrane, typically caused by tearing of bridging veins.

- Subarachnoid Hemorrhage: Bleeding into the subarachnoid space is the area between the arachnoid membrane and the pia mater where cerebrospinal fluid circulates. Most cases result from ruptured aneurysms or trauma. It presents with sudden severe headaches and can be life-threatening.

- Intraparenchymal Hemorrhage: Bleeding within the brain tissue (parenchyma). Can be the result of hypertension, trauma, vascular malformations or tumours.

- Intraventricular Hemorrhage: Bleeding can enter the brain's ventricular system, where cerebrospinal fluid is made and flows. This may occur due to the spread of intraparenchymal hemorrhage, injury, or blood vessel issues. It can block CSF flow.



The primary objective of this dissertation is to validate whether the AI is working as expected in flagging these pathologies, using the radiologist's verdict as the reference point. We will be assessing the diagnostic capabilities of the model using the standard including Sensitivity, Specificity, Accuracy, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Area Under the Receiver Operating Characteristic Curve (AUC and ROC). This research supports the careful use of AI in neuroimaging by offering solid evidence on model accuracy, pointing out areas that need improvement, and creating a clear validation process that meets ethical and clinical standards.

2. REVIEW OF LITERATURE

The integration and adoption of artificial intelligence (AI) in neuroimaging has rapidly expanded and changing, with increasing research dedicated to validating and monitoring its diagnostic accuracy and clinical benefits. Several key studies have shown the possibilities and limitations of AI in reading non-contrast CT brain scans. This is especially

true for identifying stroke and intracranial hemorrhage (ICH).

Mair et al. (2022)

Mair and colleagues performed an external validation of the AI model named e-Aspect software designed to assess early ischemic changes in brain and hemorrhage under CT scans. Their study involved multiple radiological centers and focused on the tool's performance. The AI system demonstrated strong agreement with expert assessment in terms of ASPECTS scoring and hemorrhage detection. One of the key strengths of this study was its focus on real-world variability, encompassing different scanners and clinical environments**.** However, the study did not delve into performance for specific subtypes of hemorrhages, limiting insights into how well the tool differentiated between types such as subarachnoid, subdural, or intraparenchymal hemorrhages. This is an area addressed more thoroughly in the current study.

Bar et al. (2022)

In a multicenter external validation study, Bar et al. evaluated an AI model for stroke detection on non-contrast CT scans. The study reported high sensitivity and acceptable specificity, particularly in high-resource settings. However, they found a decline in AI performance at peripheral centers due to variability in image quality and scanning protocols. Notably, the study emphasized the importance of dataset diversity and consistent preprocessing for generalizability. It also acknowledged an increase in false positives in some settings—an observation echoed in the present research, particularly for conditions such as epidural or intraventricular hemorrhage, where the AI in this study also overcalled abnormalities.

Dyer et al. (2022)

Dyer et al. validated an AI tool designed to triage non-contrast CT head scans, focusing on its ability to "rule out normal." Their model achieved high negative predictive value, enabling efficient identification of normal scans in emergency settings. This reinforced the AI's potential to streamline radiology workflows by prioritizing abnormal findings. However, the study highlighted that AI performance was not flawless and required radiologist oversight, especially for ambiguous or complex cases. In comparison, the current study further expands on this point by analyzing misclassifications due to

overlapping imaging features such as effusions, ossification, or post-operative changes—offering a richer clinical interpretation of false positives.

All these studies set a foundation for the use of AI in neuroimaging, particularly for acute pathologies. They suggested and indicated that AI systems can assist in early diagnosis, triage, and workflow optimization. However, gaps remain in the interpretations provided in AI.

This approach not only supports existing findings on AI's potential and pitfalls but also offers novel insights into the underlying causes of AI misclassifications—an area less explored in prior research.

3. METHODOLOGY

3.1 Research Question:

How accurately does the AI solution detect various brain hemorrhage pathologies on CT scans compared to radiologist-confirmed ground truth before clinical deployment?

This question aims to assess whether the AI tool meets the required diagnostic accuracy and reliability for safe and responsible integration into clinical workflows.

3.2 Objectives:

To systematically monitor and access the diagnostic performance and working of the AI solution in detecting/flagging multiple brain pathologies on CT scans against radiologist-verified ground truth performance before its integration into our platform and deployment to healthcare providers.

3.3 Research Design :

This study is a retrospective cross sectional validation study comparing AI predictions with ground truth labels extracted from radiologist reports.

3.4 Study Population:

- The patients who undergone CT scans from the health center.
- Inclusion Criteria: The CT scans should be interpreted by the Radiologist and report with the final verdict are there and the scans are in DICOM format.
- Exclusion Criteria: CT scans without Radiologists reports or incomplete verdicts and the scans whose quality as well as the standard (should be DICOM) is not supportable by the platform.

3.5 Sample Sizes and Sampling Methods:

Sample Size used: Approximately 1261 CT scans.

Sampling Method:

- Convenience Sampling – selected cases based on ease of access and availability.
- Retrospective brain CT scans with radiologist reports sourced from partner radiology center.

3.6 Research Procedure:

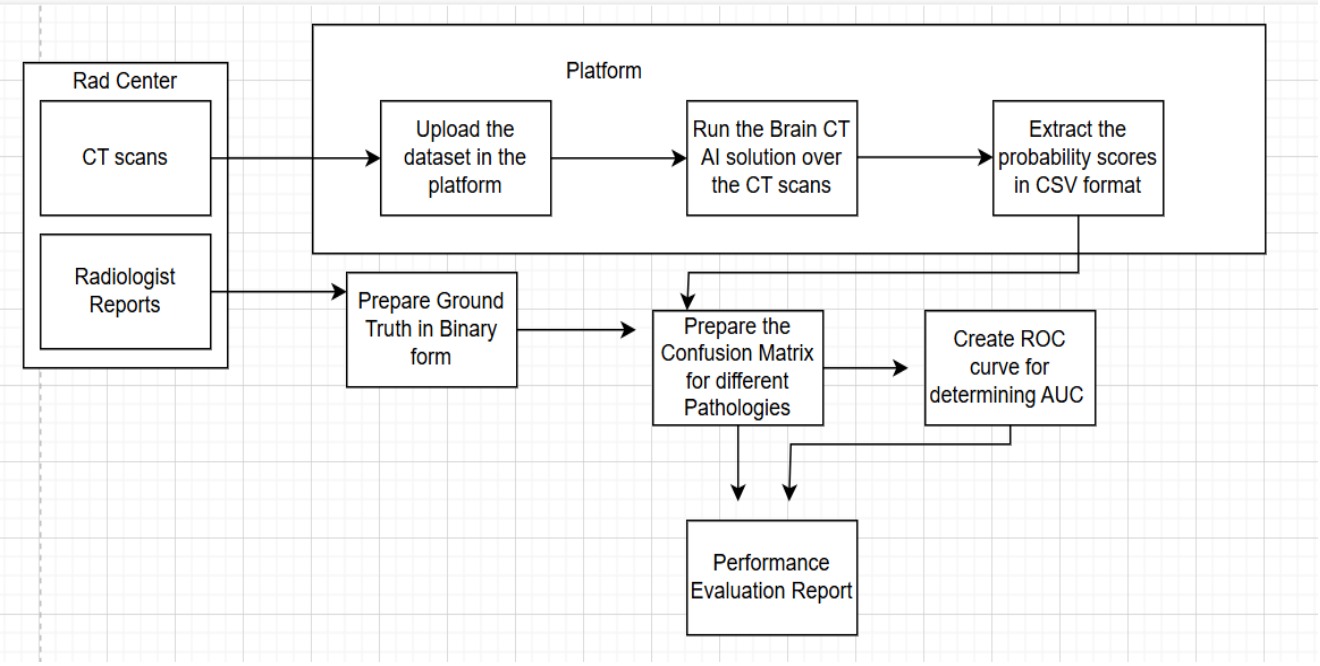


Fig 1

The following steps were undertaken to evaluate or validate the diagnostic performance of the AI model in detecting brain hemorrhages or pathologies from CT scans:

1. Data Collection

CT brain scans were obtained from a partner healthcare institution based in Brazil. A total of 1261 CT scans were received in DICOM format. Alongside these scans, radiologist-interpreted reports were provided in CSV format, detailing the presence or absence of specific findings for each case.

2. Data Compilation

All DICOMs were securely uploaded to our AI platform for further processing. Each case was mapped by a unique identifier i.e Study ID to ensure consistency between imaging and report data. The CSV reports were structured to allow direct association between clinical interpretation and AI output.

3. AI Inferencing

The brain CT scans were processed through the integrated AI algorithm deployed on our platform. Out of the 1261 cases out of which 1094 cases were processed over the platform and the AI fails to give any outcome in the remaining 167 cases. The AI model generated probability scores for six major intracranial hemorrhagic findings:

Epidural Hemorrhage (EDH)

Subdural Hemorrhage (SDH)

Subarachnoid Hemorrhage (SAH)

Intraparenchymal Hemorrhage (IPH)

Intraventricular Hemorrhage (IVH)

General Intracranial Hemorrhage (ICH)

These probability scores were exported in CSV format, creating a structured dataset for further comparison.

4. Ground Truth Creation:

To evaluate AI performance, a ground truth (GT) dataset was manually constructed by reviewing the radiologist reports. Each of the six findings was marked as '1' (present) if mentioned in the report, and '0' (absent) if not identified. This binary GT matrix served as the reference point for performance analysis.

5. Result Comparison

The AI findings were compared to the actual truth labels for each case and specific pathology. Each prediction was labeled as a True Positive (TP), True Negative (TN), False Positive (FP), or False Negative (FN). These outcomes were then used to calculate several performance metrics, including sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV). They also helped create ROC curves and AUC scores for each pathology.

	A	B	C	D	E	F	G	H
	study_id_anonymized	GT_Subdural hematon	AI_Subd	AI Biinary	TP	TN	FP	FN
	1.12.119.1357.96016.18	0	0	0	0	1	0	0
	7.16.266.6305.14334.91	1	84	1	1	0	0	0
	1.88.821.9145.73000.51	0	0	0	0	1	0	0
	3.27.643.3983.91070.85	0	0	0	0	1	0	0
	1.17.108.8330.10070.84	0	0	0	0	1	0	0
	1.18.270.6906.16617.64	0	1	0	0	1	0	0
	2.16.591.8808.19956.96	1	86	1	1	0	0	0
	7.93.550.8519.93104.44	0	2	0	0	1	0	0
0	9.94.995.1856.16989.63	0	1	0	0	1	0	0
1	3.32.450.9811.11074.91	1	99	1	1	0	0	0
2	5.59.181.4328.14691.70	0	7	0	0	1	0	0
3	1.12.598.4401.58552.20	0	0	0	0	1	0	0
4	2.15.495.9805.60625.12	1	82	1	1	0	0	0
5	8.99.872.3445.65822.65	1	97	1	1	0	0	0
6	1.14.509.4802.40034.66	0	0	0	0	1	0	0
7	1.13.100.1278.83960.45	1	27	0	0	0	0	1
8	4.66.160.7317.97197.52	0	0	0	0	1	0	0
9	1.12.769.2516.24622.49	1	100	1	1	0	0	0
0	9.46.105.3656.34784.14	1	57	1	1	0	0	0

Subarachnoid Hem | Final | Metrices | Subdural Heamatoma | Epidural Heematoma

Fig 2

4. RESULT

4.1 DATA ANALYSIS

So, our objective was to evaluate the diagnostic performance of AI algorithms in interpreting lower extremities musculoskeletal (MSK) X-rays and for this study various metrics are used:

- **Confusion Matrix:** It is a quadrant which is made for the comparison between AI provided outputs and the Radiologists verdict that are the ground truth. The quadrants help to achieve the numbers of True positives, True negatives, False positives and False negatives. This metric lays the basis for calculation of different performance metrices.

- **Sensitivity:** Sensitivity, also known as recall or true positive rate, is a measure that shows out of all the actual positive cases according to the GT are successfully predicted by the AI model.

- **Specificity:** Specificity, also known as the true negative rate, is a measure of the AI model's ability to detect the actual negative cases according to the radiologist's verified GT.

- **Precision:** Precision, also known as positive predictive value, is a measure out of all positive flagged cases by AI how many are actual positives are according to GT.

- **NPV:** Stands for the Negative Predicted Value is a measure that shows out of the total negative cases predicted by the AI how many are actual negative according to GT.

- **Accuracy:** Accuracy is a common metric for evaluating how correct a diagnostic or classification model is. It measures the number of correct predictions, including both true positives and true negatives, compared to the total number of cases assessed.
- Mathematically they are calculated by these formulas:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$= \frac{True\ Positive}{Total\ Predicted\ Positive}$$

$$\text{NPV} = \frac{\text{True negative}}{\text{True negative + false negative}}$$

$$Accuracy = \frac{TrueNegatives + TruePositive}{TruePositive + FalsePositive + TrueNegative + FalseNegative}$$

- **ROC:** The Receiver Operating Characteristic (ROC) curve is a graphical representation used to evaluate the performance of a binary classification model, such as an AI system detecting brain hemorrhages. It plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various threshold levels.

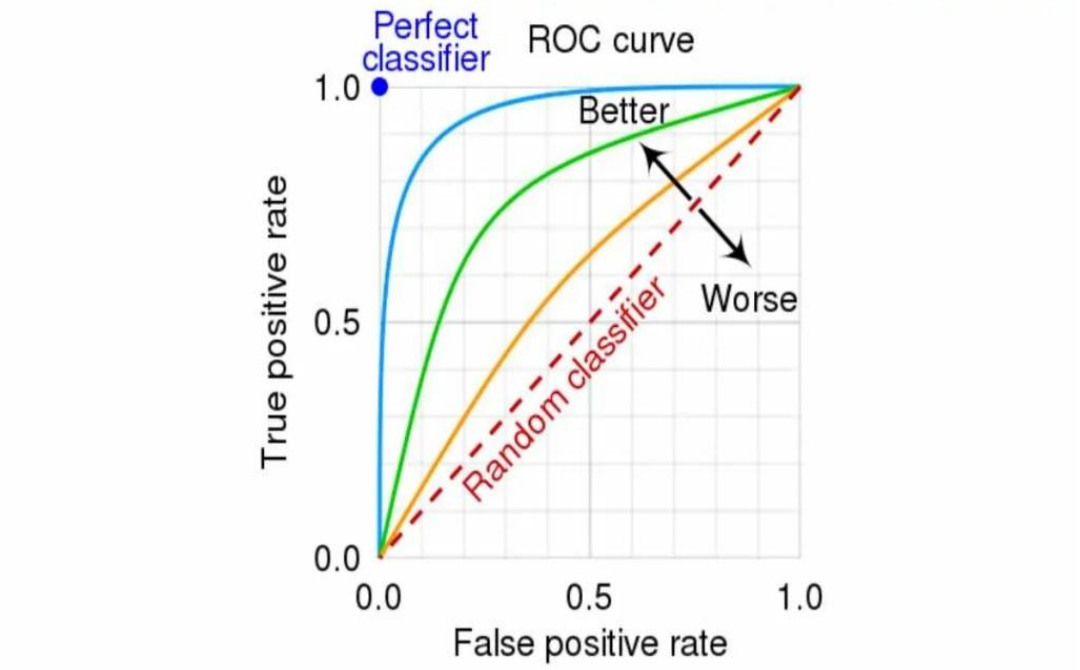


Fig 3

- **AUC:** Area Under the Curve quantifies the overall ability of the model to distinguish between positive and negative classes.
    1. AUC = 1: Perfect classifier
    2. AUC = 0.5: No better than random guessing
    3. AUC between 0.8–0.9: Good performance
    4. AUC above 0.9: Excellent performance

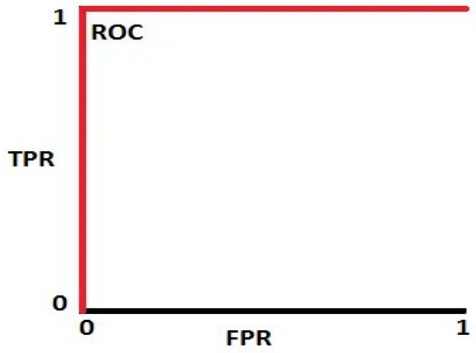


Fig 4

The AUC is 100% which means that the model has the 100% ability to distinguish between diseased and non-diseased.

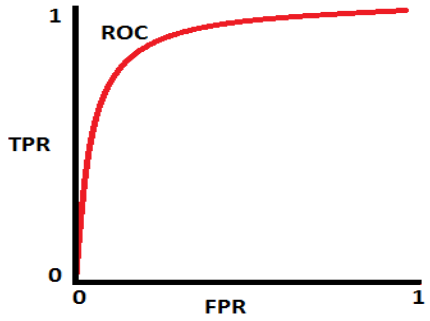


Fig 5

When the AUC is 0.7, it means there is a 70% chance that the AI model can tell the difference between diseased and non-diseased.

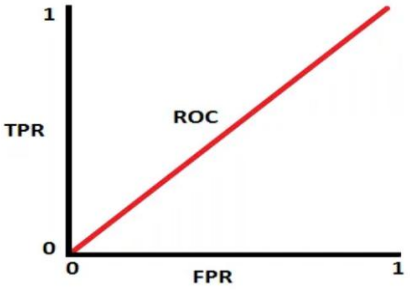


Fig 6

This is the worst situation. When AUC is approximately 0.5, the model has no ability to distinguish between positive class and negative class.

- **Construction of Confusion Matrix:**

To evaluate the AI model's ability to detect specific brain pathologies on CT scans, a structured process was followed to construct confusion matrices for each of the six findings in the dataset. The steps below outline the procedure in detail:

Step 1: Dataset Organization

For each of the six target findings an individual spreadsheet was created.

Each sheet contained case-level data with the following columns:

Study ID: A unique identifier for each case.

AI_Probability: The probability score output by the AI system for that finding, ranging from 0

to 100.

GT_Label: The ground truth label for the presence of the finding, manually annotated by expert radiologists based on the clinical report; values were 1 (present) or 0 (absent).

Step 2: Binarization of AI Output

A new column, AI binary, was added in each spreadsheet to convert AI probability scores into binary predictions.

A fixed threshold of 0.5 was used:

If AI_Probability ≥ 0.5, then AI binary = 1

If AI_Probability < 0.5, then AI binary = 0.

Step 3: Classification Outcome Assignment

For each case, the AI binary was compared against the ground truth label to classify it into one of the four confusion matrix categories:

True Positive (TP): GT_Label = 1 and AI Binary = 1

False Positive (FP): GT_Label = 0 and AI Binary = 1

True Negative (TN): GT_Label = 0 and AI Binary = 0

False Negative (FN): GT_Label = 1 and AI Binary = 0

	A	B	C	D	E	F	G	H	I	J	K
1	study_id_anonymized	GT_Intraventricular hemorrhage	AI_Intraventricular hemorrhage	AI Binary	TP	TN	FP	FN			
2	1.12.119.1357.96016.189043.	0	0	0	0	1	0	0			
3	7.16.266.6305.14334.910488.	0	7	0	0	1	0	0			
4	1.88.821.9145.73000.513089.	0	0	0	0	1	0	0			
5	3.27.643.3983.91070.851761.	0	0	0	0	1	0	0			
6	1.17.108.8330.10070.842164.	0	0	0	0	1	0	0			
7	1.18.270.6906.16617.649592.	0	1	0	0	1	0	0			
8	2.16.591.8808.19956.968780.	0	1	0	0	1	0	0			
9	7.93.550.8519.93104.447905.	0	2	0	0	1	0	0			
10	9.94.995.1856.16989.636571.	0	0	0	0	1	0	0			
11	3.32.450.9811.11074.914208.	0	29	0	0	1	0	0			
12	5.59.181.4328.14691.702132.	1	14	0	0	0	0	1			
13	1.12.598.4401.58552.204494.	0	0	0	0	1	0	0			
14	2.15.495.9805.60625.122360.	0	2	0	0	1	0	0			
15	8.99.872.3445.65822.652253.	0	5	0	0	1	0	0			
16	1.14.509.4802.40034.666172.	0	0	0	0	1	0	0			
17	1.13.100.1278.83960.451047.	0	11	0	0	1	0	0			
18	4.66.160.7317.97197.525707.	0	0	0	0	1	0	0			
19	1.12.769.2516.24622.496675.	0	9	0	0	1	0	0			
20	9.46.105.3656.34784.141516.	0	0	0	0	1	0	0			

Step 4: Aggregation of Counts

After assigning the classification outcome for each row (case) aggregation was used to count the total number of:

TP : True Positives

TN : True Negatives

FP : False Positives

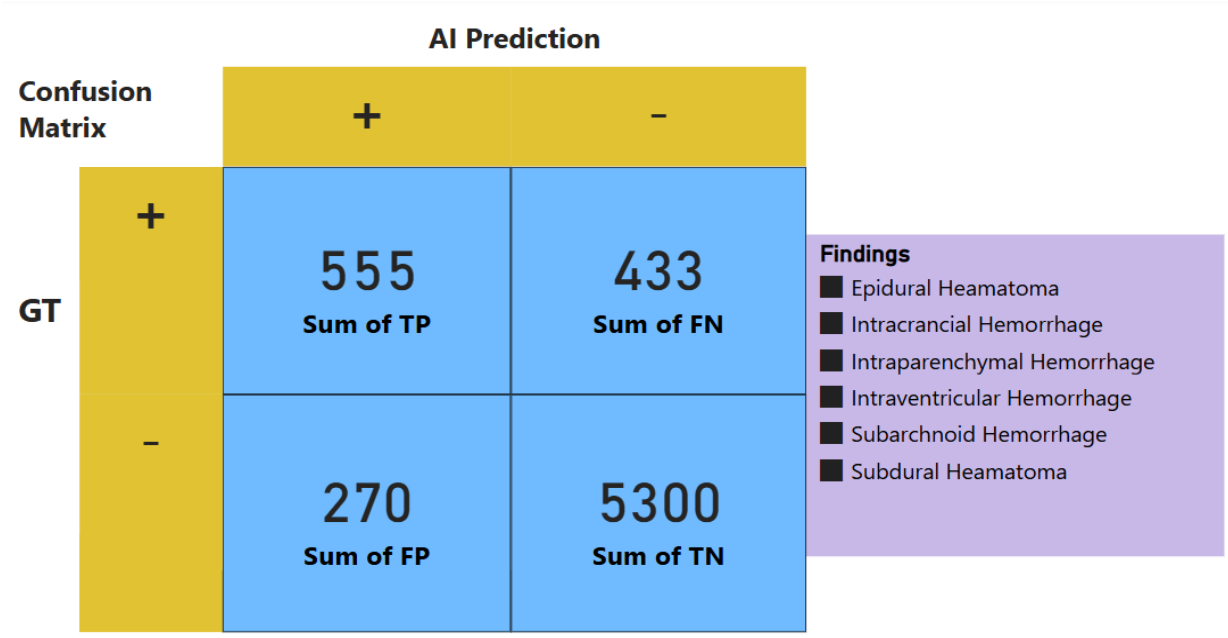
FN : False Negatives

F12

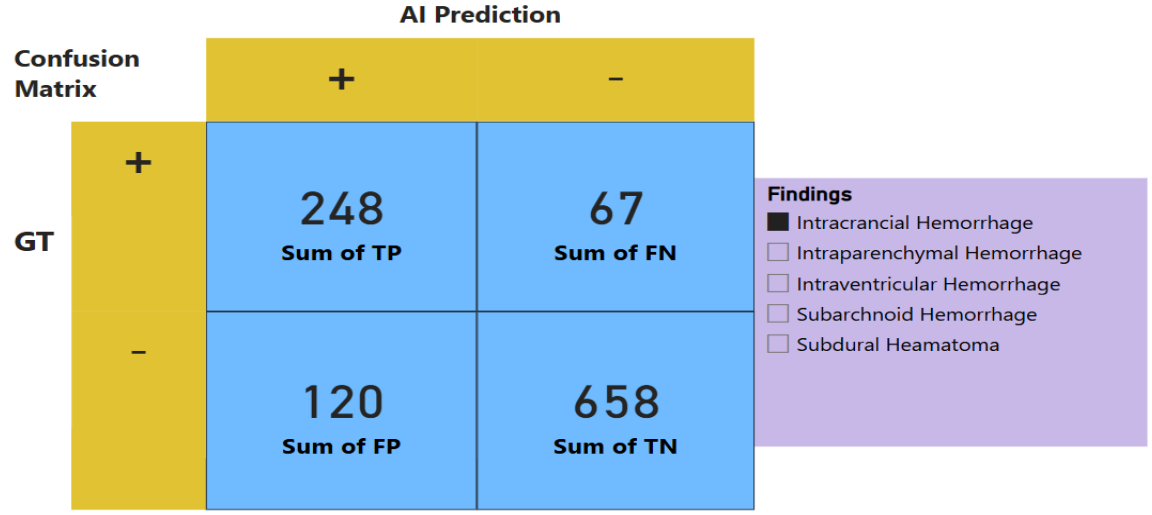
	A	B	C	D	E
1	Findings	TP	TN	FP	FN
2	Intracrancial Hemorrhage	248	658	120	67
3	Subarchnoid Hemorrhage	39	988	30	36
4	Intraventricular Hemorrhag	22	1041	14	16
5	Intraparenchymal Hemorrh	56	936	65	36
6	Subdural Heamatoma	190	627	22	254
7	Epidural Heamatoma	0	1050	19	24
8					
9					

These counts formed the basis of the 2x2 confusion matrix for each finding. The excel sheet was imported into Power BI and used for the creation of Confusion matrix for the visual representation along with the filter comprises of findings.

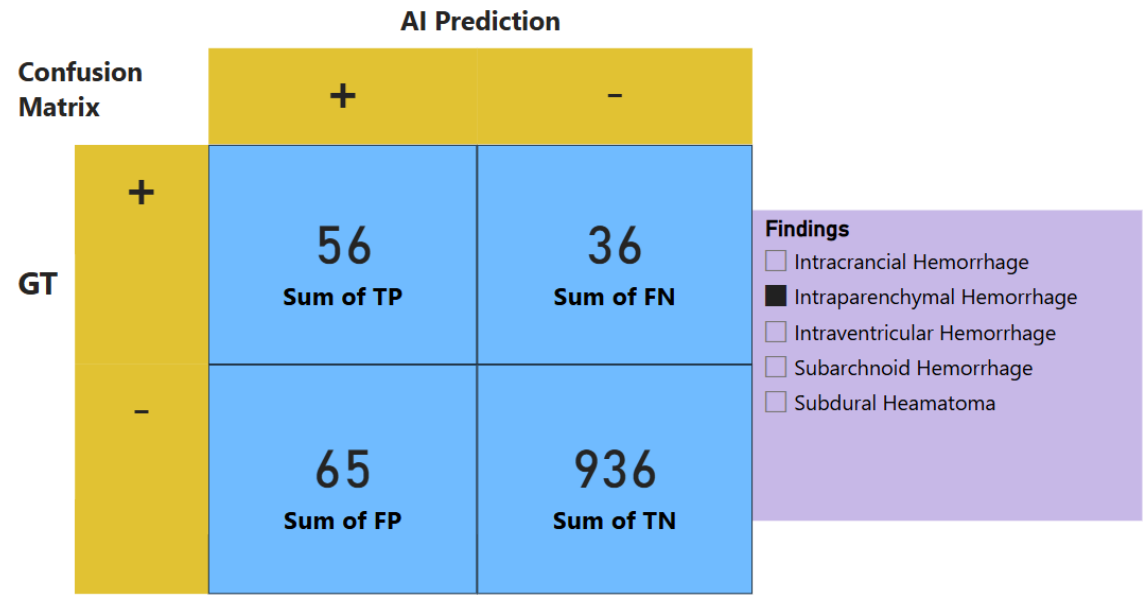
CONFUSION MATRIX FOR ALL THE FINDINGS



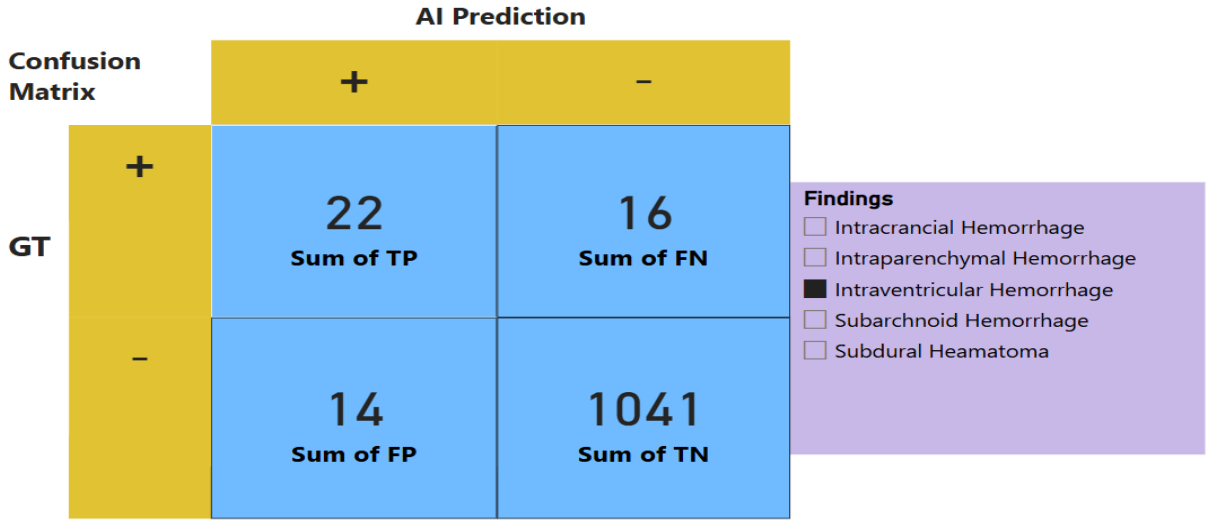
CONFUSION MATRIX FOR INTRACRANIAL HEMORRHAGE



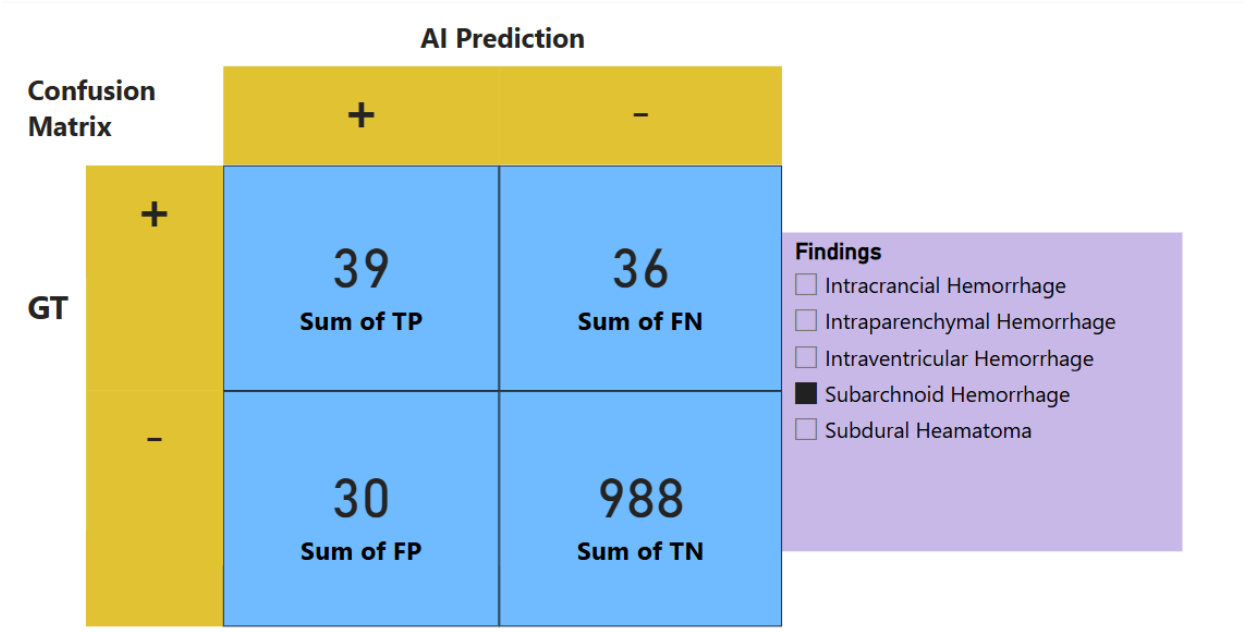
CONFUSION MATRIX OF INTRAPARENCHYMAL HEMORRHAGE



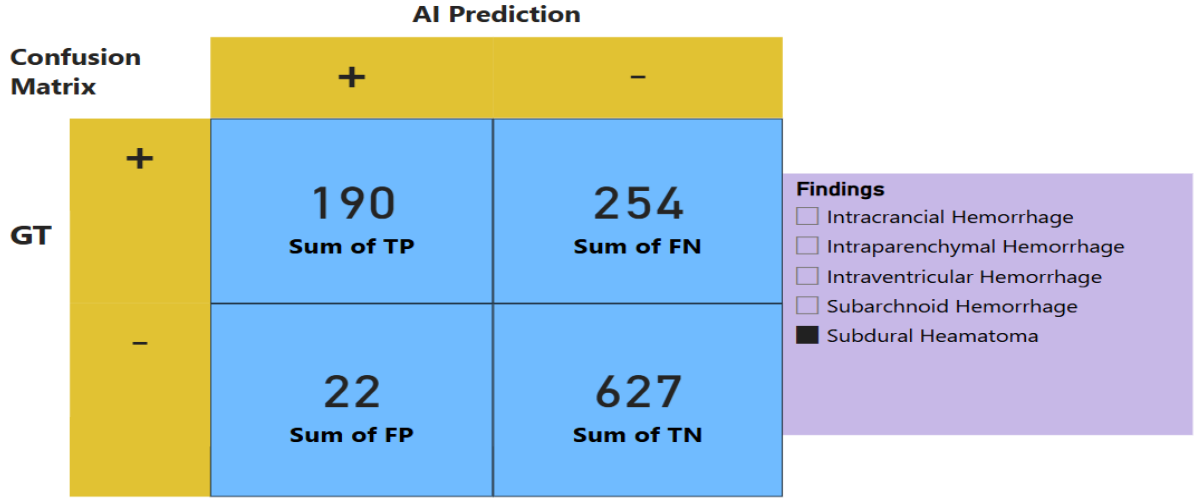
CONFUSION MATRIX OF INTRAVENTRICULAR HEMORRHAGE



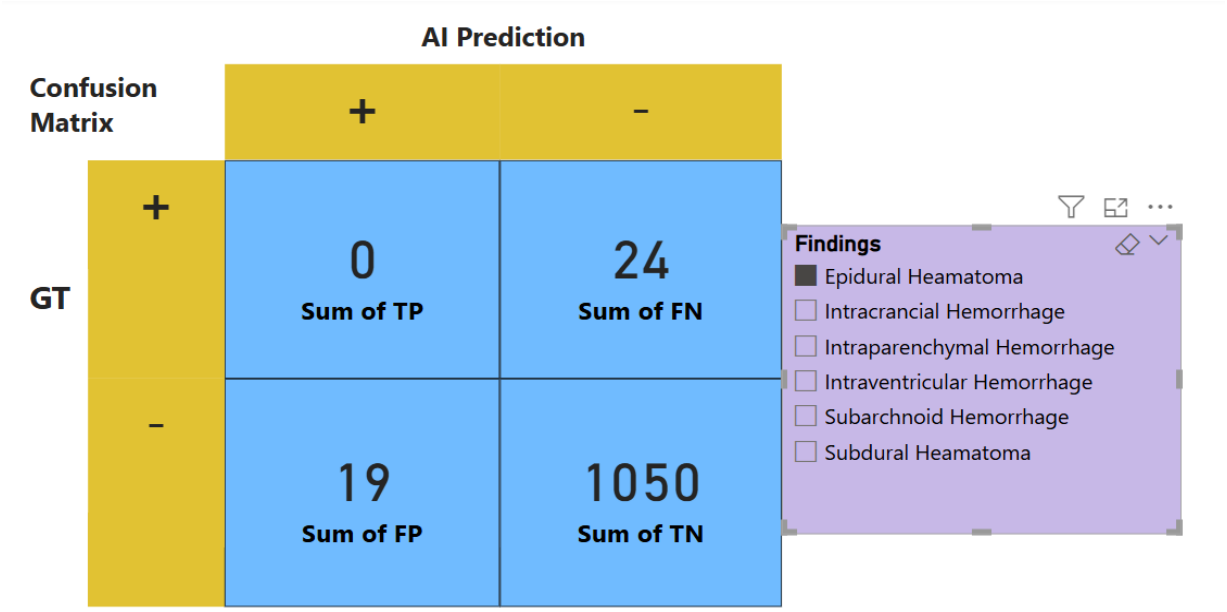
CONFUSION MATRIX OF SUBARCHNOID HEMORRHAGE



CONFUSION MATRIX OF SUBDURAL HEAMATOMA



CONFUSION MATRIX OF EPIDULAR HEAMATOMA



➢ **Calculation of Evaluation Metrics**

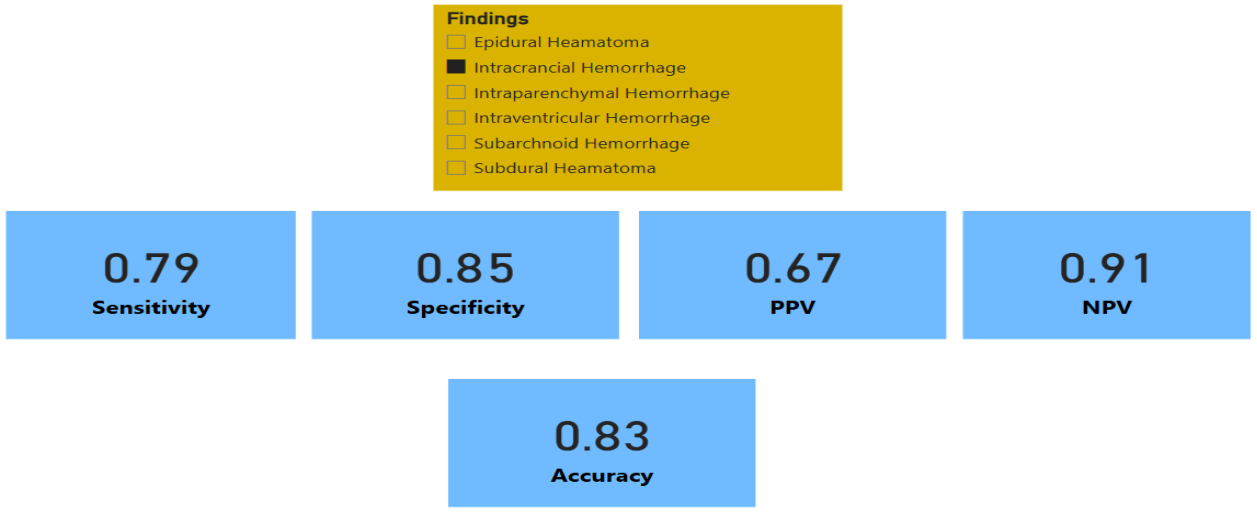
After generating confusion matrices for all six target findings, we computed essential performance metrics to assess the efficiency of the AI model in diagnosing. These indicators encompass Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Accuracy. We conducted all calculations in Microsoft Excel initially and then imported them into Power BI for interactive analysis and visualization.

Every metric was calculated using the aggregated values of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for each outcome.

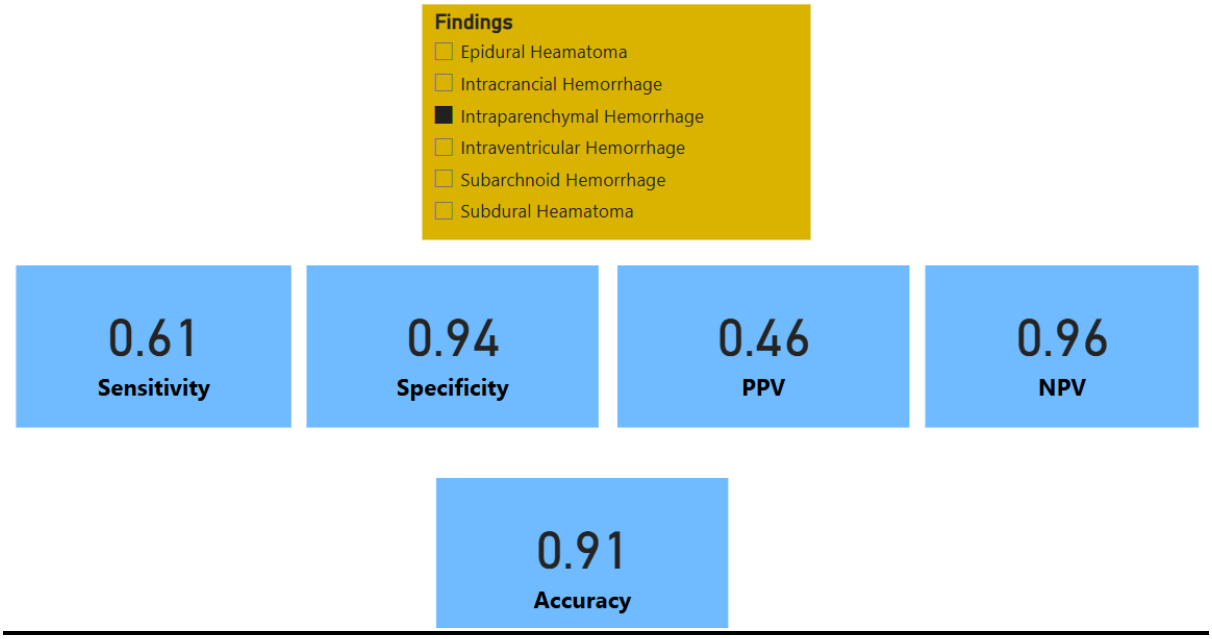
A	B	C	D	E	F
Findings	Sensitivity	Specificity	PPV	NPV	Accuracy
Intracrancial Hemorrhage	0.79	0.85	0.67	0.91	0.83
Subarchnoid Hemorrhage	0.52	0.97	0.57	0.96	0.94
Intraventricular Hemorrhag	0.58	0.99	0.61	0.98	0.97
Intraparenchymal Hemorrh	0.61	0.94	0.46	0.96	0.91
Subdural Heamatoma	0.43	0.97	0.90	0.71	0.75
Epidural Heamatoma	0.00	0.98	0.00	0.98	0.96

The metrices in Power BI interface:

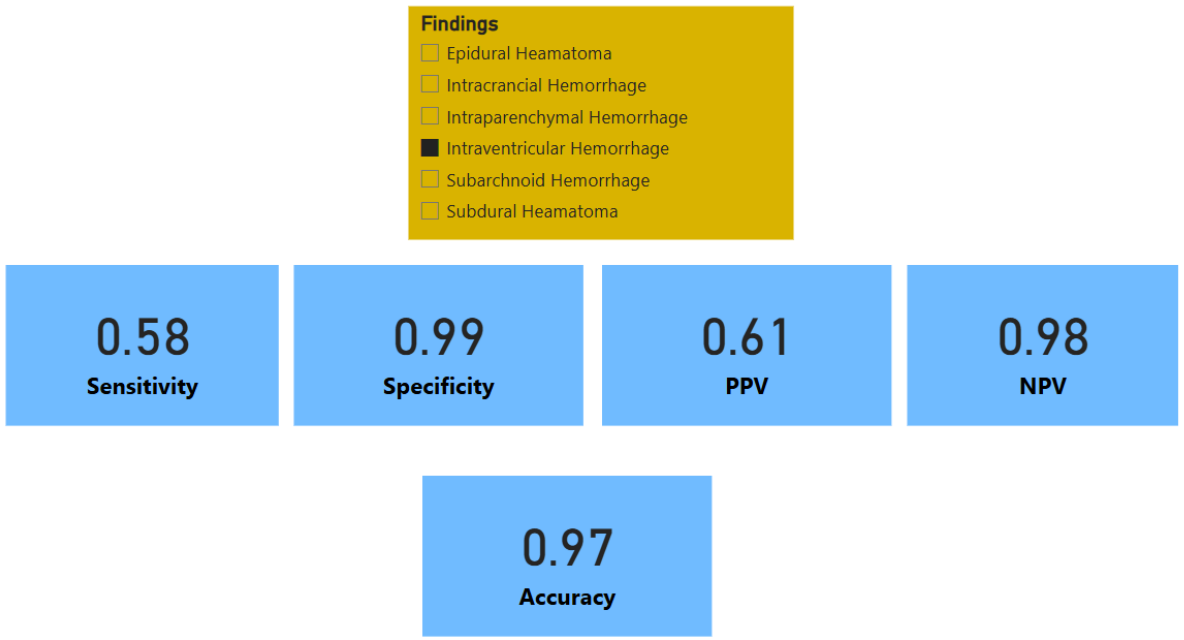
Metrices for Intracranial Hemorrhage



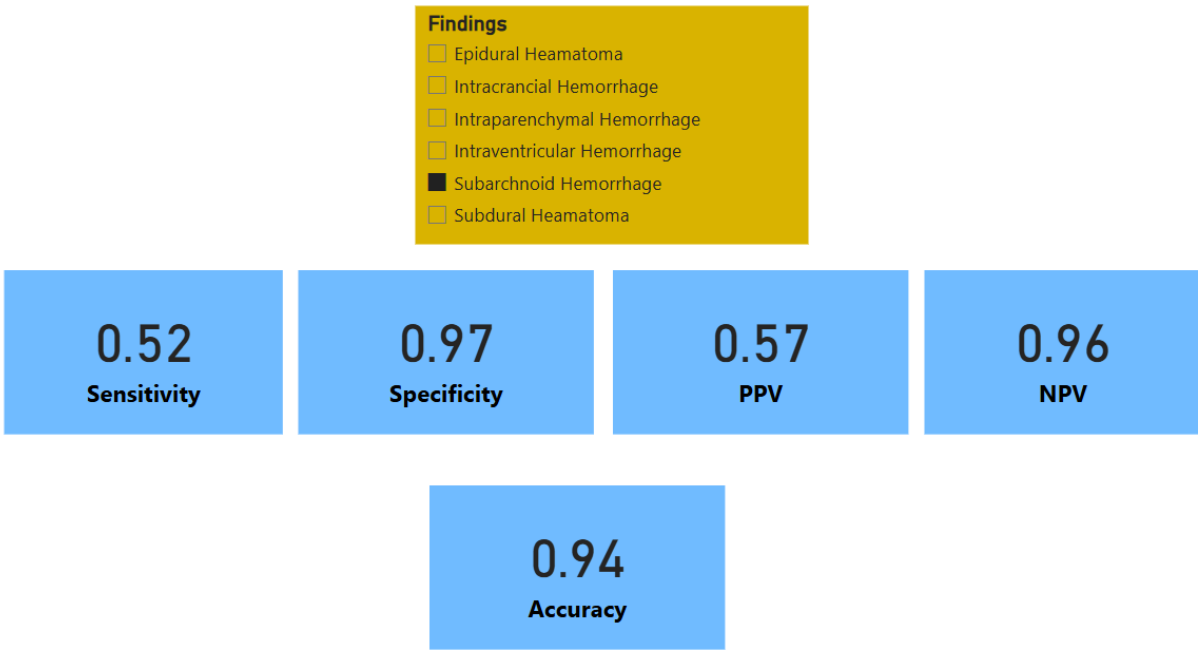
Metrices for Intraparenchymal Hemorrhage



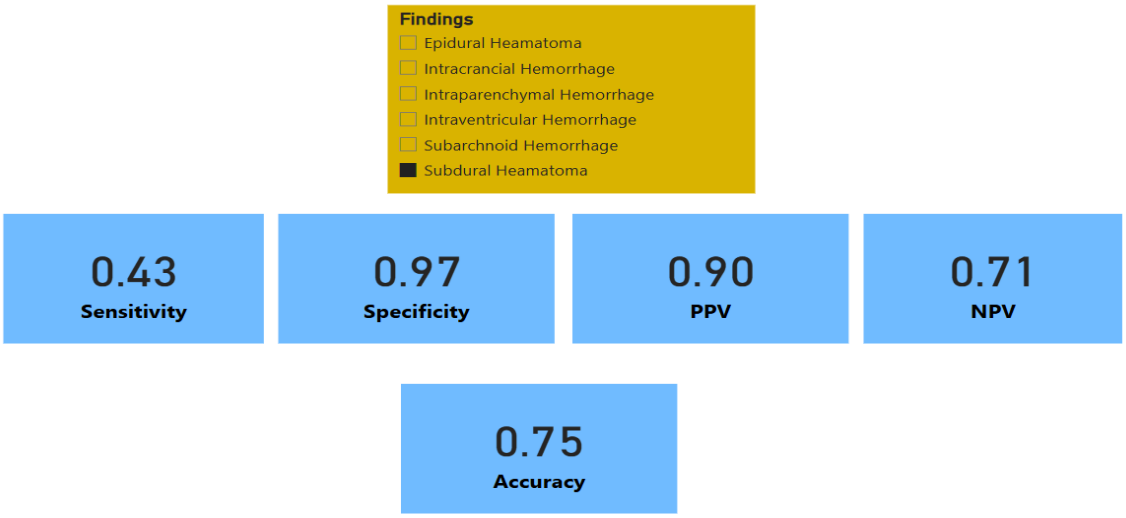
Metrices for Intraventricular Hemorrhage



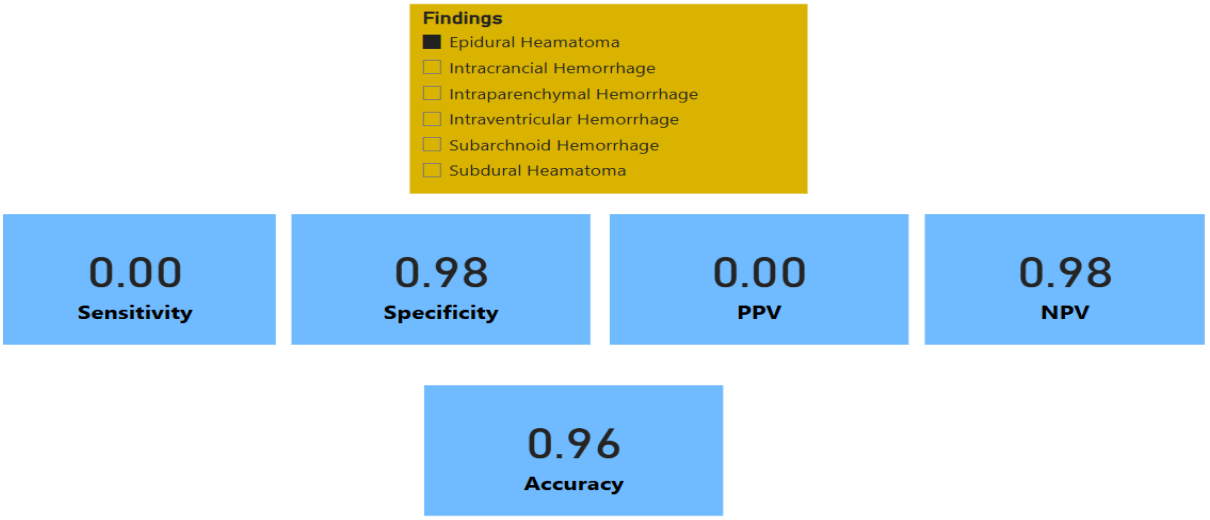
Metrices for Subarachnoid Hemorrhage



Metrices for Subdural Hematoma



Metrices for Epidural Hematoma



Comparative Analysis of the AI performance metrices for findings:

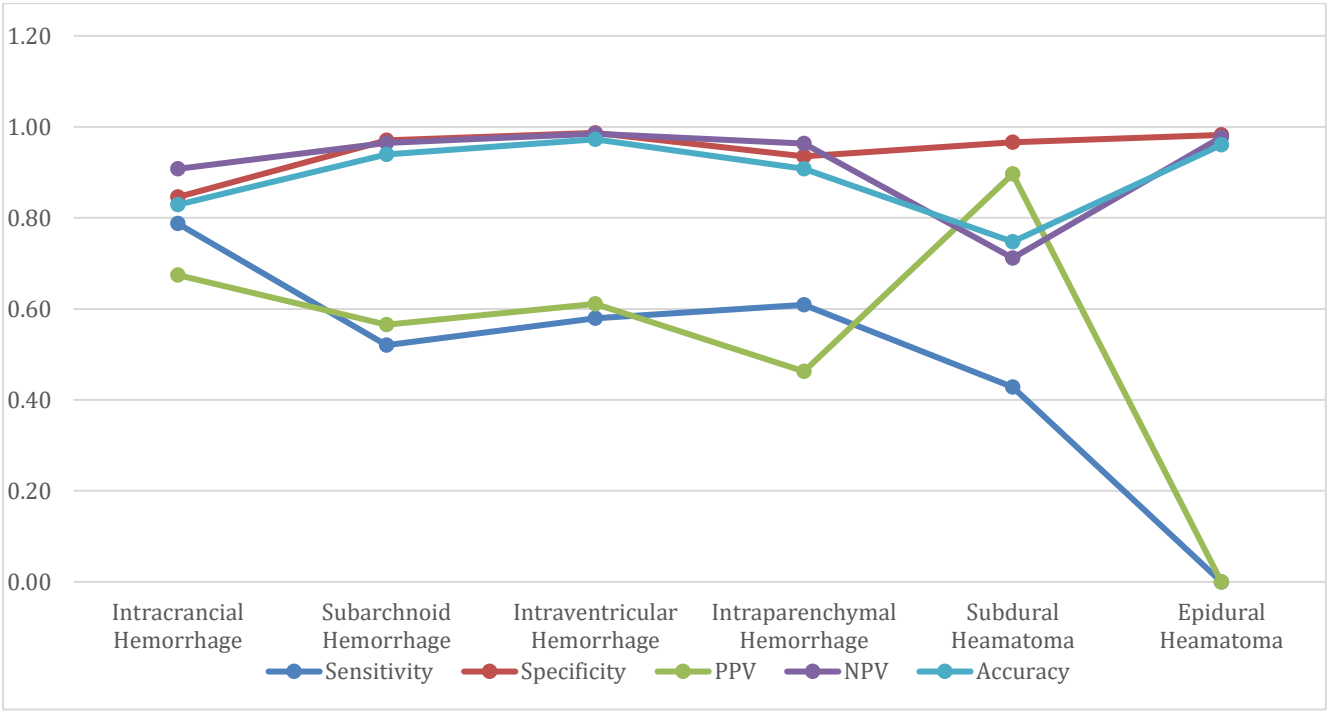


Fig 7

➢ **Plotting the Receiver Operative Characteristic Curve**

To further evaluate the ability of the AI model in classifying the findings into positive and negative classes ROCs are plotted and the area under it is a unique number which also shows the ability of discrimination at different threshold values. This analysis was conducted using Python in Google Colab, leveraging libraries such as scikit-learn, matplotlib.

The ROC curve is plotted between the True positive Rate and the False positive rates at different threshold values.

Intraparenchymal Hemmorhage

```python
from sklearn.metrics import roc_curve, auc
import matplotlib.pyplot as plt
import numpy as np

# Ground truth (1 = positive case, 0 = negative case)
y_true =[0,0,1,0,0,0,0,1,0,1,0,0,0,1,0,1,0,1,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,0,0,1,
#AI model predicted probabilities
y_scores = [0,22,61,0,0,0,35,60,1,100,35,3,2,99,0,91,0,100,8,0,0,35,99,14,83,8,94
# Compute False Positive Rate, True Positive Rate
fpr, tpr, thresholds = roc_curve(y_true, y_scores)

# Compute AUC
roc_auc = auc(fpr, tpr)

# Plot the ROC curve
plt.figure()
plt.plot(fpr, tpr, color='green', label=f'ROC curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], color='gray', linestyle='--')  # Diagonal line
plt.xlabel('False Positive Rate (1 - Specificity)')
plt.ylabel('True Positive Rate (Sensitivity)')
plt.title('ROC Curve for Intraventricular Hemorrhage finding')
plt.legend(loc='lower right')
plt.grid(True)
plt.show()
```

Fig 8

The codes which were used to plot the graph that is the ROC :

1. Importing Required Libraries

Python libraries as scikit-learn,matplotlib and numpy:

- roc_curve is used to calculate the false positive rate (FPR) and true positive rate (TPR).
- auc is used to calculate the area under the ROC curve.
- matplotlib.pyplot is used to visualize the curve.

2. Defining Ground Truth and Predicted Scores

The true labels (y_true) and predicted probability scores (y_scores) for each case were defined:
python

• y_true: Binary values that show the presence (1) or absence (0) of the finding based on radiologist reports.

• y_scores: Probability outputs from the AI model for each case.

3. Calculating ROC Curve Values
The True Positive Rate (TPR) and False Positive Rate (FPR) were calculated with the roc_curve() function.

fpr, tpr, thresholds = roc_curve(y_true, y_scores)

This function checks the model's performance at every possible classification threshold and returns:

- FPR: Proportion of negatives incorrectly classified as positives.
- TPR: Proportion of positives correctly classified.

4. Calculating AUC
The Area Under the Curve (AUC) was calculated from the ROC values:
python
roc_auc = auc(fpr, tpr)

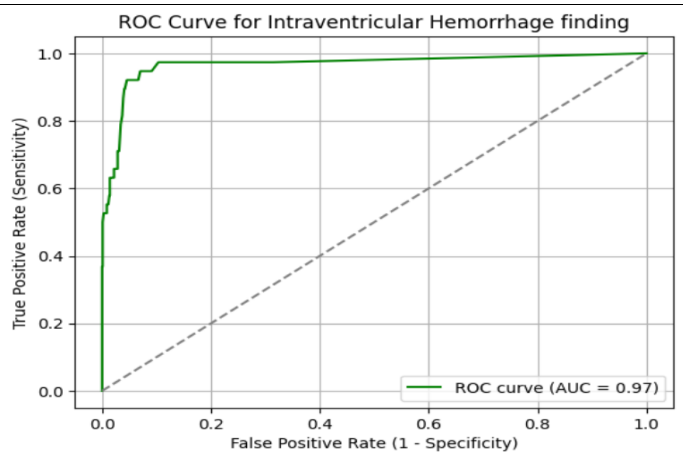
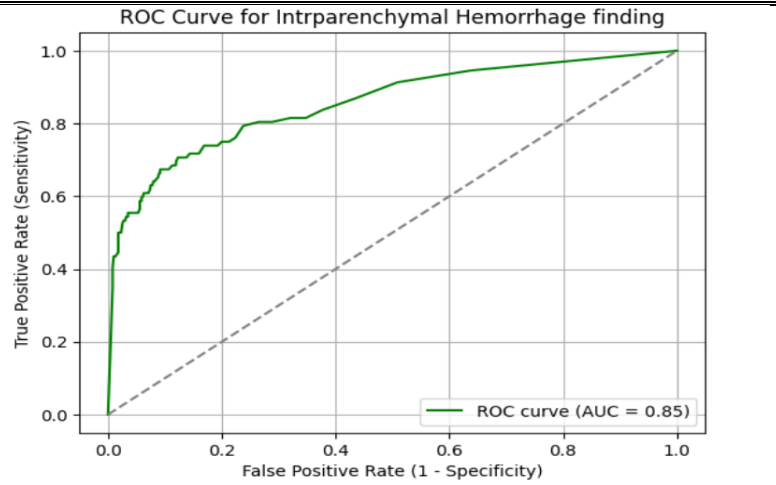
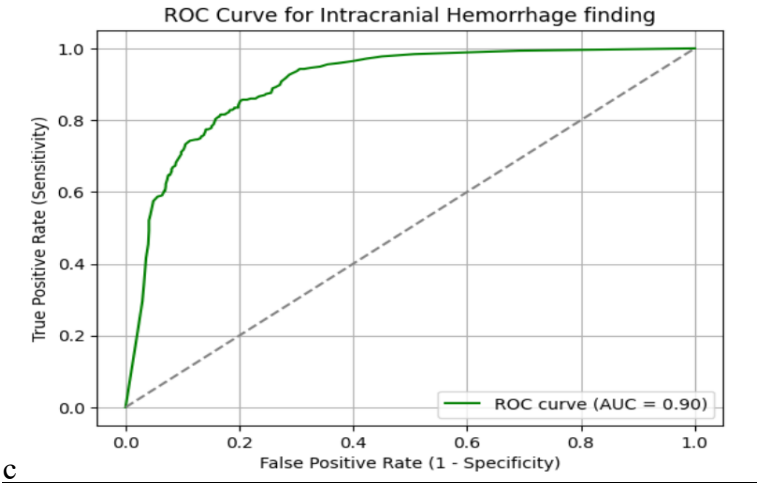
- AUC quantifies the overall ability of the model to discriminate between positive and negative cases.
- Values range from 0.5 (no discrimination) to 1.0 (perfect discrimination).

5. Plotting the ROC Curve
The ROC curve was plotted using matplotlib:
python

- The diagonal reference line represents random performance.
- The closer the ROC curve is to the top-left corner, the better the model's performance.

This same procedure was repeated for all other findings, each time using the corresponding ground truth and AI-generated probability scores. The resulting ROC curves and AUC values were stored as visual evidence of the AI model’s discriminative ability per pathology.

The ROC of different findings:

Findings	**ROC and AUC**
1. Intraventricular Hemorrhage	
2. Intraparenchymal Hemorrhage	
3. Intracranial Hemorrhage	

c

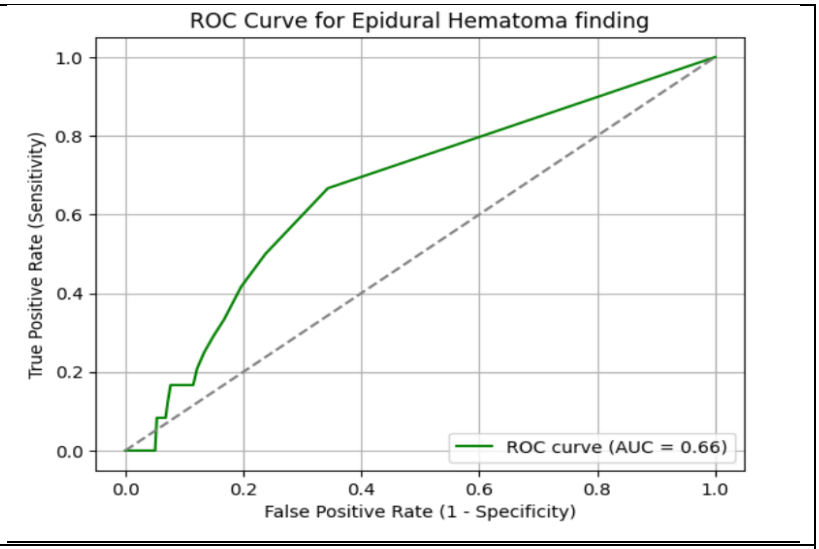
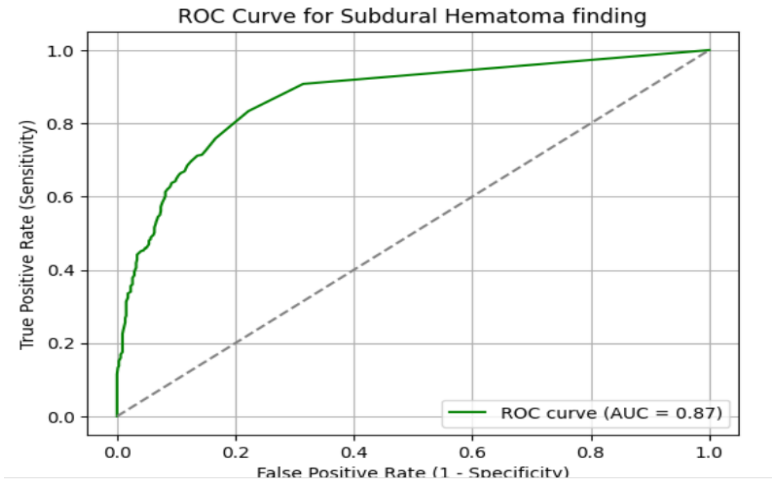
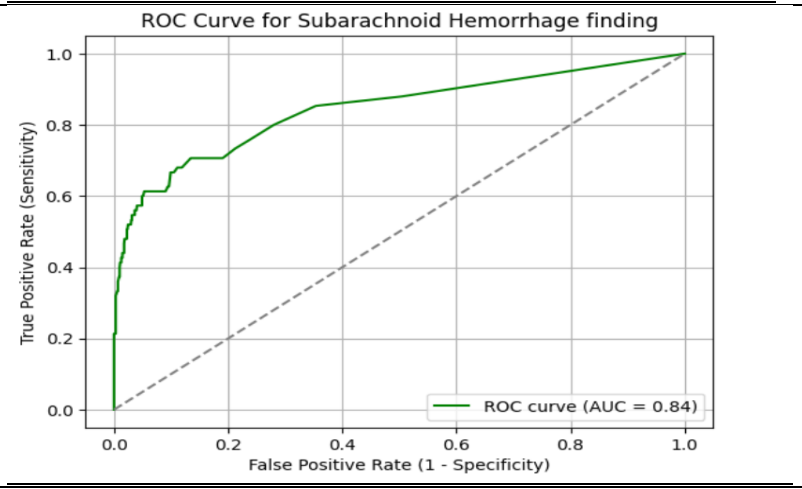
4. Epidural Hematoma	
5. Subdural Hematoma	
6. Subarachnoid Hemorrhage	

Table 1

4.2 DATA INTERPRETATIONS

The evaluation is for the six findings related to brain hemorrhages. Each finding was analyzed on the basis of the confusion matrix, Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and AUC.

Findings	TP	TN	FP	FN	Sensitivity	Specificity	PPV	NPV	Accuracy	AUC
Intracrancial Hemorrhage	248	658	120	67	0.79	0.85	0.67	0.91	0.83	0.9
Subarchnoid Hemorrhage	39	988	30	36	0.52	0.97	0.57	0.96	0.94	0.84
Intraventricular Hemorrhag	22	1041	14	16	0.58	0.99	0.61	0.98	0.97	0.97
Intraparenchymal Hemorrh	56	936	65	36	0.61	0.94	0.46	0.96	0.91	0.85
Subdural Heamatoma	190	627	22	254	0.43	0.97	0.90	0.71	0.75	0.87
Epidural Heamatoma	0	1050	19	24	0.00	0.98	0.00	0.98	0.96	0.66

Intracranial Hemorrhage

- **Sensitivity (0.79):**

  The AI model shows a high sensitivity by correctly identifying 79% of actual intracranial hemorrhage cases. This indicates a strong ability to detect positive findings, which is crucial in emergency settings where quick identification can greatly affect outcomes. However, the 21% of missed positive cases remain a concern. These false negatives mean that some hemorrhages might not be detected by the AI, which could delay diagnosis and appropriate treatment, especially in urgent cases like acute bleeds.

- **Specificity (0.85):**

  A specificity of 85% reflects the reliability of AI in the correct classification of non - hemorrhagic cases. This is valuable to reduce unnecessary alerts and research, ensuring that clinical attention is more focused. However, with 15% of negative cases marked incorrectly as positive, there is still a risk of false positives. This could lead to redundant images, an additional workload for radiologists or unnecessary anxiety of the patient due to excessive.

- **Positive Predictive Value (PPV) (0.67):**

  The AI's precision shows that 67% of the cases it predicts as positive are indeed true positives. This is moderately strong and indicates that when the AI raises a flag for

hemorrhage, it is correct often. Nevertheless, around one-third of positive predictions are incorrect, which could reduce diagnostic confidence among clinicians and result in inefficient resource allocation if not supported by radiologist review.

- **Negative Predictive Value (NPV) (0.91):**
  The high NPV of 91% signifies the model's strong reliability in confirming absence of hemorrhage. In clinical terms, when the AI indicates no hemorrhage, it is highly likely to be correct. This strengthens its role in triaging and ruling out findings in busy emergency or teleradiology workflows. However, the remaining 9% of false negatives still present a clinical risk and reinforce the need for human oversight in critical cases.
- **Accuracy (0.83):**
  A general precision of 83% shows that AI correctly classifies most cases, including both positive and negative predictions. This indicates a balanced model that performs well across different classes and improves the workflow's efficiency. However, the remaining 17% of misclassified cases should be viewed carefully. Erroneous diagnoses, whether missed hemorrhage or false alarms, can have clinical consequences.
- **AUC (0.90):**
  A strong area under the value of the ROC (AUC) curve of 0.90 shows that AI works very well to distinguish between cases of bleeding and non -bleeding. This confirms that the AI model is reliable in the evaluation of CT brain scans for intracranial hemorrhage. A high AUC is an important quality in medical AI systems because it suggests constant performance at various levels of decision, which supports its use in clinical decisions support systems.

Subarachnoid Hemorrhage

- **Sensitivity (0.52):**
  The AI demonstrates a modest sensitivity of 52%, meaning it detects just half of the actual subarachnoid hemorrhage cases.
  These false negatives show a serious clinical risk, as non-flagged SAH can rapidly increase criticality and lead to complications if not identified and managed in time.
- **Specificity (0.97):**
  This very high specificity is valuable for ruling out the condition with confidence and avoiding unnecessary alarms. Clinicians can trust that negative predictions for SAH are likely accurate, which can streamline workflow and reduce cognitive load in high-

volume emergency settings. However, this high specificity comes at the cost of sensitivity, indicating the model is conservative and potentially under-identifies actual positive cases.

- **Positive Predictive Value (PPV) (0.57):**
  A PPV of 57% suggests that slightly more than half of the AI-flagged SAH cases are accurate. While this may support the AI's use in raising red flags for further review, the remaining 43% of false positives can still result in over-investigation. In critical care settings, unnecessary interventions such as repeat scans, lumbar punctures, or specialist consultations may arise from these false alarms.
- **Negative Predictive Value (NPV) (0.96):**
  The NPV is very strong at 96%, indicating the AI is quite reliable when it predicts that SAH is not present. This helps in confidently ruling out cases and potentially de-prioritizing them for immediate radiologist review. It reduces unnecessary follow-ups for patients without hemorrhages and assists in managing clinical workload. Still, due to the low sensitivity, some actual SAH cases might slip through under the “negative” label.
- **Accuracy (0.94):**
  The overall accuracy is very high at 94%, which on the surface appears reassuring. However, because the dataset likely has an imbalance with far more negatives than positives, this high accuracy is skewed by the large number of true negatives. In such scenarios, accuracy can be misleading and may not reflect the model’s real-world reliability in identifying critical, rare findings like SAH.
- **AUC (0.84):**
  An AUC of 0.84 indicates that the model has decent ability to distinguish between SAH and non-SAH cases across various thresholds.
  While not as strong as the performance seen in intracranial hemorrhage (AUC = 0.90), this value is still acceptable and shows potential for clinical use with further calibration. The curve suggests that tuning the threshold may improve sensitivity at the cost of some specificity.

Intraventricular Hemorrhage (IVH)

- **Sensitivity (0.58):** The AI model identifies 58% of actual intraventricular hemorrhage cases. This moderate sensitivity implies that nearly 4 in 10 true cases are missed. Such

false negatives in IVH could delay the recognition of bleeding into the ventricular system, increasing the risk of complications such as obstructive hydrocephalus, elevated intracranial pressure, or deteriorating neurological status. The consequence is a tangible clinical risk in emergency and neurocritical care settings.

- **Specificity (0.99):** With near-perfect specificity, the model accurately excludes 99% of non-IVH cases. This level of accuracy is beneficial in reducing false alarms, which helps maintain radiologist confidence and prevents unnecessary neuroimaging or invasive follow-ups in patients without actual hemorrhage.
- **Positive Predictive Value (0.61):** A PPV of 61% indicates that just over half of the AI's positive IVH predictions are true positives.
- **Negative Predictive Value (0.98):** A high NPV means that when the AI predicts no IVH, it is correct in most cases. This allows clinicians to confidently rule out IVH. This is especially valuable in busy settings where triage is important.
- **AUC (0.97):** The model achieves excellent differentiating power between positive and negative IVH cases as well as suggests that the AI model performs well across thresholds and could be tweaked to reduce false negatives without significantly compromising specificity.

Intraparenchymal Hemorrhage (IPH)

- **Sensitivity (0.61):** At 61%, the model captures a mid-number of true IPH cases. However, nearly 40% remain undetected — a concerning figure, since missed parenchymal bleeds can worsen rapidly and may be associated with mass effect or midline shift. In clinical terms, these false negatives could lead to delays in neurosurgical intervention or anticoagulation reversal.
- **Specificity (0.94):** High specificity reflects the AI's strong capability to correctly identify negative cases, reducing the burden of unnecessary follow-up in patients without parenchymal bleeding.
- **PPV (0.46):** The AI demonstrates a low precision rate in detecting IPH — less than half of its positive predictions are accurate. This indicates a significant level of overcalling, which may overwhelm radiologists with false alarms, leading to fatigue or potential desensitization to alerts.
- **NPV (0.96):** A strong NPV supports the AI's value as a reliable tool to rule out IPH, giving clinicians confidence in negative reads, particularly in emergency settings

where time efficiency is critical.

- **AUC (0.85):** The overall discriminative ability is quite good, though lower than IVH or ICH. This suggests room for algorithm improvement, possibly through better feature representation of IPH or increased training samples of subtle or small-volume bleeds.

Subdural Hematoma (SDH)

- **Sensitivity (0.43):** The model detects only 43% of actual SDH cases, making it the weakest in terms of true positive identification. Over half of the true subdural hematomas are missed, presenting a substantial clinical risk — especially in elderly patients where chronic SDH may present subtly but progress rapidly. The high false negative rate undermines the model's reliability in detecting this finding autonomously.
- **Specificity (0.97):** Despite the low sensitivity, specificity remains very high at 97%. The model effectively rules out SDH in negative cases, preventing overtreatment or unnecessary patient anxiety when hemorrhage is truly absent.
- **PPV (0.90):** An excellent precision score of 90% means that when the AI flags a subdural hematoma, it is correct in most cases. This high reliability enhances clinician confidence in AI-positive cases, making it a useful supportive tool when paired with human review.
- **NPV (0.71):** The relatively low NPV signals that a negative prediction should not be fully trusted — nearly 30% of those could still be true positives. In practice, this implies that the AI should not be used to rule out SDH independently and must always be verified by expert radiologist review.
- **AUC (0.87):** A strong AUC reflects a solid overall ability to distinguish between SDH and non-SDH cases. The issue lies not in the model's capacity, but likely in its operating threshold or imbalance in the training dataset, warranting retraining or calibration to reduce missed detections.

Epidural Hematoma (EDH)

- **Sensitivity (0.00):** The AI fails to identify a single true case of epidural hematoma, resulting in 100% false negatives. This renders the model ineffective for detecting EDH and represents a critical flaw, especially because EDH requires immediate

surgical attention. The consequence is severe — undetected EDH can lead to rapid neurological deterioration and death, particularly if the classic "lucid interval" is missed.

- **Specificity (0.98):** While the model excels at correctly identifying negative cases, this is of little use when it cannot detect any actual positive findings. In other words, it confidently tells us a patient does not have EDH, but this cannot be trusted.
- **PPV (0.00):** Since there are no true positives, every positive prediction (if any) is incorrect. This completely undermines the AI's reliability for identifying EDH.
- **NPV (0.98):** Although this value appears high, it's misleading — since the model never detects any true positives, the NPV is artificially inflated. Clinically, this creates a dangerous false sense of security, and such a system must never be used to rule out EDH.
- **AUC (0.66):** With an AUC only slightly better than chance, the model has very limited ability to differentiate EDH cases. This points to a likely lack of representative EDH examples in training or failure to model the right features — and underscores the urgent need for algorithmic retraining.

Comparative Analysis

- **Subarachnoid hemorrhage showed the best sensitivity (0.89)** among all findings, indicating the AI is highly effective in detecting this critical condition. This makes it well-suited for acute emergency triage, where missing a subarachnoid hemorrhage could have severe consequences.
- **Intracranial hemorrhage (overall) had balanced performance** with high sensitivity (0.79), specificity (0.85), and an excellent AUC (0.90), suggesting the AI model is generally reliable in distinguishing between hemorrhage and non-hemorrhage cases, making it suitable for broad screening purposes.
- **Intraventricular hemorrhage (IVH) stood out for its specificity (0.99) and NPV (0.98)**, suggesting it is extremely reliable in ruling out disease when negative. However, its lower sensitivity (0.58) indicates it may miss nearly half of actual cases, highlighting a limitation in emergency diagnostic use.
- **Intraparenchymal hemorrhage (IPH) demonstrated strong negative predictive value (0.96)** and specificity (0.94), meaning false positives are low and negatives can be trusted. However, sensitivity (0.61) and precision (0.46) are only moderate,

showing that while false alarms are limited, many actual cases may be missed or misclassified.

- **Subdural hemorrhage (SDH) had the highest precision (0.90)**, implying that when the AI predicts a SDH, it is very likely correct. However, the low sensitivity (0.43) makes it poor at detecting the majority of true SDH cases, which reduces its utility as a first-line diagnostic tool.
- **Epidural hemorrhage (EDH) performed the worst across all metrics**, with sensitivity (0.00) and PPV (0.00), meaning it failed to detect any actual EDH cases. Despite high specificity (0.98) and NPV (0.98), the model is clinically unusable for EDH detection and requires significant retraining or redesign.
- **Precision was highest for SDH (0.90), followed by IVH (0.61) and ICH (0.67)**, indicating the AI is more accurate in some localized hemorrhage types. Lower precision in IPH (0.46) and SAH (0.45) means many flagged cases may be false alarms, burdening clinical workflow with unnecessary follow-ups.
- **AUC values generally remained high (>0.84) for most findings**, especially IVH (0.97) and ICH (0.90), reflecting good overall discriminative power of the model. EDH remained an outlier with an AUC of only 0.66, further emphasizing its poor classification capability.

Evaluation of AI's Differentiation Between Overlapping Pathologies

1**. Subarachnoid vs. Intraventricular Hemorrhage**

Overlap: Both involve blood in the cerebrospinal fluid (CSF) spaces and may appear similar on axial CT slices.

Performance Insight:

Subarachnoid: Sensitivity (0.52), PPV (0.57)

Intraventricular: Sensitivity (0.58), PPV (0.61)

Interpretation: The AI shows moderate difficulty in precisely distinguishing between these two CSF-space hemorrhages. Comparable sensitivity and precision indicate that some true cases may be missed or mislabeled due to image ambiguity.

2. **Subdural vs. Epidural Hematoma**

Overlap: Both are extra-axial hemorrhages but differ in shape (crescent vs. biconvex) and clinical implications.

Performance Insight:

Subdural: High PPV (0.90), but low Sensitivity (0.43)

Epidural: Sensitivity (0.00), PPV (0.00), despite Specificity (0.98)

Interpretation: The AI seems far better at detecting subdural hematomas than epidural, possibly because epidural cases are rare in the dataset. This poor generalization to epidural cases shows the AI struggles to learn the subtle but crucial morphological distinctions.

3. **Intraparenchymal Hemorrhage vs. Small Calcifications or Chronic Infarcts**

Overlap: Both may appear as hyperdensities within the brain parenchyma on CT.

Performance Insight: Moderate Sensitivity (0.61) and low PPV (0.46) suggest frequent false positives.

Interpretation: The AI likely overcalls dense lesions that mimic hemorrhage, particularly if chronic infarcts or basal ganglia calcifications are present.

4. **Global Insight on Overlap Management**

The Area Under Curve (AUC) values (mostly >0.84) suggest that while the AI can generally discriminate between positive and negative cases, specific differentiation between similar classes is still error prone.

5. DISSCUSION

The performance of the AI model evaluated in the studies that were also discussed in the ROL focuses majorly upon the validation and monitoring parts : Mair et al. (2022), Bar et al. (2022), and Dyer et al. (2022). The analysis focuses on similar strengths and limitations across studies.

Diagnostic Accuracy Across Hemorrhage Types

In this study, the AI model showed a strong overall performance in detection of intrachrinel hemorrhage (ICH) with the Sensitivity of 0.79, the uniqueness of 0.85 and AUC of 0.90. Subtractions such as Sabarachanoid Hemorrhage (SAH) and Intraventry Hemorrhage (IVH) also performed well. However, some subtypes, such as epidural hemorrhage (EDH) and intraperynthmal hemorrhage (IPH), were low accuracy due to false positivity. This analysis of sub -factories provides a clear picture of AI's abilities.

Mayor et al. (2022), recognizing e-expectes software, a high agreement was found between AI tools and radiologists for hemorrhagic and ischemic findings. However, his study did not provide the display details for each subtype. In contrast, this study provides a more detailed form on the clinical strength and weaknesses of AI for specific bleeding types. This information is particularly useful for evaluation of clinical reliability.

False Positives and Overcalling

A major challenge observed in this study was the tendency of the AI model to overcall certain hemorrhages, particularly EDH and IVH, leading to lower positive predictive values. An in-depth error analysis attributed many of these false positives to imaging features such as effusion, ossification, degenerative changes, post-surgical alterations, and normal anatomical variations that the AI misclassified as hemorrhages.

Bar et al. (2022) also highlighted concerns regarding false positives in their multicenter validation study. They noted reduced AI performance in peripheral centers due to variations in imaging protocols and scanner quality. These findings align with the current study’s conclusion that false positives stem from complex image presentations and insufficient training data diversity. However, the present study extends the discussion by identifying specific radiological features that lead to misclassification.

AI as a Workflow Tool

Despite the issue of false positives, the AI model in this study achieved high negative predictive values (e.g., NPV of 0.91 for ICH), indicating strong potential as a triage tool for ruling out negative cases. This can be valuable in high-volume emergency settings to prioritize cases that truly require immediate attention.

Dyer et al. (2022) similarly found their AI model could safely triage nearly half of non-contrast CT head scans as normal, allowing radiologists to focus on critical findings. Compared to their findings, this study supports the use of AI as a workflow support tool, particularly in flagging ICH and SAH cases. However, it emphasizes that AI alone should not be relied upon for comprehensive diagnosis, especially in complex or ambiguous cases.

Generalizability and Real-World Validity

The dataset used in this study comprises 1261 real-world CT brain cases from a Brazilian healthcare institution. The diversity of presentations—including chronic pathologies, surgical changes, and anatomical variants—makes this dataset highly representative of clinical complexity. This contributes to the external validity of the results and offers a rigorous test of AI robustness.

While Mair et al. and Bar et al. focused on multicenter validations, they did not elaborate on the clinical or pathological diversity within their datasets. This study adds value by not only reporting performance metrics but also investigating the clinical underpinnings of misclassifications, offering practical insights for AI deployment in routine practice.

Benchmarking Against Radiologists

Although this study did not include a direct head-to-head performance comparison between the AI and radiologists, it used radiologist-verified ground truth as the benchmark. This aligns with the methodology of Bar et al. (2022), where radiologist consensus was used for validation. However, Dyer et al. (2022) took a step further by reporting cases where the AI matched or outperformed radiologist accuracy.

To further strengthen the current study, future work could incorporate a more formal comparative framework, evaluating discrepancy patterns between AI outputs and individual radiologist interpretations.

6. CONCLUSION

- The study successfully validated the AI model's performance across 1261 real-world brain CT scans, focusing on six major types of intracranial hemorrhages
- Overall findings show the AI achieved strong sensitivity, specificity, and AUC for general ICH detection, with high NPV making it a reliable tool for ruling out critical cases.
- Subtype-level analysis revealed areas of excellence (e.g., SAH, ICH) and limitations (e.g., EDH, IPH), providing clear direction for refinement before deployment.
- The model demonstrated good generalizability across complex clinical cases, reinforcing its potential use in diverse radiology environments.
- Based on the results, the AI model shows promising utility as a triage and decision-support tool. However, targeted improvements and continuous feedback loops are recommended before full clinical onboarding.

7. REFERENCES

- Mair G, Morris Z, Lindley RI, Tsang A, Curry M, Cheah C, et al. External validation of e-ASPECTS software for interpreting brain CT in stroke. Ann Neurol. 2022;91(4):462–72. doi:10.1002/ana.26354

- Bar A, Saban M, Nagar H, Yehezkel G, Lichter I, Konen E, et al. Artificial intelligence model for stroke detection using non-contrast CT: multicenter external validation study. Insights Imaging. 2022;13:77. doi:10.1186/s13244-022-01188-y

- Dyer T, Chawda S, Alkilani R, Naunton Morgan T, Hughes M, Rasalingham S. Validation of an artificial intelligence solution for acute triage and rule-out normal of non-contrast CT head scans. Neuroradiology. 2022;64(4):735–43. doi:10.1007/s00234-021-02826-4

- Pham M, Quach V, Phan J, Shieh J, Gupta S, Sanghvi A, et al. Multicenter validation of an AI model for acute intracranial hemorrhage detection on non-contrast head CT. Front Neurol. 2023;14:1177723. doi:10.3389/fneur.2023.1177723.

- Abrigo JM, He W, Cheung KY, Kwok ML, Chan CP, Mak DKY, et al. Diagnostic accuracy of artificial intelligence in identifying intracranial hemorrhage and large vessel occlusion on computed tomography. Hong Kong Med J. 2023;29(1):14–21.

Swastik Mitra D report

ORIGINALITY REPORT

10%	8%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	www.frontiersin.org Internet Source	1%
3	medium.com Internet Source	1%
4	Submitted to University of Nottingham Student Paper	1%
5	core.ac.uk Internet Source	<1%
6	www.omicsdi.org Internet Source	<1%
7	www.researchgate.net Internet Source	<1%
8	Cobbina, Akosua. "A Lightweight Deep Learning Framework for Real-Time Detection of Industrial Traumatic Brain Injuries.", Southern Illinois University at Edwardsville Publication	<1%
9	Submitted to National Economics University Student Paper	<1%
10	www.ijert.org Internet Source	<1%
11	archive.rsna.org	