Computer Vision for Schizophrenia Detection

1. Introduction

Medical imaging with artificial intelligence (AI) is a promising field of technology. A thorough comprehension of the fundamentals and applications of deep learning (DL), machine learning (ML), and magnetic resonance imaging (MRI) is essential to creating AI- based algorithms that are both excellently efficient and meet clinical diagnosis requirements. An individual's emotional, psychological, and social well-being are considered aspects of their mental health. It can be negatively impacted by a number of mental health issues, which also have an adverse effect on an individual's emotional state, social interactions, and mental capacity. It is estimated that 450 million people worldwide suffer from mental health disorders, which include depression, schizophrenia, attention-deficit hyperactivity disorder (ADHD), autism spectrum disorder (ASD), and other conditions.

The most severe psychological illness, schizophrenia (SZ), has a catastrophic impact on a patient's brain and day-to-day functioning. It results in anomalies in the brain's early development, which can cause a variety of symptoms like disorders, hallucinations, and issues with motivation and cognition. Although the exact cause of this neurological disorder is unknown, neuroscientists speculate that it may be primarily caused by the interaction of multiple environmental factors and genes. Because SZ is a highly heterogeneous mental disorder, it can be difficult to diagnose due to the lack of reliable biomarkers. A few clinical symptoms, such as physical, mental, and psychological indicators, must be assessed in order to diagnose SZ. A clinical examination consists of a number of tests, including medical imaging and blood tests. The patient may be referred to a psychiatrist, psychologist, or other specialist in the field if the doctors are unable to determine a physical explanation for the symptoms they believe are caused by SZ.

In our project we leveraged the power of Convolutional Neural Networks (CNNs) to detect Schizophrenia patients among Healthy individuals. We used a 2D CNN, 3D CNN and a time distributed 3D CNN.

2. Preprocessing and dataset preparation

2.1 Dataset description

For this project we will make use of the UCLA dataset, under the revision 1.0.5. [1] This dataset consists of 138 healthy individuals, 58 patients diagnosed with schizophrenia, 49 patients diagnosed with bipolar disorder and 45 patients diagnosed with ADHD. For the purpose of our project we will make use of the healthy controls (HC) and the 49 patients diagnosed with schizophrenia (SCHZ). The dataset consists of both functional MRI (fMRI) and anatomical MRI (sMRI) images. Under the revision 1.0.5, the samples have been processed with the fMRIPrep pipeline, a preprocessing pipeline for functional MRI (fMRI). [2]

The UCLA dataset consists of the preprocessed 4D fMRI in NIfTI format aligned to MNI152NLin2009cAsym template space (MNI152NLin2009cAsym _preproc.nii.gz), the brain mask file in NifTI format, which identifies the brain region in the 3D image (MNI152NLin2009cAsym _brainmask.nii.gz) and the tab-separated values (TSV) file which contains the confound regressors. These are time series data corresponding to various sources of noise and artifacts identified during preprocessing, such as motion parameters, physiological noise (e.g. heartbeat) and other sources of variability that can be regressed out during statistical analysis to improve the signal-to-noise ratio of the fMRI data.

In the whole dataset, the BOLD (Blood Oxygen Level Dependent) method was used to observe the brain activity. When neurons in the brain become more active, they consume more oxygen. The local response to this increased consumption is an increase in blood flow to that area, bringing more oxygenated hemoglobin. The fMRI machine can detect changes in the ratio of oxygenated to de-oxygenated hemoglobin, which allows researchers to infer neural activity in different regions of the brain over time.

All the patients undergone an MRI scan when simultaneously did different experimental tasks in order to trigger different parts of their brain and help researchers investigate various aspects of cognitive and neural functioning. The data consisted of 8 different tasks:

- **Rest** (resting state): During a resting-state fMRI scan, participants are asked to relax and not focus on any particular task. This allows researchers to study the brain's default mode network and

intrinsic connectivity, revealing how different brain regions communicate with each other when not engaged in a specific task.

- **BART** (Balloon Analogue Risk Task): This task assesses risk-taking behavior. Participants inflate a virtual balloon to increase potential rewards, knowing that the balloon could burst at any moment, resulting in the loss of accumulated rewards. It measures the balance between risk and reward.
- **SCAP** (Spatial Context Associative Processing Task): This task investigates spatial memory and associative processing, requiring participants to remember the spatial location of objects or associate spatial contexts with specific events or stimuli.
- **BHT** (Breath-Holding Task): This task is used to study cerebrovascular reactivity and the brain's response to changes in carbon dioxide levels. Participants are asked to hold their breath for short periods, allowing researchers to assess the brain's blood flow regulation.
- **Stopsignal** (Stop-Signal Task): This task measures response inhibition and the ability to control impulsive behavior. Participants are instructed to respond quickly to a go signal but must inhibit their response if a stop signal appears shortly after the go signal.
- **Pamenc** (Prospective Memory Encoding Task): Similar to the prospective memory retrieval task, this task focuses on the encoding phase of prospective memory. Participants encode intentions to perform specific actions in the future, which they must remember and execute later.
- **Pamret** (Prospective Memory Retrieval Task): This task evaluates prospective memory, which involves remembering to perform an intended action in the future. Participants are typically asked to remember to do something in response to specific cues while engaged in an ongoing activity.
- **Taskswitch** (Task Switching): This paradigm is used to study cognitive flexibility and the ability to switch between different tasks or mental sets. Participants typically alternate between two or more tasks, requiring them to switch their cognitive focus and adapt to new rules or conditions.

The confounds used in the dataset are variables that represent sources of noise or artifacts in fMRI data. They are used in the preprocessing

and analysis of fMRI data to improve the accuracy and reliability of the results. The confounds that we used in our analysis are the following:

- **White Matter**: This confound represents the average signal from white matter regions of the brain. White matter does not typically show the same level of activation changes as gray matter. By including this confound, researchers can account for and remove non-neuronal signal fluctuations originating from white matter.
- **Global Signal**: This confound represents the average signal across the entire brain. Global signal regression can help to remove widespread fluctuations that may not be related to specific brain activity (e.g., due to respiration or scanner drifts).
- **Framewise Displacement (FD)**: Framewise displacement is a measure of the head movement from one frame (time point) to the next during the scan. High FD values indicate significant head motion, which can introduce artifacts into the data. FD is often used to identify and possibly exclude time points with excessive motion from the analysis.
- **X, Y, Z**: These confounds represent the translational movements of the head in the three spatial dimensions (left-right, anterior-posterior, and superior-inferior). By including these motion parameters as regressors, researchers can account for and remove the effects of head motion on the fMRI signal.
- **RotX, RotY, RotZ**: These confounds represent the rotational movements of the head around the three axes (pitch, yaw, and roll). Similar to the translational motion parameters, these are included to correct for the effects of rotational head motion on the fMRI data.

2.2 Preprocessing procedure

The preprocessing procedure we followed in this endeavor is the following. First we downloaded the already processed fMRI dataset. The raw UCLA dataset was processed using the fMRIPrep pipeline and were made available for download. Then we applied a straightforward procedure on the processed 4D data. We took the brain mask images and applied them on the 4D processed images and then we regressed out the confounds, found in the TSV file.

The brain mask is a 3D array that identifies the voxels within the brain by marking them with ones, while non-brain voxels are marked with

zeros. This mask is multiplied element-wise with each time point in the 4D fMRI data (spatial dimensions x, y, z and temporal dimension t), effectively setting all non-brain voxels to zero across all time points. This operation ensures that subsequent analyses focus exclusively on brain regions, thereby improving the signal-to-noise ratio by eliminating irrelevant signals from outside the brain.

Following this, we perform confound regression to remove noise and artifacts that can obscure true neural signals. This step involves loading confound variables such as head motion parameters (translations and rotations), white matter signal, global signal, and framewise displacement from a confound file. These confounds are structured as time series data and are used to build a regression model. For each voxel in the brain (each spatial coordinate i, j, k), we fit a linear regression model using the confound time series to predict the voxel's fMRI signal over time. The predicted signal, which represents the portion of the voxel's signal attributable to these confounds, is then subtracted from the original fMRI signal at that voxel. This process is repeated for every voxel, resulting in a cleaned 4D fMRI dataset where the temporal dimension now reflects neural activity with reduced influence from motion, physiological noise, and other confounding factors.

By applying the brain mask and regressing out confounds, we produce a refined fMRI dataset that retains its original 4D structure but is now more representative of true brain activity. This preprocessing is crucial for minimizing artifacts and enhancing the validity of subsequent neuroimaging analyses.

Below the preprocessed 4D fMRI image of a schizophrenia patient and a healthy individual are being illustrated alongside with the same image after we have applied the brain mask and regressed out the confounds. In order to correctly visualize the image, we have set the time point to 0 visualizing the image as 3D.

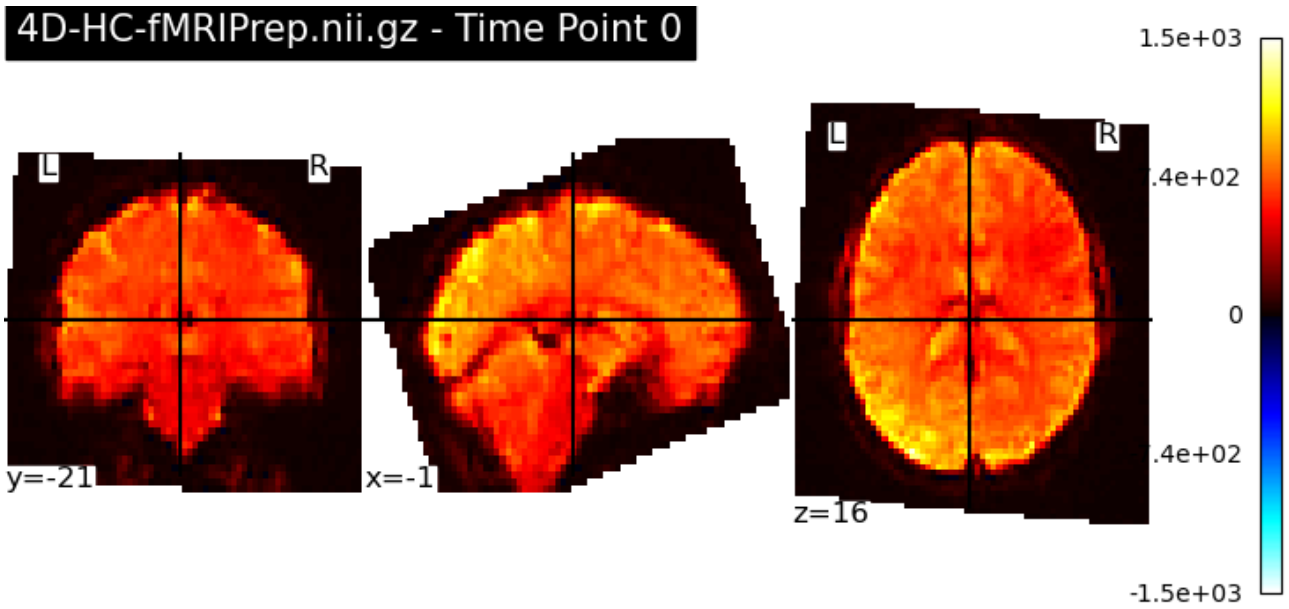


Figure 1: Processed 4D HC image

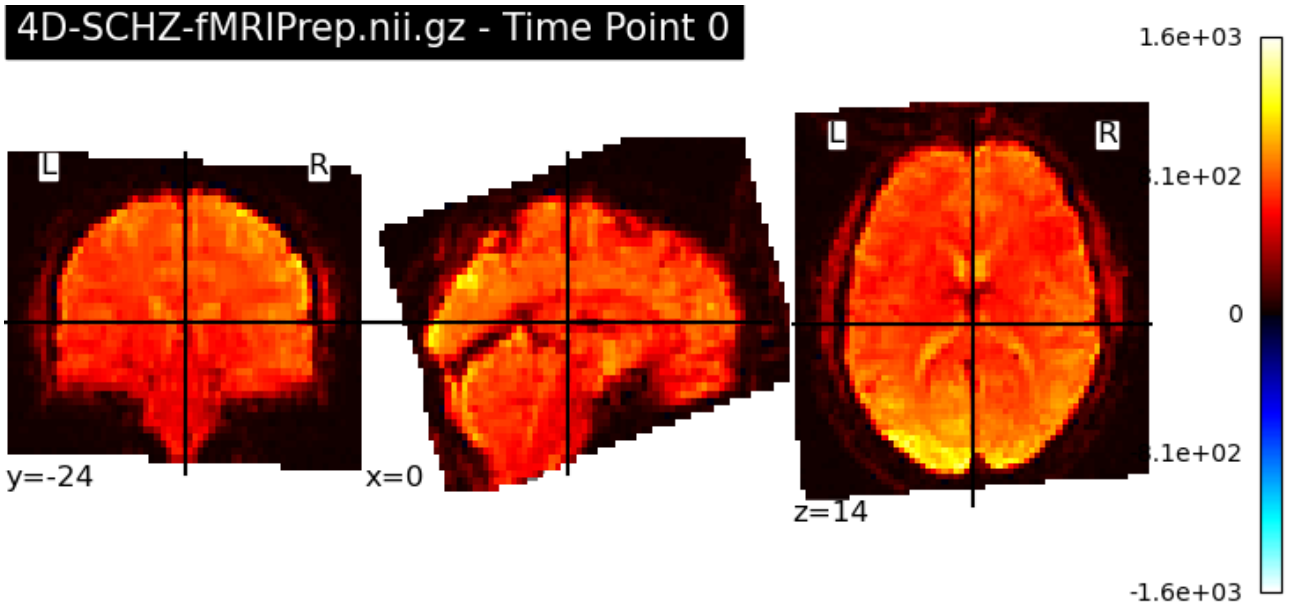


Figure 2: Processed 4D SCHZ image

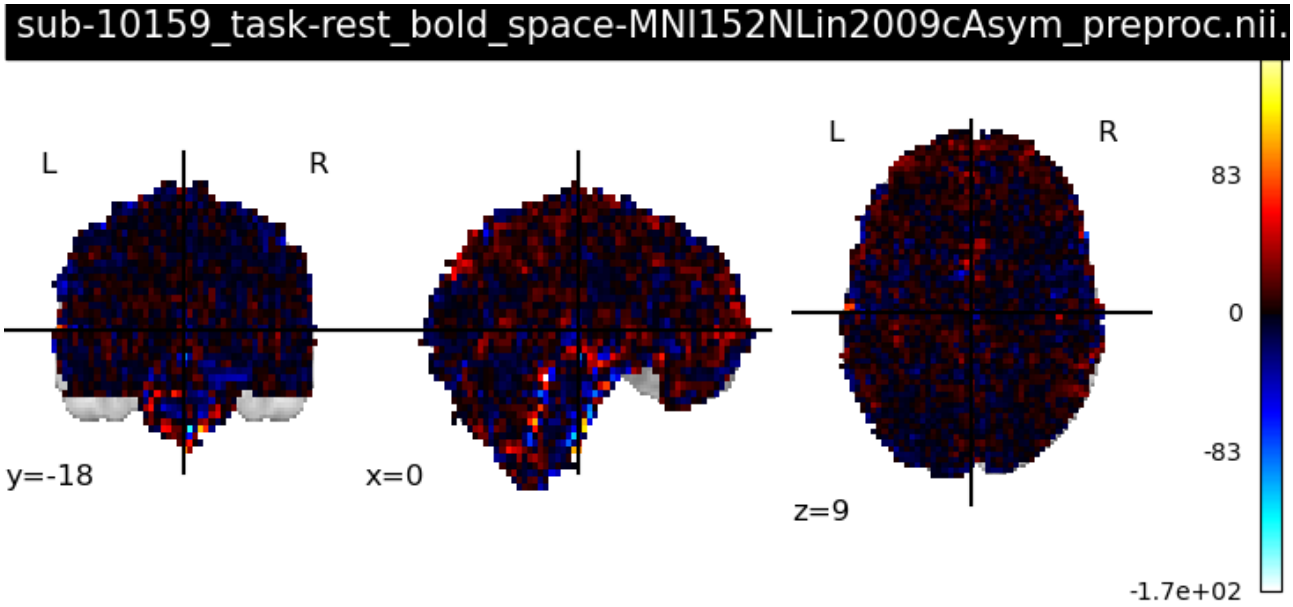


Figure 3: Processed 4D HC image_brain mask and confounds

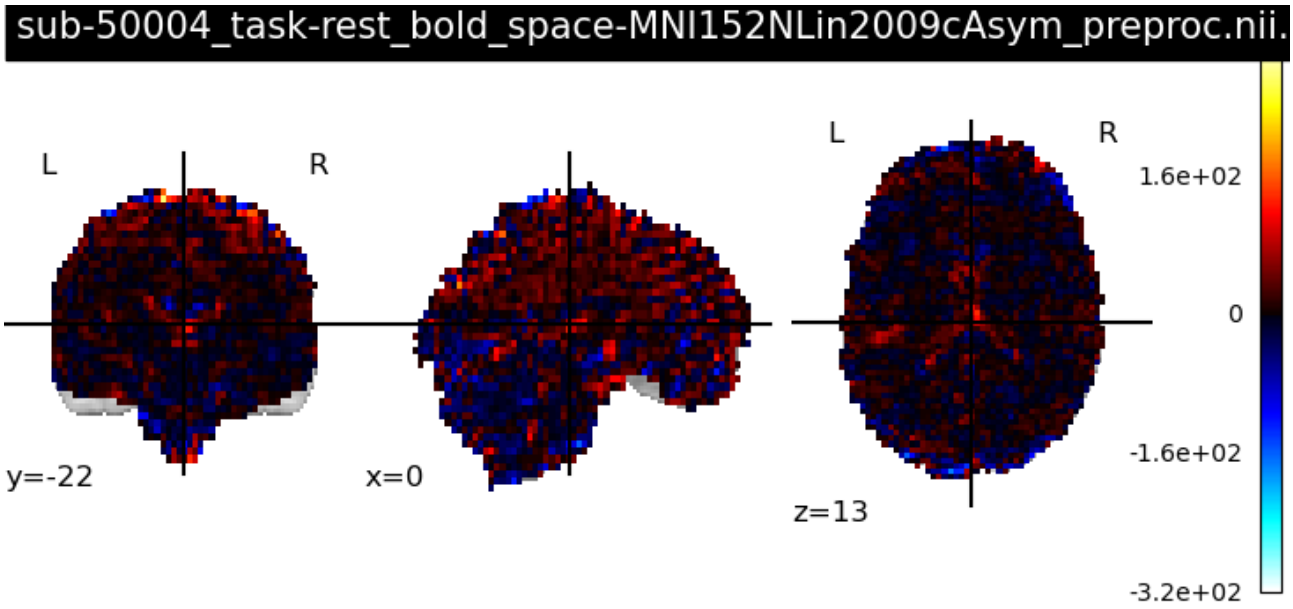


Figure 4: Processed 4D SCHZ image brain mask_confounds

3. Convolutional Neural Networks

3.1 2D CNNs

To convert fMRI data from 4D to 3D and then to 2D, the process begins by loading the original 4D data, which comprises three spatial dimensions (x, y, z) and one temporal dimension (t). The first step in this

transformation involves collapsing the temporal dimension to convert the 4D data into a 3D format. This is achieved by calculating the mean signal intensity across all time points for each voxel, resulting in a 3D volume where each voxel (x, y, z) represents the average signal over time. This 3D data encapsulates the spatial information of the brain's activity while eliminating the temporal component. Next, the 3D data is saved as a new NIfTI file to preserve this intermediate form for further analysis. Following this, the 3D volume is converted into a 2D image by averaging the signal intensities along one of the spatial dimensions, typically the z-axis. This process reduces the data to two dimensions (x, y), where each pixel represents the average signal across the selected spatial axis. The resulting 2D image is then saved as a PNG file, facilitating easy visualization and inspection. This stepwise reduction from 4D to 3D to 2D allows for the complex, multidimensional fMRI data to be simplified progressively, making it more manageable and interpretable while retaining the essential information at each stage.

The converted 2D images for the healthy individuals and schizophrenia patients are illustrated below:

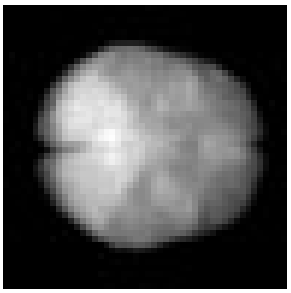


Figure 5: 2D HC Image

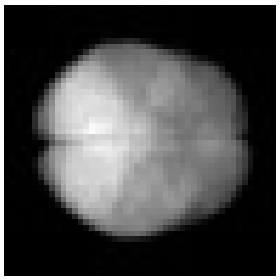


Figure 6: 2D SCHZ Image

After converting all 4D images to 2D images, we applied a straightforward method to accurately train and evaluate a 2D CNN. Firstly, we noticed a class imbalance in the data. To address this issue, a

custom augmentation approach similar to Synthetic Minority Over-sampling Technique (SMOTE) is used. Instead of creating synthetic samples, the script augments existing images of the minority class (SCHZ) by applying random rotations and shifts. This process increases the number of minority class samples, balancing the dataset and improving the model's ability to learn from both classes effectively.

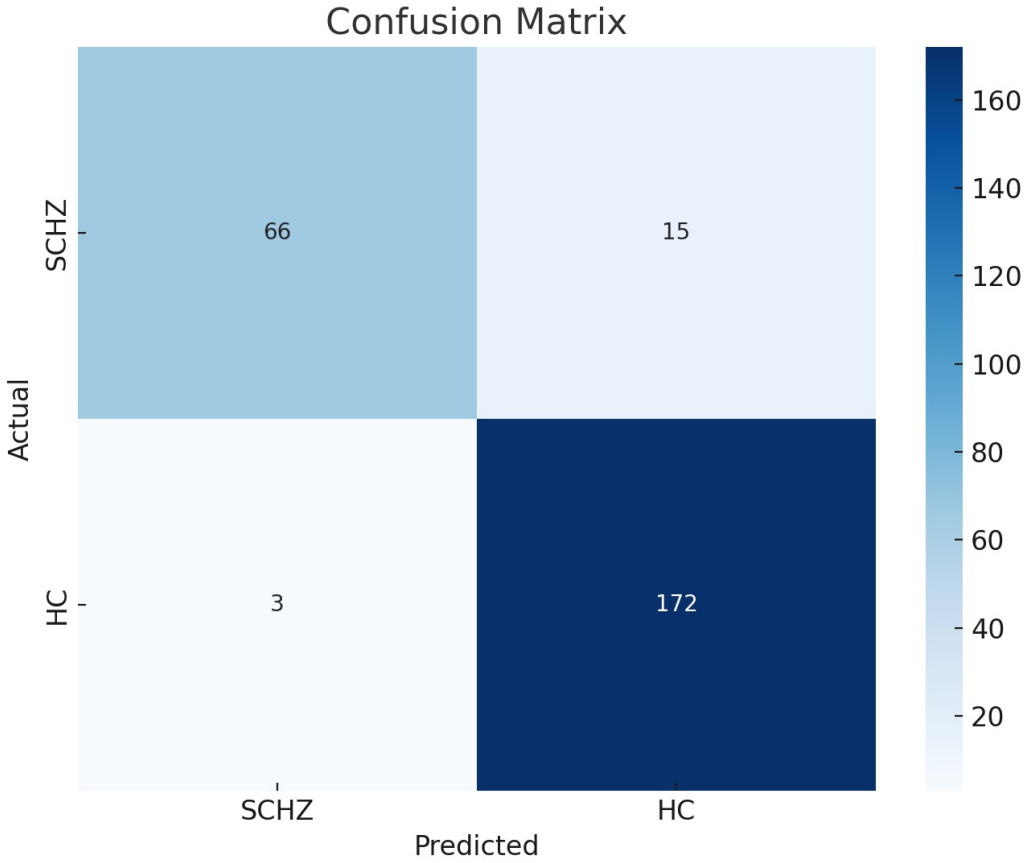
The architecture of the 2D CNN we used is minimal and straightforward, focusing on extracting essential features from the 2D images. It starts with an input layer that accepts images with a shape of 48x48 pixels and a single channel (gray-scale). The first layer is a convolutional layer with 32 filters, a kernel size of 3x3, ReLU activation, and same padding, which helps in learning spatial hierarchies in the image data. Following the convolutional layer, a max pooling layer with a pool size of 2x2 is used to down sample the feature maps, reducing dimensionality and computation while retaining important features. The feature maps are then flattened into a 1D vector, which is passed to a dense layer with softmax activation to produce the final class probabilities.

The model is trained using the Adam optimizer with a learning rate of 0.001 and categorical cross-entropy loss. To prevent over-fitting, early stopping is applied, monitoring the validation loss and restoring the best weights. A model checkpoint saves the best model during training based on validation performance. The training and validation data are fed into the model using TensorFlow data generators, which yield batches of images and labels, applying augmentations dynamically during training. After training, the best model is evaluated on the validation set, and performance metrics such as accuracy, confusion matrix, and classification report are calculated. The results, including training history and confusion matrix, are visualized and saved for further analysis. This setup ensures that the model can effectively learn to classify 2D fMRI images, even in the presence of class imbalance, by leveraging data augmentation and a simple yet effective CNN architecture.

The classification report and the confusion matrix after the 2D CNN evaluation are illustrated below:

Classification Report:

	precision	recall	f1-score	support
SCHZ	0.96	0.81	0.88	81
HC	0.92	0.98	0.95	175
accuracy			0.93	256
macro avg	0.94	0.90	0.92	256
weighted avg	0.93	0.93	0.93	256



The 2D CNN achieved an accuracy of 93% and a confusion matrix of:

True Positives (**TP**): 172 – Correctly predicted HC

True Negatives (**TN**): 66 – Correctly predicted SCHZ

False Positives (**FP**): 15 – Incorrectly predicted as HC

False Negatives (**FN**): 3 – Incorrectly predicted as SCHZ

3.2 3D CNNs

To convert fMRI data from 4D to 3D, the process begins by loading the original 4D data, which comprises three spatial dimensions (x, y, z) and one temporal dimension (t). The first step in this transformation involves collapsing the temporal dimension to convert the 4D data into a 3D format. This is achieved by calculating the mean signal intensity across all time points for each voxel, resulting in a 3D volume where each voxel (x, y, z) represents the average signal over time. This 3D data encapsulates the spatial information of the brain's activity while eliminating the temporal component. Next, the 3D data is saved as a new NIfTI file to preserve this intermediate form for further analysis.

The converted 3D images for the healthy individuals and schizophrenia patients with applied brain mask and regressed out confounds are illustrated in Figures 3 and 4.

After converting all 4D images to 3D, we applied a straightforward procedure to accurately train and evaluate a 3D CNN. Firstly, we noticed a class imbalance in the data. To address this issue, a custom augmentation approach similar to Synthetic Minority Over-sampling Technique (SMOTE) is used. Instead of creating synthetic samples, the script augments existing images of the minority class (SCHZ) by applying random rotations and shifts. This process increases the number of minority class samples, balancing the dataset and improving the model's ability to learn from both classes effectively.

For model creation, the architecture of the CNN begins with an input layer that accepts 3D volumes with dimensions 65x77x49 and a single channel (gray-scale). The first block of the model consists of a 3D convolutional layer with 32 filters, followed by batch normalization to standardize the output, a ReLU activation function to introduce non-linearity, and a max-pooling layer to reduce the spatial dimensions. This structure is repeated in a second block with 64 filters. The output from the convolutional layers is then flattened into a 1D vector, passed through a fully connected layer with 256 units and ReLU activation, and subjected to dropout to prevent overfitting. The final layer is a dense layer with 2 units and a softmax activation function for binary classification.

During training and evaluation, the data is split into training and validation sets using K-Fold cross-validation. The data is converted into TensorFlow datasets and fed into the model for training. Early stopping

is employed to halt training if the model performance stops improving, and model checkpoints save the best-performing model. After training each fold, the model's performance is evaluated on the validation set, calculating metrics such as accuracy, precision, recall, F1 score, and confusion matrices. These metrics are averaged across all folds to provide a comprehensive performance assessment. Additionally, visualizations of the training process and confusion matrices are generated and saved.

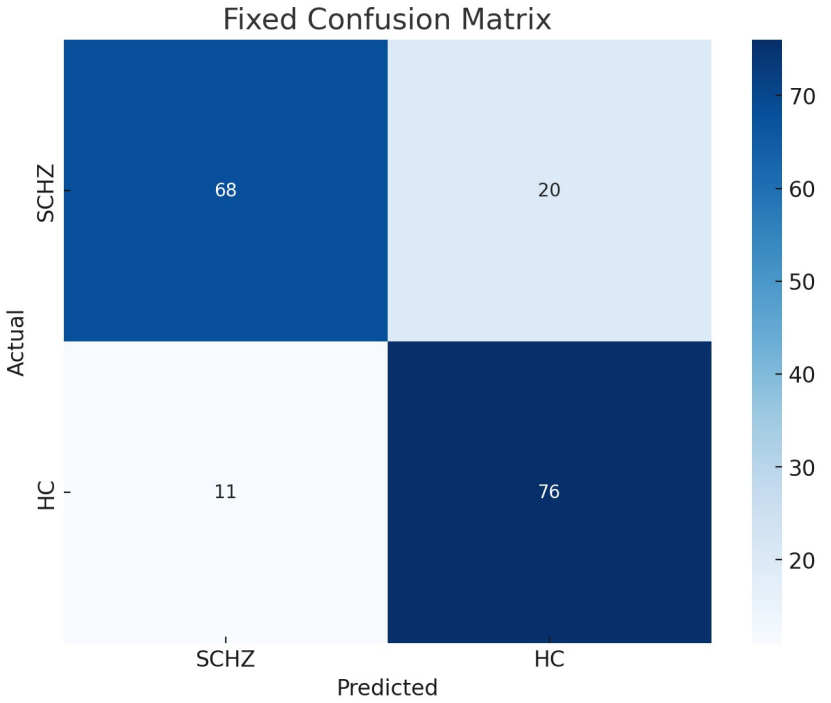
In summary, the full model architecture starts with an input layer for 3D volumes, followed by two blocks of convolutional, batch normalization, activation, and pooling layers. This is followed by a flattening layer, a fully connected layer with dropout, and a final output layer with softmax activation. The architecture is compiled with the Adam optimizer and categorical cross-entropy loss, and it is trained on augmented and balanced datasets, with performance metrics evaluated and visualized for comprehensive assessment.

The classification report and the confusion matrix after the 3D CNN evaluation are illustrated below:

Average Classification Report:

Label	Precision	Recall	F1-Score	Support
SCHZ	0.87	0.78	0.81	882
HC	0.80	0.87	0.83	878
Accuracy			0.82	
macro avg	0.84	0.82	0.82	1760
weighted avg	0.84	0.82	0.82	1760

Number of files after SMOTE: 1776
Number of files used for testing: 177



The 3D CNN achieved an accuracy of 82% and a confusion matrix of:

True Positives (**TP**): 76 – Correctly predicted HC

True Negatives (**TN**): 68 – Correctly predicted SCHZ

False Positives (**FP**): 20 – Incorrectly predicted as HC

False Negatives (**FN**): 11 – Incorrectly predicted as SCHZ

3.3 Time-distributed 3D CNN

For the time-distributed 3D CNN model, we didn’t apply any dimension modification.

The converted 3D images for the healthy individuals and schizophrenia patients with applied brain mask and regressed out confounds are illustrated in Figures 3 and 4.

Firstly, we noticed a class imbalance in the data. To adress this issue we applied data augmentation techniques. The augmentation involves rotations and zooming to the fMRI volumes. Rotations are performed at angles of 5, -5, 10, and -10 degrees along the spatial axes, creating variability in the orientation of the data. Zooming is applied with factors of 0.9 and 1.1, effectively scaling the data in and out. To ensure uniformity in the time dimension, the augmented data is padded or trimmed to maintain a consistent length of 300 time points. This padding is achieved by adding zeros to the end of the time dimension if the augmented data is shorter, or by trimming the excess if it is longer. The augmented and padded data samples are then saved as new NIfTI files, ensuring a well-prepared dataset with a balanced number of samples for each class. This comprehensive augmentation and padding process enhances the robustness of the dataset, making it suitable for training a neural network model.

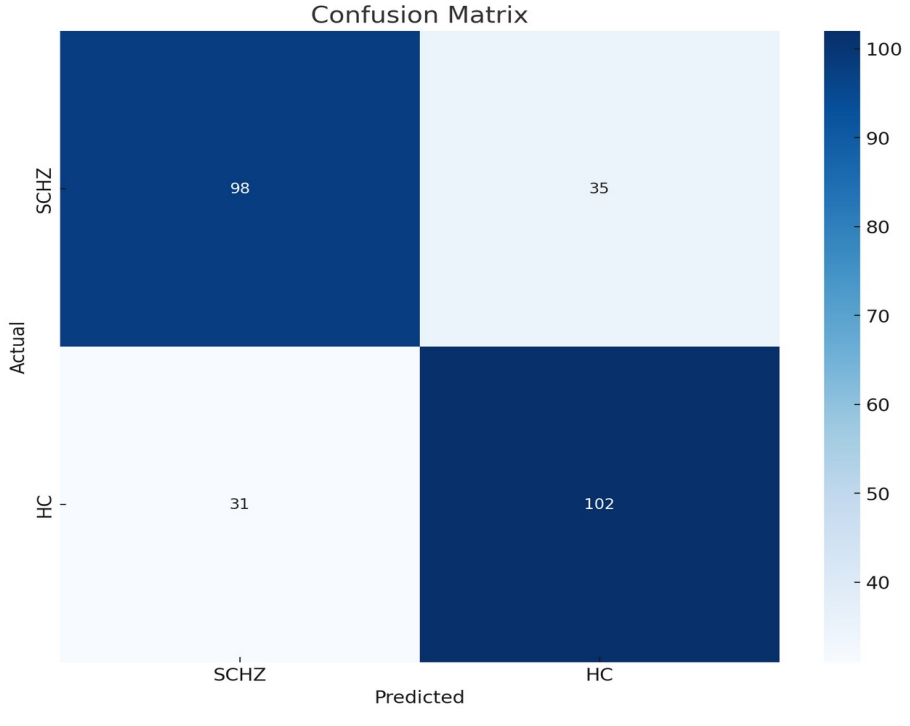
As far as the model is concerned, we implemented a a Time-Distributed 3D Convolutional Neural Network (CNN) for classifying 4D fMRI data, beginning with a setup that enables mixed precision to improve computational efficiency and ensures GPU memory is allocated dynamically. The data, sourced from specified directories, is loaded in NIfTI format, normalized, and reshaped to add a channel dimension, with the time dimension moved to the first position to accommodate TimeDistributed layers. The CNN architecture starts with an input layer designed to handle data of shape (300, 65, 77, 49, 1), representing 300

time points, the three spatial dimensions and one channel for gray-scale images. The network includes two TimeDistributed convolutional blocks: the first applies a 3D convolution with 8 filters, followed by ReLU activation and max-pooling, while the second uses 16 filters with similar operations. These layers extract spatial features independently for each time point. After the convolutional blocks, a TimeDistributed global average pooling layer summarizes the spatial information, and the output is flattened into a 1D vector. This vector is then passed through a dense layer with softmax activation, producing the final classification probabilities. The model is compiled with the Adam optimizer and categorical cross-entropy loss. Training involves creating TensorFlow datasets from the data, setting up a model checkpoint to save the best model based on validation loss, and running the training process while tracking accuracy and loss metrics, which are visualized and saved for analysis. This architecture and training process ensure a robust and efficient model for classifying fMRI data into SCHZ and HC categories.

The classification report and the confusion matrix after the time distributed 4D CNN evaluation are illustrated below:

Classification Report:

	precision	recall	f1-score	support
SCHZ	0.76	0.74	0.75	133
HC	0.74	0.77	0.76	133
accuracy			0.75	266
macro avg	0.75	0.75	0.75	266
weighted avg	0.75	0.75	0.75	266



The time-distributed 3D CNN achieved an accuracy of 75% and a confusion matrix of:

True Positives (**TP**): 102 – The number of HC cases correctly classified as HC.

True Negatives (**TN**): 98 – The number of SCHZ (Schizophrenia) cases correctly classified as SCHZ.

False Positives (**FP**): 35 – The number of SCHZ cases incorrectly classified as HC (Healthy Controls).

False Negatives (**FN**): 31 – The number of HC cases incorrectly classified as SCHZ.

4. Results & discussion

This project highlighted the importance and the power of Computer Vision models that play a crucial role in healthcare field. We developed and implemented a variety of Convolutional Neural Networks (CNNs) in order to correctly predict schizophrenia among healthy individuals. These models can detect subtle patterns and anomalies in brain activity that might not be visible to human experts, enabling earlier and more accurate diagnosis of schizophrenia. Early detection is vital as it allows for timely intervention and treatment, which can significantly improve patient outcomes. Moreover, such models contribute to the broader healthcare field by enhancing diagnostic precision, reducing human error, and enabling personalized treatment plans. Below we present a comparison of our models and their results, with other results in the literature.

	Reference	Dataset	Modalities	Method	Performance (Accuracy %)
2D scans					
	Hu et al	NUSDAST	sMRI	2D CNNs	72.41
	Li et al.	Clinical	sMRI	2D CNNs	99.72
	Our work	**UCLA**	**fMRI**	**2D CNNs**	**92.9**
3D scans					
	Hu et al	Nusdast	sMRI	3D CNNs	79.27
	Hu et al	IMH	sMRI	3D CNNs	70.98
	Qureshi et al	COBRE	fMRI	3D CNNs	98.09
	Pominova	OpenNeuro	fMRI	3D CNNs	82.3
	Our work	**UCLA**	**fMRI**	**3D CNNs**	**82.3**
4D scans					
	Our work	**UCLA**	**fMRI**	**4D CNNs**	**75**

References

[1]: https://www.nature.com/articles/sdata2016110#article-info

[2]: https://www.nature.com/articles/s41592-018-0235-4