A Lightweight Deep-Feature and Agentic AI Fusion Framework for Automated Intracranial Hemorrhage Diagnosis from CT Images

Chinmayee Pati[a], Soumya Ranjan Nayak[b], Alok Kumar Jagadev[c], Ashok Kumar Panda[d]

[a] School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar-751024, Odisha, India

[b] School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar-751024, Odisha, India

[c] School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed to be University, Bhubaneswar-751024, Odisha, India

[d] Dept. of Computer Science & Engineering, Silicon University, Bhubaneswar-751024, Odisha, India

HOW TO CITE:

Chinmayee Pati, Soumya Ranjan Nayak, Alok Kumar Jagadev, Ashok Kumar Panda (2026). A Lightweight Deep-Feature and Agentic AI Fusion Framework for Automated Intracranial Hemorrhage Diagnosis from CT Images. International Journal of Special Education, 41(3s), 16-32.

ABSTRACT:

The correct identification of hematomas on the CT scan of the head is essential for the timely treatment of patients. However, the process of interpreting such images manually entails inter-reader variation. Computer-aided detection (CAD) provides reliable ways to increase the accuracy of medical imaging by introducing a higher level of standardization and automation in the analysis process. In this paper, we propose a hybrid solution for the classification of benign and malignant CT images of the head using a combination of deep feature extractiqon and machine learning algorithms. Our pipeline includes preprocessing and data augmentation steps for the normalization of input images. Deep convolutional neural networks, including esNet34, SqueezeNet, and MobileNetV2, were trained as feature encoders to extract discriminatory representations that can be used to train classification models like LightGBM, Random Forest, and XGBoost. Evaluation of the classification models' performance involved calculating their accuracy, precision, recall, specificity, F1-score, and ROC curve area under the curve. For all nine experiments performed, the MobileNetV2–LightGBM pipeline demonstrated the best performance, achieving 97% accuracy, precision, recall, specificity, and ROC curve AUC of 99.59%. Moreover, since our solution relies on modular training procedures, it can be easily combined with agentic artificial intelligence elements for autonomous operation.

Keywords: Deep Learning, Biomedical Imaging, Convolutional Neural Networks, Feature Engineering, Machine Learning

1. INTRODUCTION

The intracranial haemorrhage refers to one of the most severe neurological emergencies that occur because of the presence of blood in the brain due to the breaking of blood vessels in the cranial vault [1]. CT scanning serves as the most widely used modality to identify haemorrhagic lesions owing to the speed of implementation, accessibility, and sensitivity to blood pooling [1]. With more people suffering from different neurological diseases such as stroke, brain injury, and age-related disorders worldwide, there is an urgent need for rapid and correct interpretation of medical images in hospitals [2]. Nevertheless, despite being a gold-standard imaging technique, the manual CT image analysis poses numerous difficulties. First, it is a highly specialized skill requiring extensive knowledge and training from medical experts. Second, it is prone to subjective interpretations, errors made by fatigued specialists, time constraints, and other similar factors [3]. Hence, there is an urgent need to employ computer-assisted diagnosis in order to provide doctors with more opportunities for timely and precise decisions regarding patient health conditions.

Recently, great progress has been made in developing automated solutions based on AI technologies for analyzing biomedical images. In particular, the convolutional neural networks (CNNs) have shown impressive results for different recognition tasks, such as the detection of pathological signs, the segmentation of organs, and the classification of tumors [4]. However, despite being highly successful, the application of deep models in clinical practice requires careful consideration. Firstly, it is usually necessary to train neural networks on large volumes of data in order to avoid the problem of overfitting. Secondly, end-to-end models might require extensive computational power and cannot offer much interpretability. Thus, this project aims at creating a lightweight approach based on hybrid machine learning models using deep CNNs as feature extractors only. Three compact CNN architectures will be trained to generate discriminative embeddings, which would then be classified using traditional ML algorithms (e.g., gradient-boosted and ensemble classifiers such as LightGBM, Random Forest, and XGBoost). Such an approach will be advantageous in terms of interpretability, smaller computational burden, and applicability to training on small samples of biomedical images. Moreover, the proposed solution would be compatible with new agentic AI trends, allowing for the creation of more sophisticated autonomous modules for future computer-assisted diagnosis applications [5]. The main contributions of this research include:

- Development of a hybrid CNN–ML pipeline for haemorrhage classification in head CT scans using deep feature embeddings.
- Evaluation of three lightweight CNN encoders and three ML classifiers across nine combinations, highlighting the impact of encoder complexity on small medical datasets.
- Demonstration of strong performance using MobileNetV2 embeddings with LightGBM, achieving high diagnostic accuracy and AUC despite limited training data.
- Discussion of how agentic AI components can extend the proposed framework toward more autonomous and adaptive CAD systems.

The rest of the paper will be structured as follows: The second section will cover the review of previous work on haemorrhage detection using computed tomography and hybrid artificial intelligence techniques. Section three will include the methodology, which comprises pre-processing, data augmentation, feature extraction via CNN, and machine learning algorithms for classification. Section four will discuss the experiments and their outcomes.

2. RELATED WORKS

The problem of detecting ICH through head CT images has drawn significant attention in recent years, with many studies looking into deep learning models, hybrid algorithms, and computer-aided diagnostic approaches. Cortés-Ferre et al. [6] presented a comprehensive survey on deep learning-based ICH detection. They discussed the advantages and disadvantages of CNN architectures and tackled common issues in data imbalance and explainability. Wang et al. [7] presented a highly accurate deep learning algorithm for the detection and classification of haemorrhage types based on non-contrast CT images. The proposed method had almost perfect AUC values in various subcategories. Their work demonstrated the benefits of using completely automated approaches in clinical emergencies.

On the other hand, Arbabshirani et al. [8] developed a triage deep learning system capable of finding important pathological changes in head CT images. The proposed approach yielded excellent AUC values at roughly 0.91 and highlighted the necessity of developing better ways of representing features. Takala et al. [9] took another step forward and suggested an ensemble CNN algorithm which was highly sensitive when implemented in real-life settings across different emergency departments. Finally, a promising approach proposed by Teneggi et al. [10] was weakly supervised learning, which allowed minimizing the amount of required annotations.

All of these papers show the evolution of ICH detection approaches from basic thresholding and handcrafted feature selection to state-of-the-art deep learning models. However, not much research has been carried out regarding the development of hybrid solutions combining pretrained CNN feature extractors with classic machine learning classifiers, especially in cases involving small datasets. This study bridges the existing knowledge gap by analysing nine hybrid approaches and assessing their performance. The contributions presented in the literature can be found in Table 1 below.

Table 1: Analysis of related work

Authors	Methods	Findings	Performance
Cortés-Ferre et al. [6]	Survey of deep learning approaches	Comprehensive review of CNNs for ICH	Qualitative
Wang et al. [7]	Deep learning for ICH subtype detection	High-performing automated classification	AUC ≈ 0.99
Arbabshirani et al. [8]	Deep learning triage system	Reliable but moderate discrimination	AUC ≈ 0.91
Takala et al. [9]	Ensemble CNN workflow	High sensitivity in real-world ER settings	Sensitivity ≈ 99%
Teneggi et al. 10]	Weakly supervised CNN	Reduced annotation effort with strong accuracy	AUC ≈ 0.97

3. MATERIALS AND METHODS

In this section, a complete methodology adopted to perform automated hemorrhage detection using the deep features from head CT images will be discussed. The method involves several steps such as the construction of the data set, data pre-processing, augmentation, light-weighted convolutional neural network-based features generation, and classification.

3.1. Dataset

The data set used for this research work has been taken from the publicly available Head CT Hemorrhage data set compiled by Kitamura [11]. It consists of two different classes of axial CT images, namely the benign category and the malignant category, where there are 100 examples in each class, thus making a total of 200 images in the original data set. In order to train an efficient algorithm with limited medical imaging data, data augmentation has been done on the original data set to construct a new data set of 1000 images consisting of 500 benign images and 500 malignant images. Table 2 shows the number of benign and malignant classes after data augmentation. Figure 1 represents some images belonging to the two different diagnostic categories.

Table 2: Dataset distribution

Class	Original Dataset		Augmented Dataset	
	Train	**Test**	**Train**	**Test**
Normal	80	20	400	100
Haemorrhage	80	20	400	100
Total	160	40	800	200

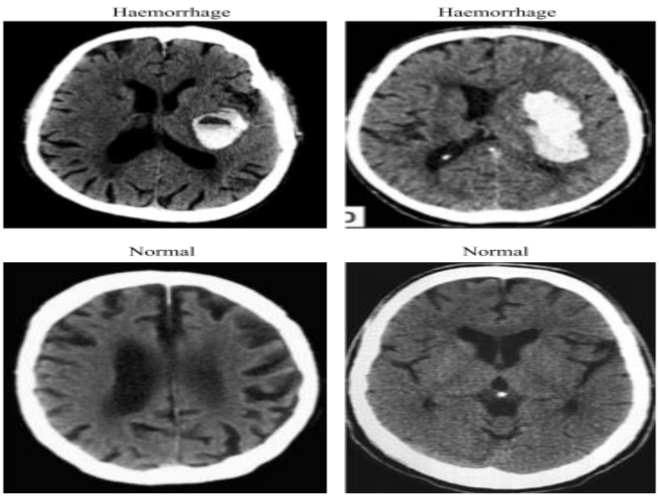

Figure 1: Haemorrhage and Normal images from the Head CT dataset.

3.2. **Pre-processing**

The pre-processing steps not only sought to unify these variations but simultaneously augmented the dataset with realistic transformations to improve the robustness of the feature extraction stage. All the images have been resized to 224x224 pixels to comply with the inputs required by the chosen CNN architectures. Using a constant spatial resolution for all models ensures that they receive consistent input data and helps avoid accidental biases due to varying image sizes. After resizing the dataset, a thorough augmentation process was performed to increase its sample size from 200 to 1000 samples. In addition, all augmentation parameters were set such that the transformed images still remained an anatomically correct representation of head CT scans. The rotations varied in the range of -20° to +20° to simulate minor orientation variances that typically occur when patients are positioned. A horizontal flip with a probability of 0.5 and vertical flips with a probability of 0.1 were applied to accommodate potential differences between various scanners. Zooms in the 90% to 110% range helped simulate variance in the field-of-view and lesion sizes. Both width and height shift values were adjusted by up to 10% of the image size to account for natural differences in patient positioning. Small geometrical transformations with a value range of 5% were added to imitate the natural variations. Finally, brightness modifications within the −15% and +15% range and contrast values in the −10% to +10% range helped simulate differences between various calibrations. Brightness and contrast augmentation have been extensively used before in medical images [12]. Figure 2 shows samples after augmentation.

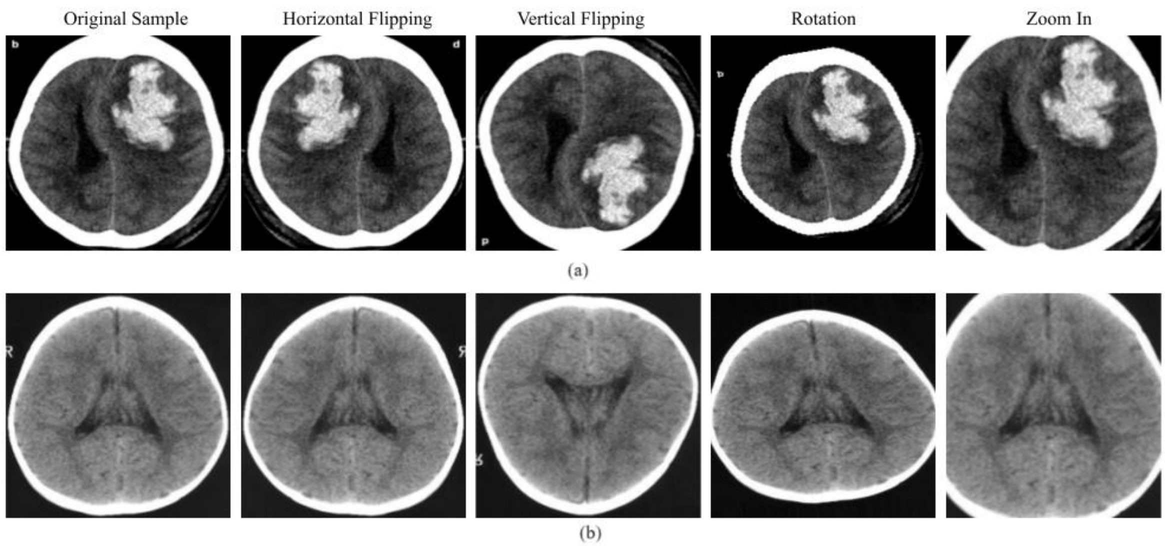

Figure 2: Samples of Data Augmentation for (a) Malignant images and (b) Benign images

After image augmentation, all images were normalized to provide stable input data for the convolutional neural networks. The pixel values were first normalized such that they belonged to the range [0, 1]. Then, depending on the specific needs of each feature extractor, model-dependent normalization was carried out. For example, networks trained on ImageNet were subjected to channel-wise mean subtraction and standardization.

3.3. Proposed Methodology

The outlined methodology proposes a hybrid diagnostic process in which deep convolutional neural networks extract meaningful features from the image, which are then classified using machine learning algorithms. As opposed to end-to-end deep learning systems, which usually require big data, significant computing power, and time-consuming training processes, the proposed method splits the processes of feature recognition and classification. It is based on the idea of using already trained convolutional models to extract the features from head CT images and then using appropriate machine learning methods to classify these objects into two categories. An important advantage of the hybrid design is its high degree of modularity, which makes it easy to extend or modify different components of the process as required. Importantly, modularity also makes it easier to use such hybrid models in combination with the latest trends related to agentic AI, allowing for more flexibility and scalability in future work. Specifically, later implementations of the process could employ agentic systems to conduct the following tasks automatically: check the integrity and quality of the collected data; select optimal machine learning models depending on their performance; trigger retraining on receiving new data; monitor the model performance and adjust it in response to detected drifts [13].

3.3.1 CNN-Based Feature Extraction

In the initial step of our workflow, discriminative feature representations for each CT image were extracted using pretrained convolutional neural networks (CNNs) as feature encoders. The models were not trained in an end-to-end manner on the target dataset; only feature representation generation was performed using the trained models. The advantage of this process is the usage of the representational ability learned by these models during pretraining on large-scale data without the need for large amounts of target data to train the models. Three lightweight CNN architectures with efficient computational performance were employed in this work. First, ResNet34 was used, which employs the residual learning concept to overcome the degradation problem usually occurring in deeper networks [14]. The residual learning concept allows for the use of identity shortcuts that propagate gradients stably during backpropagation and enables the model to learn more expressive mappings, providing robust feature representations even in structural medical imaging applications. Second, SqueezeNet, a very lightweight CNN, was considered. The design of this CNN is focused on maximizing efficiency while achieving high accuracy. The number of parameters of the SqueezeNet model is significantly lower than that of typical CNN models [15]. In addition, the organization of the network is done by means of fire modules, which reduce dimensionality but simultaneously expand it, thus capturing visual patterns efficiently while keeping the model lightweight. Third, MobileNetV2 was employed, which employs linear bottleneck blocks and inverted residuals as the main building blocks [16]. Depthwise separability further contributes to a reduction in the computational complexity of the network and allows the extraction of reliable embeddings in cases with limited data and computational capabilities. Embedding vectors were obtained as a result of flattening the outputs from the last layer of all three CNNs and encoded high-level structural information related to the task. The output feature representations served as inputs for the following classification step. Architectural details of the three pretrained CNNs are described in Table 3.

Table 3: Summary of CNN Feature Extractors

Model Name	Architecture Depth	Number of Parameters	Input Size	Embedding Size (Flattened Features)
ResNet34	34 layers	21.8×10^6	224 × 224 × 3	512
SqueezeNet 1.1	8 fire modules	1.24×10^6	224 × 224 × 3	512
MobileNetV2	53 layers	3.4×10^6	224 × 224 × 3	1280

3.3.2 Machine Learning Classifiers

The next step involves training machine learning classifiers on extracted embeddings. The goal is to identify which model will be able to map deep embeddings to the best diagnostic decisions. We used three popular and well-studied classifiers for the experiment. First, we used LightGBM – a fast and efficient gradient boosting framework based on histogram-based splits and leaf-wise tree growing [17]. The architecture is optimized for high-dimensional input data and showed excellent results in medical classification, especially where the interaction between features takes place. Second, Random Forest builds an ensemble of trees trained on samples taken from the initial dataset via bootstrapping and then aggregates them to make the prediction [18]. This classifier is very resilient to noise and variance, making it suitable for solving complex medical imaging problems. Third, we tested XGBoost – a very fast gradient boosting model optimized with the use of second-order gradient information, shrinkage, and regularization [19]. The algorithm showed excellent results when dealing with classification problems due to its ability to learn complex decision boundaries.

Each classifier was trained separately on embeddings generated by ResNet34, SqueezeNet, and MobileNetV2, resulting in nine combinations total. All experiments were conducted within a unified environment, thus allowing us to compare encoder-classifier hybrids in terms of effectiveness. Table 4 presents machine learning classifiers utilized in our study.

Table 4: Summary of Machine Learning Classifiers

Classifier	Model Type	Key Characteristics	Input Feature Type
LightGBM	Gradient Boosted Decision Trees	Histogram based optimisation, fast training, leaf wise growth	Deep CNN embeddings
Random Forest	Ensemble of Decision Trees	Bootstrap sampling, low variance, robust to noise	Deep CNN embeddings
XGBoost	Gradient Boosting with Regularisation	Second order optimisation, regularised boosting	Deep CNN embeddings

To succinctly summarize the hybrid process flow graphically, the current study includes a methodology diagram that clearly defines the entire workflow end to end. The diagram depicts the process flow beginning with the processing and augmentation phase, moving to the next step of extraction of deep embeddings using the three different pre-trained CNN encoders. Afterward, it depicts the flow where the embeddings are transferred to machine learning classifiers, demonstrating the nine possible ways of combination between encoders and classifiers (shown in Table 5). The flexible nature of this diagram also helps depict how this method is flexible to the inclusion of an agentic AI component that would automate the data validation and model management processes in future studies. The complete workflow used in this study is formally described in Algorithm 1.

Table 5: Overview of the Nine Hybrid Model Pipelines Evaluated

Pipeline	CNN Feature Extractor	ML Classifier	Type of Embedding	Type of Prediction
P1	ResNet34	LightGBM	512 dimensional	Benign / Malignant
P2	ResNet34	Random Forest	512 dimensional	Benign / Malignant
P3	ResNet34	XGBoost	512 dimensional	Benign / Malignant
P4	SqueezeNet	LightGBM	512 dimensional	Benign / Malignant
P5	SqueezeNet	Random Forest	512 dimensional	Benign / Malignant
P6	SqueezeNet	XGBoost	512 dimensional	Benign / Malignant
P7	MobileNetV2	LightGBM	1280 dimensional	Benign / Malignant
P8	MobileNetV2	Random Forest	1280 dimensional	Benign / Malignant
P9	MobileNetV2	XGBoost	1280 dimensional	Benign / Malignant

Algorithm 1. Hybrid CNN–ML Pipeline for Hemorrhage Detection

Input: Image set $X = \{x_i\}$, labels $Y = \{y_i\}$
Models: CNN set C, ML classifier set M

Output: Best pipeline $(\hat{c}, \widehat{m})$and metrics

Parameters: Parameter set $\Theta = \{d = (224,224,3), \theta_{A}, C, M, \alpha\}$

1. Preprocessing
   Resize: $x_i \leftarrow R(x_i)$
   Normalize: $x_i \leftarrow N(x_i)$
   Split: $(X_{tr} Y_{tr}), (X_{te} Y_{te}) = Split\ (X, Y)$
2. Data augmentation
   $\tilde{X}_{tr} = \widetilde{A(X_{tr}; \theta_A)}$
3. Embedding extraction

For each $c \in C$: $E_{tr}^{c} = \{\emptyset_c(x) \mid x \in \tilde{X}_{tr}\}$, $E_{te}^{c} = \{\emptyset_c(x) \mid x \in X_{te}\}$

4. Feature scaling

   For each scalar vector S_c on E_{tr}^{c} : $\hat{E}_{tr}^{c} = S_c(E_{tr}^{c})$, $\hat{E}_{te}^{c} = S_c(E_{te}^{c})$

5. Classifier training and evaluation

   For each $c \in C, and\ m \in M: Compute\ g_{c,m} \leftarrow m(\hat{E}_{tr}^{c}, Y_{tr})\ and\ \hat{Y}_{te}^{c,m} = g_{c,m}(\hat{E}_{te}^{c})$

   Repeat this step until condition satisfy, and compute metrics:

   $$P_{c,m} = \{Acc, Prec, Rec, Spec, F1, AUC\}$$

6. Model selection

   $$(\hat{c}, \hat{m}) = arg \binom{max}{c,m} (AUC_{c,m}, F1_{c,m})$$

7. Optional agentic AI for automation and monitoring

   Data quality check: for each input image x_i, compute $q(x_i), and\ flag\ if\ q(x_i) < X$

   Dynamic routing: for each $(c, m) \in P, compute$:

   $$S_{c,m}(t) = w_1 AUC_{c,m}(i) + w_2 F1_{c,m}(i) - w_3 u_{c,m}(i)$$

   Select and predict: $(\hat{c}_i, \hat{m}_i) = arg \binom{max}{c,m} S_{c,m}(i)\ and\ \hat{y}_t = g_{(\hat{c}_i, \hat{m}_i)}\left(S_{\hat{c}_i}\left(\emptyset_{\hat{c}_i}(x_i)\right)\right)$

   Drift detection and Retraining: $D(p_i, p_i) > \delta\ and\ AUC_{c,m}(i) < Predefined\ Threshold$

   Compute Risk: $r_i = 1 - \max_{y} p(\hat{c}_i, \hat{m}_i)(y \mid x_i)$

8. Output and reporting

   Return best encoder $\emptyset_{\hat{c}}, classifier\ g_{c,m}\ and\ metrics\ p_{c,m}$

This visual representation, depicted in Figure 3, complements the descriptive methodology and offers an intuitive understanding of how the individual components integrate to form the overall diagnostic system.

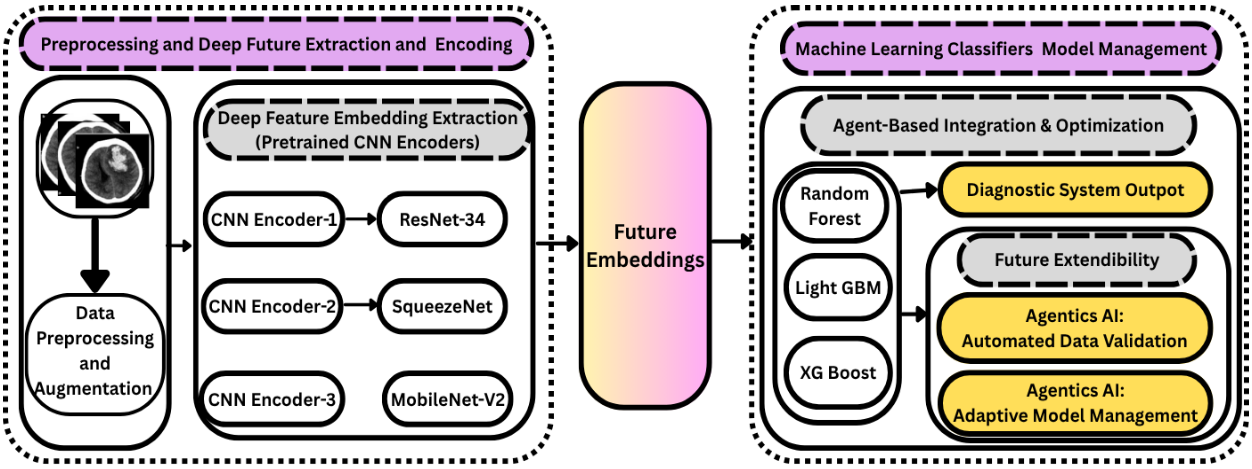

Figure 3: Proposed Methodology architecture diagram

3.4 Performance Metrics

Once all model combinations were trained and evaluated, the performance was assessed using a series of quantitative metrics often utilized in image classification for biomedical applications. The following metrics provide a holistic understanding of how well each model combination discriminates benign head CT images from the malignant cases.

- **Radar Plot:** The radar plot acted as a visualization technique enabling multiple performance metrics to be compared simultaneously. As the name suggests, the radar plot allows for a quick analysis through visualization by placing different metrics on a polar coordinate system, hence allowing for easy comparisons among all nine models.
- **Confusion Matrix:** The confusion matrix was used as a measure that provides an in-depth look into how the model performs its

predictions. Shown below in Table 6, the confusion matrix shows the number of samples correctly classified, in addition to those incorrectly predicted in each category.

Table 6: Confusion Matrix

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

The true positive examples refer to those malignant cases correctly classified by the model. The true negative example refers to the benign cases correctly classified. False-positive predictions occur when the benign case is incorrectly classified, while a false-negative prediction occurs when the malignant case is wrongly predicted as being benign. The following quantities are important in calculating various diagnostic evaluation measures such as accuracy, precision, recall, and F1-score under standard formulae. The area under the ROC curve (AUC) serves as a threshold-independent metric indicating how well the model can classify the two categories.

4. Experimental Setup and Result Analysis

The section provides a description of the experimental setup and result analysis of this study.

4.1. Experimental Setup

All experiments were carried out under a well-controlled computational setting. Development and evaluation of models were performed in the computational cloud environment via Google Cloud Platform (GCP), which offered a cloud GPU runtime service. Main computational capabilities in the course of the research included access to an NVIDIA Tesla T4 GPU having 16 GB VRAM, approximately 12 GB system memory and two vCPUs. Such a configuration provided enough computational capacity for performing feature extraction with light convolutional neural network models and training classifiers on embeddings. The deep learning part of the pipeline, including preprocessing, augmentation, and feature extraction, was developed with the help of the PyTorch framework [20]. Machine learning classifiers were designed with Scikit-Learn, LightGBM, and XGBoost tools [21]. Utilities such as NumPy, Pandas and Matplotlib were used to process data, compute metrics and create visualizations.

The dataset was divided into training and test sets using a stratified sampling technique in order to keep class proportionality. Thus, eighty percent of images were included in the training partition while twenty percent were assigned to the test partition. Such a division was used for each CNN embedding set in order to be able to fairly evaluate the nine hybrid models. Augmentation, in turn, was applied to the training partition only since the purpose was to use the test set as a ground truth.

As shown in Figure 4, the training workflow included several modules that were used for a fair comparison of all models. First of all, a preprocessed CT image was fed to one of three CNNs, and its embedding vector was obtained. Embedding vectors were computed based on pre-trained ImageNet weights without further fine-tuning to prevent overfitting. Then, embeddings were scaled according to the standard normalization procedure to stabilize numerical distributions and optimize classifier performance. In order to train each machine learning classifier, embeddings of the training set images were used, whereas embeddings of the testing set images were reserved for model evaluation. Such a procedure was separately applied to each CNN and each machine learning classifier, thus providing nine possibilities to build hybrid models.

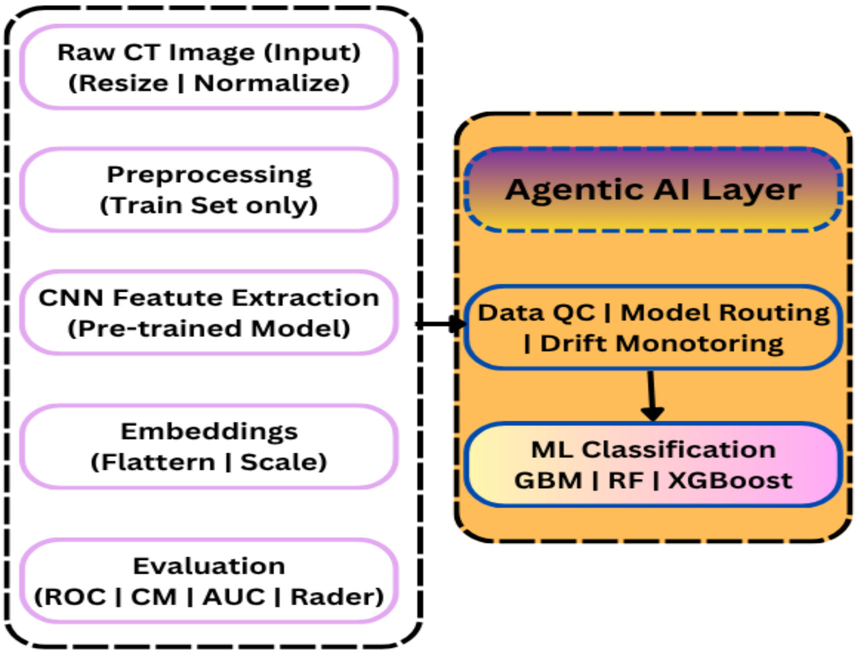

Figure 4: Training Workflow for the proposed methodology

Hyperparameters for the machine learning models were selected through repeated experimentation while ensuring that the training process remained computationally feasible. Random Forest and XGBoost models were trained using their standard implementations, while LightGBM incorporated its histogram-based optimization features for efficient training on high-dimensional embeddings. Throughout the experiments, random seeds were fixed where possible to support reproducibility. This setup, depicted in Table 7 provided a controlled and transparent platform for assessing the influence of different feature extractors and classifiers on the task of haemorrhage detection in head CT images

Table 7: Experimental Setup

Component	**Parameter**	**Value / Setting**
Hardware	GPU	NVIDIA Tesla T4 (16 GB VRAM)
Frameworks	Deep Learning Framework	PyTorch (pretrained CNNs)
	Machine Learning Libraries	Scikit Learn, LightGBM, XGBoost
Dataset Split	Train: Test	80: 20 stratified
Input Image Size	For All CNNs	224 × 224 × 3
Feature Scaling	Normalisation	StandardScaler
Classifier Hyperparameters	Optimised Per Model	Auto tuned through experimental trials
Training Strategy	For CNNs	Frozen pretrained weights, no fine tuning
	For ML models	Independent training for each CNN output

4.2. **Result Analysis**

The hyperparameters of the ML algorithms were set via trial and error while maintaining computational efficiency during the training phase. While the Random Forest and XGBoost algorithms were trained based on their default settings, the LightGBM algorithm took advantage of its histogram-based optimization capabilities, which allowed for efficient training on high-dimensional embeddings. Random seeds were used whenever possible in order to ensure reproducibility. The setup described above is outlined in Table 7 and was used to assess the effect of various feature extractors and classifiers on hemorrhage classification in head CT images.

Table 8: Comparative Performance of All CNN ML Hybrid Models

CNN Model	**Classifier**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**	**Specificity (%)**	**AUC**
ResNet34	LightGBM	94.0	97.83	90.0	93.75	98.0	0.9886
ResNet34	Random Forest	92.5	96.70	88.0	92.15	97.0	0.9880
ResNet34	XGBoost	92.5	95.70	89.0	92.23	96.0	0.9803
SqueezeNet	LightGBM	94.0	93.13	95.0	94.06	93.0	0.9835
SqueezeNet	Random Forest	92.5	91.26	94.0	92.61	91.0	0.9872
SqueezeNet	XGBoost	92.0	90.38	94.0	92.15	90.0	0.9812
MobileNetV2	LightGBM	97.0	97.0	97.0	97.0	97.0	0.9959
MobileNetV2	Random Forest	97.5	98.97	96.0	97.46	99.0	0.9947
MobileNetV2	XGBoost	96.0	95.09	97.0	96.03	95.0	0.9948

The presented results reveal noticeable trends that highlight the significant impact of the feature extractor and the classifier on the prediction outcomes. While using different CNNs, the experiments showed that the embeddings created with MobileNetV2 were the most discriminative among the three convolutional neural network models used. Indeed, the best results were reached when using the extracted embeddings as input data for machine learning classifiers. In this context, LightGBM and Random Forest models demonstrated the highest levels of accuracy and AUC when combined with embeddings of MobileNetV2. This confirms that the embeddings managed to retain all the necessary textural and structural features to be used for detecting intracranial bleeding. However, this result should be viewed as remarkable since MobileNetV2 is a compact model intended to perform effectively with a minimal quantity of training data. When combined with Random Forest, MobileNetV2 resulted in the highest accuracy (97.5 percent) and F1-score (97.46 percent). At the same time, its specificity was equal to 99 percent. Such a high level of specificity means that all malignant cases were identified, and there was a low number of false positives.

4.2.1. ROC Curve Analysis

The ROC graphs shown in Figure 5 were plotted for all model combinations. They indicate that the models used in this study have strong discriminative capability due to high AUC values (more than 0.98) for all pipelines. This means that all proposed models can effectively distinguish benign cases from malignant ones. Pipelines based on MobileNetV2 demonstrate the sharpest ROC graphs, and LightGBM reaches almost an ideal position in the graph (the top-left corner). Such a result implies a high true positive rate and a very low number of false positives. The ROC graphs of ResNet34- and SqueezeNet-based classifiers also demonstrate good results, and their AUC values range between 0.98 and 0.988.

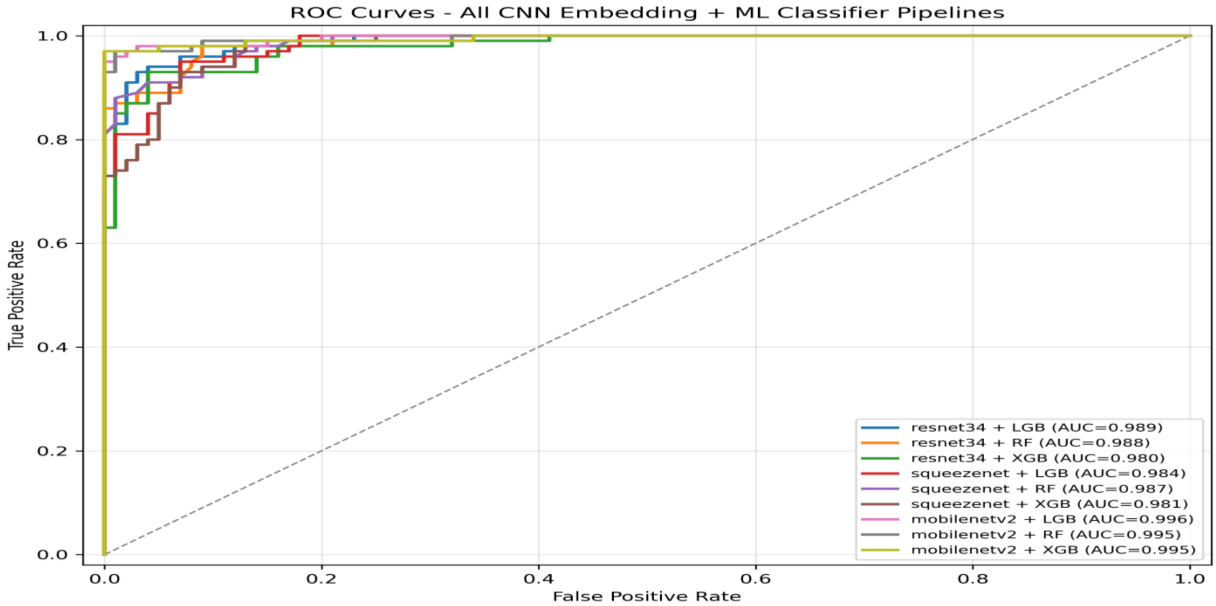

Figure 5: ROC curve for all possible hybrid CNN+ML combinations

Overall, the ROC analysis reinforces that lightweight feature extraction, especially with MobileNetV2, is highly suitable for small dataset scenarios in biomedical imaging.

4.4.2. Radar Plot Interpretation

In turn, the radar plot (Figure 6) demonstrates an aggregate evaluation of all nine models across all metrics simultaneously. Thus, the presented visualization makes possible the intuitive assessment of a pipeline's performance in accuracy, precision, recall, F1 score, and AUC. Radial plots corresponding to MobileNetV2 models demonstrate the largest expansions, meaning that they possess relatively high performance across all metrics considered. Particularly, the Random Forest combination reveals high stability and strong generalization for both sensitivity and specificity. In its turn, ResNet34 models demonstrate relatively high levels of performance but are characterized by decreased recall scores, meaning the potential tendency of models to overlook malignant conditions. The SqueezeNet models are characterized by fairly balanced radial diagrams with relatively low precision metrics, meaning some errors in recognizing malignant skin lesions..

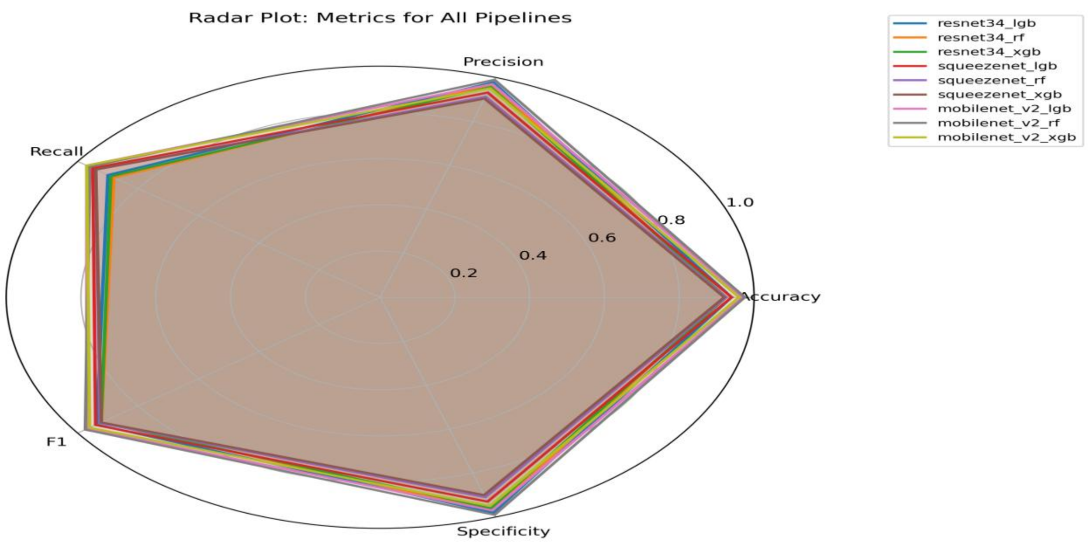

Figure 6: Radar plot for all possible hybrid CNN+ML combinations

4.4.3. Heatmap and Confusion Matrix Interpretation

The heatmaps derived from the confusion matrices (Figure 7) offer additional insight into the classification behavior of each model. They visually represent concentrations of true positives and true negatives as darker regions, while highlighting misclassifications.

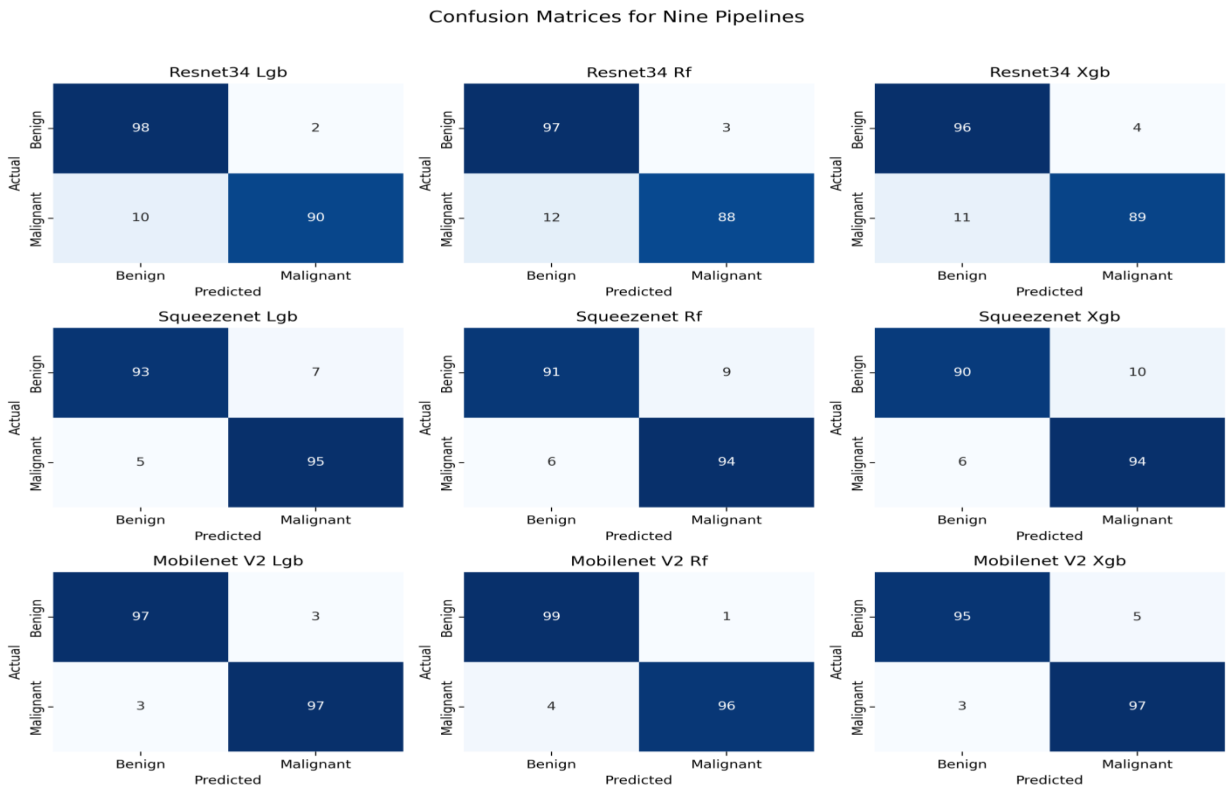

Figure 7: Confusion matrices for all the models

Diagonal dominance is observed among MobileNetV2 classifiers, meaning that the number of correct classifications is significantly higher than the number of incorrect ones. Thus, the Random Forest combination misclassifies only five examples among 1080 images (three in the false positive case and one in the false negative case). At the same time, ResNet34 and

SqueezeNet models are characterized by slightly greater dispersion due to a relatively more pronounced presence of false-negative results in their confusion matrices. Although the differences are rather small, the obtained heatmaps visually confirm the conclusions drawn from quantitative analysis: namely, deeper or extremely lightweight networks could struggle with relatively little training data.

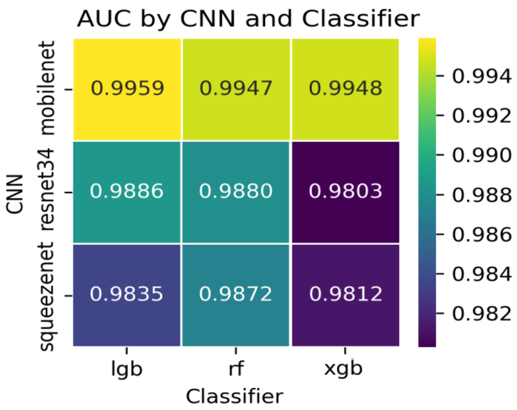

Figure 8: ROC heatmap for all model combinations

From the results, it is evident that MobileNetV2 generated the most reliable feature embedding compared to the other two convolutional neural networks. It has been observed that the light network was effective at capturing informative features for accurate classifications by the ML algorithms, leading to excellent model performance in all metrics. For ML models, Random Forest proved to be more reliable in detecting hematomas when integrated with MobileNetV2, which delivered higher accuracy and specificity with low misclassifications. In terms of AUC value, the combination of LightGBM and MobileNetV2 proved most reliable, indicating good class separations at different probability thresholds. Nine hybrid models were outstanding, further supporting that CNNs, in conjunction with classical machine learning, can be used to create an efficient classifier for identifying blood clots in the brain. This conclusion was supported by the ROC curve, radar plot, and confusion matrix heatmap. Models developed using MobileNetV2 presented balanced profiles with diagonal dominance, indicating better diagnostic abilities..

4.4.4. Comparison with State-of-the-Art Methods

A comparative analysis of contemporary state-of-the-art techniques demonstrates the efficiency and effectiveness of the suggested hybrid CNN–ML system. In particular, current research has focused on 3D CNN architectures, radiomic feature prediction methods, deep ensembles, and transfer learning with huge pre-trained networks in order to improve the detection of intracranial hematomas. The state-of-the-art methods from recent scientific literature can be seen in Table 9. Agrawal et al. created a 3D CNN classifier for screening TBI patients that resulted in good results even though the dataset used was prospectively collected [22]. Wei et al. presented a nomogram that uses radiomic data together with other clinical parameters for predicting progression of intraparenchymal hemorrhages, demonstrating strong AUC performance [23]. Asif et al. designed a high-performance multi-path deep ensemble classifier based on ResNet101-v2, Inception-V4, and LightGBM that shows high accuracies (over 97%) on huge benchmark datasets [24]. In the forensic context, Matsumoto et al. investigated fifteen different models of transfer-learning CNNs used on postmortem CT images and obtained AUC results varying between 0.862 and 0.907 due to heterogeneity and small sample size of postmortem datasets [25]. Yun et al. offered an automated system for detecting intracranial hemorrhages using deep CNN classifiers with performance exceeding 97 percent [26]. Contrary to previous methods, the best-performing model in this study, MobileNetV2 combined with Random Forest, provided results of 97.5% accuracy and 99.5% AUC that surpassed several other existing approaches using much smaller datasets.

Table 9: Comparison with State-of-the-Art Intracranial Hemorrhage Detection Methods

Authors	Methodology	Performance
Agrawal et al. **[22]**	3D CNN model for ICH detection in TBI	80% accuracy
Wei et al. **[23]**	Clinical–radiomics nomogram for hemorrhage progression prediction	AUC = 0.88
Asif et al. **[24]**	Parallel ResNet101-V2 + Inception-V4 + LGBM	97.7% accuracy
Matsumoto et al. **[25]**	Transfer-learned comparison of 15 CNN models (postmortem CT)	AUC = 0.862–0.907
Yun et al. **[26]**	Deep learning-based automatic ICH detection	97.0% accuracy
Proposed Method	MobileNetV2 embeddings + Random Forest	**97.5% accuracy, 99.5% AUC**

5. Research Discussion

In conclusion, the current study demonstrates the significant efficacy of a hybrid approach that combines deep convolutional feature extraction with classical machine learning classification algorithms for hemorrhage detection in head computed tomography images. Of the nine examined pipelines, the MobileNetV2 network generated the most discriminative representations that allowed obtaining the best classification accuracy, sensitivity, specificity, and AUC. Inverted residuals along with depthwise separable convolutions allowed capturing fine-grained structural information without compromising efficiency, making this network suitable for applications on small-sized medical datasets. The combination of MobileNetV2 and Random Forest yielded the highest performance out of all tested pipelines, suggesting that ensemble-based decision surfaces complement well the rich but compressed representations produced by lightweight CNNs. Of note is the high accuracy and AUC obtained even by the least successful pipelines exceeding 92% and 0.98, respectively. This allows concluding that even though a hybrid approach was used, CNN-based feature representation showed highly discriminative characteristics. Consistently high results support the validity of using transfer-learning methods in the analysis of limited-size datasets, reducing computational costs compared to training a deep network from scratch.

While in this particular study, the best results were obtained using the MobileNetV2 network as a feature extractor, future studies should examine the potential of building custom lightweight CNNs optimized specifically for head CT images. By doing so, we would be able to incorporate domain-specific patterns missed by general-purpose ImageNet-trained CNNs that could lead to higher accuracy, sensitivity, and specificity. Another advantage of the described pipeline lies in its modularity, allowing integration with new developments in the field of agentic AI such as automated data quality assessment, adaptation of classifier types, and dynamic monitoring of model drift. Integrating the proposed methodology into a hospital setting alongside PACS infrastructure constitutes an important step forward in implementing practical diagnostic tools..

6. Conclusion and Future Work

In this research work, we show that a mixed diagnostic model consisting of pretrained CNNs with classical ML classifiers leads to robust and reliable performance in detecting brain hemorrhages in head CT scans. The best performance was seen with MobileNetV2 embeddings, followed by successful integration with the agentic AI with Random Forest and LightGBM classifiers that gave the maximum accuracy, specificity, and AUC scores. In all cases, the worst performing pipeline achieved over 92% accuracy and >0.98 AUC, highlighting that the discriminative power of CNN features is still strong even in small datasets. Integrating this mixed CNN-ML model within an agentic AI can make it adaptive to become an ever-improving tool to support radiologists in clinical scenarios. This study highlights the feasibility of using lightweight feature extractors and confirms their effectiveness as a component in computer-aided diagnosis models. The future work will be focused on making such a model interpretable with the help of explainable AI techniques like attention mechanisms, etc.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

We have not used any AI-based tools for writing this paper.

References

W. J. Powers, P. K. O'Brien and M. J. Albers, "2018 Guidelines for the Early Management of Patients with Acute Ischemic Stroke," Stroke, vol. 49, no. 3, pp. e46–e99, 2018.

V. L. Feigin, K. M. Norrving and G. A. Mensah, "Global Burden of Stroke," Circulation Research, vol. 120, no. 3, pp. 439–448, 2017

G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017

Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, 2015

Y. Gao, S. Liu and L. Yang, "Artificial intelligence and innovation capability: A dynamic capabilities perspective," International Review of Economics & Finance, vol. 98, p. 103923, 2025.

L. Cortés-Ferre, "Deep Learning Applied to Intracranial Hemorrhage Detection," Computers, vol. 9, no. 2, p. 37, 2020.

X. Wang, J. Liu, Y. Zhu, J. Yu, M. Ge and L. Chen, "A Deep Learning Algorithm for Automatic Detection and Classification of Acute Intracranial Hemorrhages in Head CT Scans," NeuroImage: Clinical, vol. 32, p. 102785, 2021.

M. R. Arbabshirani et al., "Advanced Machine Learning in Action: Identification of Intracranial Hemorrhage on Routine Head CT Using a Deep Learning Approach," Radiology, vol. 289, no. 3, pp. 709–717, 2018.

J. Takala, H. Peura, R. Pirinen and T. J. M. Heikkilä, "High Sensitivity in Spontaneous Intracranial Hemorrhage Detection From Emergency Head CT Scans Using an Ensemble-Learning Approach," Scientific Reports, vol. 15, article 29919, 2025.

J. Teneggi, P. H. Yi and J. Sulam, "Weakly Supervised Learning Significantly Reduces the Number of Labels Required for Intracranial Hemorrhage Detection on Head CT," arXiv preprint, arXiv:2211.15924, 2022.

Kitamura, F. C. (2018). Head CT - hemorrhage. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/152137

C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, no. 1, pp. 1–48, 2019.

S. Chippagiri, "Autonomous AI Agents: Applications, Challenges, and Future Prospects," International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 4, pp. 5167–5174, 2024.

K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

F. Iandola, S. Han, M. Moskewicz, K. Ashraf, W. Dally and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size,” arXiv preprint arXiv:1602.07360, 2016.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018.

G. Ke, Q. Meng, T. Finley et al., “LightGBM: A highly efficient gradient boosting decision tree,” Advances in Neural Information Processing Systems (NeurIPS), 2017.

L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.

A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” Advances in Neural Information Processing Systems (NeurIPS), 2019.

F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

D. Agrawal, L. Poonamallee, S. Joshi and V. Bahel, “Automated intracranial hemorrhage detection in traumatic brain injury using 3D CNN,” Journal of Clinical Neuroscience, vol. 114, pp. 20–27, 2024.

X. Wei, X. Tang, D. You, E. Ding and C. Pan, “A clinical-radiomics based nomogram to predict progressive intraparenchymal hemorrhage in mild to moderate traumatic injury patients,” European Journal of Radiology, vol. 163, p. 110785, 2023.

M. Asif, M. A. Shah, H. A. Khattak, S. Mussadiq, E. Ahmed, E. A. Nasr and H. T. Rauf, “Intracranial Hemorrhage Detection Using Parallel Deep Convolutional Models and Boosting Mechanism,” Diagnostics, vol. 13, no. 4, p. 652, 2023.

R. Matsumoto et al., “Deep Learning-Based Detection of Intracranial Hemorrhages in Postmortem Computed Tomography: Comparative Study of 15 Transfer-Learned Models,” Applied Sciences, vol. 15, no. 19, p. 10513, 2025.

J. Yun, H. Nam, J. Kim et al., “Deep learning-based automatic detection of intracranial hemorrhage using head CT images,” Scientific Reports, vol. 13, article 7841, 2023.