Neuro Alert: Efficient Deep Learning-Based Detection of Intracranial Hemorrhage from CT Images

Amna Sajid[1], Anam Taskeen[1], Faryal Hayat[1], Mohsin Abbas[2], Aimen Ashraf[1], Misbah Noor[3]
[1]National University of Modern Languages, Islamabad, Pakistan
[2]National University of Modern Languages, Lahore, Pakistan
[3] Zixel Technologies, Islamabad, Pakistan
*__Correspondence__: amna.sajid@numl.edu.pk

Citation | Sajid. A, Taskeen. A, Hayat. F, Abbas. M, Ashraf. A, Noor. M, "Neuro Alert: Efficient Deep Learning-Based Detection of Intracranial Hemorrhage from CT Images", IJIST, Vol. 8 Issue. 2 pp 632-647, April 2026
Received | March 11, 2026 **Revised |** April 13, 2026 **Accepted |** April 17, 2026 **Published |** April 23, 2026.

Importance of Study: Quick and accurate detection of intracranial hemorrhage (ICH) is crucial, as delays can lead to higher risks of death, disability, and more complex treatment. While non-contrast CT scans are the main tool for diagnosing ICH, fast interpretation relies on radiologists, who may not always be available in emergencies or in places with limited resources.
Novelty Statement: This study introduces NeuroAlert, a lightweight convolutional neural network (CNN) that can detect different types of intracranial hemorrhage from non-contrast CT images. Unlike more complex deep learning systems, NeuroAlert maintains high accuracy while running efficiently with TensorFlow Lite.
Methodology: A framework named NeuroAlert is presented using the RSNA Intracranial Hemorrhage dataset, which includes about 750,000 CT images labeled for five types of hemorrhage and an extra category for any hemorrhage. The images were preprocessed with Hounsfield Unit windowing, normalization, resizing, and data augmentation to highlight important features. NeuroAlert uses a compact CNN with four convolutional blocks, a fully connected layer, and a softmax output layer. To handle class imbalance, weighted categorical cross-entropy was used during training. The final model was optimized with TensorFlow Lite for faster, lightweight use.
Results and Discussion: NeuroAlert achieved a macro-averaged accuracy of 93.7%, an F1-score of 0.926, and an AUC of 0.985, demonstrating strong overall discrimination across all hemorrhage categories. Class-wise analysis showed the highest performance for the any hemorrhage class, with an F1-score of 0.95 and an AUC of 0.992, while Intraparenchymal Hemorrhage (IPH) also showed strong recognition with an F1-score of 0.94 and an AUC of 0.990. After TensorFlow Lite optimization, the model size was reduced by 61%, and inference time remained below 1 second per image, indicating near real-time performance with efficient computational requirements.
Conclusion: NeuroAlert demonstrates that a carefully designed lightweight CNN can provide reliable multi-class intracranial hemorrhage detection while maintaining deployment-oriented efficiency. The proposed framework offers a promising solution for AI-assisted CT interpretation in time-sensitive and resource-limited healthcare settings.

Keywords: Intracranial Hemorrhage; Convolutional Neural Network (CNN); Computed Tomography (CT); Deep Learning; Machine Learning

Introduction:

Intracranial hemorrhage (ICH) constitutes a dangerous neurological emergency that requires immediate and precise medical assessment to reduce patient fatalities, long-term disabilities, and treatment delays. The medical community uses non-contrast computed tomography (CT) as its primary imaging modality to identify ICH because it provides quick results and is readily available at most medical facilities. It is also effective in diagnosing early-stage bleeding patterns. The accuracy of scan results depends on radiologists' ability to interpret them, particularly in urgent medical cases, where rapid emergency scan evaluations are critical [1][2][3].

Medical image analysis has advanced through advances in artificial intelligence, enabling deep learning networks to process images automatically. Convolutional neural networks, transfer learning models, transformer-based frameworks, and 3D deep architectures have all achieved effective results at detecting ICH through analysis of CT images [4][5][6][7][8]. Clinical studies, together with real-world research, have demonstrated that AI-assisted systems help radiologists in emergency departments by improving triage efficiency, reader confidence, and diagnostic workflow performance [2][3][9][10]. Meta-analyses and large-scale reviews have confirmed that deep learning methods achieve strong diagnostic performance in detecting ICH across various research environments [11][12][13][14].

The practical world still faces one major limitation because of these advancements. The majority of existing ICH detection systems use complex architectures based on 3D CNNs, transformer-based models, and ensemble frameworks, which all require extensive computational power, high memory capacity, and specialized server infrastructure [4][5][15][16]. The research methods serve their purpose in academic environments, but their complexity limits their use in clinical settings that require rapid decision support without access to advanced computing resources. The healthcare systems in developing countries and resource-limited areas need AI tools that are simple and quick to implement, as these deliver better practical outcomes than complex models that require extensive operational resources [15][17][18][19].

Recent academic studies show increasing interest in creating explainable systems that operate in real-world settings and can be used at scale. The research studied explainable deep learning techniques for hemorrhage detection tasks while developing external validation algorithms through optimized clinical workflow implementation [20][21][17][22][23]. The academic fields of TinyML development, privacy-aware AI research, and federated learning advancement indicate that upcoming medical AI systems will need to deploy compact, efficient systems rather than systems that prioritize only accuracy [18][19]. The research area has not yet produced enough studies to explore how multi-class ICH detection interacts with both lightweight CNN design and mobile-oriented model optimization in a single integrated framework.

NeuroAlert is a lightweight convolutional neural network that enables automated detection of intracranial hemorrhage from non-contrast CT images. The framework maintains high classification accuracy through its design, which combines a compact architecture with TensorFlow Lite optimizations to reduce processing requirements. The proposed method achieves practical AI-assisted hemorrhage detection results that maintain prediction accuracy while enabling efficient operation in time-sensitive, resource-constrained settings.

Research Objectives:

This study has the following main objectives:

The study aims to develop a lightweight CNN-based framework to detect multiple types of intracranial hemorrhage from non-contrast CT images.

The study utilizes Hounsfield Unit windowing and preprocessing normalization to create clinically relevant data, which enhances feature representation capabilities.

The study uses weighted optimization to address class imbalance during model training. The model evaluation uses standard classification metrics, which include accuracy, precision, recall, F1-score, and AUC.

The study aims to make the trained model suitable for lightweight deployment via TensorFlow Lite optimization.

Contributions: The study makes several key contributions, which can be summarized as follows:

The study presents NeuroAlert, a compact CNN architecture that detects six types of hemorrhage in CT images.

The preprocessing pipeline, which uses HU windowing along with normalization, resizing, and augmentation, provides a better image representation for hemorrhage-related visual content.

The weighted loss function improves learning when class distributions are uneven during training.

The trained model uses TensorFlow Lite optimization to achieve two objectives: reducing model size and enabling fast operational performance.

The research study proposes a practical method to enable movement between deep learning models that deliver optimal performance and lightweight systems that deploy for intracranial hemorrhage detection.

The paper has the following structure for its remaining sections. Section 2 provides an overview of related studies, which focus on automated systems for detecting intracranial hemorrhage. Section 3 details the research methods, including dataset preparation and preprocessing activities, model architectural design and optimization techniques, and the evaluation strategy used in the research. Section 4 presents the experimental results and their discussion. The document concludes in Section 5, which also describes upcoming research priorities.

Related Work:

The development of artificial intelligence systems for detecting intracranial hemorrhage in CT scans has become increasingly important, as emergency medicine requires quick, precise diagnostic results to reduce patient mortality. Deep learning models have demonstrated strong diagnostic capabilities for detecting hemorrhage and its subtypes in non-contrast CT scans, according to recent research [1][24][25]. AI systems have been demonstrated in clinical workflow research to assist radiologists by reducing the time required to understand scans and enhancing their ability to make emergency room triage decisions [2][3][9].

The main effort in this field focuses on using advanced deep learning architectures to achieve better classification results. The ICH analysis process has shown effective results using three techniques: convolutional neural networks, transfer learning models, and multi-type hemorrhage detection frameworks [7][8][6][26]. Spatial context understanding and subtype discrimination performance have both improved through the development of transformer-based and 3D convolutional methods. Strongly annotated datasets, model ensembles, and multicenter validation studies have established the credibility of automated systems for detecting hemorrhage in medical imaging [21][27][28]. Deep learning systems show potential for detecting ICH, though most existing models have been built for high predictive accuracy rather than lightweight systems.

Various reviews and meta-analyses tracked developments in individual model creation while surveying the entire research field. Systematic reviews have shown that deep learning models can achieve high sensitivity and specificity for detecting ICH on CT scans [14][13][11][12]. Technical survey papers have identified important current trends, including subtype classification, lesion localization, explainability, and real-world validation [15][29][30].

Table 1. Comparative Analysis of Existing Studies

Study	Method	Key contribution	Limitation
[1]	Deep learning for ICH detection	Demonstrated the effectiveness of deep learning for hemorrhage identification from CT scans	Focused mainly on detection performance; limited emphasis on lightweight deployment
[4]	Dual-task Vision Transformer	Improved CT image classification through transformer-based learning and dual-task modeling	Transformer models are often computationally demanding for constrained settings.
[5]	3D CNN	Captured richer spatial context across CT slices for improved hemorrhage detection	Higher computational complexity and heavier deployment requirements
[6]	Multi-type hemorrhage detection framework	Addressed the detection and quantification of multiple hemorrhagic lesions in head CT images	More focused on diagnostic capability than compact deployment
[11]	Systematic review and meta-analysis	Confirmed the strong diagnostic accuracy of deep learning methods for ICH detection	Highlighted accuracy trends, but not a practical lightweight implementation
[15]	Survey of deep learning techniques for ICH	Summarized recent architectures, frameworks, and challenges in ICH diagnosis	Identified deployment and scalability issues as continuing challenges
[9]	Deep learning-assisted clinical validation	Showed that AI can improve reader performance and support clinical workflow	Validation was clinically meaningful, but not centered on mobile-efficient design.
[17]	Optimized deep learning algorithm	Reported strong diagnostic performance and confidence for optimized ICH detection	Still oriented toward high-performance systems rather than minimal-resource deployment.
[10]	Real-world external validation	Demonstrated practical integration of automated ICH detection in emergency settings	Real-world utility shown, but lightweight edge-oriented implementation remains limited.
[18]	TinyML and federated learning for medical devices	Highlighted the importance of compact and privacy-aware AI for resource-constrained healthcare systems	General medical AI perspective; not specifically focused on multi-class ICH detection
Proposed Work (NeuroAlert)	Lightweight CNN with TensorFlow Lite optimization	Provides multi-class intracranial hemorrhage detection with strong classification performance, reduced model size, and near	Current evaluation is based on benchmark data; broader external validation and multi-

		real-time lightweight inference	hardware benchmarking remain future directions.

AI systems have been shown to operate successfully in clinical environments, as evidenced by external validation studies and real-world evaluations, but researchers need to address three main challenges: model robustness, interpretability, and generalizability [17][10][22][23].

The existing models, which demonstrate strong performance, fail to deliver effective systems that work well in resource-constrained environments. Existing methods require intensive computational resources, large memory capacity, and centralized server systems to function [5][15][16]. Low-resource healthcare environments need tools that provide quick, lightweight decision support rather than complex models that require complex deployment processes. Recent discussions on TinyML, hardware-aware medical AI, and privacy-preserving learning further emphasize the growing need for compact, efficient, and scalable solutions.

Studies are developing new systems that enable workflow management while providing explanation features. Explainable deep learning has been explored for differentiating specific hemorrhage-related conditions and improving model transparency [20]. AI tools that enhance clinical usability and reader confidence have been studied through real-world integration research. Multi-class ICH detection systems have not been investigated using a lightweight architecture suitable for practical deployment environments. Most published systems prioritize either diagnostic accuracy or clinical validation, while efficient mobile-compatible implementation remains less extensively explored.

Recent studies have made significant progress in using deep learning to automatically detect intracranial hemorrhage. Still, most systems focus on either high diagnostic accuracy or clinical validation, with less emphasis on lightweight, deployment-ready multi-class frameworks. Table 1 illustrates this gap by comparing key studies with the proposed NeuroAlert system. NeuroAlert stands out by combining a compact CNN architecture, clinically relevant preprocessing, class-imbalance-aware optimization, and TensorFlow Lite-based compression into a single framework. This makes it better suited for situations where time and resources are limited.

The present study investigates a lightweight CNN-based framework that maintains strong classification results while reducing computational demands. NeuroAlert provides a practical solution for ICH detection by integrating CT preprocessing, class imbalance optimization, and TensorFlow Lite model compression. This study contributes to the development of AI solutions for neuroimaging applications by shifting from accuracy-only model development toward the creation of efficient and deployment-ready AI systems.

Material and Methods:

This study presents NeuroAlert, a lightweight convolutional neural network (CNN)-based framework designed to automatically detect intracranial hemorrhage (ICH) in non-contrast CT images. NeuroAlert employs clinically relevant image preprocessing, efficient deep feature extraction, class-imbalance-aware optimization, and TFLite model compression deployment. The workflow includes dataset preparation, Hounsfield Unit (HU) windowing, image normalization, CNN-based classification, model optimization, and deployment-focused implementation.

Dataset:

We used the RSNA Intracranial Hemorrhage Detection dataset for our experiments. This public benchmark includes about 750,000 DICOM CT slices from nearly 25,000 head CT exams collected at several medical centers. Each slice is labeled for five hemorrhage subtypes: epidural (EDH), subdural (SDH), subarachnoid (SAH), intraparenchymal (IPH), and intraventricular (IVH). There is also an 'Any Hemorrhage' label that indicates the presence of hemorrhage regardless of subtype.

We split the dataset into training, validation, and test sets at 80:10:10. This split was performed at the examination level to prevent information leakage between sets and to ensure a fair evaluation. Since some hemorrhage types were more common than others, we used class weighting during training to help reduce bias toward the more frequent categories.

Image-Preprocessing:

We preprocessed the images to highlight key diagnostic details and ensure consistent input for the CNN. First, we converted the DICOM images to grayscale and normalized their intensity values to the range $[0,1]$ using min-max normalization:

$$I_{\text{norm}} = \frac{I - I_{\min}}{I_{\max} - I_{\min}}, \quad (1)$$

where I denotes the original image intensity, while $I_{\min}$ and $I_{\max}$ represent the minimum and maximum pixel values in the image, respectively.

The Hounsfield Unit windowing was applied with standard diagnostic settings to improve radiological measurements of clinically important features. Multiple HU windows highlight distinct properties of brain tissues, revealing structural and regional differences. The network used three-channel images, each formed by stacking windowed images to learn from all radiological perspectives simultaneously. The multi-window approach enhances the detection of subtle hemorrhagic patterns that remain invisible when using a single window.

All images were resized to a uniform spatial resolution before being passed to the network. The training process used data augmentation techniques to enhance model performance while preventing overfitting. The augmentation process included standard image transformations such as rotation, flipping, and small geometric perturbations while preserving the diagnostic content of the CT slices.

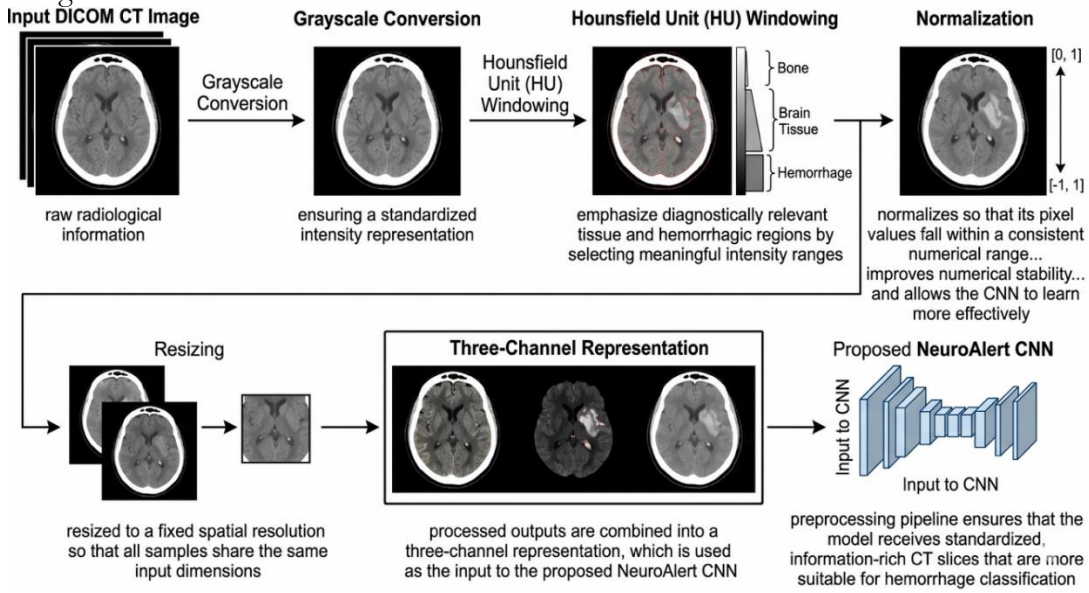

Figure 1. Preprocessing pipeline of the proposed NeuroAlert framework. The input DICOM CT slice is converted into a grayscale intensity representation, subjected to Hounsfield Unit windowing, normalized, resized, and transformed into a three-channel input for CNN-based classification.

Workflow of Proposed System:

The NeuroAlert pipeline has five main stages: CT image acquisition, preprocessing, CNN-based feature extraction, multi-class classification, and lightweight deployment. Once preprocessing is complete, the windowed CT image is fed into the CNN, which uses convolutional blocks to extract features and then maps these features to class probabilities using fully connected layers and a Softmax classifier. After training, the model is converted into TFLite format for lightweight inference.

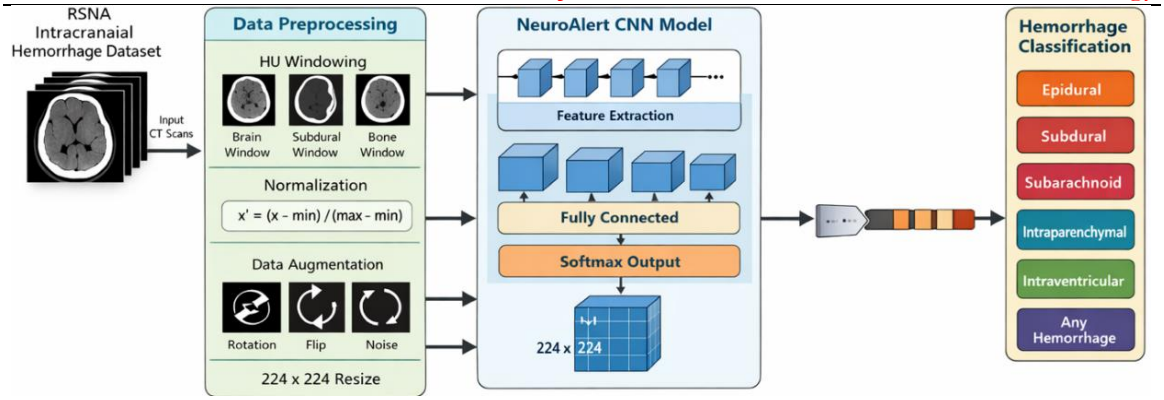

Figure 2. Overall workflow of the proposed NeuroAlert framework for intracranial hemorrhage detection. The pipeline consists of CT input, preprocessing, feature learning through a lightweight CNN, Softmax-based classification, and deployment-oriented optimizatio optimization using TensorFlow Lite.

Proposed NeuroAlert CNN Architecture:

The NeuroAlert model introduces a lightweight CNN architecture that achieves both high discriminative power and low computational cost. The architecture leverages a compact design to enable deployment in environments with limited computing resources.

The network consists of four convolutional blocks. Each block contains a convolutional layer, batch normalization, ReLU activation, and max pooling as its essential components. Batch normalization stabilizes the distribution of intermediate activations, ReLU introduces non-linearity, and max pooling reduces spatial resolution while preserving the most salient local responses. The hierarchical structure of the system enables the extraction of increasingly abstract image features as the data progresses through the network.

Mathematically, the convolution operation at layer l can be written as

$$x_{i,j,k}^{(l)} = \sum_{m=1}^{M} \sum_{u=1}^{U} \sum_{v=1}^{V} w_{u,v,m,k}^{(l)} x_{i+u,j+v,m}^{(l-1)} + b_k^{(l)} \quad (2)$$

where $x_{i,j,k}^{(l)}$ denotes the output feature value at spatial location (i, j) in channel k of layer $l, w_{u,v,m,k}^{(l)}$ represents the learnable convolutional kernel weight connecting input channel m to output channel $k, b_k^{(l)}$ is the bias term of the k-th output channel, and $U \times V$ denotes the kernel size. The summation over M accounts for all input channels.

Once the convolutional features are extracted, the resulting feature maps are flattened and passed to a fully connected layer with 256 neurons, using ReLU activation. A Softmax layer performs the final classification, producing probabilities for each of the six output categories. The probability for class c is given by

$$\hat{y}_c = \frac{\exp(z_c)}{\sum_{r=1}^{C} \exp(z_r)} \quad (3)$$

where z_c is the logit associated with class c, C is the total number of classes, and $\hat{y}_c$ is the predicted probability of class c.

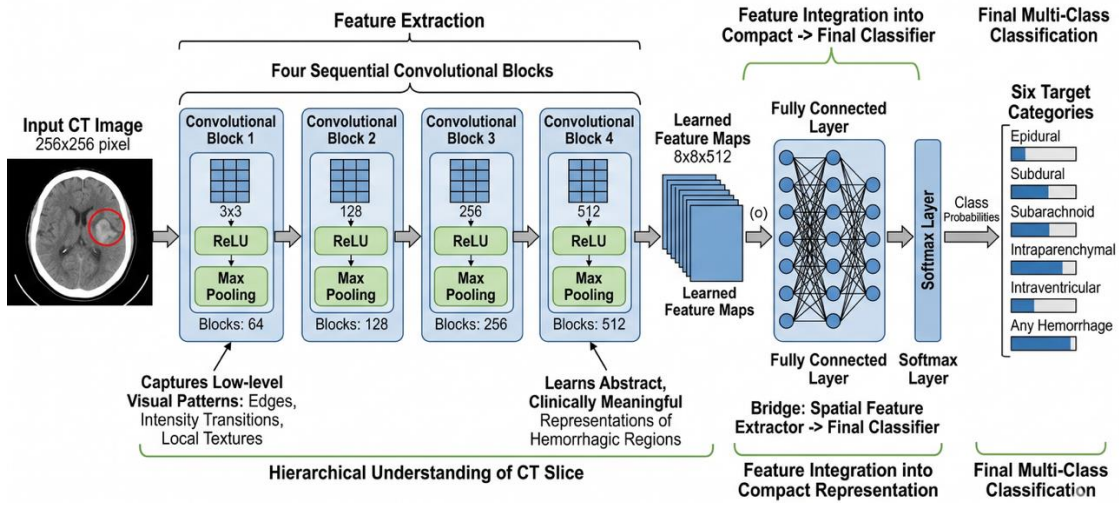

Figure 3. Architecture of the proposed NeuroAlert CNN. The network consists of four convolutional blocks followed by a fully connected layer and a SoftMax output layer for six-class hemorrhage classification.

Loss Function and Optimization:

Due to class imbalance in the dataset, the model was trained using weighted categorical cross-entropy. This loss function imposes higher penalties for mistakes in less common classes, helping the network find a more balanced decision boundary. The loss function is defined as

$$\mathcal{L}_{\text{WCE}} = -\sum_{c=1}^{C} w_c y_c \log(\hat{y}_c) \quad (4)$$

where C is the number of classes, y_c is the ground-truth label for class, $c, \hat{y}_c$ is the predicted probability for class c, and w_c Is the class-specific weight inversely related to class frequency. In this way, classes with fewer training examples contribute more strongly to the optimization objective.

To reduce overfitting and improve generalization, L2 regularization was incorporated into the training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{WCE}} + \lambda \|\theta\|_2^2 \quad (5)$$

where λ is the regularization coefficient, and θ represents the trainable parameters of the network.

We used the Adam optimizer with a learning rate of 1×10^{-4}, a batch size of 32, and trained for up to 25 epochs. Training was stopped early when the validation loss stopped improving, helping avoid unnecessary training and reducing overfitting.

Early Stopping Strategy: The study used early stopping to monitor the validation loss during training. If the validation loss did not get better for a set number of epochs, we stopped training and kept the model with the best validation results. This approach helped the model converge steadily and perform better on new test data.

Algorithm 1: Early stopping strategy used during training
Initialize best_loss ← ∞
Initialize patience counter ← 0
for each epoch do
Train the model on the training set
Compute validation loss
if validation loss < best_loss then
best_loss ← validation loss

Save model weights
Reset the patience counter to 0
else
Increment patience counter
end if
If the patience counter reaches a predefined threshold, then
Stop training
end if
end for
Restore the best saved model weights

TensorFlow Lite Optimization and Deployment Pipeline:

The trained CNN was converted to TensorFlow Lite format for lightweight inference after training finished. TFLite conversion enables mobile and low-resource platforms to operate more efficiently while using less storage space. The NeuroAlert pipeline received an optimized model, enabling it to perform well in real-world deployment scenarios.

The process began by exporting the train, which was then converted to TFLite to create a lightweight inference model. The compressed model maintained high classification accuracy while significantly reducing model size.

Algorithm 2 TensorFlow Lite Conversion and Deployment Workflow
Train the proposed NeuroAlert network on the preprocessed dataset
Identify the best model checkpoint using validation accuracy and loss
Export the selected trained model in TensorFlow SavedModel format
Initialize the TensorFlow Lite converter from the exported model
Enable optimization to improve memory efficiency and reduce computational cost
Convert the model from TensorFlow format to TensorFlow Lite format
Save the converted model as a file
Deploy the TensorFlow Lite model on the target embedded platform
Run inference on unseen medical images and obtain predicted class probabilities

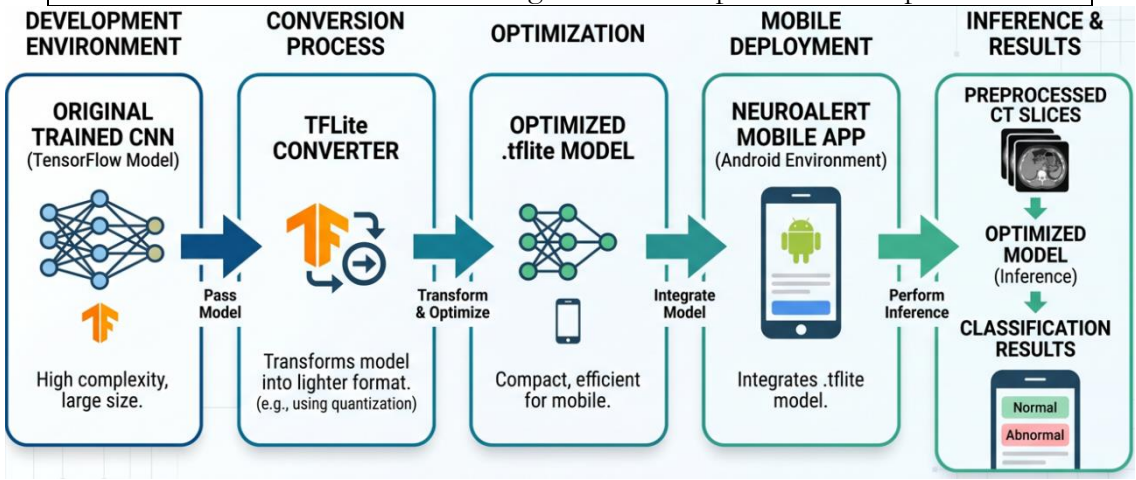

Figure 4. Deployment-oriented optimization pipeline of NeuroAlert. The trained CNN is converted into TensorFlow Lite format and integrated into a lightweight inference framework for efficient CT-slice classification.

Evaluation Metrics:

The proposed framework used standard multi-class classification metrics derived from confusion matrices to evaluate the system's performance, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC). Accuracy calculates the percentage of correctly identified samples, while precision and recall measure

the accuracy of class predictions and the ability to identify actual cases, respectively. The F1-score provides a harmonic mean of precision and recall, making it particularly useful in imbalanced classification settings.

Precision, recall, and the F1-score together establish the performance metrics for a particular class.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

The AUC metric was used to assess how well the predicted class probabilities separated the different classes. Positive and negative instances can be better distinguished by an increasing AUC.

In this study, we evaluated performance on a single held-out test split using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC. Because the subtype-level results were obtained from a single trained model on a fixed test set, we did not compute formal statistical significance or confidence intervals. With this setup, comparing subtype metrics is not robust without repeated runs, cross-validation, or bootstrap resampling. As a result, we interpret the class-wise results descriptively. For future work, more rigorous analysis, such as confidence intervals and repeated experiments, is needed.

Implementation Environment:

The NeuroAlert framework was built using TensorFlow-based deep learning tools, which provided model development capabilities alongside TensorFlow Lite for deployment optimization. The project environment included Android-based application integration, which supported the framework's lightweight design requirements. Design requirements. Python provided the environment for model training and experimentation, while the optimized inference model was developed for future use in a mobile application.

The proposed methodology creates a unified framework that combines clinically informed CT preprocessing, lightweight CNN-based representation learning, class-imbalance-aware training, regularized optimization, and deployment-oriented model conversion. NeuroAlert achieves accurate multi-class intracranial hemorrhage detection through its design, which enables efficient operation in resource-constrained environments.

Results & Discussion:

The section evaluates the effectiveness of the NeuroAlert framework by assessing its classification accuracy, class-based performance, computational power, and ability to operate in real-world settings. The findings are organized into four major sections: model performance assessment, confusion matrix error assessment, efficiency assessment, and a comprehensive assessment of system benefits and constraints.

Model Performance Analysis:

The NeuroAlert CNN was evaluated on a test set comprising 75,200 CT slices from the RSNA Intracranial Hemorrhage dataset. Table 2 presents the class-specific metrics, which include precision, recall, F1-score, AUC, and accuracy for the five hemorrhage categories and the Any Hemorrhage category.

The model achieved 0.935 macro-averaged precision, 0.918 macro-averaged recall, 0.926 macro-averaged F1-score, 0.985 AUC, and 93.7% accuracy, demonstrating strong multi-class differentiation across all target categories. The best performance among all classes occurred in the Any Hemorrhage category, which attained an F1-score of 0.95, an AUC of

0.992, and 96.2% accuracy. The Intraparenchymal Hemorrhage (IPH) category demonstrated strong detection performance through an F1-score of 0.94 and AUC value of 0.990. The model demonstrates superior performance in detecting major hemorrhagic patterns that exhibit distinct radiological features.

The preprocessing and training approach used in this research influences class-specific performance. The Hounsfield Unit (HU) windowing method provided important diagnostic intensity data via its multi-channel output, enabling improved training outcomes through weighted categorical cross-entropy, which reduces the negative impact of class imbalance. The model achieved high recall across all hemorrhage types, highlighting its importance in time-sensitive screening.

Table 2. Performance metrics of the proposed Neuro Alert CNN architecture on the held-out RSNA test set.

Class	Precision	Recall	F1-score	AUC	Accuracy (%)
Epidural (EDH)	0.94	0.91	0.93	0.983	92.7
Subdural (SDH)	0.92	0.93	0.93	0.986	94.1
Subarachnoid (SAH)	0.91	0.90	0.90	0.978	92.3
Intraparenchymal (IPH)	0.95	0.93	0.94	0.990	95.4
Intraventricular (IVH)	0.93	0.89	0.91	0.981	91.6
Any Haemorrhage	0.96	0.95	0.95	0.992	96.2
Macro Average	0.935	0.918	0.926	0.985	93.7

Confusion Matrix and Error Analysis:

The normalized confusion matrix of the proposed model is shown in Figure 4. The main diagonal contains most predictions, indicating that the network correctly classified most samples across all hemorrhage categories. The class sensitivity remained constant across all labels, as recall remained above 0.89 across all classes.

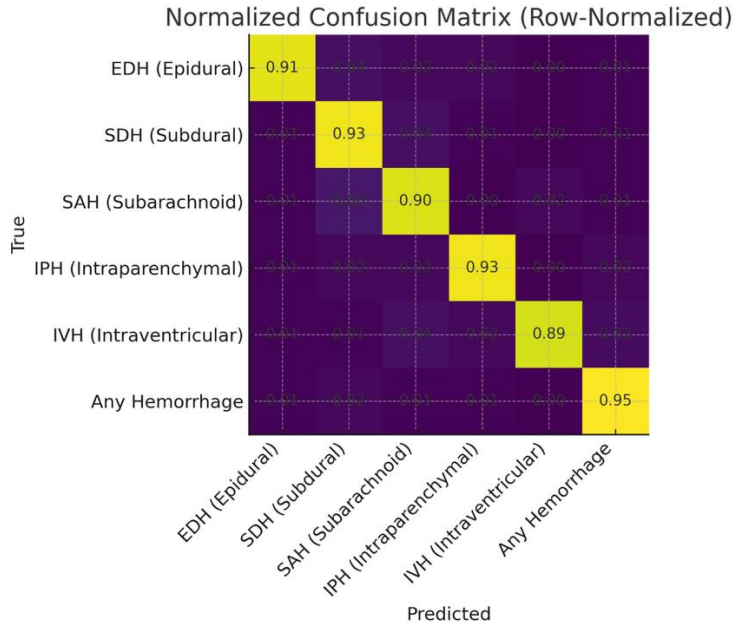

Figure 5. Normalized confusion matrix of the proposed NeuroAlert model on the held-out test set. Rows represent ground-truth classes and columns represent predicted classes.

Darker diagonal entries indicate stronger class-wise recognition, while off-diagonal entries represent inter-class confusion. The most visible confusion occurs between SAH and SDH. The greatest instance of confusion happened between Subarachnoid Hemorrhage (SAH) and Subdural Hemorrhage (SDH). This pattern occurs in clinical practice because the two types of hemorrhage can exhibit overlapping density features, while their axial CT images

display similar anatomical structures. The two conditions showed distinct separation, as evidenced by their higher F1 Scores and AUC values.

The findings demonstrate that the model effectively detects hemorrhagic cases at the triage level. The remaining subtype-level confusion shows that lightweight slice-based CNNs still struggle with visually similar hemorrhage patterns when they lack volumetric context.

Efficiency Evaluation:

NeuroAlert achieved strong classification results while being developed as a lightweight solution for environments with limited computational resources. The model achieved a 61% size reduction after TensorFlow Lite optimization, while maintaining its predictive capabilities with a macro-averaged accuracy of 93.7%, F1-score of 0.926, and AUC of 0.985. The system could conduct near-real-time screening because its inference time remained under 1 second per image.

The system achieves an effective combination of compact design and comprehensive diagnostic capability according to the research results. Mobile and edge-oriented applications need model size reduction because storage constraints, latency challenges, and energy-efficiency requirements limit their operational capabilities. The model maintained its classification accuracy because optimization processes did not significantly reduce its ability to differentiate between classes. Table 3 provides the major performance metrics of NeuroAlert, which describe its complete predictive capabilities and deployment features.

Table 3. Overall performance and deployment-oriented characteristics of NeuroAlert.

Metric	Value
Macro Accuracy	93.7%
Macro F1-score	0.926
AUC	0.985
Model size reduction after TFLite optimization	61%
Inference time per image	< 1 s
Deployment format	TensorFlow Lite
Target platform	Android-based mobile application

NeuroAlert demonstrates strong efficiency. Unlike other systems that require complex, resource-intensive setups, NeuroAlert uses a lightweight CNN that still delivers strong hemorrhage detection. This design makes it practical for mobile use, which is especially helpful in healthcare settings where time and resources are limited and quick decision support is needed.

Discussion:

The results show that NeuroAlert achieves accurate detection of intracranial hemorrhage and meets efficient computational requirements. The method supports rapid decision-making when clinicians need to identify intracranial hemorrhage but lack access to advanced computational technologies. The proposed framework combines clinically meaningful preprocessing, balanced optimization, and lightweight deployment within a single end-to-end system. The model achieved high precision, recall, and AUC across all hemorrhage classes. The network demonstrates strong diagnostic reliability, correctly distinguishing normal from abnormal slices for screening. This is evidenced by its high performance on Any Hemorrhage tests. The confusion between SAH and SDH reflects the visual similarity of certain hemorrhage patterns. It highlights the difficulty of fine-grained subtype discrimination in slice-based analysis.

The framework achieves its lightweight deployment goals through TensorFlow Lite optimization which reduces model size. The computationally intensive 3D CNN and transformer-based models provide additional representational capacity, but they require far

more memory and processing power to operate. NeuroAlert uses a compact design approach, which enables practical operation while maintaining its essential performance metrics.

The study requires acknowledgment of several existing research boundaries. The evaluation used a benchmark dataset for testing, since clinical settings and the model's slice-level operation prevented full volumetric CT analysis. The hardware platform benchmarks need expansion, as current results cover only two platforms. These limitations present critical research paths for upcoming studies. Future work will focus on multi-center validation, volumetric modeling, and hardware-aware optimization.

The results show that NeuroAlert functions as an accurate multi-class hemorrhage classification model. It serves as a computationally efficient framework for time-critical operations in low-resource settings.

Conclusion:

The study introduced NeuroAlert, a lightweight convolutional neural network that automates the analysis of non-contrast CT images to detect intracranial hemorrhage. The framework was designed to deliver both exceptional diagnostic accuracy and computational performance, enabling it to function in resource-constrained environments. NeuroAlert achieved 93.7 percent macro-averaged accuracy, 0.926 F1 score, and 0.985 AUC through when evaluated on the RSNA Intracranial Hemorrhage dataset, which demonstrated its ability to accurately classify all hemorrhage types.

The framework derives its primary value from its compact design, which enables deployment across various environments. The model size decreased by 61 percent after TensorFlow Lite optimization, while the system maintained the ability to process images in less than 1 second. The framework maintains its strong predictive capabilities while becoming better suited for environments that require quick analysis and limited resources. NeuroAlert demonstrates that lightweight CNN-based models can serve as effective alternatives for more resource-intensive architectures in hemorrhage detection.

NeuroAlert provides practical value as a preliminary screening tool and decision-support system. The framework provides a solution for situations requiring immediate results but lacking access to advanced computing resources by delivering precise classification results with low processing requirements. The model becomes essential in healthcare environments operating with limited resources through its ability to provide fast AI-assisted analysis, which helps deliver essential clinical care.

The work demonstrates several crucial limitations that researchers must acknowledge. The evaluation used a benchmark dataset rather than testing in actual clinical settings, and the model processed only slice-level data rather than complete volumetric information from CT studies. The optimization results show positive outcomes, but requires further validation across across various hardware configurations and multiple external datasets to provide proof of actual deployment.

The framework will expand through multi-center validation, including Grad-CAM and SHAP as explainability techniques, and the development of hybrid 2D-3D architectures to enhance volumetric information collection. The system would achieve greater translational value through advancements in hardware-aware optimization, thereby improving its practical imaging workflow capabilities.

NeuroAlert shows that a well-structured, lightweight deep learning system enables high-performance classification while providing efficient implementation for detecting intracranial hemorrhage. The proposed framework presents a valuable solution that enables fast, efficient, and accessible AI support for neuroimaging studies.

Acknowledgement:

The author would like to acknowledge the use of the publicly available RSNA Intracranial Hemorrhage Detection dataset, provided by the Radiological Society of North

America (RSNA), which made this research possible. The author also appreciates the support of colleagues from the National University of Modern Languages (NUML).
Author's Contribution: Amna Sajid and Mohsin Abbas conceptualized the study, wrote the manuscript, and supervised the research. Aimen Ashraf and Misbah Noor handled the model implementation and experimental setup. Anam Taskeen and Faryal Hayat helped organize the manuscript and reviewed it for intellectual content.
Conflict of interest: Authors are advised to explain that there is no conflict of interest in publishing this manuscript in IJIST.
References:
[1] Luis Cortés-Ferre, Miguel Angel Gutiérrez-Naranjo, "Deep Learning Applied to Intracranial Hemorrhage Detection," *J. Imaging*, vol. 9, no. 2, p. 37, 2023, doi: https://doi.org/10.3390/jimaging9020037.
[2] Muhannad Seyam, Thomas Weikert, "Utilization of Artificial Intelligence-based Intracranial Hemorrhage Detection on Emergent Noncontrast CT Images in Clinical Workflow," *Radiol. Artif. Intell.*, vol. 4, no. 2, p. e210168, 2022, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35391777/
[3] So Yeon Choi, Ji Hoon Kim, Hyun Soo Chung, Sona Lim, Eun Hwa Kim & Arom Choi, "Impact of a deep learning-based brain CT interpretation algorithm on clinical decision-making for intracranial hemorrhage in the emergency department," *Sci. Rep.*, 2024, [Online]. Available: https://www.nature.com/articles/s41598-024-73589-0
[4] Jialiang Fan, Xinhui Fan, Chengyan Song, Xiaofan Wang, Bingdong Feng, Lucan Li & Guoyu Lu, "Dual-task vision transformer for rapid and accurate intracerebral hemorrhage CT image classification," *Sci. Rep.*, 2024, [Online]. Available: https://www.nature.com/articles/s41598-024-79090-y
[5] Bargava Subramanian, Naveen Kumarasami, Praveen Shastry, "3D Convolutional Neural Networks for Improved Detection of Intracranial bleeding in CT Imaging," *arXiv:2503.20306*, 2025, [Online]. Available: https://arxiv.org/abs/2503.20306
[6] A. Phaphuangwittayakul, Y. Guo, F. Ying, A. Y. Dawod, S. Angkurawaranon, and C. Angkurawaranon, "An optimal deep learning framework for multi-type hemorrhagic lesions detection and quantification in head CT images for traumatic brain injury," *Appl. Intell. (Dordrecht, Netherlands)*, vol. 52, no. 7, pp. 7320–7338, May 2022, doi: 10.1007/S10489-021-02782-9.
[7] Jayesh Soni, "Toward the detection of intracranial hemorrhage: a transfer learning approach," *Art Int Surg*, 2025, doi: 10.20517/ais.2024.46.
[8] S. Kaur and A. Singh, "A New Deep Learning Framework for Accurate Intracranial Brain Hemorrhage Detection and Classification Using Real-Time Collected NCCT Images," *Appl. Magn. Reson. 2024 556*, vol. 55, no. 6, pp. 629–661, Jun. 2024, doi: 10.1007/S00723-024-01661-Z.
[9] Dong Wan Kang, Museong Kim, "Deep learning-assisted detection of intracranial hemorrhage: validation and impact on reader performance," *Neuroradiology*, vol. 67, no. 6, pp. 1511–1519, 2025, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40116947/
[10] Ronald Antulov, Martin Weber Kusk, "Real-World Integration of an Automated Tool for Intracranial Hemorrhage Detection in an Unselected Cohort of Emergency Department Patients-An External Validation Study," *Diagnostics (Basel, Switzerland)*, vol. 16, no. 2, p. 282, 2026, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/41594258/
[11] Armin Karamian, Ali Seifi, "Diagnostic Accuracy of Deep Learning for Intracranial Hemorrhage Detection in Non-Contrast Brain CT Scans: A Systematic Review and Meta-Analysis," *J. Clin. Med.*, vol. 14, no. 7, p. 2377, 2025, doi:

https://doi.org/10.3390/jcm14072377.

[12] L. Goelz *et al.*, "Machine-learning algorithms for detecting intracranial hemorrhage on head computed tomography," *Cochrane Database Syst. Rev.*, vol. 2026, no. 2, p. CD016266, Feb. 2026, doi: 10.1002/14651858.CD016266.

[13] P. Hu *et al.*, "Deep learning-assisted detection and segmentation of intracranial hemorrhage in noncontrast computed tomography scans of acute stroke patients: a systematic review and meta-analysis," *Int. J. Surg.*, vol. 110, no. 6, p. 3839, Jun. 2024, doi: 10.1097/JS9.0000000000001266.

[14] Gul Cihan Habek, Fatih Basciftci, "Slice-level and scan-level performance of deep learning models for intracranial hemorrhage detection and subtype classification: a systematic review and meta-analysis," *Eng. Sci. Technol. an Int. J.*, vol. 77, p. 102368, 2026, doi: https://doi.org/10.1016/j.jestch.2026.102368.

[15] Alireza Golkarieh, Mohammad Sassani, "Intracranial Hemorrhage Diagnosis Using Deep Learning: A Survey of Techniques, Frameworks, and Challenges," *Comput. Decis. Mak. An Int. J.*, vol. 3, pp. 780–804, 2026, doi: 10.59543/comdem.v3i.15543.

[16] Omid Mirzaei, Sedra Alialsagher Mohammed, "Comparison of Intracranial Hemorrhages Detection Performances of Deep Learning Models on CT Images," *Procedia Comput. Sci.*, vol. 258, pp. 3194–3202, 2025, doi: https://doi.org/10.1016/j.procs.2025.04.577.

[17] Franziska Tombach, Kristina Krompaß, "Diagnostic performance and confidence of an optimized deep-learning algorithm for the detection of intracranial hemorrhages," *Insights Imaging*, vol. 17, no. 1, p. 73, 2026, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/41843310/

[18] P. Fusco, G. P. Rimoli, and M. Ficco, "TinyML and Federated Learning for Resource-Constrained Medical Devices," pp. 113–126, 2025, doi: 10.1007/978-3-031-70775-9_7.

[19] Md Wahidur Rahman, Mais Nijim, "Privacy-Preserving Federated IoT Architecture for Early Stroke Risk Prediction," *Electronics*, vol. 15, no. 1, p. 32, 2026, doi: https://doi.org/10.3390/electronics15010032.

[20] Kai Cheng Yang, Yunzhi Xu, "Explainable deep learning algorithm for identifying cerebral venous sinus thrombosis-related hemorrhage (CVST-ICH) from spontaneous intracerebral hemorrhage using computed tomography," *eClinicalMedicine*, vol. 81, p. 103128, 2025, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40093990/

[21] A. Nada *et al.*, "External validation and performance analysis of a deep learning-based model for the detection of intracranial hemorrhage," *Neuroradiol. J.*, vol. 38, no. 3, pp. 312–321, Jun. 2025, doi: 10.1177/19714009241303078.

[22] Mohammadreza Chavoshi, Aawez Mansuri, Wasif Bala, "Real-world performance evaluation of a commercial deep learning model for intracranial hemorrhage detection," *npj Digit. Med.*, 2026, [Online]. Available: https://www.nature.com/articles/s41746-025-02244-3

[23] H. Kavandi, K. Costenbader, S. Yazbek, P. Kamel, N. Yahyavi-Firouz-Abadi, and J. Jeudy, "Performance Evaluation of a Commercial Deep Learning Software for Detecting Intracranial Hemorrhage in a Pediatric Population," *J. imaging informatics Med.*, 2026, doi: 10.1007/S10278-026-01857-8.

[24] Tommaso D'Angelo, Giuseppe M. Bucolo, "Accuracy and time efficiency of a novel deep learning algorithm for Intracranial Hemorrhage detection in CT Scans," *Radiol. Med.*, vol. 129, no. 10, pp. 1499–1506, 2024, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/39123064/

[25] S. Ahmed, "Exploring Deep Learning and Machine Learning Approaches for Brain

Hemorrhage Detection," *IEEE Access*, vol. 12, pp. 45060–45093, 2024, doi: 10.1109/ACCESS.2024.3376438.

[26] Kasra Davoodi, Mohammad Hoseyni, Javad Khoramdel, Reza Barati, Reihaneh Mortazavi, "Hemorica: A Comprehensive CT Scan Dataset for Automated Brain Hemorrhage Classification, Segmentation, and Detection," *arXiv:2509.22993*, 2025, [Online]. Available: https://arxiv.org/abs/2509.22993

[27] D. W. Kang *et al.*, "Strengthening deep-learning models for intracranial hemorrhage detection: strongly annotated computed tomography images and model ensembles," *Front. Neurol.*, vol. 14, 2023, doi: 10.3389/FNEUR.2023.1321964/PDF.

[28] Jingjing Liu, Weijie Fan, "Deep learning-based identification and localization of intracranial hemorrhage in patients using a large annotated head computed tomography dataset: A retrospective multicenter study," *Intell. Med.*, vol. 5, no. 1, pp. 14–22, 2025, doi: https://doi.org/10.1016/j.imed.2024.11.002.

[29] P. Raguraman, M. Kumaresan, and S. Ramesh, "A Comprehensive Survey of AI-Driven Biomedical Image Processing for Intracerebral Hemorrhage Detection and Classification: Current Trends, Challenges, and Future Directions," *8th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2024 - Proc.*, pp. 1649–1653, 2024, doi: 10.1109/ICECA63461.2024.10801000.

[30] Qian Gao, Yujia Jin, "Transforming Intracerebral Hemorrhage Care with Artificial Intelligence: Opportunities, Challenges, and Future Directions," *Diagnostics*, vol. 16, no. 5, p. 752, 2026, doi: https://doi.org/10.3390/diagnostics16050752.