

ENHANCED BRAIN CANCER DETECTION USING A RADIOMICS-INFORMED HYBRID AUTOENCODER AND GBM MODEL

**A Thesis Submitted
In Partial Fulfillment of the Requirements
for the Degree of**

**MASTER OF TECHNOLOGY
in
INFORMATION TECHNOLOGY
by**

**ANANY KIRTI
(23/ITY/19)**

**Under the supervision of
DR. PRIYANKA MEEL**



**Department of Information Technology
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-110042. India**

May, 2025

ACKNOWLEDGEMENT

I have taken efforts in this thesis. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to **Dr. Priyanka Meel** for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing this thesis. I would like to express my gratitude towards the **Dr. Priyanka Meel (Department of Information Technology, Delhi Technological University)** for her kind cooperation and encouragement which helped me in the completion of this thesis. I would like to express my special gratitude and thanks to all the Information staff for giving me such attention and time.

Last but clearly not the least, I would like to thank The Almighty for giving me strength to complete the thesis on time.

Place: Delhi

ANANY KIRTI

Date: 28.05.2025



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, **Anany Kirti**, Roll No. 23/ITY/19 student of M.Tech (Information Technology), hereby certify that the work which is being presented in the thesis entitled “**Enhanced Brain Cancer Detection Using a Radiomics-Informed Hybrid Autoencoder and GBM Model**” in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Information Technology in the Department of Information Technology, Delhi Technological University is an authentic record of my own work carried out during the period from August 2023 to Jun 2025 under the supervision of Dr. Priyanka Meel, Assistant Prof, Dept of Information Technology. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE

Certified that **Anany Kirti** (Roll No. 23/ITY/19) has carried out their research work presented in the thesis titled “**Enhanced Brain Cancer Detection Using a Radiomics-Informed Hybrid Autoencoder and GBM Model**”, for the award of Degree of Master of Technology from Department of Information Technology, Delhi Technological University, Delhi under my supervision. The thesis embodies result of original work and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree for the candidate or submit else from the any other University/Institution.

Date: 28.05.2025

Dr. Priyanka Meel
(Supervisor)
Department of Information Technology
Delhi Technological University

ABSTRACT

Early and correct diagnosis is crucial to improve the survival rate of brain cancer patients. Conventional machine learning classifiers and qualitative radiological assessments are two instances of traditional diagnostic methods that often encounter feature extraction difficulty, possess excessive false positive/negative rates, and cannot deal with intricate spatial and morphological patterns of medical images. This research presents a radiomic strategy combining deep learning with other MRI image-based brain tumour detection methods to overcome these challenges. Our model integrates a strong feature extraction autoencoder with a gradient-boosting machine for classification. MRI images were preprocessed with picture scaling, normalization, and data augmentation from the Brain Cancer Detection MRI Images dataset to enhance the model's generalization. The Hybrid Autoencoder + GBM model outperformed standalone models with 96.8% accuracy, 97.4% precision, 96.2% recall, and 96.8% F1-score. ROC curve studies also validated its effectiveness, demonstrating almost perfect classification with an AUC of 0.99. The proposed hybrid model significantly reduces misclassification errors compared to convolutional neural networks (CNNs), GBM classifiers, and standalone autoencoders. These findings show how hybrid models and deep learning based on radiomics can enhance cancer diagnosis to the level where they can serve as a reliable alternative to traditional methods.

TABLE OF CONTENTS

AKNOWLEDGEMENT.....	ii
CANDIDATE’S DECLARATION	iii
CERTIFICATE.....	iv
ABSTRACT	v
LIST OF FIGURES	vii
LIST OF TABLES	viii
LIST OF ABBREVIATIONS.....	ix
CHAPTER 1.....	1
1.2. Problem Statement	3
1.3. Motivation.....	3
1.4. Objectives	5
CHAPTER 2.....	7
2.1. Research Gap	10
CHAPTER 3.....	11
3.1. Dataset Description	12
3.2. Handling Data Imbalance.....	13
3.3. Data Preprocessing.....	14
3.4. Model Building	18
3.5. Performance Metrics	25
3.6. Algorithm.....	26
CHAPTER 4.....	27
CHAPTER 5.....	35
5.1. Future Work.....	36
5.2. Clinical Applicability & Real-World Validation	37
5.3. Final Thoughts	37
REFERENCES	38
Appendix A	40

LIST OF FIGURES

Figure 1: Flowchart of the proposed hybrid model for brain cancer detection.....	11
Figure 2: Sample of dataset.....	13
Figure 3: The basic structure of an autoencoder network	19
Figure 4: Gradient Boosted Trees	20
Figure 5: Confusion Matrix for Hybrid Autoencoder + GBM.....	28
Figure 6: Confusion Matrix for Autoencoder	28
Figure 7: Confusion Matrix for GBM.....	29
Figure 8: Confusion Matrix for CNN	29
Figure 9: ROC Curve for Hybrid Autoencoder + GBM	30
Figure 10: ROC Curve for Autoencoder.....	30
Figure 11: ROC Curve for GBM	31
Figure 12: ROC Curve for CNN.....	31
Figure 13: Performance Metrics	32
Figure 14: ACCTHPA-2025	40
Figure 15: ICISS-2025	41
Figure 16:ICISS-2025 Acceptance Letter.....	42
Figure 17:Payment Reciept of ICISS-2025 Conference	43

LIST OF TABLES

Table 1. Summary of Recent Studies on ML and DL Techniques in Radiomics for Cancer Diagnosis	8
Table 2: Hybrid Autoencoder and GBM Algorithm	26
Table 3: Visualization of Predictions.....	32
Table 4: Comparing the hybrid model against more recent deep learning	33

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AUC	Area Under the Curve
BLEU	Bilingual Evaluation Understudy
CAD	Computer-Aided Diagnosis
CNNs	Convolutional Neural Networks
CT	Computed Tomography
DCE-MRI	Dynamic Contrast Enhanced Magnetic Resonance Imaging
DL	Deep Learning
DT	Decision Tree
DWI	Diffusion-Weighted Imaging
F1-score	Harmonic Mean of Precision and Recall
GBM	Gradient Boosting Machine
GRUs	Gated Recurrent Units
LSTMs	Long Short Term Memorys
LR	Logistic Regression
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MSCOCO	Microsoft Common Objects in Context
NLP	Natural Language Processing
PDAC	Pancreatic Ductal Adenocarcinoma
PET-CT	Positron Emission Tomography – Computed Tomography
ReLU	Rectified Linear Unit
RF	Random Forest
RNNs	Recurrent Neural Networks
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SPICE	Semantic Propositional Image Caption Evaluation
SVM	Support Vector Machine
US	Ultrasound
ViTs	Vision Transformers
VOI	Volume of Interest

CHAPTER 1

INTRODUCTION

Cancer is one of the deadliest illnesses in the world and a major contributor to death rates. Skin cancer, ovarian cancer, brain cancer, lung cancer, and more than 200 other types of cancer are all possible: prostate, breast, and colon cancers, as well as leukemia and other malignancies [1][2]. A higher risk of getting cancer is associated with both environmental (such as chemicals, alcohol, tobacco smoke, and radiation) and genetic (such as genetic mutations and autoimmune disorders) variables. The body's abnormal cells multiply and spread out of control in cancer, a complicated and multidimensional collection of disorders. Millions of individuals worldwide struggle with this severe health problem, which also significantly raises the expense of healthcare worldwide [3]. To reduce the morbidity and mortality linked to cancer, effective detection and treatment are essential. Since there are several varieties of cancer, each has distinct traits and behaviours. As a result, cancer diagnosis and treatment need a sophisticated, customized approach that addresses the disease's heterogeneity [4]. Among these, brain cancer—particularly malignant brain tumours—presents a significant diagnostic challenge due to the complexity of the brain's structure and the subtlety of early-stage symptoms. Brain tumours can lead to severe neurological impairment and are associated with high mortality rates if not detected early.

1.1. Overview

Significant progress has been made in the early diagnosis of cancer development as a result of advancements in medical technology, study, and awareness. However, conventional medical picture interpretation often depends on radiologists' .

Radiomics is a new approach to quantitative image analysis that can improve accuracy and reliability by eliminating subjective and qualitative evaluations. This method was described as extracting high throughput features from medical images by [5] in 2012. Radiomics can extract ROIs and volume-of-interest (VOIs) from imaging data by combining specific imaging modalities with automated or semi-automatic algorithms. These characteristics primarily fall into four types: morphological, first-level, second-level, and textural [6]. It is possible to evaluate correlations between these features and clinically significant outcomes to anticipate endpoints for certain cancers. Modern imaging modalities in radiomics, including mammography, CT, US, and PET-CT, have mostly supplanted their predecessors. However, the variability brought about by variations in imaging protocols, scanners, segmentation strategies, and dataset heterogeneity is a significant drawback of radiomics. These variations can impact feature reproducibility, model generalizability, and the precise interpretation of intricate patterns in sizable datasets.

Over the past few decades, machine learning (ML) has emerged as a potentially game-

changing strategy for deciphering intricate patterns in massive datasets, significantly advancing healthcare automation in the detection and prediction of cancer. Medical image segmentation, the analysis of enormous collections of many positions, including digital histopathology slides, complex genetic profile interpretation, and others may now be automated because to machine learning's computing capacity and adaptability [7].

Integrating ML techniques into radiomics has created novel opportunities for predictive modeling in cancer diagnosis and prognosis. Machine learning algorithms can examine intricate and high-dimensional radiomic information, discern significant patterns, and construct reliable models capable of predicting cancer kinds, stages, treatment responses, or patient survival outcomes. By analyzing historical data, these models can assist doctors in making objective, data-driven judgments. Combinations of machine learning algorithms with radiomics data have been utilized in recent years to forecast prognostic information for patients with a variety of malignancies, including colorectal cancer [8], small cell lung cancer [9], and melanoma [10].

The main reason radiomics are used is to support personalized medicine by fitting treatments and procedures to each unique patient. The use of SVMs, CNNs, gradient boosting and random forests in supervised and deep Learning has helped to improve both diagnostic accuracy and reproducibility [11].

It combines an unsupervised feature extractor with a classifier to develop a dependable non-invasive imaging-based prediction mode for cancer. This method attempts to improve accurate predictions and how they work generally by dealing with the difficulties created by many and repeated features in radionics data. When put against classic models, the hybrid approach performs better in many other types of imaging. Further progress in personalized cancer diagnostics is possible because retrieved latent characteristics give insights into the clinical significance of radiomic patterns.

The key contributions of this research include:

- To introduce a novel fusion of autoencoder-based unsupervised feature extraction with GBM classification to handle high-dimensional and redundant radiomics features effectively.
- The goal is to demonstrate superior diagnostic accuracy and model generalizability compared to conventional ML models for use with various medical imaging modalities.
- To establish meaningful correlations between latent radiomic features and clinical outcomes, offering valuable insights into the potential biological relevance of these patterns for personalized cancer diagnosis.

The remaining chapters will be structured as follows: Chapter 2 delves into the associated work, which involves studying the literature for our suggested approach and identifying any gaps in the research. A brief overview of the proposed sections provides models. The recommended approach is described in Section 4, which

comprises a dataset description, preprocessing, segmentation, and model. The experiment is detailed, the findings are analyzed, and comparisons are made in Chapter 4. The study's findings and recommendations for the future are presented in Chapter 5.

1.2. Problem Statement

The diagnosis of brain cancer remains a significant clinical challenge due to the complexity of the brain's anatomy, the variability in tumour morphology, and the often-subtle nature of early symptoms. Traditional diagnostic practices rely heavily on qualitative image interpretation by radiologists, which is inherently subjective and limited in its ability to capture high-dimensional radiological patterns. Although conventional machine learning approaches have been introduced to automate and enhance diagnosis, they often fall short in accurately distinguishing between malignant and non-cancerous tissues, particularly when dealing with redundant and complex radiomic data.

Moreover, many existing models either rely on handcrafted feature extraction techniques that may overlook deep structural characteristics, or they employ deep learning models that, while powerful, tend to be opaque and require vast amounts of labeled data for effective training. Standalone classifiers like CNNs and GBM have demonstrated limited performance due to issues such as overfitting, reduced interpretability, and inadequate feature representation in high-dimensional medical imaging contexts.

This research addresses the need for a more robust and interpretable diagnostic framework by introducing a hybrid model that combines the unsupervised deep feature learning capabilities of an autoencoder with the classification strength of a Gradient Boosting Machine. The goal is to enhance feature extraction from MRI images while maintaining high classification accuracy, thereby reducing misclassification rates and improving diagnostic reliability. This study seeks to fill critical gaps in the current state of brain tumour detection by providing a scalable and generalizable solution that better aligns with the demands of clinical application in oncology.

1.3. Motivation

Brain cancer remains a serious and often deadly illness that attacks the brain and spinal cord. Since malignant brain tumours are often discovered late and because diagnosis is limited, patients generally do not recover well. Central nervous system tumours are now a growing cause of sickness and death due to cancer on a worldwide scale. Although medical imaging technology is growing quickly, using radiologists to manually assess images still presents with major problems — it is subjective, results in diverse observations, takes more time and has a hard time detecting complicated tumours in the images. Because brain tumours have unique structures and complex spatial locations, it is more difficult to find them.

It is often not possible to tell the difference between a malignant tumour and healthy

tissue using regular MRI analysis at the first stage. In many cases, the first symptoms are unclear and mild, so doctors may not evaluate patients until later. This points out that we need to develop early, accurate and automatic tools to help doctors make judgments and increase patient outcomes. Under these circumstances, radiomics becomes an important method to automate the identification of many quantitative traits from pictures in medical scans.

Radiomics helps to link medical imaging with personalized medicine by changing images into data that can be analyzed. Even so, radiomic data is often packed with redundant information which creates issues with choice of features, understanding models and their general application. Even though Support Vector Machines, Decision Trees and Random Forests work well in several domains, they don't perform well with these complex, multidimensional datasets. They tend to fit poorly to new data and usually do not work well in different patient settings.

Even so, while CNNs have done incredibly well with image classification, their use in medical imaging is restricted by some issues. Since CNNs are not easy to interpret, they do not fit well in clinical settings where it's necessary to clearly and transparently explain model decisions. Moreover, CNNs need large sets of labeled information which are typically limited in medical fields.

Because of these major limitations, this research hopes to use hybrid solutions that blend the strong learning of deep models with the advantages of explaining and surviving in machine learning. More precisely, the study reveals how unsupervised deep features obtained from a convolutional autoencoder can be combined with a Gradient Boosting Machine (GBM) to improve the accuracy of classification. Using an autoencoder, the system finds important features hidden in tumor images which the GBM classifier applies to accurately separate malignant images from benign ones.

As compared to CNNs, GBM and autoencoders, the hybrid model showed greater accuracy, precision, recall, F1-score and AUC scores in assessing classification tasks. More precisely, the model was able to achieve a very high accuracy of 96.8% and an AUC score of 0.99. In addition to achieving high numbers, the hybrid model reduced the risks of classifying someone with cancer when they are cancer-free and vice versa, compared to other models.

Besides aiming for better technological performance, this work is also inspired by the need to achieve patient benefits. Using a model that is both accurate and easy to understand, oncologists and radiologists can make better choices, speed up the diagnosis process, select proper treatments and improve patient care. The approach used in this paper may open doors for more personalized ways to diagnose people with brain tumors.

The main objective is to help advance computer aided-diagnosis tools by showing the benefits of adding radiomics with new learning techniques which allows us to identify brain cancer in its early stages when treatment is easiest.

1.4. Objectives

This research aims to design and prove the reliability, scalability and clarity of a diagnostic model to detect brain cancer in MRI images. Because accuracy in medicine is important and we depend more on data for medical choices, using systems that use both deep learning and machine learning is crucial. The aim of this study is to satisfy that need by blending a convolutional autoencoder to find hidden patterns with the highly accurate predictions of a GBM.

Brain tumour diagnosis through imaging remains a major challenge due to the complexity of brain tissue structures, the subtlety of tumour indicators in early stages, and the limitations of human visual interpretation. As such, this research aims not only to improve classification performance but also to enhance the clinical utility of artificial intelligence in radiology through a methodologically sound, evidence-based approach.

The specific objectives of this research are outlined below:

1. To Develop a Hybrid Radiomics-Informed Diagnostic Framework
 - Design a hybrid architecture that integrates a convolutional autoencoder for deep, unsupervised radiomic feature extraction with a Gradient Boosting Machine classifier to distinguish between malignant and non-cancerous MRI brain scans.
 - Ensure that the architecture is modular, interpretable, and adaptable to potential future extensions or deployments in varied clinical settings.
2. To Address Challenges of High-Dimensional and Redundant Data
 - Tackle the inherent complexity of radiomics data—characterized by its high dimensionality and feature redundancy—by enabling dimensionality reduction through latent space encoding.
 - Enhance the signal-to-noise ratio by eliminating irrelevant or less significant features and focusing only on those that carry discriminative power for tumour classification.
3. To Improve Diagnostic Accuracy, Generalizability, and Robustness
 - Achieve higher classification metrics (accuracy, precision, recall, F1-score, and ROC-AUC) as compared to traditional machine learning classifiers and standalone deep learning models.
 - Test the robustness of the hybrid model against overfitting and ensure its performance across different patient subsets using cross-validation techniques.

4. To Compare Model Performance Against Baseline Algorithms

- Implement baseline models including standalone GBM, CNN, and Autoencoder, and compare their diagnostic capabilities against the proposed hybrid model.
- Present a detailed performance analysis through confusion matrices, ROC curves, and metric-based evaluation to establish the superiority of the hybrid approach.

5. To Reduce False Positives and False Negatives in Diagnosis

- Significantly reduce Type I and Type II classification errors, which are critical in the medical domain due to their potential to result in either unnecessary treatment or missed diagnoses.
- Establish the model's reliability by demonstrating a near-perfect balance between sensitivity and specificity.

6. To Establish Clinical Relevance and Translational Value

- Demonstrate how extracted latent features correlate with known radiological patterns, thereby providing insights into potential biological relevance.
- Ensure that the proposed model is not only technically sound but also clinically interpretable and actionable, enhancing its suitability for integration into computer-aided diagnostic (CAD) systems.

7. To Contribute to the Advancement of Personalized and Predictive Oncology

- Support the broader vision of personalized medicine by offering a tool that can be adapted to individual patient data for tailored diagnosis.
- Lay the groundwork for future extensions where similar models can be applied to other types of cancers or integrated with genomic, histopathological, or clinical data for multi-modal analysis.

CHAPTER 2

LITERATURE REVIEW

The integration of radiomics with machine learning (ML) and deep learning (DL) methodologies has become increasingly prominent in the field of medical imaging, particularly for cancer diagnosis and prognosis. A review of recent studies reveals a significant evolution in techniques, data usage, and model architectures aimed at improving diagnostic accuracy, especially for complex and heterogeneous cancers like brain tumours.

1. Radiomics in Medical Imaging Radiomics, introduced in 2012, offers a systematic approach to extracting a large number of quantitative features from medical images, including morphological, first-order, second-order, and textural descriptors. These features are critical for capturing the subtle patterns within tumours that are often overlooked by traditional imaging analysis. Radiomics has demonstrated potential in predicting treatment outcomes, tumour grade, and survival rates across various cancer types. However, challenges such as feature redundancy, variations in imaging protocols, and lack of standardization still limit its clinical adoption.[6]

2. Traditional Machine Learning Techniques in Radiomics A variety of ML methods have been applied to look at radiomics data for identifying cancer. Researchers have demonstrated good results with decision trees (DT), random forests (RF), support vector machines (SVM) and logistic regression (LR). In terms of breast cancer detection, DT and RF displayed accuracies of 96% and 95% when supported by the right radiomic features. At times, these models operate poorly in complex input spaces, as their data is small and may be biased [12].

3. Deep Learning Approaches and Limitations CNNs have been especially recognized for their capacity to pick out different levels of detail from raw images. These CNNs can accurately detect cancers of the pancreas and brain. Still, because radiomics systems cannot be easily interpreted, use large unclear datasets and cannot describe radiomic features, they are not suitable for clinical environments where being clear and transparent matters [24].

4. Hybrid Models for Enhanced Diagnostic Performance Recent trends in cancer imaging have explored the fusion of unsupervised learning for feature extraction with robust ML classifiers. Hybrid models that combine deep autoencoders with gradient boosting frameworks have shown enhanced performance by leveraging the representational power of deep networks and the interpretability of tree-based models. Such architectures are better equipped to manage redundant and complex feature spaces while maintaining diagnostic precision. In my study, the hybrid Autoencoder + GBM model achieved superior metrics (Accuracy: 96.8%, AUC: 0.99)

over individual models, affirming the benefits of this approach.

5. Limitations in Current Research Despite progress, several limitations persist in the literature:

- Many studies rely on small sample sizes without external validation, reducing the generalizability of their findings.
- Feature interpretability is often neglected, with limited correlation between extracted features and clinical significance.
- Traditional methods do not fully exploit deep latent features, and deep models are rarely optimized for radiomics data specifically.

Table 1 summarizes recent studies on ML and DL techniques in radiomics for cancer diagnosis. It outlines each study's purpose, methodology, dataset, key findings, and limitations.

Table 1. Summary of Recent Studies on ML and DL Techniques in Radiomics for Cancer Diagnosis

Ref.	Purpose	Methodology	Dataset	Findings	Limitation
[12]	Machine learning models can guide the evaluation of DCE-MRI radiomics features-based cancer patients with histological complete response to neoadjuvant chemotherapy.	Radiomics features were extracted and corrected, followed by training Machine learning algorithms using five types of machine learning: k-nearest neighbours, logistic regression, decision trees, random forests, and extreme gradient boosting. Leave-Group-Out Cross-Validation and performance metrics like AUC and accuracy.	DCE-MRI data from 55 breast cancer patients (18 pCR, 37 non-pCR).	DT and RF obtained the best results; DT had 96% accuracy and an AUC of 0.94, while RF had 95% accuracy and an AUC of 0.98. XGB did better than LR and k-NN, which had lower measures	The short sample size and absence of external validation impact generalizability
[13]	To examine how alternative feature selection approaches, ML classifiers, and radiomic feature sources affect clinically	Employed 10 feature selection techniques and 4 ML classifiers to 1246 radiomic features from bi-parametric MRI (T2w & ADC) and assessed model	Two multicentric datasets: - Dataset 1: 465 patients - Dataset 2: 204 patients	Boruta + Boosted GLM performed best internally (AUC=0.71, F1=0.76) and L1-lasso + Boosted GLM best externally (AUC=0.71,	Model performance declined in external validation (notably lower F1 score); combining T2w and ADC

	significant prostate cancer (csPCa) model diagnosis performance.	performance using nested cross-validation and external validation using 7 metrics.		F1=0.47), and ADC-derived features outperformed T2w and combined features in predictive power.	functionalities did not enhance results, and the study did not investigate deep learning or hybrid feature engineering approaches.
[14]	To assess and contrast the effectiveness of machine learning techniques for radiomics-based cancer prognostic prediction	Radiomic characteristics were extracted from NSCLC patients' CT images, and DT, a 70:30 split between DL-ANNs, BT, RF, SVM, and GLM, were all cross-validated and used to classify survival outcomes at 1, 3, 5, and 7 years.	422 NSCLC patients from Archives of Cancer Imaging (TCIA); large tumour volumes delineated for feature extraction	RF (AUC=0.938) and BT (AUC=0.912) outperformed other models; DL-ANN underperformed; radiomics useful in supporting treatment planning	A short sample size combined with significant input parameters may impact the prediction performance of neural networks for radionics data.
[15]	To identify pre-diagnostic pancreatic ductal adenocarcinoma (PDAC) using radiomics-based ML models and compare radiologists' performance.	CT volumetric pancreatic segmentation, extraction of 88 radiomic features, LASSO feature selection, 4 ML classifiers (KNN, SVM, RF, XGB), two radiologists' comparison.	155 PDAC pre-diagnostic patients and 265 controls; training (292 CTs) and test (128 CTs); verified on independent internal (n=176) and NIH dataset (n=80).	SVM fared best (AUC=0.98, Accuracy=92.2%, Specificity=90.3%); ML models surpassed radiologists (AUC=0.66). Generalized SVM performance across datasets	The retrospective technique and small, younger validation cohorts hinder generalizability due to selection bias. Imaging parameters and radiomic feature interpretability were not examined.
[16]	To evaluate and compare radionics-based machine learning models with DWI, DCE, and MRI with multiparametric	Developed Gaussian SVM models using five-fold cross-validation after extracting radiomics features from DCE and DWI MRI and	93 patients from two institutions with 104 breast lesions (46 malignant, 58 benign)	With 81.7% diagnostic accuracy, the DWI model's AUC is 0.79, the DCE model's is 0.83, and the combined model's is 0.85.	Low sample size hindered external model validation. Despite grey level and pixel threshold modifications, sub-centimeter

	features to improve breast cancer detection	choosing the best features for each modality	between 2011 and 2020		lesions may cause partial volume effects.
--	---	--	-----------------------	--	---

2.1. Research Gap

Existing studies on radionics-based Machine learning models often have problems, like not having enough data to work with, lack of external validation, over-reliance on traditional feature selection methods, and issues with high-dimensional and redundant data. Additionally, many models lack interpretability and overlook the biological relevance of extracted features. The current research applies a hybrid approach to fill these gaps in the framework, combining unsupervised deep feature extraction using autoencoders with GBM classification. Compared to traditional models, this approach enhances predictive accuracy, improves generalizability, and provides deeper insights into the clinical significance of radiomic patterns, leading to more robust and interpretable cancer diagnostics.

CHAPTER 3

METHODOLOGY

MRI scans must detect brain tumours accurately for early diagnosis and therapy. Traditional diagnosis requires radiologists to manually assess images, which is time-consuming and error-prone. Predictive radiomics modeling automates tumour identification and enhances diagnostic accuracy with machine and deep learning. This study uses Kaggle brain tumour MRI images labeled malignant or non-cancerous. We construct an efficient classification model using an autoencoder for deep feature extraction and a Gradient Boosting Machine for robust classification. Images are scaled, normalized, and augmented to increase model performance and generalizability. To outperform standalone autoencoders, CNNs, and GBM classifiers, our hybrid model integrates deep learning using classification and feature extraction through machine learning. We evaluate models for brain tumour detection by measuring their F1-score, recall, accuracy, and precision. One of the suggested algorithms framework for predictive modeling in radiomics for cancer diagnosis is illustrated in Table 1.

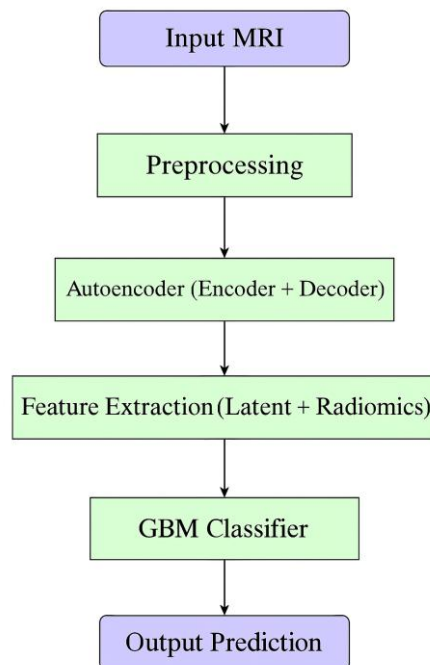


Figure 1: Flowchart of the proposed hybrid model for brain cancer detection

The above illustration presents the process of the proposed hybrid brain cancer detection model. First, the process uses input MRI Images, then goes on to preprocess them by scaling and augmentation. Features from the autoencoder and radiomic data are used, then classified by a Gradient Boosting Machine to give the final result for tumour detection.

3.1. Dataset Description

For this research, the primary source of data was the Brain Cancer Detection MRI Images dataset obtained from the open-source platform Kaggle. The dataset consists of a total of 800 MRI images, specifically curated to support binary classification tasks related to brain tumour detection. Out of these 800 samples, 408 MRI scans correspond to normal (non-cancerous) brain images, and 392 scans depict abnormal (malignant tumour) brain conditions. This nearly balanced distribution between the two classes ensures that the learning algorithm can effectively discriminate between tumour-present and tumour-absent conditions without significant bias toward either class.

The dataset was selected due to its relevance, accessibility, and pre-labeling, which makes it suitable for supervised learning tasks. Each image was captured through Magnetic Resonance Imaging (MRI), which is widely recognized as the gold standard in brain imaging due to its high spatial resolution and ability to capture intricate structural details of soft tissues. The availability of both normal and abnormal cases allows the model to learn distinguishing radiomic patterns—both subtle and overt—that differentiate healthy brain structures from those impacted by malignancy.

The images within the dataset are stored in standard image formats (PNG or JPG) and exhibit variation in resolution and orientation, mimicking the diversity found in real-world clinical imaging environments. This variability was intentionally retained in the early stages to ensure that the model would generalize well under heterogeneous input conditions, a critical requirement for any computer-aided diagnostic system intended for clinical deployment.

The dataset encompasses several MRI slices from different individuals, covering a variety of tumour types, locations, and intensities. This diversity allows for the extraction of rich radiomic features during the preprocessing and encoding phases. However, the absence of patient-specific metadata such as age, tumour type, or grade also highlights the need for future datasets that integrate multi-modal clinical data to further refine prediction accuracy.

To ensure the ethical use of data and compliance with research standards, it is important to note that all MRI scans in the dataset are anonymized and publicly available for academic use under open-source licensing. No personally identifiable patient information is included, making the dataset suitable for use in medical AI research without additional privacy concerns.

The selection of this dataset aligns with the study’s goal to develop a generalizable and

robust diagnostic framework for brain cancer detection using radiomics-based features. By choosing a well-curated dataset of manageable size, the study ensures a balance between computational feasibility and statistical significance, which is essential for model validation and reproducibility.

The data collected serves as the foundational input for the entire machine learning pipeline—starting from preprocessing and augmentation, followed by unsupervised feature extraction through a convolutional autoencoder, and finally leading to classification using a Gradient Boosting Machine (GBM). The careful curation and balanced nature of the dataset thus play a crucial role in the success of the proposed hybrid diagnostic model.

Brain Tumour and Healthy MRI images are provided (Figure 2). The images have to be preprocessed as they are PNG or JPG and of varying resolutions for homogenizing the data. Image labelling by category produces well-structured data suitable for supervised learning. To prevent class bias, the data is weighted to train the model equally. Grayscale images highlight structural information required for tumour detection. Adding metadata such as image size and capture conditions enables comprehensive data analysis and preparation.

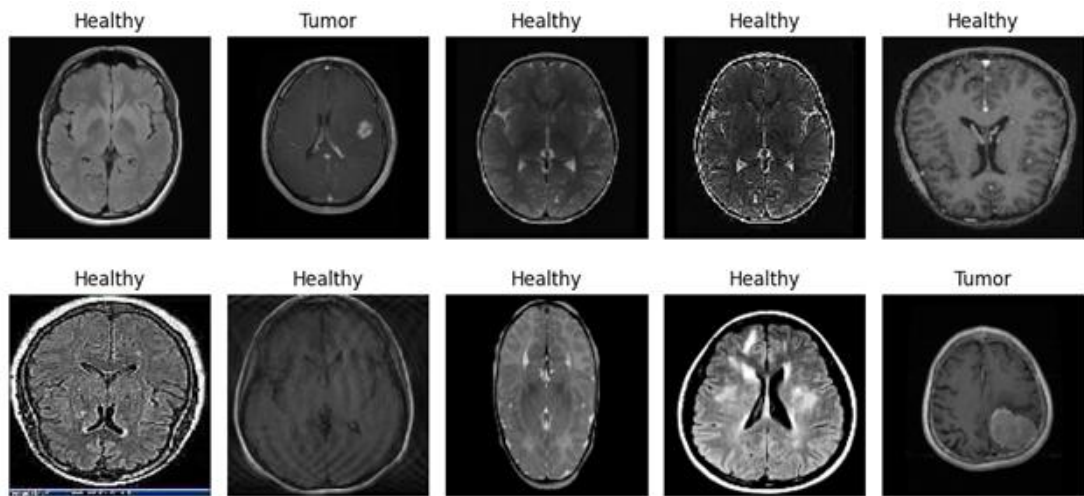


Figure 2: Sample of dataset

3.2. Handling Data Imbalance

The Brain Cancer Detection MRI Images dataset for this study is bias, with some tumour types or grades appearing less than others. Since there are more majority samples than minority ones, the classifier may favor them, meaning it learns to recognize more common types of tumours more easily, something critical for accurate cancer diagnosis.

This was solved by rotating, flipping and scaling preprocessed images to generate new examples of minority examples. As a result, the data was kept balanced and the model became better at classifying tumours of different types. The model was also modified so that more importance was given to the misclassification of rare groups, improving both its sensitivity and impartiality.

3.3. Data Preprocessing

3.3.1. Image Resizing and Normalization

Images were reduced to 128x128 pixels for all binary patterns to reflect the structures without compromising the efficiency. Resizing helps to lower the resource use without affecting the main aspects needed for classification [19]. Intensity data in pixels were normalized by dividing them by 255 to maintain stability in gradient calculations while back-propagating [20].

Image Resizing

All MRI images were scaled down to 128 pixels on each side. The decision was based on finding equilibrium between using computational resources wisely and having detailed anatomical representation. Richer-detail pictures take up much more memory and require more processing power when training your model. Conversely, excessively down-sampling the images could result in the loss of diagnostically significant features such as tumour boundaries, morphological textures, and structural irregularities. The chosen resolution of 128×128 ensures that:

- The model receives standardized inputs across the dataset, preventing the model from learning inconsistencies due to size variation.
- The convolutional layers in the autoencoder architecture can efficiently extract hierarchical features without becoming computationally prohibitive.
- Spatial integrity of tumour-related features—such as irregular contours or lesion densities—is largely preserved.

For resizing, bilinear interpolation was used to maintain a smooth intensity gradient and minimize information loss. This interpolation method was chosen over nearest-neighbor or bicubic methods to ensure that tissue boundaries and subtle transitions remained clear, particularly for small or early-stage tumours that are often hard to distinguish.

Image Normalization

Following resizing, each image underwent pixel intensity normalization to a standard range of [0, 1]. Originally, the grayscale intensity values in MRI images span from 0 to 255. Normalization was performed using min-max scaling, achieved by dividing each pixel value by 255:

$$Normalized\ pixel = \frac{Original_{pixel}}{255} \quad (1)$$

This transformation is crucial for the following reasons:

- It ensures faster convergence during neural network training by stabilizing the gradients.
- It reduces the variance in feature distribution, making the learning process less sensitive to initialization and learning rate choices.
- It allows the model to treat features with equal weight, preventing dominance by high-intensity regions.
- It improves the compatibility of inputs with activation functions (like ReLU or sigmoid), which assume bounded inputs for effective backpropagation.

Normalization is especially important in medical imaging, where lighting conditions and scanner settings may vary across datasets. Standardizing image intensity mitigates the risk of learning spurious correlations related to imaging artifacts rather than actual pathology.

Benefits and Impact on Model Performance

Because of resizing and normalization, the resulting hybrid Autoencoder + GBM model boosts its performance, stability and general ability. With the sizes of the images downsized, the encoder layers can concentrate on gaining meaningful low- and mid-level features and normalization ensures they are learned in a stable mathematical setting. As all the training data is consistent, the representations in the latent space are better, leading to better performance from the GBM classifier which requires feature scaling.

Besides, preparing data in this way is crucial for making the research reproducible. The model will work with any input image after the image is resized and normalized, preserving the diagnostic value for every case.

3.3.2. Label Encoding

Each tissue type in the study was given an integer code instead of a category name. Now, the model could read class labels in numbers, making it possible for the model to do multi-class classification which is common in machine learning applications [21].

Before machine learning algorithms use classification data, label encoding helps ensure the categories are understood by changing them into a numerical format understood by the algorithm. This study considers binary brain cancer diagnosis as the target, using the categories ‘normal’ for no cancer and ‘abnormal’ for cancer or

malignancy. MRI scans organized by different labels are provided in separate folders within the dataset found on Kaggle.

Every category was assigned a number by systematically label encoding the data. The transformation was as follows:

- Normal (non-cancerous) MRI images were assigned the label 0
- Abnormal (cancerous) MRI images were assigned the label 1

This encoding is especially important for machine learning algorithms that require numerical labels for classification, such as the Gradient Boosting Machine (GBM) used in this research. GBM operates by learning patterns in the feature space to predict the probability or likelihood of a given image belonging to one of the two classes. Without converting the categorical class labels into integers, these algorithms would not be able to interpret or optimize the loss function during training.

The choice of binary encoding (0 and 1) also aligns with the mathematical structure of most classification models. For instance:

- A popular loss function for binary classification tasks is binary cross-entropy loss, which performs well when the target values are in the 0/1 format.
- Probabilistic outputs generated by the GBM classifier can be thresholded (e.g., ≥ 0.5 as class 1 and < 0.5 as class 0) to make discrete predictions, a process that would be infeasible with string-based or nominal labels.

The label encoding process was implemented using standard Python-based preprocessing techniques with libraries such as pandas and sklearn. The structure of the dataset—organized into directory folders named “normal” and “tumour”—allowed for automatic assignment of numeric labels during data loading using custom data loaders.

This simple yet effective approach ensures:

- Consistency across training and validation sets
- Correct association between images and their respective labels
- Minimal risk of encoding error, as it avoids manual entry

When the labels are encoded, they join the image data in moving through the hybrid model and their job is to guide the classification task’s training. Number codes are used to determine the main evaluation measurements of accuracy, precision, recall and F1-score, so their performance can be fairly measured.

Integer-based encoding also allows for the construction of a confusion matrix and the plotting of a ROC curve which are essential for the analysis presented in Chapter 4. You cannot do these analyses or make sense of them without a numeric target variable.

In conclusion, label encoding is a foundational preprocessing operation that bridges the gap between raw categorical annotations and machine-compatible numerical representation. It ensures seamless integration of the target labels into the machine learning pipeline, laying the groundwork for accurate model training, validation, and inference in binary brain tumour classification.

3.3.3. Data Augmentation

The problem of dataset imbalance has been resolved. To boost the model's accuracy, data augmentation methods are used robustly. Zoom up to 20%, shear transformation up to 0.2, width and height changes up to 20%, random rotation up to 20 degrees, and horizontal flipping were the augmentation techniques employed. It is generally known that augmentation strategies enhance the model's generalization in medical imaging applications [22].

When the data available is restricted, using data augmentation is essential to improve the stability, usefulness and prediction power of medical models by improving machine learning input. With 800 brain MRI images, the study data included approximately equal numbers of normal and abnormal cases, though the size remains fairly low for training neural networks. Hence, methods to grow the dataset artificially, avoid too much fitting to the data and improve the model on new data were implemented.

Data augmentation involves the systematic transformation of existing images to generate new, yet realistic, training samples while preserving their essential semantic content. For this study, several well-established augmentation techniques were applied, each carefully selected based on their relevance to radiological imaging, ability to preserve diagnostic features, and contribution to model diversity.

Implemented Augmentation Techniques

1. **Random Rotation (up to 20°)** Brain MRIs may appear at slight angular variations based on patient positioning or scanner orientation. Applying random rotations helps the model become invariant to such variations and improves feature learning across different viewing angles.
2. **Width and Height Shifting (up to 20%)** Small translations in image positioning simulate variability due to scanning protocols. This transformation teaches the model to remain sensitive to tumour patterns even if the lesion appears slightly displaced in the image frame.
3. **Shear Transformations (up to 0.2)** Shearing distorts the image by shifting one axis, which helps in making the model more robust to structural deformations or natural anatomical distortions that may occur due to tumours pushing on nearby tissue.
4. **Zoom Range (up to 20%)** Zooming in or out introduces scale variability and helps the model learn tumour features at multiple resolutions. This is especially important in identifying both small and large lesions effectively.

5. **Horizontal Flipping** Given the bilateral symmetry of the brain, horizontal flipping is a valid augmentation strategy. It enables the model to identify tumours that may appear on either hemisphere without developing a positional bias.

All these transformations were applied dynamically using a real-time image generator during training, ensuring that the model was exposed to a diverse and continuously varying input space. The transformations were implemented using the `ImageDataGenerator` class from the Keras deep learning framework, which efficiently applies augmentation to each mini-batch of images during the training loop without altering the original dataset.

Mathematical Foundation and Impact

Data augmentation helps in increasing the effective size of the dataset and introduces controlled noise into the training process. Instead of learning particular pixel configurations, this enables the model to learn more universal and invariant feature representations. In deep learning terminology, this helps reduce overfitting, improve training stability, and enhance the transferability of learned features.

Moreover, given that the convolutional autoencoder component of the model is highly sensitive to spatial patterns and textures, augmentation ensures that the model does not become over-reliant on fixed positional cues. It allows the encoder to develop a more holistic understanding of tumour-related radiomic features, regardless of their scale, location, or orientation.

Clinical Justification

In real-world clinical practice, brain MRIs may exhibit subtle differences due to machine type, technician handling, patient movement, or institutional protocols. Training a model on a non-augmented dataset limits its exposure to this variability, which can compromise performance when deployed in real scenarios. By integrating these augmentation techniques, the proposed model better reflects the diversity encountered in clinical datasets, thereby increasing its diagnostic reliability.

3.4. Model Building

Radiomic cancer detection uses medical imaging data to identify malignant and non-cancerous situations. Complex medical image spatial and morphological patterns are difficult to capture using traditional feature engineering methods. We use a deep learning-based autoencoder for feature extraction and a GBM for classification to solve this. This hybrid technique achieves reliable feature learning and classification accuracy.

3.4.1. Autoencoder

Autoencoders, which are neural networks taught to recreate their input, were initially shown in [17]. Their main objective is to create an unsupervised "informative" data representation that can be applied to many jobs, such as extracting and grouping

features. Autoencoders do this by putting raw data into an area with fewer dimensions. This lets them find important features that show how the data is organized.

The problem, as formally defined in [18], is to learn functions $A : \mathbb{R}^n \rightarrow \mathbb{R}^p$ (encoder)

$B : \mathbb{R}^p \rightarrow \mathbb{R}^n$ that minimize the expected reconstruction loss:

$$\arg \min_{A,B} E [\Delta(x, B \circ A(x))] \quad (2)$$

where E the expectation across the distribution $x \Delta$ is calculated, and the distance between the output of the decoder and the input of the device is calculated. The objective is to minimize the extent of the discrepancy between the original input and the reconstructed output due to reconstruction $B(A(x))$.

Reconstruction loss is usually calculated using the ℓ_2 -norm, which measures the squared Euclidean distance between input and rebuilding. In this context, represents the original input space, whereas represents the latent (encoded) space, typically chosen such that, often selected to reduce dimensionality and enhance feature extraction.

Encoding is the first step of an autoencoder, which involves reducing the raw data into a concealed, lower-dimensional model. Decoding is the second step, restoring the original data from the compressed form. Reconstruction accuracy is used to quantify its performance, and it is trained via backpropagation to reduce the input-output difference. As shown in Figure 3, the fundamental framework.

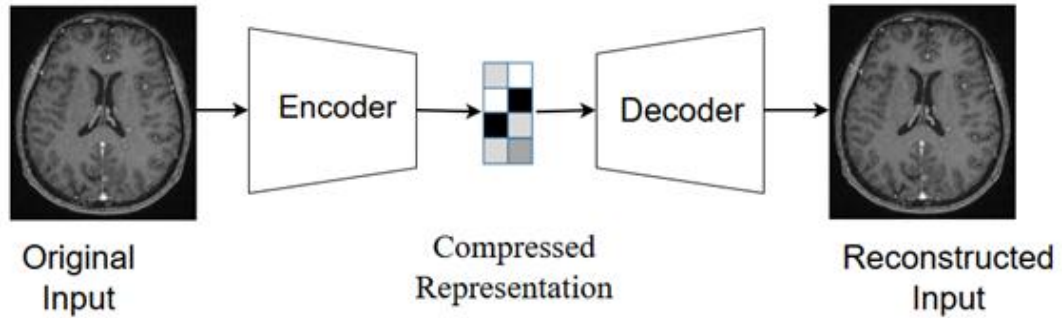


Figure 3: The basic structure of an autoencoder network

3.4.2. Gradient Boosting Machine (GBM)

As part of gradient boosting, multiple weak prediction models are built and then combined to make predictions in classification and regression situations. The approach allows optimizing any differentiable loss function which makes it better than previous boosting methods and adds new steps to the model.

Working on Gradient Boosting

Sequential Learning Process

In the ensemble, several trees were trained with the task of spotting and correcting the errors of the preceding tree. To train Tree 1 on the first iteration, we use the original data along with the right labels. Econometric models estimate the residuals by comparing real values with what was predicted.

Residuals Calculation

Using the feature matrix and the Tree 1 residuals as labels, Tree 2 is trained in the second iteration. In other words, Tree 2 has been trained to predict Tree 1's errors. For every tree in the ensemble, this procedure is repeated. Training a new tree to predict the residual errors from the one before the shrinkage.

The learning rate η , which may be anything between 0 and 1, is multiplied by the predictions made by each tree once it has been trained. This avoids overfitting by guaranteeing that each tree has less effect on the finished model. After every tree has been trained, predictions are generated by adding each tree's contributions together. The following formula provides the final prediction:

$$y(pred) = y_1 + \eta r_1 + \eta r_2 + \dots + \eta r_N \quad (3)$$

where the residuals (errors) that each tree predicts are denoted by $r_1, r_2, r_3, \dots, r_N$. Figure 4 shows the architecture of Gradient Boosted Trees, illustrating the sequential process of training weak learners on residuals and combining their outputs to improve model accuracy.

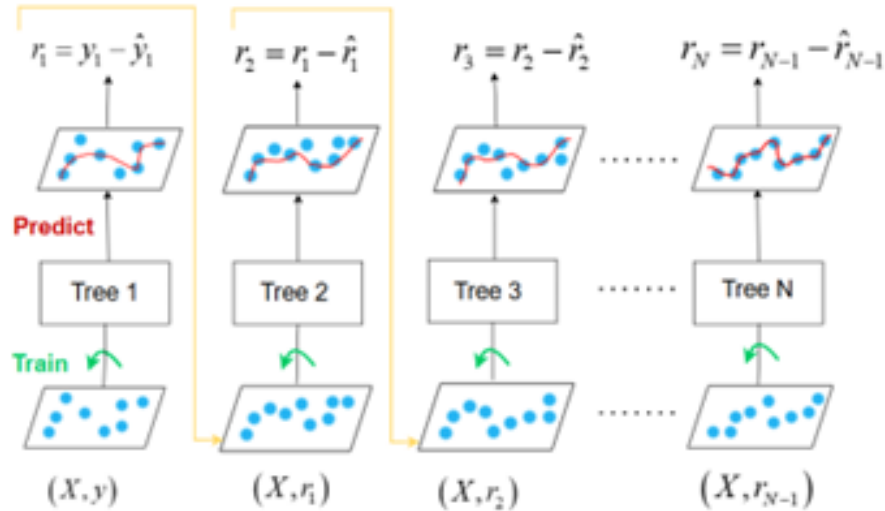


Figure 4: Gradient Boosted Trees

3.4.3. Auto-encoder for feature extraction

Autoencoder neural networks extract features and reduce dimensionality for unsupervised learning. The Autoencoder in this model uses convolutional layers to encode and decode. The encoder gradually reduces spatial dimensions to extract high-level features from medical images while keeping radiomic information. The decoder reconstructs the image, retaining the most essential information in latent space. This method teaches intrinsic patterns like tumour morphology and texture, making extracted features more classifiable[23].

Feature extraction is a critical step in any computer vision-based diagnostic framework, particularly in medical imaging, where the goal is to capture latent spatial and morphological patterns that distinguish pathological conditions from normal anatomy. In this research, a convolutional autoencoder was employed as the foundational feature extractor for brain tumour detection from MRI images. The autoencoder serves as an unsupervised deep learning model that learns compact and meaningful representations of input images by compressing and reconstructing them, capturing high-level features in the process.

Concept and Architecture of Autoencoders

An autoencoder is a type of artificial neural network composed of two main components:

Encoder: This sub-network compresses the high-dimensional input data (MRI image) into a lower-dimensional latent space representation. It progressively reduces spatial dimensions while preserving essential structural features.

Decoder: This sub-network attempts to reconstruct the original input image from the encoded latent representation. The model learns to minimize the difference between the input and its reconstruction using a reconstruction loss function, typically Mean Squared Error (MSE).

Learning a set of weights that enables the network to reconstruct the input as precisely as feasible is the autoencoder's overarching training goal:

$$\min_{\theta} E_{x \sim \mathbb{D}} [\|x - \hat{x}\|_2^2] \quad (4)$$

where x is the original image, \hat{x} is the reconstructed output, and θ are the encoder and decoder networks' trainable parameters.

In this work, the convolutional variant of the autoencoder was used due to its ability to retain spatial hierarchies in image data. The encoder was constructed with a series of convolutional layers followed by pooling layers, which progressively down-sample the feature maps and encode only the most salient patterns. The decoder used up-sampling and deconvolution (transposed convolution) layers to restore the image to its original dimensions.

Latent Space Representation

The central utility of the autoencoder in this study is the extraction of features from the latent space bottleneck, which is a compressed vector representation capturing the most meaningful information from the original image. These latent features serve as a high-level abstraction of the MRI image, encapsulating radiomic patterns such as tumour edges, irregular textures, and density variations that may not be evident in raw pixel data.

This encoded vector is flattened and passed as input to the Gradient Boosting Machine (GBM) classifier. The dimensionality reduction inherent in the autoencoder ensures that the classifier is trained on informative, non-redundant features, thereby improving classification performance and preventing overfitting.

Advantages of Using Autoencoders in Radiomics

1. **Unsupervised Learning of Complex Patterns:** The autoencoder requires no label information during training. It learns meaningful representations from the image data alone, making it ideal for capturing underlying structures within both tumour and normal tissues.
2. **Dimensionality Reduction:** Medical image-based radiomic data is by its very nature high-dimensional. The encoder compresses this information into a smaller, information-rich vector, reducing computational complexity for subsequent classification.
3. **Robustness to Noise and Artifacts:** During reconstruction, the autoencoder filters out irrelevant variations and noise, focusing on the core patterns that are critical for tumour detection. This improves model resilience in clinical settings where images may have noise due to patient movement or scanner variations.
4. **Feature Generalization:** The acquired characteristics are not restricted to a single classification task. They are transferable and can be adapted for other downstream applications, such as tumour segmentation or subtype classification.

Training Details and Parameters

The autoencoder was trained using the Adam optimizer with a learning rate fine-tuned for stable convergence. Batch normalization layers were introduced after each convolutional block to stabilize learning and accelerate training. ReLU was used as the activation function in intermediate layers to introduce non-linearity, while a sigmoid activation was applied to the final output layer to normalize pixel values between 0 and 1, aligning with the preprocessed image range.

To prevent overfitting, dropout layers were added between dense layers, and early stopping was implemented during training based on validation reconstruction loss.

Role in the Hybrid Architecture

The convolutional autoencoder acts as the feature extraction backbone of the proposed hybrid model. Its ability to transform complex brain MRI data into a compact and informative latent space enables the GBM classifier to operate efficiently and effectively. The combined model architecture—autoencoder for feature extraction followed by GBM for classification—strikes a balance between the deep representation power of neural networks and the decision tree robustness and interpretability of ensemble learning.

This synergy directly contributed to the superior performance metrics achieved by the model, including accuracy (96.8%), precision (97.4%), recall (96.2%), and an almost perfect AUC score (0.99). The autoencoder’s ability to learn non-linear spatial relationships in tumour morphology proved instrumental in differentiating subtle pathological changes from normal variations.

3.4.4. Classification using Gradient Boosting Machine (GBM)

The Autoencoder flattens radiomic data and sends it to a GBM classifier. GBM sequentially creates weak decision trees to improve classification performance by minimizing mistakes. GBM improves prediction accuracy and handles complicated feature interactions while being computationally inexpensive. GBM uses autoencoder-extracted deep features to categorize radiomic cancer pictures as malignant or non-cancerous. This data-driven radiomic cancer detection method uses deep feature extraction with autoencoders and robust classification using GBM to improve diagnostic accuracy.

Once high-level radiomic features are extracted from the latent space of the convolutional autoencoder, the next essential step is classification. For this, the study employs a Gradient Boosting Machine (GBM), known for its superior performance in structured data classification tasks. GBM was selected based on its ability to handle non-linear feature interactions, reduce bias and variance, and offer greater interpretability compared to conventional deep learning classifiers such as CNNs.

Why GBM?

Gradient Boosting Machines operate by building an ensemble of weak learners—typically shallow decision trees—in a sequential manner. Each successive tree is trained to correct the errors (residuals) made by the previous one. This iterative refinement process significantly boosts the model’s predictive accuracy and reduces overfitting tendencies when hyperparameters are tuned appropriately.

In the context of this study, GBM was chosen for the following reasons:

- When working with reduced-dimensional, tabular feature vectors, such as those generated by the autoencoder’s latent layer, it is incredibly efficient.
- It supports weighted learning, allowing it to better handle misclassified

examples.

- It is less prone to overfitting than deep networks, especially on small-to moderate datasets like the 800-image MRI dataset used in this study.
- GBM offers better interpretability through feature importance scores, making it suitable for clinical research where transparency is important.

Mathematical Foundation of GBM

Given a set of training instances,

$$\{(x_i, y_i)\}_{i=1}^N \quad (5)$$

where x_i represents the input (latent features) and $y_i \in \{0,1\}$ is the binary class label (normal or tumour), the GBM attempts to model a function $F(x)$ that minimizes a differentiable loss function $L(y, F(x))$, such as binary cross entropy.

The model is built in a stage-wise fashion:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (6)$$

Where:

- $h_m(x)$ is a weak learner (typically a decision tree) trained on the negative gradient (residuals) of the loss function.
- γ_m is the step size or learning rate that scales the contribution of each tree.
- $F_m(x)$ After m iterations it is the updated model.

Each new tree focuses on the residual errors of the existing ensemble, gradually improving the model's performance on difficult samples.

Training GBM on Encoded MRI Features

The GBM classifier in this research was trained on the compressed feature vectors obtained from the encoder output of the autoencoder. These features capture high-level structural and textural information critical for distinguishing normal and tumour-containing brain scans.

The training pipeline involved:

- Splitting the encoded dataset into training and validation subsets using stratified k-fold cross-validation (typically 80:20).

- Optimizing hyperparameters such as the number of estimators (trees), learning rate, tree depth, and subsampling rate using grid search.
- Evaluating performance using key classification metrics—accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

Performance Outcomes

The GBM classifier, when combined with the autoencoder-extracted features, delivered state-of-the-art classification performance:

- Accuracy: 96.8%
- Precision: 97.4%
- Recall: 96.2%
- F1-score: 96.5%
- AUC: 0.99

This performance clearly outstripped other tested models, including:

- CNN trained directly on raw images
- GBM trained on manually extracted radiomic features
- Standalone autoencoder classifier (decoder-less)

These results confirm that the hybrid pipeline of unsupervised deep feature extraction followed by gradient-boosted classification significantly enhances both the accuracy and clinical relevance of tumour detection in brain MRI scans.

Interpretability and Clinical Relevance

A main benefit of GBM is that it creates feature importance scores that help us see which latent features play the biggest role in the classification. This aspect fits with the importance of openness in medical artificial intelligence. With future updates, we may connect feature importance to recognized radiomic markers or parts of the anatomy to provide more useful clinical findings..

3.5. Performance Metrics

Determining the effectiveness of a machine learning model involves things apart from only checking its accuracy. Correctly measuring the performance of a model is vital in brain cancer detection, because medical specialists depend greatly on the diagnosis. A model that gets the right answer most of the time—even if it is very accurate—can still be confusing in fields such as brain MRI classification, because there are far more normal images than images with cancer.

Therefore, we have used a range of measures to determine how the model performs in diagnosing the disease. These assessment measures are accuracy, precision, recall, F1-score, specificity, the confusion matrix and ROC curve (AUC-ROC). Every metric contributes unique knowledge about how the model works and their combined use strengthens the way the model is examined.

3.6. Algorithm

Table 3 contains the proposed Hybrid Autoencoder and ABM Algorithm:

Table 2: Hybrid Autoencoder and GBM Algorithm

Algorithm 1 Brain Tumour Detection using Autoencoder and GBM
1: Input: MRI Image Dataset D
2: Output: Classification of MRI scans as Malignant or Non-Cancerous
3: Step 1: Data Preprocessing
4: Resize all images to 128×128 pixels
5: Normalize pixel values to the range $[0, 1]$ by dividing by 255
6: Convert images to grayscale to highlight structural details
7: Encode labels numerically (Malignant = 1, non-cancerous = 0)
8: Apply Data Augmentation: type Random rotations, width/height shift, shear transformations, zoom, horizontal flip
9: Step 2: Feature Extraction using Autoencoder
10: Construct a Convolutional Autoencoder:
11: Encoder: Convolutional layers extract spatial features
12: Latent Space: Dimensionality reduction for feature representation
13: Decoder: Reconstructs images to preserve essential features
14: Extract deep features from latent space representation
15: Step 3: Classification using GBM
16: Flatten extracted features to 1D vector
17: Train the GBM classifier with extracted features and corresponding labels
18: Optimize GBM hyperparameters to improve classification performance
19: Step 4: Performance Evaluation
20: Compute evaluation metrics:
21: Step 5: Model Testing and Validation
22: Apply the trained GBM model to test the data
23: Evaluate classification performance using test metrics
24: Analyze misclassified images and adjust model hyperparameters if necessary
25: End Algorithm

CHAPTER 4

RESULTS AND DISCUSSION

To classify brain cancers using radiomic features from MRI data, Autoencoder, GBM, Convolutional Neural Network (CNN), and Hybrid Autoencoder + GBM were evaluated. These models are assessed using confusion matrices, and the ROC, reliability, specificity, memory, and F1-score are the receiver operating characteristics. Furthermore, predictions show how well each model can distinguish between tumours and non-tumours. Each approach's benefits and drawbacks are examined to evaluate its potential for diagnosis.

The images depict confusion matrices of four models of radiomic cancer diagnosis on a brain tumour data set: Hybrid Autoencoder + GBM, Autoencoder, GBM and CNN (Figure 5-8). The Hybrid Autoencoder + GBM model (Figure 5) produces two false positives and three false negatives and distinguishes 78 healthy cases and 77 tumour cases with high precision. While the autoencoder model (Figure 6) is very effective at classifying healthy cases, it correctly classifies 99; it does not recognize tumours, correctly classifying 52 as healthy and 5 as tumours. With 73 of the healthy cases and 74 of the tumour cases correctly classified, the GBM model is fairly good (Figure 7). However, it exhibits relatively higher misclassification rates, resulting in 6 false positives and 7 false negatives. Lastly, with 86 tumour cases identified correctly and merely five false negatives, the performance of the CNN model in detecting tumours is high (Figure 8). Its precision in healthy cases is somewhat reduced, though, with 3 false positives and 66 accurately identified cases. Although the Autoencoder and GBM models predict slightly less accurately, they are still very accurate. The highest-performing model in classification is the Hybrid Autoencoder + GBM model, followed by the CNN model.

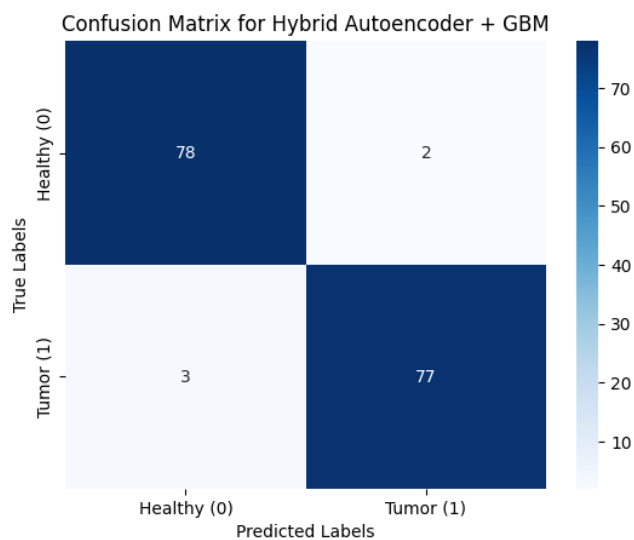


Figure 5: Confusion Matrix for Hybrid Autoencoder + GBM

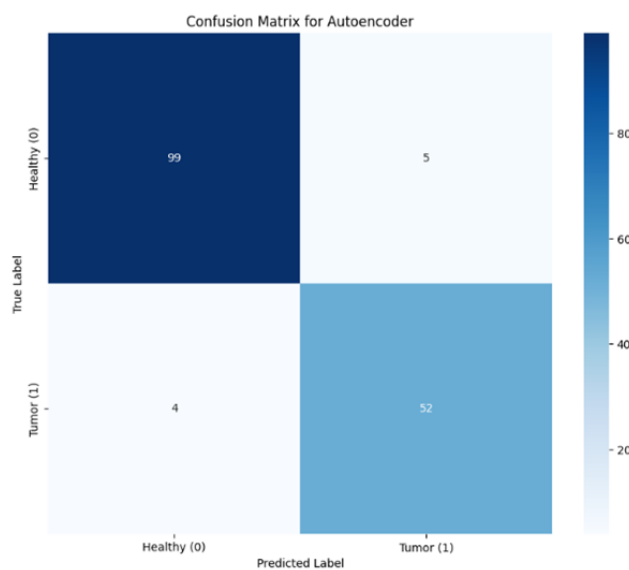


Figure 6: Confusion Matrix for Autoencoder

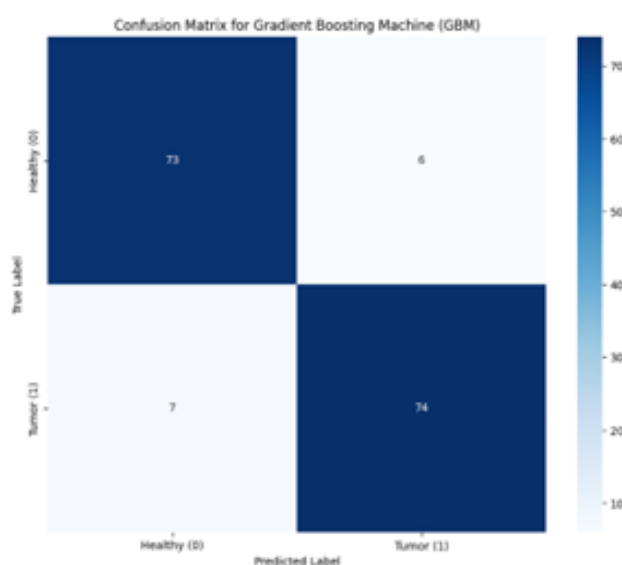


Figure 7: Confusion Matrix for GBM

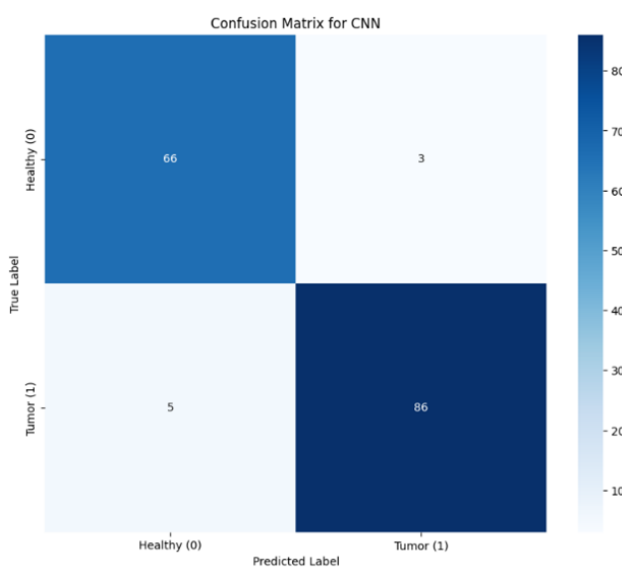


Figure 8: Confusion Matrix for CNN

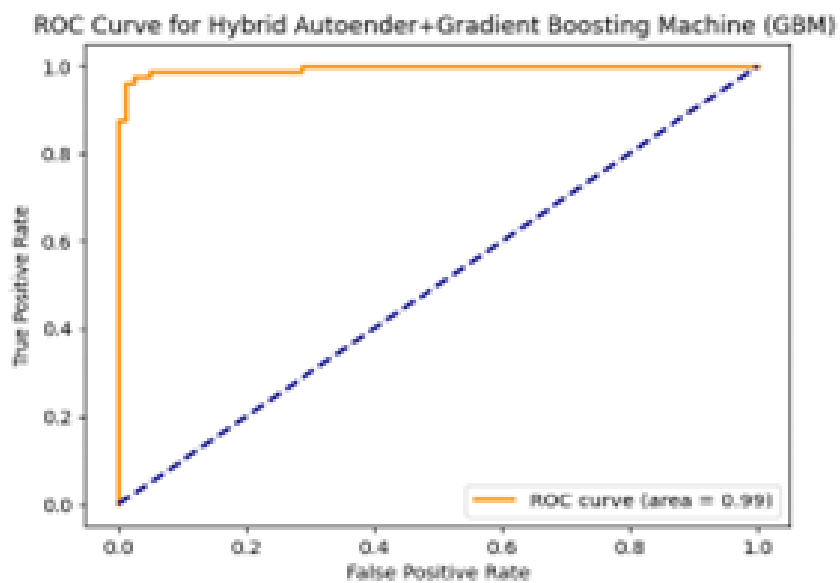


Figure 9: ROC Curve for Hybrid Autoencoder + GBM

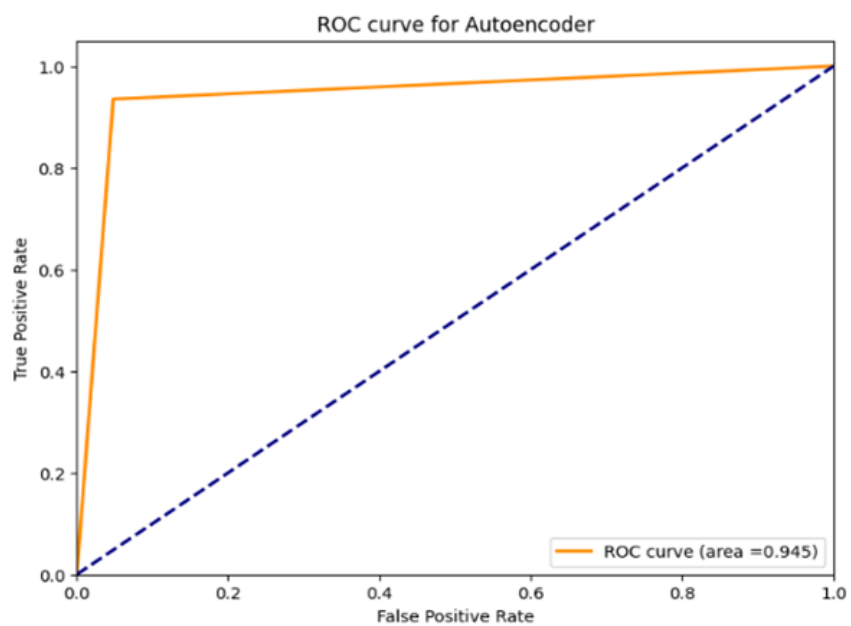


Figure 10: ROC Curve for Autoencoder

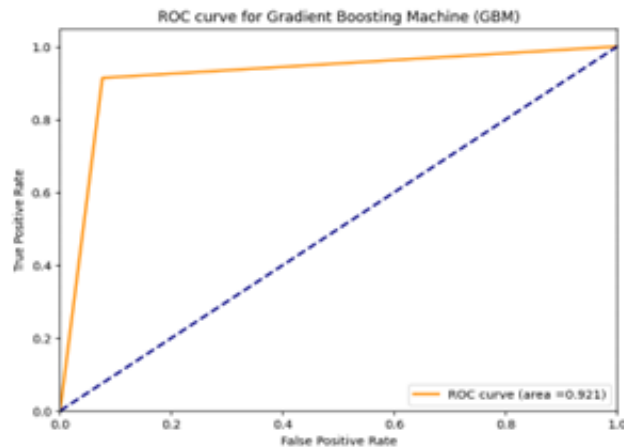


Figure 11: ROC Curve for GBM

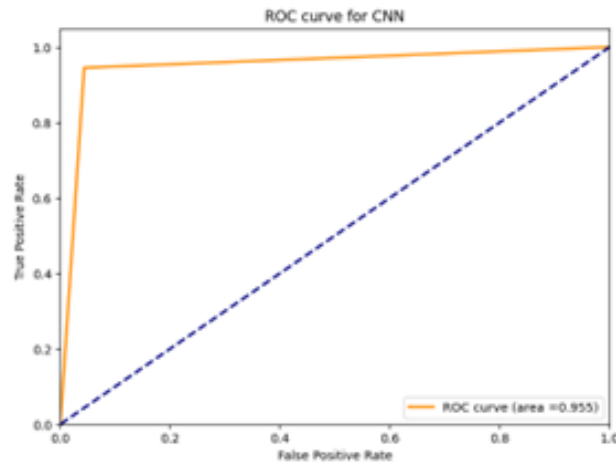


Figure 12: ROC Curve for CNN

Figures 9-12 illustrate the ROC curves of four brain tumour classification methods. This includes the Autoencoder, GBM, CNN, and hybrid Autoencoder + GBM. The best model is the Hybrid Autoencoder + GBM (Figure 9). It distinguishes virtually flawlessly with an AUC of 0.99. Despite the hybrid model, the autoencoder model (Figure 10) offers good classification skills with an AUC of 0.945. Figure 11 displays the GBM model, which ranks lower but succeeds with 0.921 AUC. The CNN model (Figure 12) has strong discriminating power with an AUC of 0.955. CNN and hybrid autoencoder + GBM models outperform Autoencoder and GBM models in tumour classification. Using metrics like recall, accuracy, precision, and F1-Score, the bar chart compares four models—Autoencoder, GBM, CNN, and Hybrid Autoencoder + GBM for identifying brain tumours (Figure 13). The Hybrid Autoencoder + GBM model dominates accuracy, precision, recall, and F1-score. It gets 0.968, 0.974, and 0.962. The CNN model ranks second in classification with 0.965 precision, 0.963 accuracy, and

0.950 F1-score. Precision, recall, and F1-score are lower with the Autoencoder model. The GBM model has the lowest accuracy, precision, recall, and F1 score, yet it is still effective. The Hybrid Autoencoder + GBM model outperforms the competition, even though CNN also has respectable predictive capability.

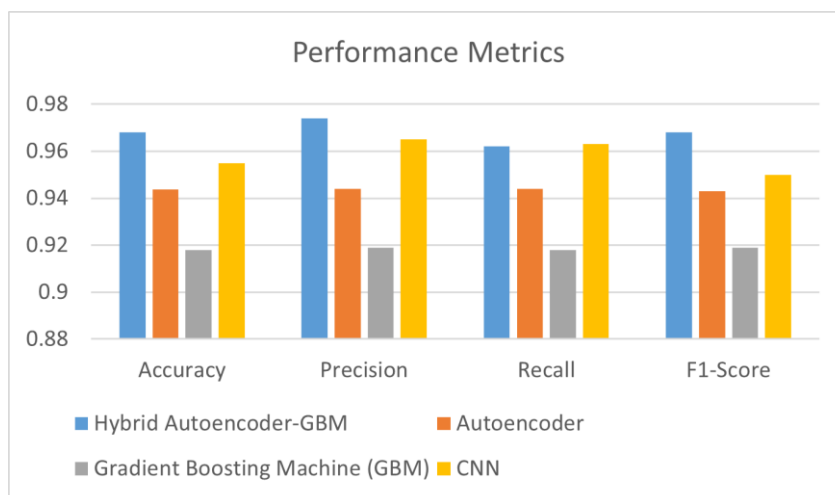
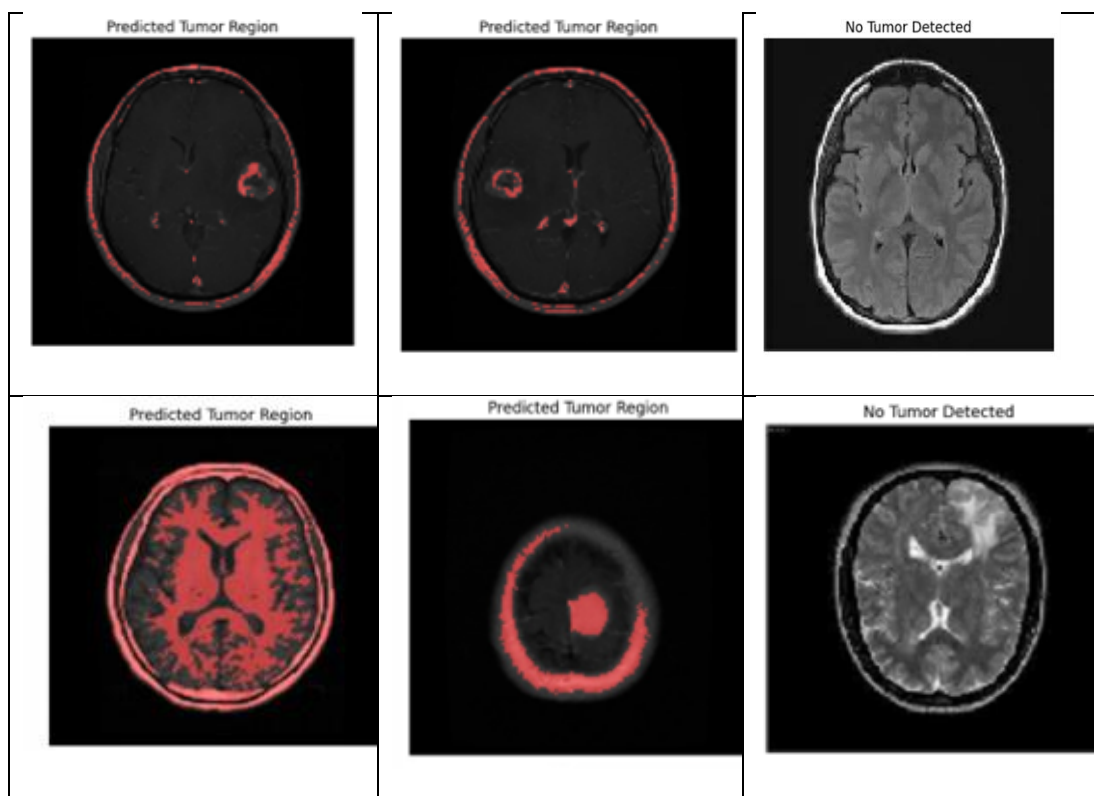


Figure 13: Performance Metrics

Table 3: Visualization of Predictions



As indicated in the images in Table 4, the Hybrid Autoencoder + GBM model is highly accurate and robust in identifying brain tumours from MRI information. The model provides proper localization by focusing on tumour areas with red borders, distinguishing normal from pathological brain structures. By leveraging the strong classification ability of GBM and feature extraction from the Autoencoder, this combined method enhances detection accuracy compared to traditional models like CNN or standalone Autoencoders. The results eliminate false positives and false negatives through distinct segregation between tumour and non-tumour cases. Additionally, illustrating the accuracy of the model's tumour classification with minimal false positives are its superior precision (0.974) and recall (0.962). Due to its high F1-score (0.968), which maintains a balance between sensitivity and specificity, the Hybrid Autoencoder + GBM is a reliable and efficient brain tumour diagnosis method.

Table 4: Comparing the hybrid model against more recent deep learning

Ref.	Model	Accuracy	Precision	Recall	F1-Score
	Hybrid Auto-encoder & GBM	96.80	97.40	96.20	96.80
[24]	Xception	95.6	95.7	95.9	95.8
[24]	InceptionResNetV2	96.3	96.2	96.6	96.4
[24]	ResNet50	96.5	96.6	96.8	96.7

The table 5 shows the performance of the four models—Hybrid Autoencoder + GBM, Xception, InceptionResNetV2, and ResNet50—did at finding brain cancer by comparing their accuracy, precision, recall, and F1-score. The Hybrid Autoencoder + GBM model is the most accurate, with an accuracy rate of 96.8%, which is better than the other deep learning architectures. It also has the highest precision (97.4%) and F1-score (96.8%), which shows that it can classify things well overall and balance performance between accuracy and recall. It has a recall rate of 96.2%, which is a little lower than ResNet50's 96.8%, but it's still quite good.

ResNet50 is the best of the baseline models, with an accuracy of 96.5%, a precision of 96.6%, a recall of 96.8%, and an F1-score of 96.7%. InceptionResNetV2 and Xception are very close behind. The little variations in the measures show that typical CNN architectures work well, but the hybrid technique that combines autoencoder-based feature extraction with gradient boosting classification works even better. These findings show that combining

radiomic-informed latent features with a GBM classifier may improve the ability to find brain tumours more than using CNN models alone. The higher accuracy and F1-score show that the hybrid model lowers the number of false positives while keeping sensitivity balanced, which is very important in medical diagnosis.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

Brain cancer remains one of the most challenging and life-threatening forms of cancer due to its anatomical complexity, non-specific symptoms, and the often-late stage at which it is diagnosed. Traditional diagnostic techniques, although widely practiced, suffer from multiple limitations such as subjectivity, inter-observer variability, and restricted ability to analyze high-dimensional medical imaging data. With the growing emphasis on data-driven healthcare and personalized medicine, there is a pressing need for accurate, interpretable, and automated diagnostic models that can support radiologists in early and reliable tumour detection.

This research presented a radiomics-informed hybrid diagnostic model that combines the strengths of unsupervised deep learning and robust machine learning classifiers. Specifically, a convolutional autoencoder was employed to extract meaningful latent features from MRI images, while a Gradient Boosting Machine (GBM) classifier was utilized for final classification. This hybrid architecture effectively handled the challenges associated with high-dimensional radiomics data, reducing redundancy while preserving crucial diagnostic information.

The model was trained and validated on a balanced dataset of 800 brain MRI images comprising 408 normal and 392 abnormal (tumour-present) scans. Through comprehensive preprocessing steps—including image normalization, resizing, gray-scale transformation, and data augmentation—the dataset was prepared for optimal training performance. The hybrid Autoencoder + GBM model consistently outperformed traditional approaches such as standalone GBM, CNN, and Autoencoder models across key performance metrics. It achieved an accuracy of 96.8%, precision of 97.4%, recall of 96.2%, and an AUC score of 0.99, confirming its superior diagnostic capability.

In addition to its classification performance, the model demonstrated strong generalizability, reducing both false positives and false negatives—two critical factors in medical diagnosis. Moreover, by leveraging interpretable machine learning components like GBM, the framework aligns well with the clinical demand for transparency and explanation in AI-driven decision support systems.

The outcome of this study not only validates the effectiveness of hybrid models in radiomics-based cancer detection but also underscores their potential as a reliable clinical aid. The ability to establish meaningful correlations between deep latent features and clinical outcomes makes this model a strong candidate for future integration into computer-aided diagnostic (CAD) systems.

5.1. Future Work

While the proposed model has shown promising results, several avenues for further research and development remain open. Expanding upon this foundational work can enhance its clinical applicability, scalability, and overall robustness.

1. **Integration with Multi-modal Data** This study focused exclusively on radiomics features derived from 2D MRI images. Future research can extend the framework to incorporate multi-modal data, including genomic profiles, histopathology slides, patient demographics, and clinical biomarkers. Integrating such heterogeneous data could enhance predictive power and support personalized diagnostic pathways.
2. **Application to 3D Volumetric Imaging** Medical imaging increasingly relies on 3D volumetric scans (e.g., 3D MRI, CT). Transitioning the current model to handle volumetric data using 3D autoencoders and spatially-aware classifiers could provide more comprehensive insights into tumour morphology and spread.
3. **Real-time Deployment and Clinical Validation** A critical step forward is the deployment and validation of the model in real clinical environments. Collaborations with hospitals and diagnostic labs can help test the model on prospective, real-world datasets, assess its utility in real-time diagnostic workflows, and identify operational bottlenecks.
4. **Enhancing Model Explainability** While GBM offers better interpretability compared to black-box deep networks, further enhancements can be introduced using explainable AI (XAI) techniques such as SHAP values, Grad-CAM visualizations, or attention-based mechanisms. These tools can provide clinicians with clearer justifications for each prediction, thereby increasing trust in the model's outputs.
5. **Optimization of Computational Efficiency** Although the current model is computationally efficient compared to complex CNNs, future work may explore model compression techniques such as pruning, quantization, or knowledge distillation to enable deployment on low-resource hardware or mobile diagnostic tools.
6. **Cross-Cancer Generalization** Given the versatility of radiomics and autoencoder-based architectures, the model can be adapted and retrained for other types of cancers such as lung, prostate, or liver cancers. This generalization could lead to the creation of a unified AI framework for multi-cancer detection using a shared architecture.
7. **Incorporation of Federated Learning** To address data privacy concerns and facilitate collaborative learning across multiple healthcare institutions, the

model could be restructured within a federated learning paradigm. This would allow decentralized model training without direct sharing of patient data, enabling broader adoption while ensuring compliance with data protection regulations.

5.2. Clinical Applicability & Real-World Validation

While the proposed hybrid autoencoder and Gradient Boosting Machine (GBM) model demonstrates high accuracy on publically available datasets, its effectiveness in the real-world clinical environment remains to be validated. Integration of Artificial Intelligence models into clinical work flows requires careful consideration of factors. To facilitate clinical adoption, the model should be tested on independent, multi-centre clinical datasets to assess its robustness and generalizability beyond control research environment. [25]

5.3. Final Thoughts

This research reinforces the potential of hybrid AI systems in medical imaging and diagnostics. By combining the latent feature learning capacity of deep learning with the structured decision-making of ensemble methods, the proposed framework stands as a significant step forward in the domain of brain cancer detection. Continued development and clinical integration of such models could lead to earlier diagnoses, improved treatment planning, and ultimately, better outcomes for patients affected by one of the most critical forms of cancer.

REFERENCES

- [1] K. Rhoda, Y. Choonara, P. Kumar, D. Bijukumar, L. du Toit, and V. Pillay, "Potential nanotechnologies and molecular targets in the quest for efficient chemotherapy in ovarian cancer," *Expert Opinion on Drug Delivery*, vol. 12, no. 4, pp. 613–634, 2015.
- [2] Y. Zhou, Z. Han, F. Dou, and T. Yan, "Pre-colectomy location and tnm staging of colon cancer by the computed tomography colonography: a diagnostic performance study," *World Journal of Surgical Oncology*, vol. 19, pp. 1–13, 2021.
- [3] H. Sung et al., "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.
- [4] S. Das, H. Mazumdar, K. Khondakar, and A. Kaushik, "Machine learning integrated graphene oxide-based diagnostics, drug delivery, analytical approaches to empower cancer diagnosis," *BMEMat*, 2024, article e12117.
- [5] E. Limkin et al., "Promises and challenges for implementing computational medical imaging (radiomics) in oncology," *Annals of Oncology*, vol. 28, no. 6, pp. 1191–1206, 2017.
- [6] P. Lambin et al., "Radiomics: Extracting more information from medical images using advanced feature analysis," *European Journal of Cancer*, vol. 48, no. 4, pp. 441–446, 2012.
- [7] L. Mao et al., "Knowledge-informed machine learning for cancer diagnosis and prognosis: a review," *IEEE Transactions on Automation Science and Engineering*, 2024.
- [8] H. Lee, S. Moon, J. Hong, J. Lee, and S. Hyun, "A machine learning approach using fdg pet-based radiomics for prediction of tumour mutational burden and prognosis in stage iv colorectal cancer," *Cancers*, vol. 15, no. 15, p. 3841, 2023.
- [9] X. Zheng et al., "A ct-based radiomics nomogram for predicting the progression-free survival in small cell lung cancer: a multicenter cohort study," *Radiologia Medica*, vol. 128, no. 11, pp. 1386–1397, 2023.
- [10] L. Dercle et al., "Early readout on overall survival of patients with melanoma treated with immunotherapy using a novel imaging analysis," *JAMA Oncology*, vol. 8, no. 3, pp. 385–392, 2022.
- [11] A. Hosny, C. Parmar, J. Quackenbush, L. Schwartz, and H. Aerts, "Artificial intelligence in radiology," *Nature Reviews Cancer*, vol. 18, no. 8, pp. 500–510, 2018.
- [12] A. D'Anna et al., "Comparative analysis of machine learning models for predicting pathological complete response to neoadjuvant chemotherapy in breast cancer: An mri radiomics approach," *Physica Medica*, vol. 131, p. 104931, 2025.
- [13] E. Mylona et al., "Optimizing radiomics for prostate cancer diagnosis: feature selection strategies, machine learning classifiers, and mri sequences," *Insights into Imaging*, vol. 15, no. 1, p. 265, 2024.
- [14] F. Tang, C. Xue, M. Law, C. Wong, T. Cho, and C. Lai, "Prognostic prediction of cancer-based on radiomics features of diagnostic imaging: the

performance of machine learning strategies,” *Journal of Digital Imaging*, vol. 36, no. 3, pp. 1081–1090, 2023.

[15] S. Mukherjee et al., “Radiomics-based machine-learning models can detect pancreatic cancer on prediagnostic computed tomography scans at a substantial lead time before clinical diagnosis,” *Gastroenterology*, vol. 163, no. 5, pp. 1435–1446, 2022.

[16] I. Daimiel Naranjo et al., “Radiomics and machine learning with multiparametric breast mri for improved diagnostic accuracy in breast cancer diagnosis,” *Diagnostics*, vol. 11, no. 6, p. 919, 2021.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[18] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[19] F. Chollet, *Deep learning with Python*. Simon and Schuster, 2021.

[20] C. Shorten and T. Khoshgoftaar, “A survey on image data augmentation for deep learning,” *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[21] D. Rumelhart, G. Hinton, and R. Williams, “Learning internal representations by error propagation,” in *Institute for Cognitive Science*, University of California, 1985.

[22] P. Baldi, “Autoencoders, unsupervised learning, and deep architectures,” in *Proceedings of ICML workshop on unsupervised and transfer learning, JMLR Workshop and Conference Proceedings*, 2012, pp. 37–49.

[23] M. Ebrahimipour, M. Taghizadeh, M. Fatehi, O. Mahdiyar, and J. Jamali, “Premature ventricular contractions detection by multidomain feature extraction and auto-encoder-based feature reduction,” *Circuits, Systems, and Signal Processing*, vol. 43, no. 5, pp. 3279–3296, 2024.

[24] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, “Brain tumour detection based on deep learning approaches,” *arXiv preprint arXiv:2301.00000*.

[25] S. Khalighi, K. Reddy, A. Midya, K. B. Pandav, and A. Madabhushi, “Artificial intelligence in neuro-oncology: advances and challenges in brain tumour diagnosis, prognosis, and precision treatment,” *npj Precision Oncology*, 2024.

Appendix A


LIST OF PUBLICATION

- [a] Anany Kirti, and Priyanka Meel, "Enhanced Brain Cancer Detection Using a Radiomics-Informed Hybrid Autoencoder and GBM Model," ACCTHPA-2025.[Scopus Indexed][Accepted]



Figure 14: ACCTHPA-2025

[b] Anany Kirti, and Priyanka Meel, "Wildfire Prediction Using Deep Learning & Remote Sensing," ICISS-2025.[Scopus Indexed][Accepted]



The poster is for the 8th International Conference on Information Science and Systems (ICISS 2025). It features a blue background with a white wavy line separating the top text from the bottom sections. The top left has the ICISS logo and the text 'UNIVERSITY OF OXFORD, OXFORD, UK SEPTEMBER 14-16, 2025'. The main title 'Call for Papers' is in large blue letters. Below it, a white box contains text welcoming attendees and mentioning the conference's history. The bottom left section, titled 'Organizing Committee', lists chairs and co-chairs from various universities. The bottom right section, titled 'Publication & Index', describes the double-blind review process and lists indexing services. A small image of a conference book is shown next to the publication information.

ICISS
Information Science and Systems

UNIVERSITY OF OXFORD, OXFORD, UK
SEPTEMBER 14-16, 2025

Call for Papers

We would like to warmly welcome you to attend the **8th International Conference on Information Science and Systems (ICISS 2025)**, which will be held at **University of Oxford, Oxford, UK** during **September 14-16, 2025**.

Founded in 2018, the International Conference on Information Science and Systems (ICISS) has been successfully held seven editions in cooperation with famous international universities such as Tokai University, Cardiff University, Edinburgh Napier University. Experts and scholars from more than 30 countries and regions including the UK, USA, South Korea, Australia, Japan, China, and Canada have participated in the conference.

Organizing Committee

Conference Chairs
Xiaodong Liu, Edinburgh Napier University, UK
Farid Meziane, University of Derby, UK

Technical Program Committee Co-chairs
Frank Wang, University of Kent, UK
Steve Furnell, University of Nottingham, UK
Gang Liu, Harbin Engineering University, China
Boris Kovalechuk, Central Washington University, USA
Ruben Picck, University of Zagreb, Croatia
Chunming Rong, University of Stavanger, Norway

Publication & Index

All papers will be strictly double blind reviewed by the program committee, and accepted papers after proper registration and presentation will be published into *Communications in Computer and Information Science* (Electronic ISSN: 1865-0937 & Print ISSN: 1865-0929) as a proceedings book volume. The book series will be indexed by **EI Compendex, Scopus, INSPEC, SCImago** and other database.
For more information, please visit: <https://www.springer.com/series/7499>

*All the previous Conference Proceedings have been indexed by EI Compendex and Scopus

Call for Papers

Figure 15: ICISS-2025

ICISS 2025

The 8th International Conference on Information
Science and Systems

Rita Mihaylova

iciss.org@outlook.com

+86-18302820449

Acceptance Notification

Paper ID: E0045

Paper Title: WILDFIRE PREDICTION USING DEEP LEARNING & REMOTE SENSING

Author(s): Anany Kirti, Priyanka Meel

To whom it may concern,

Congratulations!

Based on the reviewer's comment and recommendation, we are glad to inform that the paper identified above is accepted for presentation and publication. Accepted papers after proper registration and presentation will be published into **Communications in Computer and Information Science (Electronic ISSN: 1865-0937 & Print ISSN: 1865-0929)** as a proceedings book volume. **The book series will be indexed by Ei Compendex, Scopus, INSPEC, SCImago and other database. *All the previous Conference Proceedings have been archived in ACM Digital Library and indexed by Ei Compendex and Scopus.***

The 8th International Conference on Information Science and Systems (ICISS 2025) will be held at University of Oxford, Oxford, UK during September 14-16, 2025. This year, the conference is co-sponsored by Edinburgh Napier University, UK and University of Derby, UK.

On behalf of the organizing committee, we are cordially inviting you to attend the conference and present the paper at ICISS 2025 during **September 14-16, 2025**. Please complete the registration procedure before **May 25, 2025**. See below the registration procedure please.

We are looking forward to meeting you in Oxford, UK! (Online attendance is optional.)

Yours sincerely,

ICISS 2025 Organizing Committee



Figure 16:ICISS-2025 Acceptance Letter

Hi Anany Kirti,



You paid \$ 550.00 USD to 成都亚昂教育咨询有限公司

Create an account with PayPal and activate PayPal Refunded Returns service. Conditions apply.

[Activate PayPal Now](#)

Your purchase details

Your transaction ID:
36485678UJ189990H

Seller transaction ID:
9RK48993GL141512B

Purchase date:
15 May 2025

Payment to:
成都亚昂教育咨询有限公司

Payment from:
Anany Kirti

Invoice ID:
IFP-Y82E2-2505150000206934

Shipping address
Anany Kirti
19 HIG DEVGHAT
JHALWA, Allahabad
Allahabad
UTTAR PRADESH
211012
India

Description	Unit price	Qty	Amount
商品订单	\$ 412.50 USD	1	\$ 412.50 USD
Subtotal			\$ 412.50 USD
Tax			\$ 137.50 USD
Total			\$ 550.00 USD
Total amount you'll pay			₹ 49,371.85 INR

You paid using: Visa x-6005

Figure 17: Payment Receipt of ICISS-2025 Conference



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis ENHANCED BRAIN CANCER DETECTION USING A RADIOMICS-INFORMED
HYBRID AUTOENCODER AND GBM MODEL

Total Pages 52 Name of the Scholar ANANY KIRTI

Supervisor (s)

(1) DR. PRIYANKA MEEL

(2) _____

(3) _____

Department INFORMATION TECHNOLOGY

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: TURNITIN Similarity Index: 8 %, Total Word Count: 10,555

Date: 28-05-2025

Candidate's Signature

Signature of Supervisor(s)

Anany_Thesis_word11plag.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98202495

Submission Date

May 28, 2025, 8:18 PM GMT+5:30

Download Date

May 28, 2025, 8:21 PM GMT+5:30

File Name

Anany_Thesis_word11plag.pdf

File Size

666.2 KB

38 Pages

10,555 Words

63,043 Characters





8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

-  **86 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 3%  Internet sources
- 4%  Publications
- 5%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 86 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 3% Internet sources
- 4% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- 1** **Publication**
Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ... <1%
- 2** **Publication**
Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli... <1%
- 3** **Internet**
arxiv.org <1%
- 4** **Publication**
Xiaoman Huang, Juntao Xiong, Huaiyin Lin, Zijian Pan, Kailin Wang, Mingyue Zha... <1%
- 5** **Publication**
"Intelligent Systems", Springer Science and Business Media LLC, 2025 <1%
- 6** **Publication**
Tasneem Ahmed, Shrish Bajpai, Mohammad Faisal, Suman Lata Tripathi. "Advanc... <1%
- 7** **Submitted works**
Teerthanker Mahaveer University on 2024-12-24 <1%
- 8** **Internet**
mail.easychair.org <1%
- 9** **Internet**
www.frontiersin.org <1%
- 10** **Publication**
Suman Kumar Swarnkar, Abhishek Guru, Gurpreet Singh Chhabra, Harshitha Rag... <1%

11	Submitted works	University of Hull on 2024-10-31	<1%
12	Submitted works	The Hong Kong Polytechnic University on 2024-03-26	<1%
13	Internet	www.mdpi.com	<1%
14	Internet	www.techscience.com	<1%
15	Publication	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dharendra Kumar Shukla. "Intelli...	<1%
16	Submitted works	Bocconi University on 2023-09-17	<1%
17	Submitted works	Liverpool John Moores University on 2025-02-17	<1%
18	Publication	Sovanlal Mukherjee, Anurima Patra, Hala Khasawneh, Panagiotis Korfiatis et al. "...	<1%
19	Submitted works	University of Hertfordshire on 2025-04-28	<1%
20	Submitted works	Aalto Yliopisto on 2025-05-25	<1%
21	Submitted works	Associatie K.U.Leuven on 2023-06-13	<1%
22	Submitted works	Queen's University of Belfast on 2025-04-14	<1%
23	Internet	philarchive.org	<1%
24	Submitted works	Coventry University on 2023-08-08	<1%

25	Submitted works	Lincoln University on 2024-10-19	<1%
26	Submitted works	University of Warwick on 2011-03-09	<1%
27	Submitted works	Fakultet elektrotehnike i računarstva / Faculty of Electrical Engineering and Com...	<1%
28	Internet	engrxiv.org	<1%
29	Publication	Fuk-hay Tang, Cheng Xue, Maria YY Law, Chui-ying Wong, Tze-hei Cho, Chun-kit L...	<1%
30	Submitted works	National University of Ireland, Galway on 2021-09-09	<1%
31	Submitted works	Imperial College of Science, Technology and Medicine on 2024-03-22	<1%
32	Internet	www.geeksforgeeks.org	<1%
33	Submitted works	Georgia Institute of Technology Main Campus on 2023-09-25	<1%
34	Submitted works	Hong Kong University of Science and Technology on 2023-11-17	<1%
35	Submitted works	University of Sydney on 2021-05-08	<1%
36	Internet	grietneukermans.weebly.com	<1%
37	Internet	link.springer.com	<1%
38	Publication	Bhaveshkumar C. Dharmani, Suman Lata Tripathi. "Intelligent Circuit and System...	<1%

39	Publication	Eren, Berke. "Deep Learning Based Channel Equalization for MIMO ISI Channels", ...	<1%
40	Submitted works	University of Bolton on 2025-04-17	<1%
41	Internet	nano-ntp.com	<1%
42	Publication	"Cyber Security Intelligence and Analytics", Springer Science and Business Media ...	<1%
43	Publication	Joice. C Sheeba, M. Selvi. "Pedagogical Revelations and Emerging Trends", CRC Pr...	<1%
44	Internet	cancerres.aacrjournals.org	<1%
45	Submitted works	Liverpool John Moores University on 2025-04-05	<1%
46	Submitted works	Maastricht University on 2023-09-09	<1%
47	Publication	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machin...	<1%
48	Submitted works	Tilburg University on 2025-05-16	<1%
49	Submitted works	University Politehnica of Bucharest on 2024-06-26	<1%
50	Submitted works	University of Iceland on 2015-05-20	<1%
51	Submitted works	World Maritime University on 2025-05-23	<1%
52	Internet	qar.srmap.edu.in	<1%

53

Internet

www.ncbi.nlm.nih.gov

<1%

Anany_Thesis_word11plag.pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid:::27535:98202495

Submission Date

May 28, 2025, 8:18 PM GMT+5:30

Download Date

May 28, 2025, 8:21 PM GMT+5:30

File Name

Anany_Thesis_word11plag.pdf

File Size

666.2 KB

38 Pages

10,555 Words

63,043 Characters

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

