AUTOMATED DETECTION AND EVALUATION OF ISCHEMIC STROKE ON CT BRAIN IMAGING USING MACHINE LEARNING TECHNIQUES

NUR AMIRAH ATIKAH BINTI SHARUDDIN

SCHOOL OF HEALTH SCIENCES

UNIVERSITI SAINS MALAYSIA

2025

AUTOMATED DETECTION AND EVALUATION OF ISCHEMIC STROKE ON CT BRAIN IMAGING USING MACHINE LEARNING TECHNIQUES

by

NUR AMIRAH ATIKAH BINTI SHARUDDIN

Dissertation submitted in partial fulfilment of the requirement for the degree of Bachelor of Health Sciences (Honours) (Medical Radiation)

JUNE 2025

CERTIFICATE

This is to certify that the dissertation entitled "**AUTOMATED DETECTION AND EVALUATION OF ISCHEMIC STROKE ON CT BRAIN IMAGING USING MACHINE LEARNING TECHNIQUES**" is the bona fide record of research work done by NUR AMIRAH ATIKAH BINTI SHARUDDIN during the period from October 2024 to June 2025 under my supervision. I have read this dissertation and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation to be submitted in partial fulfilment for the degree of Bachelor of Health Science (Honours) (Medical Radiation).

Main Supervisor:	Co-Supervisor:	Field-Supervisor:
………………………….	………………………….	……………………….
Pn. Siti Aishah Abd Aziz Senior Lecterur School of Health Science Universiti Sains Malaysia Health Campus Kubang Kerian 16150 Kubang Kerian Kelantan, Malaysia	Dr. Nur Asma Sapiai Lecterur/Neuroradiologist School of Medical Sciences Universiti Sains Malaysia Health Campus Kubang Kerian 16150 Kubang Kerian Kelantan, Malaysia	Pn. Norfataha Mohd Daud Radiographer Universiti Sains Malaysia Health Campus Kubang Kerian 16150 Kubang Kerian Kelantan, Malaysia

Date: June 2025

DECLARATION

I, NUR AMIRAH ATIKAH BINTI SHARUDDIN hereby declare that the dissertation entitled "**AUTOMATED DETECTION AND EVALUATION OF ISCHEMIC STROKE ON CT BRAIN IMAGING USING MACHINE LEARNING TECHNIQUES**" is the result of my own investigations, except where otherwise stated and duly acknowledged. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at Universiti Sains Malaysia or other institutions. I grant Universiti Sains Malaysia the right to use the dissertation for teaching, research and promotional purposes.

....

NUR AMIRAH ATIKAH BINTI SHARUDDIN

Date: June 2025

ACKNOWLEDGEMENT

Alhamdulillahirabbil ‘alamin, all praise and gratitude be to Allah SWT, the Most Gracious and the Most Merciful, for granting me the strength and patience to complete this thesis. Without His endless blessings and guidance, none of this would have been possible.

I would like to extend my deepest appreciation to my supervisor, Pn. Siti Aishah Abd Aziz, for their unwavering support, guidance, and encouragement throughout this journey. Their thoughtful advice and insightful feedback have been invaluable. I am truly grateful for their mentorship and the time and effort they have invested in me.

To my beloved parents Azizah Binti Sulaiman and Sharuddin Bin Othman, no words can truly express how thankful I am for your unconditional love and constant prayers. You have been my pillars of strength and the reason I never gave up. Thank you for believing in me even when I doubted myself. To my family and friends, thank you for your understanding, and continuous encouragement. Your support has been a quiet yet powerful force behind my progress, especially during the most challenging times.

A special note of gratitude to the Institute for Artificial Intelligence and Big Data (AIBIG), Universiti Malaysia Kelantan (UMK) for their kind guidance and valuable assistance in the data analysis phase of this study. Your expertise, and support have been crucial in helping me navigate and interpret complex data with confidence and clarity

Lastly, I would like to extend my heartfelt gratitude to myself for enduring sleepless nights, and for pushing through moments of doubt and fatigue. This thesis is not only a symbol of academic achievement, but also of personal growth anad strength. This journey has been one of great learning academically, emotionally, and spiritually. I am humbled by the experience and grateful to everyone who has been a part of it.

TABLE OF CONTENTS

LIST OF FIGURE

LIST OF TABLE

LIST OF ABBREVIATIONS

AIBIG - Institute for Artificial Intelligence and Big Data

AISD - Acute Ischemic Stroke Detection

AUC - Area Under the Curve

B-spline - B-spline Interpolation

CAD - Computer-Aided Diagnosis

CT - Computed Tomography

DICOM - Digital Imaging and Communications in Medicine

DL - Deep Learning

EEG - Electroencephalography

fMRI - Functional Magnetic Resonance Imaging

F1-score - Harmonic Mean of Precision and Recall

GLCM - Gray-Level Co-occurrence Matrix

HPUSM - Hospital Pakar Universiti Sains Malaysia

HRF - Hemorrhagic Stroke

HU - Hounsfield Unit

KNN - K-Nearest Neighbours

LBP - Local Binary Pattern

LDA - Linear Discriminant Analysis

MRI - Magnetic Resonance Imaging

MNI - Montreal Neurological Institute

NCC - Non-contrast CT

PACS - Picture Archiving and Communication System

PET - Positron Emission Tomography

ROC - Receiver Operating Characteristic

RF - Random Forest

SNR - Signal-to-Noise Ratio

SVM - Support Vector Machine

SPM12 - Statistical Parametric Mapping Version 12

SMOTE - Synthetic Minority Over-sampling Technique

SVM - Support Vector Machine

TN - True Negative

TP - True Positive

UCN - Unconscious Nervous

USM - Universiti Sains Malaysia

WHO - World Health Organization

PENGESANAN DAN PENILAIAN AUTOMATIK STROK ISKEMIK PADA IMEJAN CT OTAK MENGGUNAKAN TEKNIK PEMBELAJARAN MESIN

ABSTRAK

Kajian ini menyiasat aplikasi algoritma pembelajaran mesin untuk pengesanan strok iskemik menggunakan imej CT otak. Strok, terutamanya strok iskemik, kekal sebagai penyebab utama kematian dan kecacatan di seluruh dunia. Pengesanan dan diagnosis strok iskemik pada peringkat awal adalah penting untuk mengurangkan kerosakan jangka panjang dan meningkatkan hasil pesakit. Kaedah diagnosis tradisional bergantung kepada kepakaran ahli radiologi, yang boleh memakan masa dan terdedah kepada variabiliti antara pemerhati. Penyelidikan ini bertujuan untuk membangunkan sistem automatik untuk pengesanan strok iskemik dengan menggunakan teknik pembelajaran mesin seperti Support Vector Machine (SVM), K-Nearest Neighbours (KNN), dan Random Forest (RF), yang diterapkan pada imej CT otak.

Kajian ini menggunakan set data yang terdiri daripada 397 imej CT strok iskemik dan 25 imej otak normal. Beberapa langkah pemprosesan awal, termasuk pengecilan saiz, normalisasi, dan pengurangan hingar, telah dijalankan ke atas imej CT bagi memastikan ia sesuai untuk analisis pembelajaran mesin. Ciri-ciri yang relevan telah diekstrak daripada imej, seperti intensiti, tekstur, dan bentuk, yang kemudiannya digunakan untuk melatih model pembelajaran mesin. Prestasi model dinilai menggunakan metrik seperti ketepatan, ketepatan positif, ketinggalan, skor F1, dan AUC. Model Random Forest menunjukkan ketepatan tertinggi sebanyak 92.76%, dengan AUC sebanyak 0.973, mengatasi model KNN dan SVM. Model KNN mencapai ketepatan sebanyak 93.93%

dengan AUC 0.940, manakala model SVM mencapai ketepatan sebanyak 87.87% dengan AUC 0.984.

Selain itu, masa latihan bagi setiap model dicatatkan: SVM mengambil masa 0.0152 saat, KNN mengambil masa 0.0114 saat, dan Random Forest mengambil masa 0.2083 saat. Keputusan ini menunjukkan bahawa model pembelajaran mesin, terutamanya Random Forest dan KNN, boleh memberikan pengesanan strok yang tepat dan konsisten, menawarkan potensi untuk aplikasi klinikal yang cepat dan boleh dipercayai, dengan KNN menjadi yang paling pantas dalam masa latihan.

AUTOMATED DETECTION AND EVALUATION OF ISCHEMIC STROKE ON CT BRAIN IMAGING USING MACHINE LEARNING TECHNIQUES

ABSTRACT

This study investigates the application of machine learning algorithms for the detection of ischemic stroke using CT brain images. Stroke, particularly ischemic stroke, remains a leading cause of death and disability globally. The early detection and diagnosis of ischemic stroke are crucial for minimizing long-term damage and improving patient outcomes. Traditional methods of diagnosis rely on the expertise of radiologists, which can be time-consuming and prone to inter-observer variability. This research aims to develop an automated system for ischemic stroke detection by leveraging machine learning techniques such as Support Vector Machine (SVM), K-Nearest Neighbours (KNN), and Random Forest (RF), applied to CT brain images.

The study uses a dataset consisting of 397 ischemic stroke CT images and 25 normal brain scans. A series of preprocessing steps, including resizing, normalization, and noise reduction, were performed on the CT images to ensure they were suitable for machine learning analysis. Relevant features were extracted from the images, such as intensity, texture, and shape, which were then used to train the machine learning models. The performance of the models was evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC. The Random Forest model achieved the highest accuracy at 92.76%, with an AUC of 0.973, outperforming both the KNN and SVM models. The KNN model achieved an accuracy of 93.93% with an AUC of 0.940, while the SVM model achieved an accuracy of 87.87% with an AUC of 0.984.

Additionally, the training time for each model was recorded: SVM took 0.0152 seconds, KNN took 0.0114 seconds, and Random Forest took 0.2083 seconds. The results demonstrate that machine learning models, particularly Random Forest and KNN, can provide accurate and consistent stroke detection, offering potential for rapid and reliable clinical application, with KNN being the fastest in training time.

CHAPTER 1

INTRODUCTION

1.1. Background of the Study

Strokes are a leading cause of death and disability worldwide. They occur when the blood supply to the brain is interrupted, either due to a blocked or burst blood vessel. As strokes can lead to severe consequences, including physical disabilities, brain damage, and death, early diagnosis and treatment are crucial (*Stroke Diagnosis: Ct or Mri Scan? | MedShun*, 2024). To work properly, your brain needs oxygen. Your arteries deliver oxygen-rich blood to all parts of your brain. If something happens to block the flow of blood, brain cells start to die within minutes, because they can't get oxygen and this causes a stroke *(CDC, 2024)*. According to the World Health Organization, 15 million people worldwide suffer a stroke per year. Among those, 5 million die, and another 5 million end up being permanently disabled. While brain stroke commonly occurs in people over 40 years old, it can also strike younger individuals and is the main cause of high blood pressure in those under 40**.** *(World Health Organization, 2022)*

Brain strokes can be divided primarily into two types: ischemic stroke, which is caused by blockages (such as blood clots), and hemorrhagic stroke, caused by bleeding in or around the brain. An ischemic stroke, which is the most common type, occurs when a blood clot blocks a blood vessel in your brain. Hemorrhagic strokes occur when a blood vessel in your brain ruptures *(Yetman, 2023)*. About 87% of all strokes are ischemic. Fatty deposits called plaque can also cause blockages by building up in the blood vessels. While the hemorrhagic stroke is about 13% and happens when an artery in the brain leaks blood or ruptures (breaks open). The leaked blood puts too much pressure on brain cells, which damages them ***(CDC, 2024).***

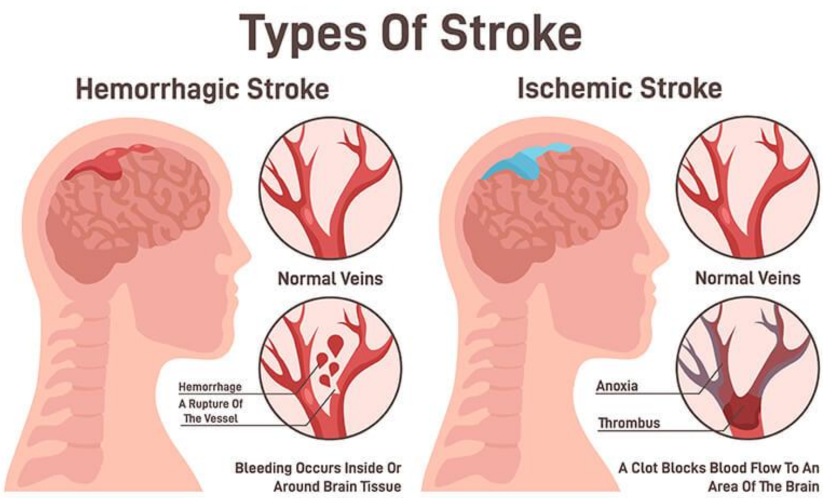

Figure 1 : Types Of Stroke (Trihealth.com, 2024)

Neuroimaging technique for stroke detection such as computed tomography (CT) has been widely used for emergency setting that can provide precise information on an obvious difference between white and gray matter (Hasanah Ali et al., 2021). CT scans are widely available, provide faster results, and are less sensitive to motion. The choice between CT and MRI scans depends on various factors, including the availability of advanced technology, time constraints, and the patient's condition. In emergency situations, CT scans are often preferred due to their speed and accessibility, while MRI scans may be more suitable for follow-up examinations or when highly detailed images are required (*Stroke Diagnosis: Ct or Mri Scan? | MedShun*, 2024). A CT scan provides detailed images of the brain, which are essential for identifying abnormalities like blockages or bleeding. However, interpretation of CT images requires expertise, as the subtle differences between stroke types or the presence of less apparent abnormalities can lead to diagnostic challenges. Given that early diagnosis is critical in stroke care, automated analysis using CT images can potentially accelerate diagnosis. Machine learning can assist in extracting relevant patterns and identifying stroke characteristics that may not be immediately apparent to the human eye.

According to Chan et al., (2023), machine learning (ML), a branch of artificial intelligence, enables computers to learn patterns from data, improving their performance on specific tasks without being explicitly programmed. In recent years, machine learning has demonstrated substantial promise in medical imaging analysis, including applications in disease detection, classification, and prediction. For stroke classification, ML techniques can analyze complex patterns in CT images, differentiating stroke types by learning from a large dataset of labeled images. The primary advantages of machine learning in stroke classification include:

- Efficiency: ML algorithms can process large volumes of images much faster than human interpretation, which is beneficial in time-sensitive medical situations.
- Consistency: Unlike human diagnosis, ML models do not suffer from fatigue and provide consistent interpretations.

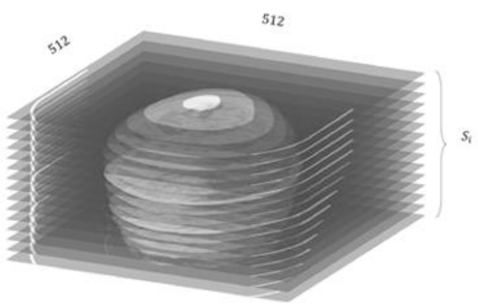

Figure 2: Brain CT Scan of a Patient (Babutain et al., 2021)

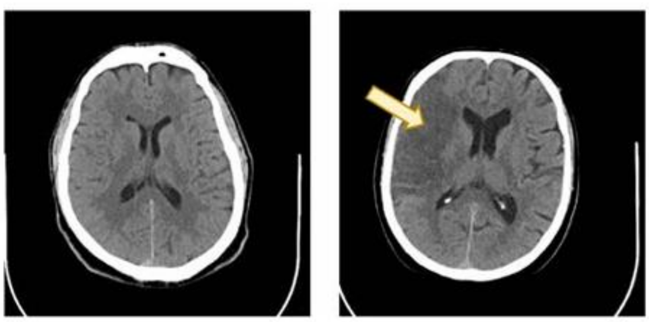

Figure 3:Normal Brain CT Slice (Left) and Acute Ischemic Stroke (Right) (Babutain et al., 2021)

1.2. Problem statement & Study rationale

Stroke remains a leading cause of mortality and long-term disability worldwide, necessitating accurate and timely diagnosis to initiate appropriate treatment and minimise brain damage (Badriyah et al., 2020). Ischemic stroke, caused by a blood clot blocking brain vessels, requires early intervention with clot-dissolving treatments, making its rapid identification crucial. While Computed Tomography (CT) is widely used for initial stroke diagnosis due to its accessibility and speed, interpreting CT images presents challenges due to subtle and overlapping imaging features, particularly in distinguishing ischemic from hemorrhagic strokes *(Gautam & Raman, 2021; Ho et al., 2019).*

Manual interpretation of CT scans, often reliant on expert radiologists, is time-consuming, prone to inter-observer variability, and difficult to standardise, especially in resource-limited settings. Existing methods for automated stroke detection using machine learning have shown promise but face limitations such as imbalanced datasets, lack of optimization, and reliance on advanced imaging modalities like MRI, which are not always available *(Badriyah et al., 2020).*

Machine learning approaches like K-Nearest Neighbours (KNN)**,** have demonstrated success in extracting complex features from medical images for accurate classification (Liu et al., 2023). Traditional models, such as Support Vector Machines (SVM) and Random Forests (RF), also offer reliable stroke classification using extracted imaging features. However, there remains a need to compare and optimize these models specifically for ischemic stroke detection using CT images to identify the most accurate and clinically applicable approach.

This research aims to address these challenges by developing and evaluating a machine learning framework using MATLAB that compares the performance of KNN, SVM, and RF models for ischemic stroke classification. By identifying the best-performing model, this study seeks to improve diagnostic accuracy and speed, supporting clinicians in making timely and informed decisions for effective stroke management.

1.3. Objective

1.3.1. General Objective:

To develop an automatic classification system that accurately identifies ischemic stroke in CT brain images, enhancing the diagnostic process and supporting clinical decision-making in stroke care.

1.3.2. Specific Objective:

1) To design an image processing loop in CT Brain segmentation for future extractions of Ischemic Stroke imaging data.
2) To apply K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest (RF) in identifying the most suitable algorithm for ischemic stroke detection
3) To evaluate the overall performance of K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forest (RF) in identifying the most suitable algorithm for ischemic stroke detection.

1.4. Hypothesis

1.4.1. Null Hypothesis (H0):

- There is no significant difference in the performance of KNN, SVM, and RF models in classifying ischemic stroke using CT brain imaging data.
- There is no correlation between the evaluation performance metrics (accuracy, sensitivity, specificity, area under curve (AUC)). F1 score& 10 fold-cross validation of the models and their suitability for ischemic stroke detection.
- There is no improvement in the accuracy or clinical suitability of the optimized best-performing model for ischemic stroke classification.

1.4.2. Alternative Hypothesis (H1):

- There is a significant difference in the performance of KNN, SVM, and RF models in classifying ischemic stroke using CT brain imaging data.
- There is a correlation between the evaluation performance metrics (accuracy, sensitivity, specificity, area under curve (AUC)). F1 score& 10 fold-cross validation of the models and their suitability for ischemic stroke detection.
- There is an improvement in the accuracy and clinical suitability of the optimized best-performing model for ischemic stroke classification.

1.5. Significant of Study

This study is significant because it aims to improve the early detection and classification of ischemic stroke using machine learning (ML) techniques applied to CT brain images. Ischemic stroke is a leading global health issue, often resulting in death or long-term disability if not diagnosed and treated promptly. While CT scans are commonly used for initial assessments due to their speed and accessibility, detecting ischemic strokes can be challenging, especially in the early stages, and is further complicated in areas with limited access to trained radiologists. This research compares K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forest (RF) models, addressing the gap in literature by providing a consistent evaluation of these models under identical conditions. The study applies a standardised preprocessing pipeline and validation techniques such as cross-validation and performance metrics (accuracy, sensitivity, specificity, F1-score, and AUC), ensuring robust, reproducible results. The findings can help guide the development of AI-assisted diagnostic tools for clinical settings, particularly in emergency and rural healthcare centers. Clinically, the optimised model can assist radiologists in diagnosing ischemic strokes more accurately and quickly, potentially reducing misdiagnosis and treatment delays, ultimately improving patient outcomes. This research also contributes to the growing field of medical AI, providing valuable insights into model selection and deployment in healthcare.

1.6. Conceptual Framework

The conceptual framework of this study, illustrated in Figure 1.6, focuses on the automatic detection of ischemic stroke using machine learning techniques applied to CT brain imaging. The CT images obtained for both training and testing sets will undergo image pre-processing, detection, and feature extraction, where relevant features will be extracted from the pixel values. These features will then be classified using machine learning classifiers to distinguish between ischemic stroke and non-stroke cases.

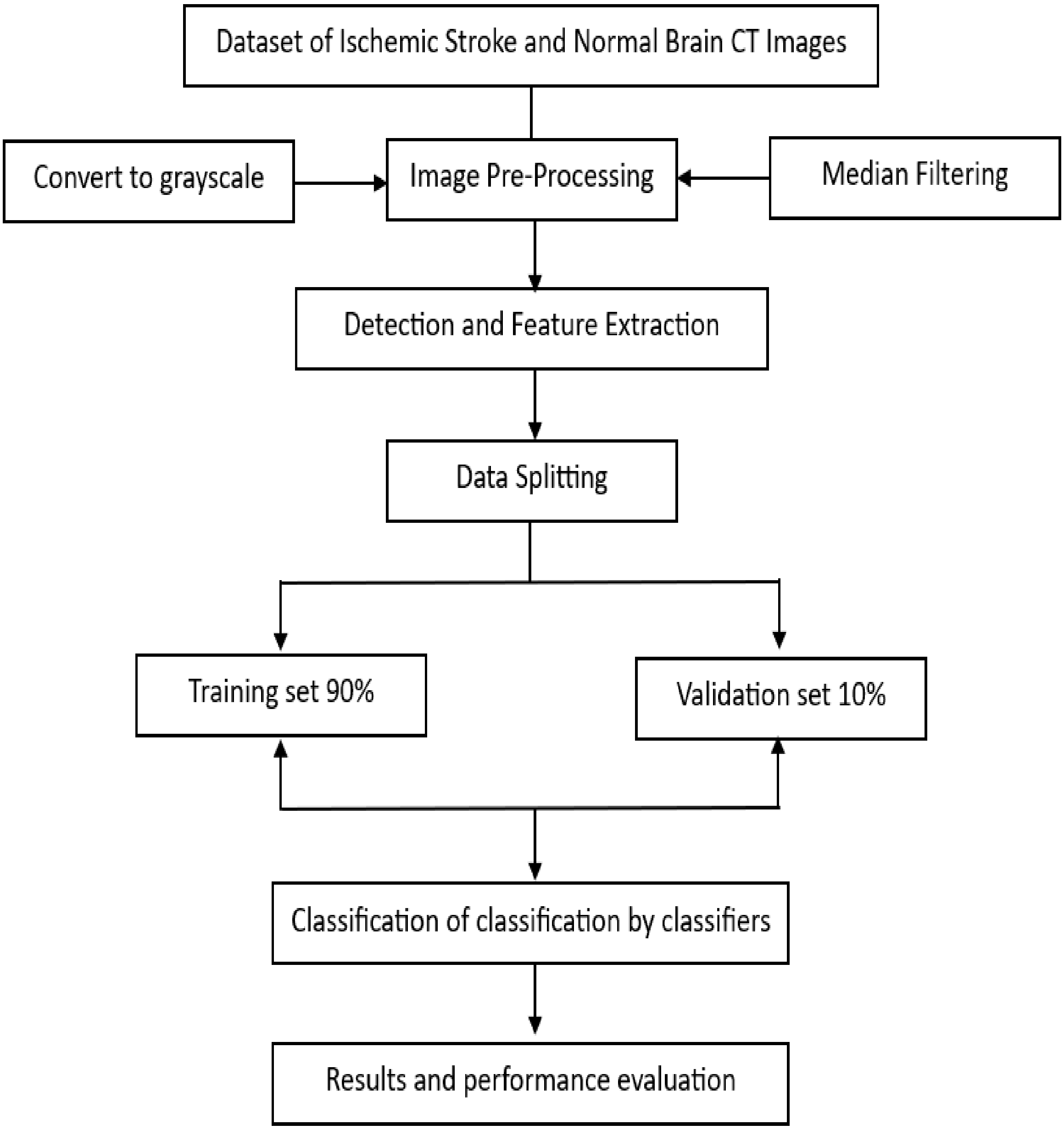

Figure 4: Conceptual Framework

CHAPTER 2

LITERATURE REVIEW

2.1. Brain CT Imaging and Calcification Detection

Stroke is a life-threatening medical emergency that continues to be one of the leading causes of death and long-term disability globally. Among its two main types which are ischemic and haemorrhage. Ischemic stroke accounts for approximately 87% of all cases, occurring due to a blockage in the blood vessels supplying the brain (Badriyah et al., 2020; Tursynova et al., 2023). One of the advantages of CT imaging over other techniques, such as magnetic resonance imaging (MRI), is its ability to provide high-resolution images at a lower cost and with a shorter acquisition time.

Early detection and precise classification of stroke type are critical to initiating timely treatment and minimizing neurological damage. In clinical settings, Computed Tomography (CT) is widely used as the first-line imaging modality because of its speed, affordability, and broad availability (Fontanella et al., 2023; Mokli et al., 2019). This makes it particularly useful in emergency settings where rapid diagnosis is essential. However, interpreting CT images accurately can be difficult due to overlapping features between stroke types, subtle lesion visibility in early phases, and a heavy dependence on the skill of the radiologist.

The detection of calcifications in brain CT images requires a combination of expert interpretation and advanced imaging algorithms. While traditional CT imaging can effectively detect large calcifications, smaller calcifications or those located in complex anatomical regions may be missed without the use of advanced computational tools. The conventional method for detecting calcifications involves radiologists visually inspecting the CT scan. However, this process can be time-consuming and subject to human error, especially when the calcifications are subtle or located in difficult-to-visualize areas (Shalikar et al., 2014)

Recent years have seen the emergence of Machine Learning (ML) and Deep Learning (DL) as powerful tools to assist radiologists in stroke detection and classification. These computational methods have the potential to improve diagnostic consistency, reduce interpretation time, and support clinical decision-making, especially in resource-limited settings. The use of machine learning (ML) and deep learning (DL) models for automated calcification detection is a growing area of research. These algorithms, particularly convolutional neural networks (CNNs), are trained on large datasets of brain CT images to detect and classify calcifications based on their size, shape, and location.

A study by Shalikar et al. (2014) introduced a computer-aided diagnosis (CAD) system using wavelet transforms and texture feature extraction, achieving 90% accuracy in classifying brain CT images as normal, ischemic, or hemorrhagic. This type of system could significantly improve the accuracy and efficiency of calcification detection, particularly in clinical settings where time is critical. This chapter reviews the current literature on ML and DL approaches applied to stroke detection and classification using CT imaging, with a focus on comparing models like KNN, SVM and RF.

2.2. Machine Learning Techniques in Stroke Detection

Machine learning (ML) has emerged as a promising approach to automate the process of ischemic stroke detection from medical imaging data, particularly Computed Tomography (CT) scans. Unlike traditional rule-based systems, ML algorithms can learn patterns from large datasets without explicit programming, making them highly effective in complex tasks like stroke detection (Gautam & Raman, 2021). Machine learning techniques can significantly reduce the time required for image analysis, improve diagnostic accuracy, and assist in making more consistent diagnoses across different clinicians. These techniques have been increasingly applied in medical diagnostics, particularly in the classification of medical images. Ischemic stroke detection has greatly benefited from advancements in machine learning (ML), particularly for automated classification and assessment of stroke severity from medical imaging data. Recent studies have focused on various machine learning models such as Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Random Forest (RF) for analying CT images. These methods provide a computational advantage by automatically learning from labelled datasets, which allows them to classify ischemic stroke lesions effectively (Gautam & Raman, 2021).

In a recent study by Xie et al. (2025), an imaging-based ML model was used to evaluate stroke severity in the middle cerebral artery (MCA) territory from CT images. The study employed Logistic Regression for predicting the severity of ischemic strokes, showing that machine learning models could correlate CT features like infarction volume with clinical stroke severity scores (NIHSS) (Xie et al., 2025). These results align with the findings of Badriyah et al. (2020), who demonstrated that ML models, particularly Random Forest, were effective for stroke classification, although accuracy can be sensitive to data quality.

These algorithms rely on manually extracted features from the CT images, such as texture, intensity, and shape, to classify images as either ischemic or normal. In ischemic stroke diagnosis, machine learning techniques have been used to classify CT scans as normal or ischemic. For instance, an SVM classifier has been applied to classify CT scans into ischemic and hemorrhagic stroke categories, achieving promising results (Tursynova et al., 2023)

2.2.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the most widely used supervised learning algorithms in medical image classification. SVM works by mapping the input data into a high-dimensional space and finding the optimal hyperplane that separates the classes. In ischemic stroke detection, SVM has been successfully applied to classify CT images into ischemic and non-ischemic categories.

In a study by Babutain et al. (2021), SVM was employed to classify ischemic stroke using pre-processed CT images. The authors reported that SVM performed reasonably well with a linear kernel, achieving an accuracy of 91.7%. However, this study highlighted the sensitivity of SVM to feature extraction and preprocessing methods, indicating that the choice of kernel and feature selection significantly impacts model performance. Additionally, Tursynova et al. (2023) showed that SVM's performance could be affected by the quality of data preprocessing, which makes it crucial to apply appropriate image enhancement and feature selection techniques to improve accuracy.

Despite its strong theoretical foundation and performance in certain settings, SVM's limitations are evident when applied to more complex datasets or when faced with noisy data. Fontanella et al. (2023) noted that while SVM could classify ischemic lesions with reasonable accuracy, its ability to handle multiple lesion types and varying sizes in real-world clinical data was suboptimal. Therefore, SVM might not always be the best choice in dynamic clinical environments that require real-time, accurate stroke detection.

2.2.2 K-Nearest Neighbours (KNN)

K-Nearest Neighbours (KNN) is a non-parametric algorithm that classifies data based on the majority class of its nearest neighbours. One of the main advantages of KNN is its simplicity and interpretability, making it a popular choice for medical image classification tasks. Gautam & Raman (2021) applied KNN to classify ischemic stroke from CT scans, achieving an accuracy of 93%. The study emphasized the algorithm's effectiveness when combined with preprocessing techniques like image segmentation and normalisation, which helped enhanced the feature representation. In stroke detection, KNN has been used to classify ischemic and normal CT brain images, based on features such as texture, intensity, and shape.

One of the key advantages of KNN is ease of interpretation. It does not require a training phase in the conventional sense, as it simply memorises the training data and classifies new samples based on their similarity to the stored instances. However, KNN is highly sensitive to the choice of the number of neighbours (k) and the distance metric used, which can make its performance highly variable, especially when dealing with large datasets or noisy images (Badriyah et al., 2020). Tursynova et al. (2023) highlighted that KNN showed the highest variability in performance among the three models tested, indicating that while it could achieve high accuracy in some cases, it might fail to generalise well when the dataset contains outliers or noise. In the context of ischemic stroke, where precise lesion identification is crucial, KNN's susceptibility to such variations might hinder its clinical applicability without additional steps like feature selection and data cleaning.

Despite these challenges, KNN's relatively low computational cost make it an attractive choice for rapid, on-site diagnosis, especially in emergency settings where speed is crucial. Furthermore, Sheth et al. (2023) showed that KNN, when combined with data augmentation techniques such as horizontal flips and rotations, could improve its robustness and reduce its sensitivity to image variations, suggesting potential improvements in clinical deployment.

2.2.3 Random Forest (RF)

Random Forest (RF), an ensemble learning method, builds multiple decision trees and aggregates their predictions to improve classification accuracy and reduce overfitting (Breiman, 2001). RF has gained significant attention in medical imaging applications, including ischemic stroke detection, due to its ability to handle large datasets with many features, as well as its robustness to noise and overfitting (Badriyah et al., 2020). Fontanella et al. (2023) reported that RF provided one of the best performances in ischemic stroke detection, with an accuracy of 94%, which was higher than that of SVM and KNN.

The strength of RF lies in its ability to manage both structured and unstructured data, providing a comprehensive model that can account for complex relationships between features. This was evident in the work of Babutain et al. (2021), where RF outperformed other algorithms in terms of precision, recall, and F1-score. The authors noted that RF's ability to aggregate predictions from multiple decision trees allowed it to generalize better across diverse datasets, making it particularly well-suited for clinical stroke detection, where variability in image quality and patient characteristics is common.

However, as Tursynova et al. (2023) pointed out, RF can sometimes produce suboptimal results when dealing with imbalanced datasets, as it may prioritize the majority class at the expense of the minority class (i.e., ischemic stroke lesions). To address this issue, the authors recommended balancing the dataset through techniques like SMOTE (Synthetic Minority Over-sampling Technique) or using class-weighted decision trees.

Moreover, Sheth et al. (2023) demonstrated that RF can be further enhanced by combining it with feature selection methods, such as recursive feature elimination (RFE), to improve performance by reducing overfitting and focusing on the most discriminative features. This makes RF a strong candidate for ischemic stroke classification, particularly when dealing with large, high-dimensional CT datasets.

The main advantage of traditional machine learning models is their interpretability, which allows clinicians to understand the reasoning behind the model's predictions. However, in a study by Radwan Qasrawi et al., (2024), the need for feature engineering and the reliance on manually selected features can limit their ability to capture the complex patterns present in medical images, which may hinder their performance in more challenging scenarios.

	Classification		Segmentation	Object Detection
	ML [1]	DL		
Models	• LR • DT • RF • SVM • NB • GB • ANN [1]	• ANN [2] • AlexNet [22] • VGG [23] • ResNet [24] • Inception [25] • DenseNet [26] • EfficientNet [27]	• FCN [28] • U-Net [29] • SegNet [30] • DeepLab [31] • Mask R-CNN [32]	• Faster R-CNN [33] • YOLO [34] • SSD [35] • RetinaNet [36] • DETR [37]
Metrics	• Accuracy • Precision • Recall • F1 Score • ROC-AUC • Confusion Matrix		• Pixel Accuracy • Intersection over Union • Dice Coefficient • Boundary Accuracy	• Precision and Recall • Average Precision • Mean Average Precision (mAP) • Intersection over Union (IoU)
Losses	• Binary Cross-Entropy Loss • Hinge Loss • Multi-Class Cross-Entropy Loss • Categorical Cross-Entropy Loss • Sparse Categorical • Cross-Entropy Loss		• Dice Loss • Cross-Entropy Loss • Jaccard Loss	• Focal Loss • Smooth L1 Loss • Combined Objectness and Class Specificity Loss • Region Proposal Network Loss

Figure 5: Overview of the most common models, metrics, and losses used in classification, segmentation, and object detection problems. (Fernandes, et al., 2024)

2.3. Preprocessing Techniques in Medical Imaging

Preprocessing in medical imaging plays a crucial role in ensuring that images are ready for accurate analysis and classification, particularly in the case of ischemic stroke detection. Raw medical images, such as those obtained from CT and MRI scans, often contain noise, varying contrast, and other artefacts that may obscure important information. Preprocessing techniques aim to enhance the relevant features of the image and reduce the impact of unwanted artefacts, making them more suitable for downstream analysis by ML algorithms. For ischemic stroke detection, effective preprocessing techniques are essential for improving the performance of automated systems designed to identify stroke-related changes in brain tissue.

The detection of ischemic stroke from brain CT images requires high sensitivity to subtle tissue changes caused by reduced blood flow and oxygenation. Raw images may contain several imperfections, such as:

- **Noise**: Variability in pixel intensity due to limitations in imaging devices.
- **Artefacts**: Unwanted patterns or distortions in the image caused by technical issues.
- **Varying Contrast**: Differences in brightness and contrast that can obscure critical stroke features.

Preprocessing aims to address these issues, improving the quality of the images for stroke detection algorithms. Techniques such as image normalisation, filtering, and enhancement are commonly employed to ensure that the data is suitable for further analysis.

Several preprocessing techniques are commonly used in the context of ischemic stroke detection, including image normalisation, noise reduction, image segmentation, and intensity adjustment. Each of these steps is designed to refine the input data before it is passed to ML models for stroke classification or lesion detection.

2.3.1. Image Normalisation

Image normalisation is a fundamental preprocessing step that ensures all images are standardised in terms of intensity values. This step is particularly important when working with CT scans, where intensity levels are measured in Hounsfield Units (HU). For stroke detection, it is crucial to adjust the images so that they have consistent intensity distributions, making it easier for the algorithm to detect changes in brain tissue associated with ischemia. Normalisation typically involves scaling the pixel values to a standard range or applying contrast-limited adaptive histogram equalization to enhance image features. For CT images, a typical range of Hounsfield Units for brain tissue can be from -1000 (air) to +3000 (bone).

By normalizing these values, the images become more comparable across different patient scans (Kang et al., 2021).

2.3.2. Noise Reduction

Medical images, particularly CT and MRI scans, can be noisy due to various factors like scanner limitations, patient movement, or interference from external sources. Noise reduction techniques, such as Gaussian smoothing, median filtering, and wavelet transforms, are applied to reduce the high-frequency noise while preserving essential details in the image. Smoothing techniques help in removing small imperfections without significantly blurring the critical structures of the brain, such as the ischemic regions that need to be detected. A median filter, which replaces each pixel with the median value of its neighbour pixels, is often used in preprocessing to smooth out noise without sacrificing the sharpness of the boundaries between healthy and ischemic brain tissue (Liu et al., 2023)

2.3.3. Image Segmentation

Segmentation is a critical step in preprocessing for stroke detection, as it isolates regions of interest (ROIs), such as the ischemic areas of the brain, from the surrounding healthy tissue. Techniques such as thresholding, region-growing, and clustering are used to identify and isolate the ischemic lesions in brain CT and MRI images. Recent advancements in deep learning have introduced more sophisticated methods for image segmentation, such as U-Net and Mask R-CNN, which can segment stroke lesions with high accuracy by learning complex patterns from labeled datasets. A deep learning-based model such as U-Net can automatically learn to segment ischemic lesions in the brain, which are then used to calculate metrics like the Alberta Stroke Program Early CT Score (ASPECTS) for stroke assessment (Chan et al., 2023).

2.3.4. Intensity Adjustment and Contrast Enhancement

Intensity adjustment is used to enhance the contrast in medical images, making it easier to distinguish between different types of tissues. This is particularly important for ischemic stroke detection, as the differences in tissue density between healthy and ischemic areas may be subtle. Contrast enhancement techniques, such as histogram equalization or adaptive histogram equalization, are employed to enhance the image's contrast, making it more suitable for analysis. Adaptive histogram equalization adjusts the contrast of local regions in the image, which is particularly useful for improving the visibility of ischemic areas in brain CT scans, where ischemic changes might be faint and hard to distinguish from normal tissue (Liu et al., 2023)

2.3.5. Advanced Preprocessing Techniques

As the field of medical imaging advances, new preprocessing techniques are emerging to further enhance stroke detection accuracy. One of these is **symmetry-based preprocessing**, which is particularly useful in stroke detection, as ischemic lesions often show distinct asymmetry in the brain. Recent research has utiliseutilised anatomical symmetry to enhance stroke detection by exploiting the differences between the left and right brain hemispheres. The Symmetry-Enhanced Attention Network (SEAN) introduced by Liang et al. (2021) uses symmetry-based alignment and attention mechanisms to preprocess brain CT images. This method aligns the brain image to a standard space to enhance the detection of acute ischemic infarcts by leveraging the inherent symmetry of the brain, thus improving lesion localisation and segmentation.

By incorporating these techniques, preprocessing in ischemic stroke detection has become a pivotal step in improving diagnostic accuracy and outcomes. Advanced techniques, such as symmetry-based preprocessing and deep learning models, are pushing the boundaries of what is possible in clinical stroke care, offering significant improvements in both speed and precision.

2.4. Feature Extraction

Feature extraction is an essential step in medical image analysis, particularly in the context of ischemic stroke detection. It involves identifying and isolating the most relevant information from raw CT or MRI images to help the model understand the key differences between healthy and ischemic brain tissue. In simple terms, feature extraction allows the model to "learn" the important patterns in the images that are indicative of a stroke. For ischemic stroke detection, the aim of feature extraction is to identify specific characteristics in the brain images that differentiate ischemic tissue from healthy tissue. This step is essential because the performance of machine learning models heavily depends on the quality and relevance of the features used for training. (Liu et al., 2023)

In ischemic stroke classification, the most commonly used features are intensity-based features, texture-based features, and shape-based features. Each of these features plays an important role in identifying ischemic lesions within the brain. Intensity-based features are derived from the pixel intensities of the CT images, which correspond to the varying levels of tissue density. Ischemic tissue typically appears darker on CT scans due to reduced blood flow and lower density compared to healthy tissue. Intensity-based features focus on identifying these variations in tissue density, which can be quantified by metrics such as mean intensity, standard deviation, and histogram-based features (Ho et al., 2019). These intensity differences are often the most direct indicators of ischemic lesions, as ischemic brain regions usually show decreased attenuation (brightness) in non-contrast CT images. For instance, Wu et al. (2020) used intensity-based features to distinguish between ischemic and healthy tissue, achieving promising results with high classification accuracy.

Texture analysis refers to the study of patterns in the spatial arrangement of pixel intensities, which helps capture structural and local features in the image. Gray-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP) are two widely used methods for texture analysis in medical image processing. GLCM, introduced by Haralick et al. (1973), calculates the frequency of occurrence of pixel pairs with specific values and in a specified spatial relationship. This allows for the extraction of texture features such as contrast, homogeneity, and entropy, which are important for identifying the heterogeneity of ischemic lesions. Similarly, LBP captures local patterns by comparing the intensity of each pixel with its neighbours, which is useful for identifying subtle texture variations in the brain tissue. These texture features have been shown to improve stroke classification by distinguishing ischemic tissue from normal tissue, as ischemic lesions typically exhibit different textural properties compared to healthy brain regions (Gautam & Raman, 2021). Fontanella et al. (2023) employed GLCM-based texture features to improve the performance of stroke detection models, highlighting their significance in ischemic lesion identification.

Meanwhile for shape-based features, it describes the geometric properties of brain structures, such as the shape, size, and boundary of the ischemic lesion. Shape-based features are important in distinguishing different types of ischemic lesions, which may vary in size and shape depending on the location and the extent of the infarction. Commonly used shape-based features include area, perimeter, and compactness, as well as more advanced shape descriptors like Hu moments, which capture the overall symmetry and elongation of the ischemic lesion (Ho et al., 2019). Shape-based features help to identify the regions of the brain affected by ischemic strokes and can be particularly useful for detecting large infarcts or identifying lesions in particular vascular territories. Ali et al. (2021) highlighted the importance of shape features in classifying ischemic strokes, noting that lesions with distinct boundaries or irregular shapes could be more easily identified through this method.

2.4.1. Importance of Feature Selection

While extracting features is crucial, not all features are equally useful for classification tasks. Therefore, feature selection is an essential step in reducing the dimensionality of the data and improving the efficiency and accuracy of the machine learning models. Feature selection helps in identifying the most relevant features for stroke classification and eliminating redundant or irrelevant ones that may add noise to the data.

For example, Badriyah et al. (2020) used recursive feature elimination (RFE), a feature selection technique, to improve the performance of Support Vector Machine (SVM) in ischemic stroke detection. RFE iteratively removes the least important features based on the model's performance, ensuring that only the most discriminative features remain for training. This technique can significantly reduce the risk of overfitting, especially when dealing with high-dimensional data like CT images, where many features may not contribute meaningfully to the classification task.

Furthermore, Fontanella et al. (2023) demonstrated that combining multiple feature extraction methods and selecting only the most important features can greatly enhance the performance of Random Forest (RF) models. By using feature importance rankings provided by RF, they were able to identify the most relevant features for ischemic stroke classification and improve the model's classification accuracy. This highlights the importance of combining both feature extraction and feature selection to optimize model performance.

Despite the advances in feature extraction, there are several challenges that need to be addressed for more accurate ischemic stroke detection. One of the primary challenges is the heterogeneity of ischemic lesions, which can vary significantly in size, shape, and location across different patients. As Xie et al. (2025) pointed out, ischemic lesions may not always be clearly visible on CT images, especially in the acute stages, which can lead to difficulty in feature extraction. Additionally, variations in imaging quality, such as differences in contrast, noise, and resolution, can complicate the feature extraction process and affect model performance.

Furthermore, ischemic lesions often have subtle features that may not be easily captured through traditional methods of feature extraction, such as texture or shape. This highlights the importance of developing more advanced feature extraction techniques, such as multiscale or multimodal feature extraction, which can capture a broader range of image characteristics at different levels of granularity.

2.5. Receiver Operating Characteristic and Area Under the Curve Analysis

Receiver Operating Characteristic (ROC) analysis and the Area Under the Curve (AUC) are fundamental in evaluating the performance of classification models, particularly in medical imaging tasks like stroke detection. ROC curves represent the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) across various threshold settings, while AUC quantifies the overall ability of the model to discriminate between classes.

In ischemic stroke detection, particularly in CT brain imaging, ROC and AUC analyses are critical for assessing model performance, as stroke lesions are often subtle and difficult to distinguish. Several studies have demonstrated the effectiveness of ROC and AUC in stroke detection using machine learning models. Hanover et al. (2020) emphasized the role of AUC in evaluating the predictive power of convolutional neural networks (CNNs) for ischemic stroke, where an AUC of 0.98 indicated excellent model performance. Similarly, in a study on SVM and RF models, AUC values were employed to compare classification efficacy for detecting ischemic and hemorrhagic strokes. A higher AUC was indicative of better model performance, particularly in distinguishing ischemic strokes from other conditions. In ischemic stroke detection, a study using Random Forest (RF) and SVM classifiers achieved an AUC of 0.91, highlighting the reliability of these models in identifying stroke lesions.

ROC and AUC analyses are also crucial for model selection, especially when comparing multiple algorithms such as CNN, SVM, and RF. Higher AUC values suggest that the model has a greater ability to correctly classify ischemic stroke cases, reducing the likelihood of misdiagnosis, which is critical in clinical settings where early stroke detection can be life-saving.

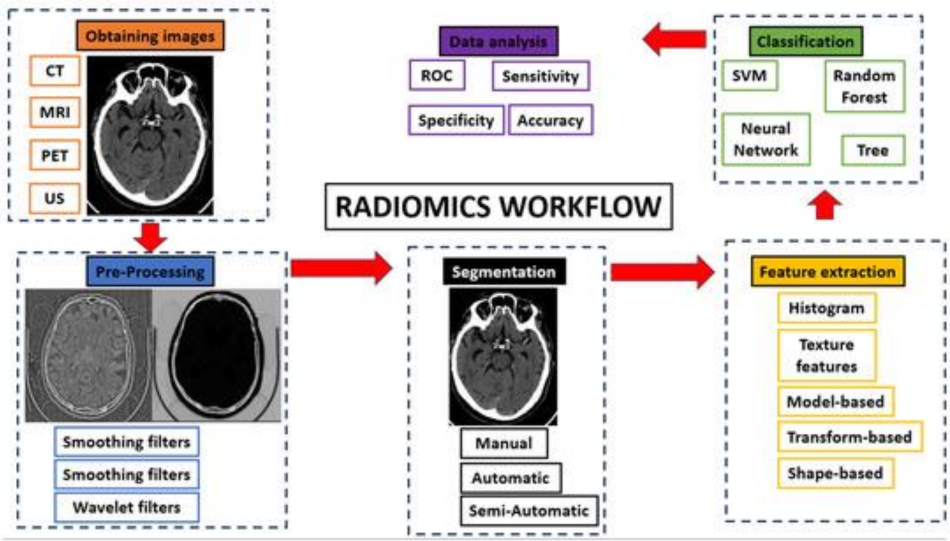

Figure 6: Radiomics workflow (Jacobo Porto-Álvarez et al., 2023)

2.6. Performance Evaluation Metrics

Performance evaluation metrics are key to determining the effectiveness of machine learning models in medical imaging tasks. Metrics such as accuracy, precision, recall, F1-score, and confusion matrix provide valuable insights into model performance. These metrics are particularly important in stroke detection, as both false positives and false negatives can have serious consequences for patient health.

Accuracy is a metric that illustrates the degree of accuracy of a model prediction across all parameters. It is measured as the percentage of correct predictions made by the model. It is especially useful in situations where all classes are of equal importance. The formula for determining it is the ratio of the number of accurate predictions to the total number of predictions made. In fact, it is the probability that a class will be predicted correctly (Tursynova et al., 2023). However, it may not be sufficient for imbalanced datasets.

$$Accuracy\ (a) = \frac{\sum_{i=1}^{N} [a(x_i) = y_i]}{N} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Here TP are true positive results, TN are true negative results, FP are false positive results, FN are false negative results.

When measured relative to absolute truth, precision and recall are often used to evaluate models in such scenarios. Precision refers to the proportion of positive predictions that are actually correct and gives an accurate indication of the validity of our positive detections. while recall (sensitivity) measures how well the model identifies all true positive cases.

$$precision = \frac{TP}{TP + FP} \quad (2)$$

While recall or sensitivity is a useful metric to use when trying to accurately describe the extent to which our optimistic predictions match the real world and measures how well the model identifies all true positive cases.

$$recall = \frac{TP}{TP + FN} \quad (3)$$

The F1-score, the harmonic mean of precision and recall, provides a balance between these two metrics and is especially useful in evaluating imbalanced datasets like