Homomorphic Encryption for EDSS Classification: Safeguarding Patient Privacy in MRI-Based Assessment of Multiple Sclerosis

Stefano Cirillo[*,†], Vincenzo Deufemia[†], Luigi Di Biasi[†], Giuseppe Polese[†], Giandomenico Solimando[†] and Genoveffa Tortora[†]

[1]*University of Salerno, Department of Computer Science, Fisciano, Salerno, Italy*

Abstract

Hospitals and healthcare organizations collect vast amounts of patient data, such as MRI scans, which hold significant potential for advancing automated clinical support systems. However, privacy concerns and the lack of robust data anonymization and protection mechanisms often hinder data sharing and collaborative research. To this end, privacy-preserving and data sanitization techniques have emerged as a promising direction. Among them, Homomorphic Encryption (HE) allows computations to be performed directly on encrypted data without requiring decryption, thereby safeguarding sensitive information throughout the analytical pipeline. In this paper, we investigate the feasibility of leveraging homomorphic encryption to enable Expanded Disability Status Scale (EDSS) classification in Multiple Sclerosis (MS). Thus, we design a dedicated neural network, namely Hybrid AHE-CNN, tailored for processing images together with homomorphically encrypted sensitive data, allowing for secure and privacy-preserving inference without exposing raw patient data. Experimental results demonstrate that our proposed method achieves classification performance comparable to that of models trained and evaluated on plaintext data, highlighting the practical applicability of HE in real-world healthcare settings.

Keywords

Multiple Sclerosis Diagnosis, Homomorphic Encryption, Secure Medical Image Processing.

1. Introduction

Human error in medicine remains a significant cause of misdiagnosis, and the rapid expansion of medical knowledge makes it increasingly challenging for physicians to keep pace. In this scenario, intelligent systems that support clinical decision-making, such as Computer-Aided Diagnosis (CAD) and Clinical Decision Support Systems (CDSS), are gaining traction in both scientific and medical circles because they can help address longstanding challenges in healthcare. Today's Decision Support Systems (DSS) often rely on Machine Learning (ML) and Deep Learning (DL) technologies, which can extract valuable insights from healthcare data. By doing so, they enhance diagnostic accuracy, expedite medical decision-making, and streamline clinical workflows.

CAD and CDSS may assist in detecting diseases, predicting conditions, such as Alzheimer's or Parkinson's, or suggesting appropriate diagnoses and treatments based on extensive clinical datasets. Moreover, these systems help reduce errors by providing objective, data-driven analyses [1]. Additionally, DSS can automate certain aspects of the diagnostic process, enabling physicians to save time and concentrate on more complex or urgent cases. Over the long term, such tools could enhance overall welfare, especially in areas with a shortage of specialists, such as rural regions or developing countries, by enabling the dissemination of digital health solutions in under-resourced settings. However, despite

TADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

*Corresponding author.

†These authors contributed equally.

scirillo@unisa.it (S. Cirillo); deufemia@unisa.it (V. Deufemia); ldibiasi@unisa.it (L. D. Biasi); gpolese@unisa.it (G. Polese); gsolimando@unisa.it (G. Solimando); tortora@unisa.it (G. Tortora)

0000-0003-0201-2753 (S. Cirillo); 0000-0002-6711-3590 (V. Deufemia); 0000-0002-9583-6681 (L. D. Biasi); 0000-0002-8496-2658 (G. Polese); 0009-0000-6627-8820 (G. Solimando); 0000-0003-4765-8371 (G. Tortora)

their high performance and clear advantages, CAD and CDSS face a significant barrier to widespread adoption: the scarcity of extensive and balanced datasets for many diseases. Although healthcare institutions possess vast amounts of data, sharing and processing this sensitive information without compromising patient privacy requires addressing several key privacy concerns. Patient records often contain critical predictive features that are also highly personal, and any accidental exposure or misuse could lead to severe ethical and legal repercussions. As a result, these challenges significantly slow down the development of collaborative diagnostic systems that depend on data from multiple institutions.

Increasing attention has been given in recent years to privacy-preserving machine learning (ML) techniques, which aim to enable predictive models to be continuously trained and applied to sensitive data without compromising its confidentiality [2]. Traditional anonymization methods are often inadequate because they either reduce the data's usefulness for machine learning or fail to prevent re-identification, especially when combined with other information thoroughly [3].

One of the most promising approaches is Homomorphic Encryption (HE), a form of encryption that allows computations to be performed directly on encrypted data without requiring decryption. This type of encryption enables encrypted inputs to undergo various arithmetic and logical operations as if they were in plaintext. This groundbreaking property makes HE particularly suitable for privacy-preserving healthcare applications. For instance, a hospital could encrypt its patient records and outsource model training or inference tasks to third-party services without revealing any sensitive information. These services can perform ML computations, such as classifying a tumor as benign or malignant, on the encrypted data and return an encrypted result. Only the hospital, with the appropriate decryption key, can then access the final output.

In the context of collaborative learning across multiple institutions, homomorphic encryption enables the development of shared models without requiring any party to expose its patient data. Each institution can encrypt its data, contribute to the training process, and receive updates without ever compromising patient confidentiality. However, the definition of these types of models is highly challenging due to the computational overhead and complexity associated with performing machine learning operations on encrypted data, as well as the fact that most existing algorithms are not naively compatible with the limited set of operations supported by most HE schemes.

Scope and contributions of this work. We propose a neural network that enables secure training and inference on encrypted multimodal medical data using approximate homomorphic encryption (AHE). As a use case, we focus on Multiple Sclerosis classification and grading problems (MSCGP), combining MRI images with sensitive clinical data to enable the privacy-preserving classification task. We introduce a new CNN adapted for encrypted computation, namely Hybrid AHE-CNN, for binary and multiclass classification tasks. To preserve patient confidentiality, we adopt the CKKS scheme, which supports computations on encrypted clinical features without revealing sensitive clinical data. Therefore, the contributions of this paper are:

- A new Hybrid AHE-CNN architecture that integrates encrypted clinical data with unencrypted MRI slices to improve diagnostic performance while preserving privacy;
- A comparative evaluation of the encrypted neural network with traditional CNN on both binary and multiclass EDSS classification tasks.

The remainder of the paper is organized as follows. In Section 2, we describe relevant studies concerning homomorphic encryption and its application in medical image classification, with a focus on MRI data. In Section 3, we provide a brief overview of the dataset used, including the characteristics of the brain MRI dataset. In Section 4, we first formalize the problem of working with encrypted MRI images, and then we present the new Hybrid AHE-CNN. In Section 5, we discuss the results achieved from the experimental evaluations, including both binary and multiclass classification performance with respect to the homomorphic encryption approach. Finally, conclusions and future directions are provided in Section 7.

2. Related Work

Computer-Aided Diagnosis (CAD) and Clinical Decision Support Systems (CDSS) have shown great potential in assisting clinicians in early diagnosis, prognosis, and treatment planning, especially in complex neurological

disorders such as Multiple Sclerosis (MS) [4, 5]. MS is characterized by heterogeneous progression patterns and multifactorial etiology and benefits from data-driven approaches that can integrate imaging, clinical history, and lab results to enhance diagnostic accuracy [6, 7, 8]. However, the adoption of such systems is limited by the sensitive nature of medical data and the lack of sufficiently large and diverse datasets, especially for rare or chronic conditions like MS [9, 10].

In this scenario, Fully Homomorphic Encryption (FHE) is emerging as a promising solution, thanks to its ability to allow inference directly on encrypted data. FHE helps preserve patient confidentiality without sacrificing model performance. In this scenario, FHE can facilitate collaborative training and inference across institutions while ensuring regulatory compliance and enabling privacy-preserving deep learning methodologies [11].

Recent studies have demonstrated the feasibility of performing deep learning inference under FHE. For instance, in [12], authors presented an FHE-based ResNet-20 using the RNS-CKKS scheme, showing that standard networks can operate on encrypted images without retraining, achieving accuracy close to the plaintext baseline. However, their solution suffers from long inference times, with hours needed per image due to computational overhead, particularly bootstrapping.

Another recent proposal is CaRENets [13], which represents a resource-aware framework that applies compact matrix packing strategies to reduce ciphertext count and latency in FHE-CNN inference for medical imaging. Their approach led to significant speedups and memory savings across synthetic and real clinical datasets, making homomorphic evaluation more practical for high-resolution inputs.

Other contributions have extended the application of FHE beyond inference to support training as well. The MORE framework [14] enables both training and inference over encrypted floating-point data by encoding plaintexts as matrices and applying operations homomorphically, including nonlinear activations through matrix functions or eigendecomposition. The framework was tested on multiple medical and synthetic tasks with accuracy comparable to non-encrypted pipelines.

Recently, more hardware-aware frameworks have emerged, such as HoRNS-CNN [15] that leverages FPGA-based encryption modules and low-degree polynomial approximations of ReLU to enable efficient end-to-end encrypted MRI classification. Their design achieves strong privacy-utility trade-offs, particularly in energy and latency metrics. Another recent framework is PervPPML [16] that integrates lightweight symmetric encryption with homomorphic methods to reduce the overhead of FHE on edge devices, showing its applicability in ECG classification with minimal accuracy loss.

Other recent studies have investigated new strategies, such as using binarized networks under encryption [17], approximating neural operations through Chebyshev polynomials [18], or combining secure multiparty computation and FHE for hybrid privacy guarantees [19]. Recent works like [20] and [21] have focused on software abstractions and optimizations to make encrypted inference more accessible to practitioners.

In our study, we propose a novel hybrid neural network that processes MRI images along with encrypted clinical data to classify EDSS points with different levels of granularity. Unlike prior works that either process plaintext data or consider only a single type of data, our approach combines MRI images with sensitive clinical data encrypted using approximate homomorphic encryption (AHE). This enables secure classification of Multiple Sclerosis patients in both binary and multiclass settings, preserving patient privacy while maintaining reasonable diagnostic performances.

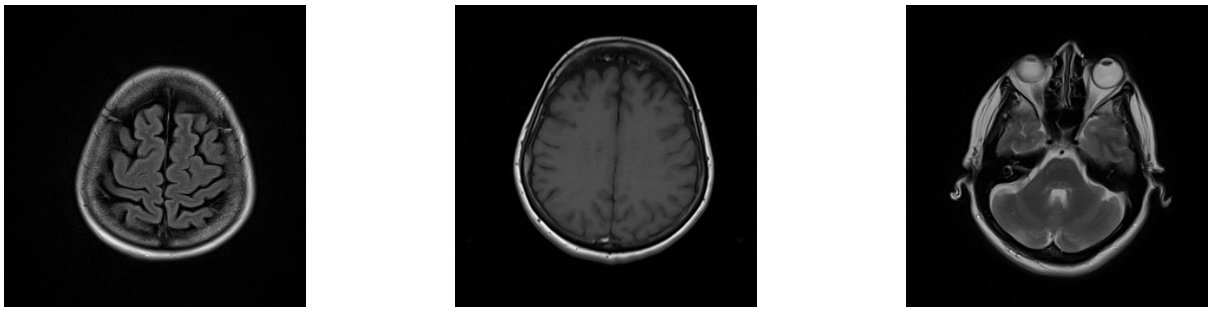
3. Materials and Methods

In this section, we describe the datasets and preprocessing steps used to evaluate our proposed Hybrid AHE-CNN for Multiple Sclerosis classification. We first outline the characteristics of the brain MRI dataset and the associated clinical metadata. Then, we provide an overview of the preprocessing steps adopted for labeling datasets and addressing data imbalance.

3.1. Brain MRI Dataset

To enable the classification of disability severity in Multiple Sclerosis (MS) patients from MRI scans, we use one multi-sequence MRI dataset related to 60 MS patients with consensus manual lesion segmentation, EDSS, general patient information, and clinical information [22]. The dataset contains 4,189 images segmented and validated by three radiologists and neurologist experts. As we can see in Figure 1, it contains representations of manual MS-lesion segmentations on three MRI sequences: T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR).

2D axial MRI slices were extracted from each patient's 3D scan for both datasets and stored as individual grayscale images. Each image was labeled according to the EDSS-based class assigned to the corresponding



(a) FLAIR (b) T1 (c) T2

Figure 1: Example of MRI images extracted from [22].

patient. The resulting datasets were stored in separate folders and loaded via a custom PyTorch `Dataset` class that records patient identifiers and slice indices, facilitating data stratification and per-patient evaluation.

This dual-dataset approach allows the exploration of both simplified and nuanced diagnostic models, enabling the evaluation of classifier performance across different levels of clinical relevance and computational complexity.

Starting from this, we consider two different sets of labels associated with the original imaging and metadata sources. These datasets differ in their target labeling strategy and correspond to two different levels of granularity in the clinical assessment of disease severity.

The first configuration aims to address a binary classification task, in which the patients are grouped into two classes based on their Expanded Disability Status Scale (EDSS) score, i.e., class_0 that includes all patients with an EDSS value less than or equal to 2.0, and class_1 with all patients with EDSS greater than 2.0. It is important to notice that a patient belonging to class_0 is a patient that does not show significant lesions and minimal or no functional impairment, whereas class_1 contains patients with evident neurological lesions and clinical symptoms.

The second configuration aims to address a multi-class classification task. In this case, EDSS scores were mapped to three categories, *normal*, *mild*, and *severe*. Specifically, scores from 0 to 2.0 were labeled as *normal*, indicating little to no disability; scores between 2.5 and 4.0 were labeled as *mild*, capturing patients with moderate impairment but preserved ambulation, and scores greater than 4.0 were labeled as *severe*, corresponding to individuals with significant motor or systemic dysfunction.

3.2. Preprocessing of MS Patient Data

Following the introduction of both binary and multiclass labeling strategies based on EDSS scores, a significant class imbalance was observed in the resulting dataset.

Table 1 summarizes the data distribution across the different configurations for both binary and multiclass classification tasks. As we can see, there is a significant class imbalance in the original dataset in both configurations. In fact, for the binary task, we have $3,765$ and 424 instances for class_0 and class_1, respectively, whereas for the multiclass we have $2,071$, $1,233$, and 885 instances labeled as *Normal*, *Mild*, and *Severe*, respectively.

Since class imbalance can negatively impact the performance of predictive models in classification tasks, we adopted several undersampling and oversampling techniques to mitigate this issue. These techniques were applied

Task	Techniques	Class_0	Class_1	Normal	Mild	Severe	Total
Binary	*Original*	3,765	424	–	–	–	4,189
	Undersampling	424	424	–	–	–	848
	Oversampling	3,765	3,765	–	–	–	7,530
	SMOTE	3,765	3,765	–	–	–	7,530
Multiclass	*Original*	–	–	2,071	1,233	885	4,189
	Undersampling	–	–	885	885	885	2,655
	Oversampling	–	–	2,071	2,071	2,071	6,213
	SMOTE	–	–	2,071	2,071	2,071	6,213

Table 1
Data distribution of the Brain MRI dataset for the binary and multiclass EDSS classification tasks under different dataset balancing techniques (original, undersampling, oversampling, SMOTE).

to the training set after a stratified split was performed to partition the data into 70% training, 15% validation, and 15% test sets. The validation and test sets were left unchanged to preserve their statistical representativeness.

The undersampling was performed by randomly reducing the number of instances in the majority class to match the minority classes. This resulted in a balanced dataset containing 848 instances for the binary task and 2,655 instances for the multiclass task, with equal representation for each class.

Instead, for augmenting data, we applied both transformations and the SMOTE (Synthetic Minority Oversampling Technique) on the original data. Concerning the transformations, each original image was first resized to 128×128 pixels and normalized to zero mean and unit variance. Augmented samples were then generated by applying a random combination of horizontal flips, small-angle rotations (up to 20°), and affine translations, within a 10% range along both axes. These transformations preserved the anatomical structure while introducing controlled variability in orientation and positioning. Conversely, the SMOTE technique allowed us to generate synthetic samples of minority classes by interpolating feature vectors [23]. These changes in the composition of the dataset let us assess the models' performance in both imbalanced and balanced conditions.

4. Homomorphic Encryption in Classification Scenarios

Homomorphic Encryption (HE) is a class of encryption schemes that allows computations to be performed directly on encrypted data, without needing to decrypt it first. This property is especially valuable in privacy-sensitive domains like medical imaging, where sensitive data must remain confidential even when outsourced to untrusted servers for analysis. Depending on the scheme, HE supports either exact arithmetic over integers or approximate arithmetic over real numbers. In both cases, the goal is to ensure that operations performed on ciphertexts correspond, up to some approximation, to operations on the original plaintexts.

In this section, we provide an overview of the techniques underlying our work, including the application of Approximate Homomorphic Encryption (AHE) on data processed with Convolutional Neural Networks (CNNs).

4.1. Processing Data with AHE

Let $\mathcal{X} = \{x_i\}_{i=1}^N$ be a set of Magnetic Resonance Imaging (MRI) scans, where each image $x_i \in \mathbb{R}^{h \times w}$ is a grayscale image of resolution $h \times w$. Let $P = \{p_1, p_2, \dots, p_k\}$ be a set of patients where each p_j is associated with a d sensitive patient-specific variables, such as age and diagnostic codes. The goal is to enable secure and privacy-preserving computation on this data by processing it with encrypted weights, aiming to output an encrypted representation of the feature maps conditioned by the related clinical information.

Let $\mathbf{f}_j \in \mathbb{R}^d$ be the clinical feature vector associated with a patient j, to preserve privacy, we encrypted $\mathbf{f}_j$ using a homomorphic encryption scheme, resulting from the encryption operation under a public key pk and denoted as:

$$\tilde{\mathbf{f}}_j = \mathsf{Encrypt}(\mathbf{f}_j, \mathsf{pk}, \mathscr{C}). \tag{1}$$

where $\mathscr{C}$ is the encryption context, whose exact parameters depend on the chosen scheme. For example, in CKKS, which is designed for approximate real arithmetic, $\mathscr{C}$ typically includes the following parameters: n, the degree of the polynomial modulus, which determines the ciphertext size and the number of available slots; $\mathcal{Q} = \{\log_2 q_0, \log_2 q_1, \dots, \log_2 q_L\}$, a modulus chain that supports a multiplicative depth L, thus affecting both precision and computational capacity; and Δ, a global scaling factor used for fixed-point encoding of real numbers.

Let $\mathbf{x}_i$ be the flattened version of an image x_i, to process privatized data together with MRI images, it is necessary to compute arithmetic operations between x_i and weights $\mathbf{w}_i$ by using an AHE scheme. Thus, given a convolutional filter $\mathbf{w} \in \mathbb{R}^{c \times k \times k}$, where c is the number of input channels and k the kernel size, the filter is flattened and encrypted as $\tilde{\mathbf{w}} = \mathsf{Encrypt}(\mathbf{w}_i, \mathsf{pk}, \mathscr{C})$.

A homomorphic element-wise multiplication is then performed between $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{f}}_j$, resulting in a patient-specific encrypted weight tensor $\tilde{\mathbf{w}}_i = \tilde{\mathbf{w}} \circ \tilde{\mathbf{f}}_j$. This operation allows for personalizing the convolutional weights with respect to encrypted patient features, without ever revealing their plaintext values. The encrypted tensor $\tilde{\mathbf{w}}_i$ is then decrypted under a private key sk to obtain the conditioned filter $\mathbf{w}_i = \mathsf{Decrypt}(\tilde{\mathbf{w}}_i, \mathsf{sk}, \mathscr{C})$. The convolution $\mathbf{x}_i * \mathbf{w}_i$ produces a feature map that conveys personalized representations of the input image $\mathbf{x}_i$, reflecting the patient-specific characteristics encoded in $\mathbf{f}_j$. This approach enables the model to adapt its processing to both image and sensitive data while preserving privacy throughout the training phase.

4.2. Model Architecture and Approximate HE Integration

We developed a conditioned CNN architecture specifically designed to classify EDSS scores from MRI brain slices, with the distinctive integration of homomorphic encryption to modulate part of the model's parameters securely.

Figure 2 shows an overview of the architecture of the proposed Hybrid AHE-CNN. A first convolutional block, denoted Conv2 and parameterized by a weight tensor $W \in \mathbb{R}^{1\times1\times3\times3}$, operates on the 128×128 single-channel input slice. Before every forward pass, the nine-dimensional clinical vector f associated with the current patient is encrypted using the CKKS scheme. TenSEAL then performs an element-wise product between the encrypted weights and the encrypted vector, after internally repeating f until the shapes match. The modulated weights are then decrypted and clipped to $[-1, 1]$ before being frozen for the remainder of the batch:

$$\widetilde{W} = \mathrm{Dec}\big(\mathrm{Enc}(W) \odot \mathrm{Enc}(f)\big), \qquad W_{\text{batch}} \leftarrow \mathrm{clip}\big(\widetilde{W}, -1, 1\big). \tag{2}$$

A parallel unconstrained 3×3 convolution, Conv2_2, is applied to the same input. The output is then concatenated with that of Conv2, yielding a two-channel feature map. The network follows a pyramidal pattern of channel expansion $2 \to 8 \to 16 \to 32$, with each block composed of a 3×3 convolution, batch normalization, a ReLU activation and 2×2 max-pooling. After flattening (8,192 units), the representation passes through a fully connected layer with 512 neurons, spatial dropout with $p = 0.4$, and a final three-way softmax classifier. Only the weight-modulation step is executed on encrypted data, whereas all subsequent convolutions, normalizations, and dense layers run in plaintext. This hybrid design preserves the privacy of sensitive clinical variables while minimizing the latency overhead associated with fully homomorphic inference.

Figure 3 shows the convolutional layer conditioning procedure. As we can see, convolutional weights are first initialized and stored under CKKS homomorphic encryption. For each MRI slice, the corresponding encrypted feature vector is retrieved and used to modulate the encrypted weights via homomorphic multiplication. The modulated weights are then decrypted for gradient-based optimization on the clear-text data, re-encrypted before the next slice, and never expose patient features in clear throughout the process.

The most distinctive aspect of Hybrid AHE-CNN is the encrypted conditioning mechanism applied to the `Conv2` layer. Clinical metadata were encrypted using the CKKS scheme from the TenSEAL library. These encrypted vectors modulated the layer's convolutional weights before inference. Specifically, the original weights were flattened and encrypted, homomorphically multiplied with the encrypted clinical features, decrypted, reshaped, and reinserted into the network. This process enables secure, privacy-preserving influence of sensitive patient information on model behavior without exposing the data in plaintext form.

5. Experimental Evaluation

In this Section, we outline the experimental evaluation performed to address the EDSS classification in Multiple Sclerosis (MS). In particular, we provide the details about the experimental setting, training protocols, and performance metrics employed to assess the effectiveness of the proposed hybrid CNN in classifying the disability severity in Multiple Sclerosis (MS). Then in Section 5.1, we describe the results achieved in the two classification tasks, i.e., binary and multiclass.

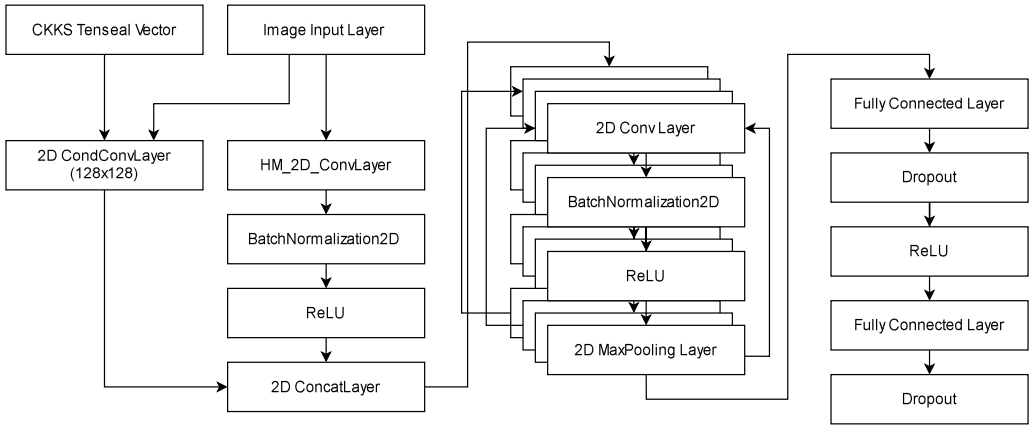


Figure 2: Architecture of the proposed CNN for EDSS classification from MRI slices. The `Conv2` layer is conditioned via homomorphically encrypted clinical features. Its output is concatenated with that of a parallel convolution (`Conv2_2`), followed by a sequence of convolutional, batch normalization, ReLU, and pooling layers. The flattened representation passes through two fully connected layers with dropout. Grad-CAM highlights relevant brain regions to enhance interpretability.

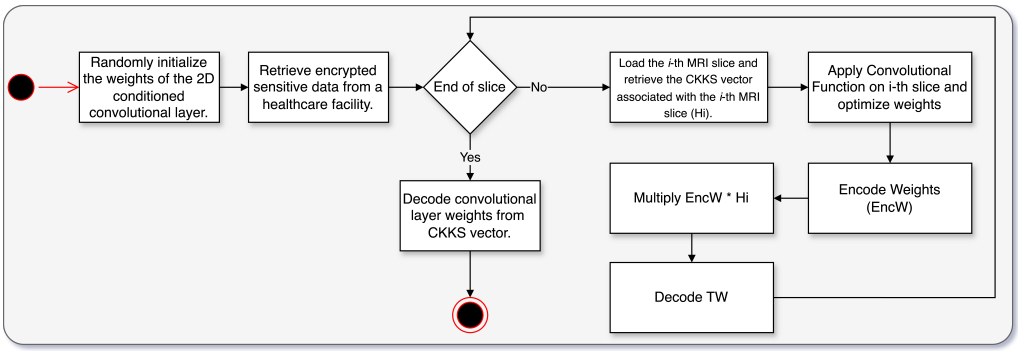


Figure 3: Hybrid training pipeline for the 2D conditioned convolution layer.

Experimental Settings The CNN has been implemented using Python version 3.9 and with the support of PyTorch 2.7.1, CUDA 12.6, Scikit-learn 1.6.1, and TensorFlow 2.19.0. All the experiments have been executed on a workstation with an Intel i9 CPU at 5 GHz, 14-core, and 64GB of memory, equipped with a NVIDIA 3060 GPU.

The CNN was trained using the AdamW optimizer with a mini-batch size of 32. To improve convergence, we employed the *Reduce Learning Rate on Plateau* strategy, which monitors training performance at each epoch and automatically reduces the learning rate if no improvement is observed for 15 consecutive epochs. Moreover, a class-weighted cross-entropy loss was adopted to address the natural imbalance in the class distributions. The weights of the classes were calculated on the basis of the inverse frequency of each class.

For the oversampling, we used the functions of the *torchvision* library for multidimensional image processing, which contains several functions and filters for multidimensional image processing. The images have been transformed through the *transform* method, which enables to combine of multiple transformations to be applied to an image. To evaluate the effectiveness of the proposed Hybrid AHE-CNN, we compare its performance with CNNs operating on unencrypted data, taking into account all the variations applied to the dataset in each experimental setup.

Concerning the homomorphic settings, we employed an approximate homomorphic encryption scheme based on the CKKS protocol, which supports arithmetic operations on encrypted floating-point numbers. The encryption context was instantiated with a polynomial modulus degree of 8192 and coefficient modulus bit sizes set to $[60, 40, 40, 60]$, providing a balance between computational efficiency and encryption depth. We set the global scale to 2^{40}, ensuring sufficient precision for encrypted tensor operations during inference.

Evaluation Metrics. To evaluate the performance of the proposed CNNs, we use Accuracy, Precision, Recall, and F1-score metrics. The latter are defined in terms of the number of True Positives (TP), i.e., when an instance of an EDSS type is identified to belong to its true class, e.g., a patient belonging to the class EDSS 1, is correctly classified as EDSS 1. False Positive (FP), i.e., when an instance is incorrectly predicted to belong to a class other than its true class, e.g., a patient belonging to the class EDSS 1, is incorrectly classified as EDSS 0. True negative (TN), i.e., an instance of the 0 class is correctly predicted as 0. False Negative (FN), i.e., a patient belonging to the class 1, is incorrectly predicted as 0.

For both binary classification, where models aim to distinguish between *EDSS* 1 and *EDSS* 0, and multiclass classification, where the goal is to differentiate between multiple severity levels such as *normal*, *mild*, and *severe*, the corresponding evaluation metrics are reported in Table 2.

5.1. Evaluation Results

In this Section, we discuss the results of the proposed Hybrid AHE-CNN in identifying both the EDSS score and the severity level, for binary and multiclass classification tasks, respectively. All models were evaluated on encrypted inputs using the AHE scheme, with intermediate feature maps decrypted before being forwarded to the final classification layers, as described in Section 4.1.

Binary Classification. In the binary task, the classification task involves distinguishing between patients with an EDSS lower than or equal to 2.0, labeled as the class indicating the absence of significant lesions and minimal or no functional impairment, labeled as the negative class, and those with an EDSS greater than 2.0, corresponding to the presence of neurological lesions and clinically relevant symptoms, labeled as the positive class. We evaluated the proposed encrypted CNN architecture under different training conditions, including

Metric	Binary	Multiclass
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$	$\frac{\sum_{i=1}^{z} TP_i + \sum_{i=1}^{z} TN_i}{\sum_{i=1}^{z} TP_i + FP_i + TN_i + FN_i}$
Precision	$\frac{TP}{TP+FP}$	$\frac{\sum_{i=1}^{z} TP_i}{\sum_{i=1}^{z} TP_i + FP_i}$
Recall	$\frac{TP}{TP+FN}$	$\frac{\sum_{i=1}^{z} TP_i}{\sum_{i=1}^{z} TP_i + FN_i}$
F1-score	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	$2 \cdot \frac{\text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}$

Table 2
Definitions of the evaluation metrics used for both binary and multiclass classification: Accuracy, Precision, Recall, and F1 score (with micro average formulation for the multiclass case).

imbalanced data, balanced datasets obtained through undersampling and oversampling, and synthetic sample generation using SMOTE.

Table 3 reports accuracy, precision, recall, and F1-score reached by the encrypted CNN and plaintext CNN in binary task classification, considering different training setups.

As we can see, for the dataset Brain MRI Dataset, without data augmentation techniques, the Hybrid AHE-CNN achieves an accuracy of 0.90, with a value of 0.32 for precision, 0.33 for recall, and 0.31 for F1-score. Instead, for Plaintext CNN, it achieves values of 0.99, 0.33, 0.33, 0.33, for accuracy, precision, recall, and F1-score.

While they have achieved higher accuracy 0.90 and 0.99, they exhibit similarly low precision and recall, suggesting that both models suffer from class imbalance in the original dataset. This suggests that both models are overfitting to the majority class, and they fail to reflect their poor ability to detect the minority class.

When class imbalance is mitigated through undersampling, both models show similar performances. The Hybrid AHE-CNN achieves an accuracy value of 0.69, a precision of 0.73, a recall of 0.59, and a F1-score of 0.65, while the plaintext CNN achieved accuracy, precision, recall, and F1-score values of 0.68, 0.70, 0.68, and 0.67. These results demonstrate that both models perform better in classifying both classes. In particular, the CNN demonstrates significantly higher accuracy than the plain CNN, indicating a tendency to classify most instances as instances with neurological lesions and clinically relevant symptoms.

In the health domain, identifying diseases or lesions is a critical challenge, as misclassifications of an image can lead to a missed early diagnosis, with potentially severe consequences for the patient. Therefore, in case of ambiguity, it is more appropriate to identify the presence of a disease so that patients and clinicians can proceed with further diagnostic tests.

On the other hand, a model that tends to classify cases as non-pathological may reduce the number of false positives, but at the cost of misclassifying actual disease cases.

Concerning the oversampling technique applied to the dataset Brain MRI Dataset, as we can see, plaintext CNN achieves higher performances than the Hybrid AHE-CNN, reaching a value of 0.82 for accuracy, 0.84 for precision, 0.82 for recall, and 0.82 for F1-score. Instead, the Hybrid AHE-CNN achieved a value of 0.67 for accuracy, 0.64 for precision, 0.76 for recall, and 0.69 for F1-score. These results indicate that, although Hybrid AHE-CNN is less accurate, it achieves a high level of recall, suggesting that it is more effective in identifying lesions. However, this leads to an increase in false positives, which is shown by its lower precision. Although the plain CNN achieves higher accuracy and a better balance between precision and recall, its lower recall suggests that it may fail to detect some pathological cases.

With SMOTE augmentation, both CNNs exhibit higher performances and a more balanced trade-off between precision and recall. In particular, Hybrid AHE-CNN performs slightly better than the plain CNN, achieving values of 0.84 for accuracy, 0.80 for precision, 0.85 for recall, and 0.82 for F1-score. Similarly, the plain CNN achieved a value of accuracy of 0.82, a precision value of 0.81, a recall value of 0.82, and an F1-score of 0.82. This result suggests that SMOTE augmentation effectively mitigates class imbalance, allowing both encrypted and unencrypted models to generalise classification instances better.

Multiclass Classification. The multiclass classification task extends the binary setting by splitting patients into three distinct clinical severity levels, i.e., *Normal* EDSS scores from 0 to 2.0, *Mild*, i.e., EDSS scores between 2.5 and 4.0, and *Severe*, i.e., EDSS scores greater than 4.0. This task allows for a more detailed categorisation of neurological impairment, providing a classification framework that closely reflects clinical practice in multiple sclerosis assessment.

Dataset	Support	Hybrid AHE-CNN				Plaintext CNN			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Original	4189	0.90	0.32	0.33	0.31	0.99	0.33	0.33	0.33
Undersampling	823	0.69	0.73	0.59	0.65	0.68	0.70	0.68	0.67
Oversampling	3982	0.67	0.64	0.76	0.69	0.82	0.84	0.82	0.82
SMOTE	3982	0.84	0.80	0.85	0.82	0.82	0.81	0.82	0.82

Table 3
Performance comparison of the Hybrid AHE-CNN versus the plaintext CNN on the binary EDSS classification task, evaluated under original, undersampled, oversampled, and SMOTE-balanced training sets.

As shown in Table 4, the plaintext CNN generally outperforms the encrypted CNN in the original setting, reaching an accuracy of 0.64, with a precision of 0.67, but exhibiting lower recall and F1-score, both at 0.50. This indicates that while the plaintext model is better calibrated in terms of correct predictions overall, it may struggle to correctly identify all severity levels, especially the minority class. While the CNN achieved slightly lower precision, a value of 0.60, and accuracy, a value of 0.58, in the encrypted setting, it reached higher recall, a value of 0.58, and F1-score, a value of 0.57, compared to the plaintext CNN. This suggests that the encrypted model, despite achieving lower performance, may offer better sensitivity to minority classes, probably due to the conditioning implemented during the training phase of the encrypted model. These results highlight a trade-off between model precision and class sensitivity, which is particularly relevant in medical contexts where the cost of misclassifying serious cases can be high.

About the results achieved by the application of the undersampling approach, the encrypted Hybrid AHE-CNN outperforms the plaintext CNN across all metrics, achieving an accuracy of 0.38, a precision of 0.39, a recall of 0.38, and an F1-score of 0.36. Instead, the plaintext CNN exhibits lower performance, achieving a value of accuracy of 0.31, a precision of 0.32, a recall of 0.33, and an F1-score of 0.31, suggesting that the encrypted model is more robust to synthetic data than the plaintext CNN and benefits more from augmentation in imbalanced scenarios.

Regarding the oversampling augmentation, the Hybrid AHE-CNN performs slightly better than the plaintext CNN, achieving values for accuracy, precision, recall, and F1-score of 0.39, 0.40, 0.39, and 0.38. For the plaintext, it achieved for all metrics a value of 0.33.

Concerning the SMOTE-based balanced setting, the encrypted model achieves lower accuracy, 0.43, compared to the plaintext CNN, which reaches 0.49. However, the encrypted CNN outperforms the plaintext model in terms of precision, exhibiting a value of 0.44 and 0.39, and an F1-score of 0.42 and 0.33, respectively. These results indicate that, with increased SMOTE, the unencrypted CNN benefits most in terms of accuracy and recall, while the Hybrid AHE-CNN shows a more balanced compromise between precision and recall, leading to a higher F1-score. This suggests that the encrypted model may be better at handling synthetic samples in order to preserve class-level discrimination, especially in multi-class scenarios. Despite the improvements introduced by augmentation strategies, overall performance across all models remains below 50% for most metrics. This can be attributed to several factors: (i) the multiclass classification task is inherently more complex than the binary case, due to the presence of three clinically adjacent severity levels with overlapping EDSS score ranges, which can complicate classification between classes; (ii) the dataset exhibits significant class imbalance, particularly affecting minority classes, i.e. “Mild” and “Severe”, which restricts the model’s ability to learn representative features for all groups; (iii) oversampling and undersampling techniques may introduce synthetic artifacts or remove informative instances, reducing the quality of the training. Indeed, as we can see from the result achieved from the original dataset, both the Hybrid AHE-CNN and the Plaintext model achieved results equal to or close to 50%.

Dataset	Support	Hybrid AHE-CNN				Plaintext CNN			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Original	4189	0.52	0.49	0.52	0.49	0.64	0.67	0.51	0.50
Undersampling	885	0.38	0.39	0.38	0.36	0.31	0.32	0.33	0.31
Oversampling	2071	0.39	0.40	0.39	0.38	0.33	0.33	0.33	0.32
SMOTE	2071	0.43	0.44	0.43	0.42	0.49	0.39	0.49	0.33

Table 4
Performance comparison of the Hybrid AHE-CNN versus the plaintext CNN on the multiclass EDSS classification task, evaluated under original, undersampled, oversampled, and SMOTE-balanced training sets.

6. Discussion and Limitations

Privacy-preserving models are essential in the clinical domain to protect sensitive patient data, with homomorphic encryption (HE) techniques playing a key role due to their ability to perform computations directly on encrypted data. However, HE presents important limitations that impact its practical application. One major challenge is the significant computational overhead introduced by homomorphic operations. Our proposed model relies on the TenSEAL (CKKS) library, which causes notable latency during both encryption and decryption phases. Frequent serialization and deserialization of encrypted data further slow down training and inference.

Although the proposed model computes only one layer on encrypted data, the end-to-end execution time is significantly increased compared to a completely plaintext version. In addition, the proposed Hybrid AHE-CNN works on 2D MRI images and considers a limited set of clinical features provided with the dataset. Nevertheless, it has the potential to be adapted to larger sets of clinical data. Moreover, it is necessary to consider that the encryption of data can lead to introducing some approximation in the data due to the encryption schemes adopted. This can lead to an approximation in the classification results, due to minor rounding errors at each operation, yielding a slight degradation of the overall performance.

In a real-world clinical scenario, clinicians need precise responses and not overly expensive infrastructure. The encryption of clinical data requires competitive hardware requirements, making it challenging to integrate these solutions into existing clinical workflows without a secure and solid client-server infrastructure. Moreover, it is necessary to consider the complexity of integrating with legacy healthcare systems. In fact, many hospital information systems and electronic health records (EHRs) were not designed with privacy-preserving machine learning in mind and may lack the necessary interfaces to support encrypted computation workflows. This creates integration challenges that require technical adaptations for the definition of support systems that comply with strict healthcare regulations and data privacy laws, such as HIPAA and GDPR.

7. Conclusion and Future Directions

In this work, we addressed the EDSS classification tasks for Multiple Sclerosis patients while preserving the privacy of sensitive clinical data. To this end, we proposed Hybrid AHE-CNN, a convolutional neural network that integrates homomorphically encrypted patient metadata with MRI images through an encrypted weight conditioning mechanism. Unlike traditional models that rely solely on plaintext inputs, our Hybrid AHE-CNN enables secure, patient-specific inference without revealing raw clinical information. Experimental results demonstrate that the proposed hybrid model achieves classification performance comparable to that of conventional plaintext CNNs. These findings confirm the practical feasibility of applying homomorphic encryption in real-world clinical scenarios, showing that privacy preservation does not necessarily come at the cost of diagnostic accuracy.

In the future, we would like to investigate methods to improve the scalability and efficiency of our encrypted computation framework for practical deployment. This includes reducing computational overhead to handle more complex data, such as high-resolution 3D MRI volumes and deeper neural networks, by optimizing the homomorphic encryption pipeline. We will explore cryptographic enhancements like ciphertext packing, faster bootstrapping, and key-switching, alongside hardware acceleration using GPUs, FPGAs, or ASICs. Additionally, we plan to study algorithmic strategies such as low-precision and quantized networks to reduce encrypted computation complexity. Research on hybrid privacy-preserving techniques combining homomorphic encryption with other methods could also balance security and performance. Finally, we would like to simplify key management and integrate encrypted models into clinical workflows will be essential to enable real-time and privacy-preserving AI in healthcare.

Acknowledgments

D3 4 Health – Digital Driven Diagnostics, Prognostics and Therapeutics for Sustainable Health Care (Project PNC0000001 – CUP: B53C22006090001), funded by the European Union – NextGenerationEU under the National Plan for Complementary Investments to the NRRP.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. J. Leming, E. E. Bron, R. Bruffaerts, Y. Ou, J. E. Iglesias, R. L. Gollub, H. Im, Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting, NPJ Digital Medicine 6 (2023) 129.

[2] M. Anisetti, C. A. Ardagna, N. Bena, E. Damiani, P. G. Panero, Continuous management of machine learning-based application behavior, IEEE Transactions on Services Computing (2024).

[3] N. Bena, M. Anisetti, E. Damiani, C. Y. Yeun, C. A. Ardagna, Protecting machine learning from poisoning attacks: A risk-based approach, Computers & Security 155 (2025).

[4] A. S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on mri, Zeitschrift für Medizinische Physik 29 (2019) 102–127.

[5] M. Vázquez-Marrufo, E. Sarrias-Arrabal, M. García-Torres, R. Martín-Clemente, G. Izquierdo, A systematic review of the application of machine-learning algorithms in multiple sclerosis, Neurología (English Edition) 38 (2023) 577–590.

[6] F. La Rosa, E. S. Beck, J. Maranzano, R.-A. Todea, P. van Gelderen, J. A. de Zwart, N. J. Luciano, J. H. Duyn, J.-P. Thiran, C. Granziera, et al., Multiple sclerosis cortical lesion detection with deep learning at ultra-high-field mri, NMR in Biomedicine 35 (2022) e4730.

[7] S. Umirzakova, M. Shakhnoza, M. Sevara, T. K. Whangbo, Deep learning for multiple sclerosis lesion classification and stratification using mri, Computers in Biology and Medicine 192 (2025) 110078.

[8] F. De Marco, L. Di Biasi, A. A. Citarella, M. Tucci, G. Tortora, Identification of morphological patterns for the detection of premature ventricular contractions, in: 2022 26th International Conference Information Visualisation (IV), IEEE, 2022, pp. 393–398.

[9] G. A. Kaissis, M. R. Makowski, D. Rückert, R. F. Braren, Secure, privacy-preserving and federated machine learning in medical imaging, Nature Machine Intelligence 2 (2020) 305–311.

[10] F. De Marco, L. Di Biasi, A. Auriemma Citarella, G. Tortora, Improving pvc detection in ecg signals: A recurrent neural network approach, in: Italian Workshop on Artificial Life and Evolutionary Computation, Springer, 2023, pp. 256–267.

[11] M. J. Khan, B. Fang, G. Cimino, S. Cirillo, L. Yang, D. Zhao, Privacy-preserving artificial intelligence on edge devices: A homomorphic encryption approach, in: 2024 IEEE International Conference on Web Services (ICWS), IEEE, 2024, pp. 395–405.

[12] J.-W. Lee, H. Kang, Y. Lee, W. Choi, J. Eom, M. Deryabin, et al., Privacy-preserving machine learning with fully homomorphic encryption for deep neural network, IEEE Access (2022).

[13] J. Chao, A. A. Badawi, B. Unnikrishnan, J. Lin, C. F. Mun, J. M. Brown, J. P. Campbell, M. Chiang, J. Kalpathy-Cramer, V. R. Chandrasekhar, et al., Carenets: Compact and resource-efficient cnn for homomorphic inference on encrypted medical images, arXiv preprint arXiv:1901.10074 (2019).

[14] A. Vizitiu, C. I. Nită, A. Puiu, C. Suciu, L. M. Itu, Applying deep neural networks over homomorphic encrypted medical data, Computational and mathematical methods in medicine 2020 (2020) 3910250.

[15] O. L. Usman, R. C. Muniyandi, K. Omar, M. Mohamad, A. A. Owoade, M. A. Kareem, Horns-cnn model: an energy-efficient fully homomorphic residue number system convolutional neural network model for privacy-preserving classification of dyslexia neural-biomarkers, Brain Informatics 12 (2025) 11.

[16] K. Nguyen, M. Budzys, E. Frimpong, T. Khan, A. Michalas, A pervasive, efficient and private future: Realizing privacy-preserving machine learning through hybrid homomorphic encryption, in: 2024 IEEE Conference on Dependable, Autonomic and Secure Computing (DASC), IEEE, 2024, pp. 47–56.

[17] B. D. Rouhani, M. S. Riazi, F. Koushanfar, Deepsecure: Scalable provably-secure deep learning, in: Proceedings of the 55th annual design automation conference, 2018, pp. 1–6.

[18] E. Hesamifard, H. Takabi, M. Ghasemi, Cryptodl: towards deep learning over encrypted data, in: Annual Computer Security Applications Conference (ACSAC 2016), Los Angeles, California, USA, volume 11, 2016.

[19] C. Juvekar, V. Vaikuntanathan, A. Chandrakasan, Gazelle: A low latency framework for secure neural network inference, in: 27th USENIX security symposium (USENIX security 18), 2018, pp. 1651–1669.

[20] F. Boemer, Y. Lao, R. Cammarota, C. Wierzynski, ngraph-he: a graph compiler for deep learning on homomorphically encrypted data, in: Proceedings of the 16th ACM international conference on computing frontiers, 2019, pp. 3–13.

[21] A. Falcetta, M. Roveri, Privacy-preserving deep learning with homomorphic encryption: An introduction, IEEE Computational Intelligence Magazine 17 (2022) 14–25.

[22] A. M. Muslim, S. Mashohor, G. Al Gawwam, R. Mahmud, M. binti Hanafi, O. Alnuaimi, R. Josephine, A. D. Almutairi, Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information, Data in Brief 42 (2022) 108139.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.