UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA

Master's degree in Biomedical Engineering

Alzheimer's disease early diagnosis using convolutional neural networks from volumetric scans

Supervisor:
Prof. Laura Burattini

Thesis by:
Giada Bernardi

Co-supervisors:
Dr. Selene Tomassini
Dr. Agnese Sbrollini

Academic Year 2021 / 2022

Abstract

Nowadays, Alzheimer's disease is becoming a major public health issue worldwide. Thus, a higher knowledge yet early diagnosis of it may be fundamental to both slow down the development of symptoms and enable better therapeutic interventions.

Clinically, neuroimaging techniques such as magnetic resonance imaging are available for the Alzheimer's disease diagnosis. Through magnetic resonance imaging, volumetric scans can be obtained, helping in the detection of structural abnormalities and tracking of the evolution of brain atrophy.

Deep learning algorithms for Alzheimer's disease diagnosis applied to volumetric scans are increasingly used in medical field, but they are still not trusted by clinicians because they lack interpretability.

This thesis has a dual purpose. Firstly, it has been performed a review of state-of-the-art studies which applied interpretability algorithms for Alzheimer's disease diagnosis in order to understand the current trends. Then, it has been conducted an analysis of volumetric magnetic resonance scans by exploiting two convolutional neural networks and comparing their performance: a pre-trained 3D convolutional neural network (C3DKeras) and an end-to-end time-distributed one. To fulfil the first task, a descriptive literature review has been performed, whereas for the second one, a Python-based implementation was conducted.

According to the literature outcomes, there is still uncertainty concerning the best interpretability technique to be applied for Alzheimer's disease diagnosis, even though attribution map approaches seem to produce the most coherent interpretations. For what concerns convolutional neural networks for volumetric data processing, the end-to-end time-distributed one resulted to be the best approach because of its higher performance and lower computational cost.

A future development of this thesis could be the addition of an interpretability module to the end-to-end time-distributed convolutional neural network in order to make a step forward in the direction of an interpretable Alzheimer's disease diagnosis.

A voi, nonni miei, che mi state guardando da lassù

A mamma, a papà e a mio fratello

Index

Introduction I

1. The nervous system 1

1.1 Basic structure of the nervous system 1

1.2 Function of the nervous system 4

1.3 The action potential 4

1.4 The central nervous system: the brain 9

1.4.1 The cerebrum 10

1.4.1.1 The memory 12

1.4.1.2 Subcortical structures: the hippocampus 13

2. Alzheimer's disease 16

2.1 Neuropathology of Alzheimer's disease 17

2.2 The stages of Alzheimer's disease 21

2.3 Diagnosis 22

3. Image basics and bioimages 25

3.1 Biomedical imaging 26

3.1.1 Magnetic resonance imaging 27

4. The importance of interpretable Alzheimer's disease diagnosis 29

4.1 Most used interpretability techniques applied to deep learning algorithms 30

4.1.1 Concept learning model 31

4.1.2 Case-based model 33

4.1.3 Counterfactual explanation 34

4.1.4 Concept attribution 35

4.1.5 Attribution map 36

4.2 Most used interpretability techniques applied to Alzheimer's disease 41

4.2.1 Literature review 42

4.2.1.1 Shahamat et al., 2020 45

4.2.1.2 Dyrba et al., 2020 ... 48
4.2.1.3 Guan et al., 2021 ... 49
4.2.1.4 Turkan et al., 2021 ... 50
4.2.2 Discussion ... 51
5. Convolutional neural networks applied on volumetric magnetic resonance scans ... 55
5.1 Data and methodology ... 55
5.1.1 Data selection ... 55
5.1.2 Environmental setup ... 55
5.1.3 3D convolutional neural network ... 56
5.1.3.1 Pre-trained C3DKeras ... 56
5.1.3.2 Data preparation ... 58
5.1.3.3 Neural network classification ... 59
5.1.4 Time-distributed convolutional neural network ... 59
5.1.4.1 End-to-end VGG16 + ConvLSTM ... 59
5.1.4.2 Data preparation ... 61
5.1.4.3 Neural network classification ... 61
5.1.5 Neural network evaluation ... 61
5.2 Results ... 62
5.2.1 3D convolutional neural network ... 62
5.2.2 Time-distributed convolutional neural network ... 63
5.3 Discussion ... 64
Conclusion ... III
Ringraziamenti ... IV
References ... VI

Introduction

Nowadays, with population aging, neurodegenerative diseases such as Alzheimer's disease (AD) are becoming a major public health issue worldwide. However, at this point of life, there is still lack of understanding of AD by patients and their families, mostly missing the optimal intervention stage. Hence, a better characterisation of this disease and early diagnosis of AD might be fundamental in order to allow a better management of patients and to slow down the development of AD. In fact, while there is no cure for AD, early diagnosis and accurate prognosis may enable therapeutic interventions that strive to improve symptoms, or at least slow down mental deterioration, thereby improving the quality of life.

Clinically, there are different forms of neuroimaging techniques available for AD diagnosis, including magnetic resonance imaging (MRI), from which volumetric data can be obtained and it is considered as a marker of AD progression since it can help to detect the structural abnormalities and track the evolution of brain atrophy. However, at present, the AD identification process is still performed manually by specialists in clinical practice, which is expensive and time-consuming. To solve this issue, thanks to the rapid development and application of artificial intelligence in the medical field, computer-aided diagnosis of AD using neuroimaging may be an auxiliary method to assist physicians in the clinical decision-making. In fact, several attempts based on deep learning techniques have been employed to analyse the MRI data by constructing models in order to avoid manually extracting features and deep learning methods have proved to be effective in the feature extraction from images. Although models based on deep learning have achieved great classification performance for AD diagnosis, have yet to achieve full integration into clinical practice mainly because deep learning models are 'black-box' algorithms, meaning that they still lack interpretability, which is a fundamental aspect. Nowadays, much attention has been given in order to try to solve this issue and so different interpretability algorithms have been proposed in literature. However, since it is a novelty, only a few studies have been published, especially regarding 3D applications which are the ones of relevance for this work. Nevertheless, still there appears to be confusion about which the most accurate and reliable interpretability technique is for an AD interpretable diagnosis.

This thesis has a dual purpose because it has been firstly performed an analysis of the studies already published in literature which apply interpretability algorithms to AD diagnosis, in order to understand the state-of-the-art. On the other hand, an analysis of the volumetric data

(MRI images) with two different convolutional neural networks has been conducted to highlight which the best approach is. In particular, these two different approaches are the implementation of a 3D convolutional neural network and of a time-distributed convolutional neural network.

In the following, the reader will find some chapters to be introduced to the main notions necessary to understand this research topic. In order to give some information about the physiology of the part of the body and the pathology of interest for this work, the first two chapters are dedicated to the nervous system (Chapter 1) and the Alzheimer's disease (Chapter 2). In Chapter 3, there are some details related to image basics and bioimages, whereas Chapter 4 is about the importance of interpretable Alzheimer's disease diagnosis. Then the reader will receive more in-depth information about the current situation in literature. In fact, a literature review was performed and presented in Chapter 5 in order to understand the state-of-the-art of the interpretability of deep neural networks for AD early diagnosis. Chapter 6 reports the details relative to the research part of this work. In fact, there is information about the dataset used, the environmental setup, it can be found the description of the two different convolutional neural network analysed and their evaluation. In Chapter 7, instead, all the results related to the two different neural networks are reported, which are then discussed in Chapter 8 in order to give an answer to the aim of this work, which is then underlined in the conclusion in Chapter 9.

1. The nervous system

The nervous system is a very complex organ system. In Peter D. Kramer's book *Listening to Prozac*, a pharmaceutical researcher is quoted as saying, "If the human brain were simple enough for us to understand, we would be too simple to understand it" (1994) [1].
In order to understand the structure of the nervous system, it is good to start with the large divisions and work through to a more in-depth understanding. However, for the purpose of all this work, only the main concepts of interest will be treated.

1.1 Basic structure of the nervous system

In order to make this complex organ system easily understandable, biologists have divided the nervous system as whole into two large portions: the Central Nervous System (CNS) and the Peripheral Nervous System (PNS). The former includes the brain and the spinal cord, while the latter is made up of the nerve tissues located in the periphery of the CNS [2], meaning beyond the brain and spinal cord (Figure 1). However, it is important to note that the concept of CNS (or *neuroaxis*) as a separate entity from the PNS is a purely didactic distinction, since the PNS consists mainly of the extensions of the nerve cells that are part of the CNS [3].

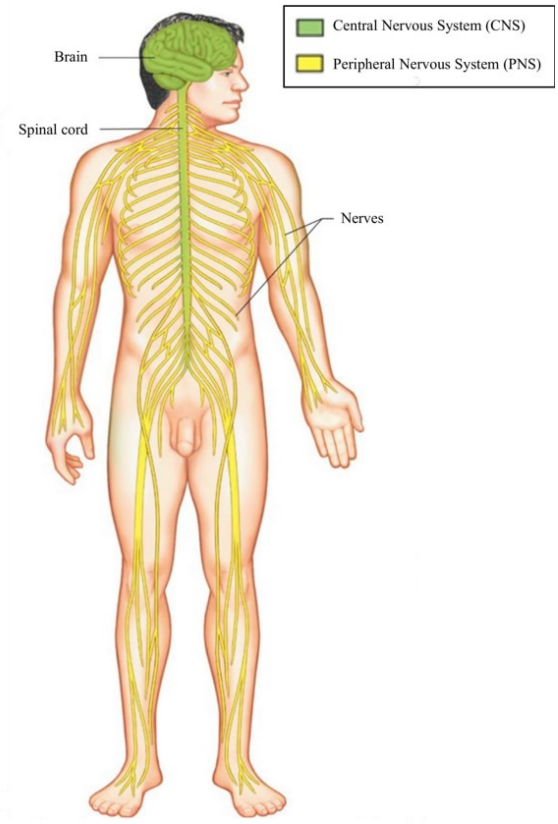


Figure 1. ***Nervous system*** *- The major anatomical structures of the human nervous system include the brain, spinal cord, and each of the individual nerves. The brain and spinal cord form the central nervous system (CNS), while all the nerves with their branches make up the peripheral nervous system (PNS)* [2].

Furthermore, nervous tissue, present in both the CNS and PNS, contains two basic types of cells: neurons and glia cells. A glial cell is one of a variety of cells that provide a framework of tissue that supports the neurons and their activities, and they usually do not conduct information [2], while the neuron is the more functionally important of the two, in terms of the communicative function of the nervous system [1].

Before describing the functional divisions of the nervous system, it is fundamental to know the structure of the neuron. The structure of a typical neuron can generally be divided into four distinct domains: the cell body, the dendrites, the axon and the presynaptic terminals (Figure 2) [5]. Since neurons are cells, they are formed by a cell body (also called *perikaryon*, or *soma*) [2, 4], which is much like that of other cells as it contains the nucleus [2, 6], but they also have extensions of the cell; each extension is generally referred to as a process [1] and there are at least two processes: an axon (or *neurite* or *cylindrass* [4]) and one or more dendrites [2]. The axon is perhaps the most remarkable feature of every neuron, and it arises from the cell body, like the dendrites, and its point of origin is a tapered region known as the axon hillock and, just distal to the cone-shaped hillock, is an untapered, unmyelinated region known as the initial segment [5]. Axon may also extend very much (even more than a meter) and it is the message-sending portion of the neuron as it is the fibre that connects a neuron with its target, such as another neuron or a muscle [1, 5]. Some axons have a special electrical insulation, called myelin, that consists of the coiled cell membranes of glial cells that wrap themselves around the nerve axon making the action potential jump from one node of Ranvier (the space between adjacent myelin segments) to another in a process called saltatory conduction, which makes the conduction faster with respect to the case of an axon not covered with myelin [5]. On the other hand, the dendrite is another type of process that branches off from the soma and they are responsible for receiving most of the input from other neurons [1]. Finally, the axon terminates in multiple endings, which are the presynaptic terminals, where there are usually several branches extending toward the target cell, each of which ends in an enlargement called a synaptic end bulb, which is what makes the connection with the target cell at the synapse [1].

Moreover, looking at nervous tissue, there are regions that predominantly contain cell bodies and regions that are largely composed of just axons. These two regions within nervous system structures are often referred to as gray matter (the regions with many cell bodies and dendrites) or white matter (the regions with many axons) [1]. In particular, the gray matter is not necessarily gray, but it can be pinkish because of blood content, while the white matter is white because axons are insulated by myelin, which is a lipid-rich substance [1]. Nevertheless,

the cell bodies of neurons or axons can be located in discrete anatomical structures which take different names depending on whether the structure is central or peripheral. Precisely, a localized collection of neuron cell bodies in the CNS is referred to as a nucleus while, in the PNS, a cluster of neuron cell bodies is referred to as a ganglion [1]. Lastly, there is also a different terminology applied to bundles of axons (or fibres) still depending on location: in the CNS, it is called a tract whereas the same thing in the PNS would be called a nerve [1].

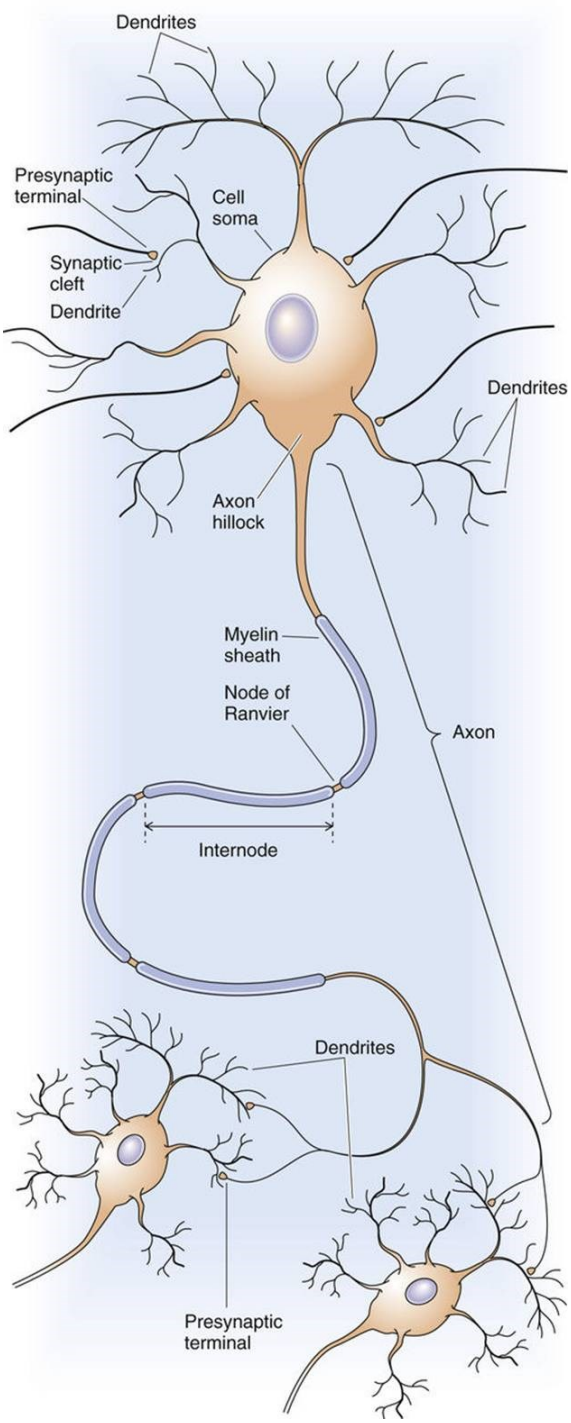


Figure 2. *Morphology of a typical neuron* [6].

1.2 Function of the nervous system

The nervous system is a complex network that enables an organism to interact with its surroundings [7]. The organism has sensors (sensitivity receptors), mostly located on the surface, which have the task of "feeling" the world around us and informing the centres which constitute the nervous system. The information occurs by sending electrical signals that will be encoded by the centres themselves and finally perceived as "sensations" [3]. In general terms, it can be said that the nervous system is involved in receiving information about the environment around us (sensation) and generating responses to that information (motor responses), thus the nervous system can be divided into regions that are responsible for sensation (sensory functions) and for the response (motor functions) [1]. However, there is a further function that needs to be included, which is the integration. In fact, the nervous system is characterized by sensory components which detect environmental stimuli, and motor components which provide skeletal, cardiac, and smooth muscle control, as well as control of glandular secretions, which are coordinated in a system to compel appropriate motor responses to the stimuli or sensory inputs that have been received, stored, and processed [7]. This means that stimuli that are received by sensory structures are communicated to the nervous system where that information is processed [1]. In fact, stimuli are compared with, or integrated with, other stimuli, memories of previous stimuli, or the state of a person at a particular time and this leads to the specific response that will be generated [1]. Thus, CNS and PNS work in such a way that the PNS, of which the previously mentioned sensitivity receptors are part, is responsible for perceiving environmental and visceral stimuli and for sending them to the CNS, and for carrying the responses generated at the level of the CNS to all the organs of the body. The CNS is instead responsible for decoding the information received from the periphery, for their processing and for the genesis of the responses [3].
Nevertheless, the nervous system is composed of vast neural networks [7]; signalling within these circuits enables thinking, language, feeling, learning, memory, and all function and sensation [7]. Thus, the nervous system (together with the endocrine system) is responsible for a vital function for the human organism: the communication [2].

1.3 The action potential

The functions of the nervous system – sensation, integration, and response – depend on the functions of the neurons underlying these pathways. To understand how neurons are able to communicate, it is necessary to describe the role of an excitable membrane in generating these

signals. The basis of this communication is the action potential, which demonstrates how changes in the membrane can constitute a signal.

At the end of the IXX century Waldeyer and Ramon y Cajal expressed the theory of the neuron, which is the foundation of modern neurophysiology, that is the anatomical and functional individuality of these cells of the CNS [4], and it can be seen as a first attempt to consider the nerve tissue as composed of distinct structural units: the neurons [3]. In fact, the CNS can be considered as a machine composed of elements (neurons) specialized for the detection, transmission and processing of information [4] and so the principal function of a neuron is to process and communicate information [8]. Therefore, the neuron is the specific anatomical unit of the nervous system, and the CNS is constituted of separated cellular elements with different shapes, dimensions and morphological characteristics, which are connected to each other by highly specialized contact zones, represented by the synapses [3]. Precisely, in order to achieve its goals of communication, each neuron integrates information across thousands of its synaptic inputs [8]. Moreover, neurons are very particular cells, since they have the ability to originate and propagate nerve impulses, which means that they have both the characteristics of excitability and conductivity [2]. The nerve impulse is nothing more than an action potential [4], which is the membrane potential of an active neuron and so of a neuron which is conducting an impulse [2]. The production of the action potential is governed by the law of all or nothing, which propagates irresistibly, that is, without decrement, up to the extremity of the excited fiber, a real electrical message which travels quickly and with constant amplitude [4].

This is just a general overview about the topic of interest of this paragraph. In fact, in the following of the current paragraph, more in-depth information will be given about this.

A potential is a distribution of charge across the cell membrane, and it is measured in millivolts (mV). The standard is to compare the inside of the cell relative to the outside, so the membrane potential is a value representing the charge on the intracellular side of the membrane based on the outside being zero, relatively speaking [1]. The resting membrane potential is measured at about -70 mV, and it describes the steady state of the cell, which is a dynamic process balanced by ion leakage and ion pumping, where leakage channels allow Na^+ to slowly move into the cells or K^+ to slowly move out, and the Na^+/K^+ pump restores them [1]. However, without any outside influence, the resting membrane potential will not change and so, in order to get an electrical signal started, the membrane potential has to change.

Excitatory input to a neuron usually generates an inward flow of positive charge (i.e., an inward current) across the dendritic membrane, which makes the membrane voltage more positive (i.e., less negative) and so it is said to depolarize the cell [9]. So, due to the higher concentration of Na^+ ions outside the cell than inside, there is an inflow current of Na^+ which increases the membrane potential. As the membrane potential reaches +30 mV, there is the opening of K^+ channel allowing the exiting of these ions due to a concentration gradient, making the membrane potential do move back towards its resting voltage and thus repolarizing the cell [1]. Repolarization returns the membrane potential to the -70 mV value that indicates the resting potential, but it actually overshoots that value because potassium ions reach equilibrium when the membrane voltage is below -70 mV, so a period of hyperpolarization occurs while the K^+ channels still open and they are slightly delayed in closing (Figure 3) [1]. All of this takes place within approximately 2 milliseconds and while an action potential is in progress, another one cannot be initiated, which is an effect referred to as the refractory period [1]. There are two phases of the refractory period: the absolute refractory period, during which another action potential will not start, and the relative refractory period [1].

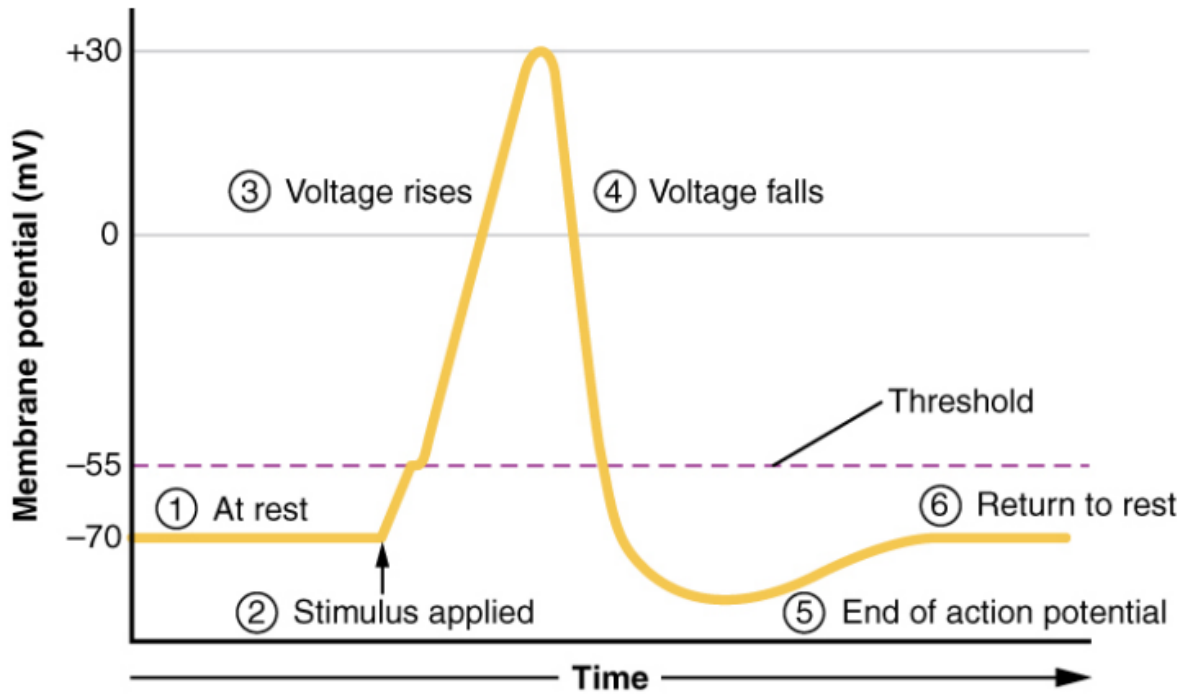


Figure 3. ***Stages of an action potential*** - *Plotting voltage measured across the cell membrane against time, the events of the action potential can be related to specific changes in the membrane voltage. (1) At rest, the membrane voltage is -70 mV. (2) The membrane begins to depolarize when an external stimulus is applied. (3) The membrane voltage begins a rapid rise toward +30 mV. (4) The membrane voltage starts to return to a negative value. (5) Repolarization continues past the resting membrane voltage, resulting in hyperpolarization. (6) The membrane voltage returns to the resting value shortly after hyperpolarization* [1].

If the neuron receives its input from a neighbouring cell through a chemical synapse, neurotransmitters trigger currents by activating ion channels. If the cell is a sensory neuron, environmental stimuli (e.g., chemicals, light, mechanical deformation) activate ion channels and produce a flow of current [9]. The change in membrane potential (V_m) caused by the flow of charge is called a postsynaptic potential (PSP) if it is generated at the postsynaptic membrane by a neurotransmitter, and a receptor potential if it is generated at a sensory nerve ending by an external stimulus [9]. In particular, PSP is the graded potential in the dendrites of a neuron that is receiving synapses from other cells, and it can be depolarizing (excitatory postsynaptic potential – EPSP) or hyperpolarizing (inhibitory postsynaptic potential – IPSP) [1]. The synaptic (or receptor) potentials generated at the ends of a dendrite are communicated to the soma, but not usually without substantial attenuation of the signal (Figure 4A) [9]. As an EPSP reaches the soma, it may also combine with EPSPs arriving by other dendrites on the cell and this behaviour is a type of spatial summation and can lead to EPSPs that are substantially larger than those generated by any single synapse (Figure 4B, 4C) [9]. On the other hand, temporal summation occurs when EPSPs arrive rapidly in succession: when the first EPSP has not yet dissipated, a subsequent EPSP tends to add its amplitude to the residual of the preceding EPSP (Figure 4D) [9]. Thanks to this summation, if the V_m change in the soma is large enough to reach the threshold voltage, the depolarization may trigger one or more action potentials between the soma and axon, as shown in Figure 4B to 4D, which are fixed in amplitude, not graded, and have uniform shape [9].

In conclusion, neurons are “excitable” cells [2, 6] conducting impulses which make all functions of the nervous system possible [2]. In fact, they are cells specialized in the rapid transmission of electrical signals which release chemical substances (neurotransmitters or neurohormones) through which the neuron communicates with other cells [3] and so they communicate via a combination of electrical and chemical signals [6]. In other words, neurons form the "conduction wires" of the information circuits of the nervous system [2].

Figure 5 illustrates the different functional regions of neurons, distinguished on the basis of their role in the reception and conduction of nerve signals. The dendrites and the cell body mainly act as an entrance area, receiving the nerve stimulus and generating the response nerve impulses [2]. The cone of emergence of the axon acts as a summation area, integrating all the nerve impulses coming from the soma and the dendrites and deciding whether to continue propagating the impulse along the neuron [2]. The axon, on the other hand, is the conduction area, since its primary purpose is to conduct the nerve impulse from the emergency cone of the axon, along its entire length, to the end of the neuron [2]. Whether the axon is myelinated

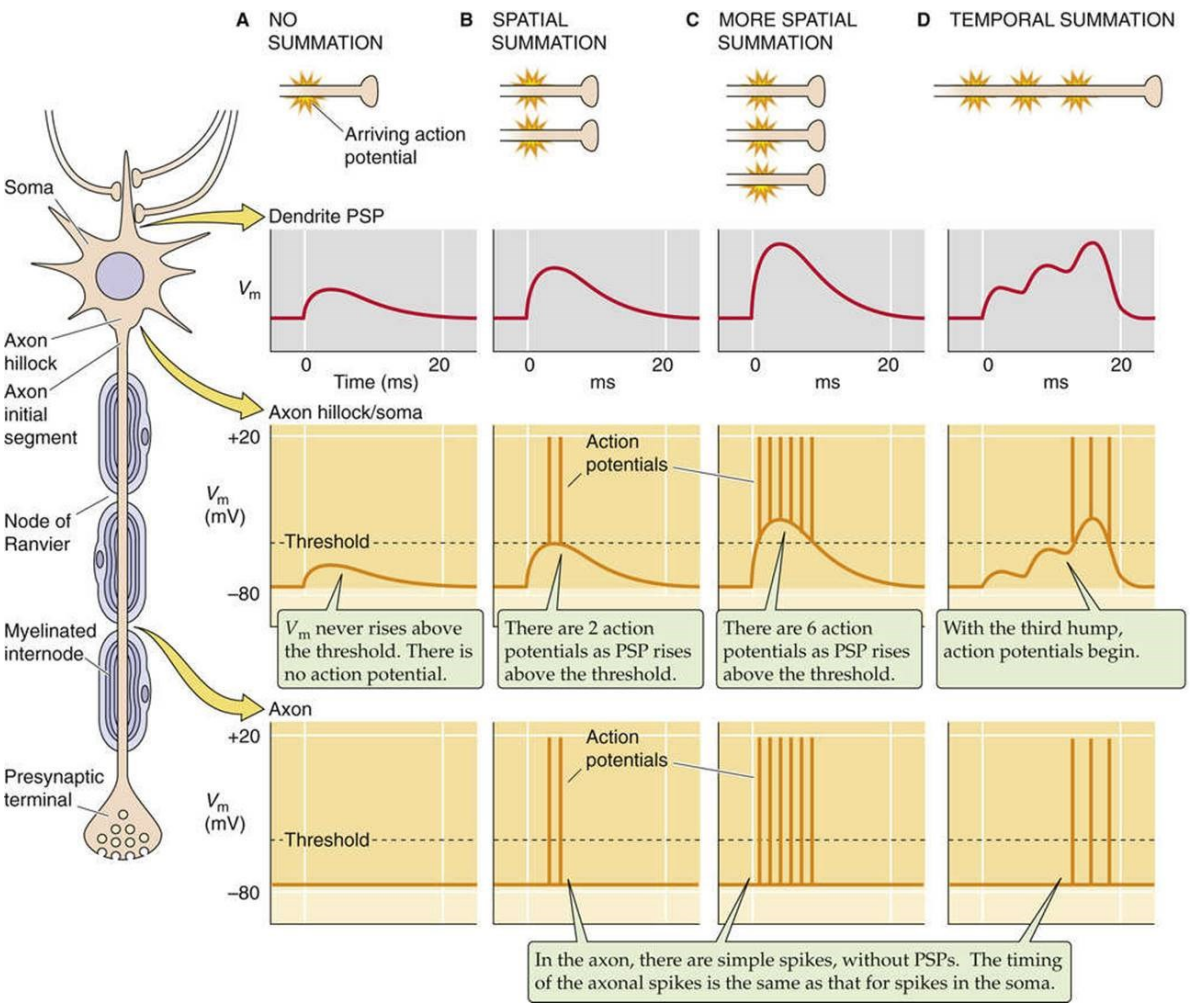


Figure 4. *Spatial versus temporal summation of excitatory postsynaptic potentials (EPSPs)* [9].

or not affects the speed of impulse conduction in the axon [2]. In particular, in axons with myelin sheaths, the propagation of the nerve impulse occurs in a saltatory way, from one Ranvier node to another [3]. Therefore, the axon has an efferent function, as it conducts the signals transmitted by the soma [4]. Finally, the distal ends of the axons form branches, the so-called telodendrites, each of which ends with a synaptic button [2], which constitute the distal end of two contiguous neurons [4]. Thus, the telodendrites, together with the synaptic buttons, act as an exit area [2]. Neuronal synapses are contact sites where signals between two neurons (pre and postsynaptic) are transferred [10]. There are two general categories of synapses: electrical synapses and chemical synapses [2, 10].

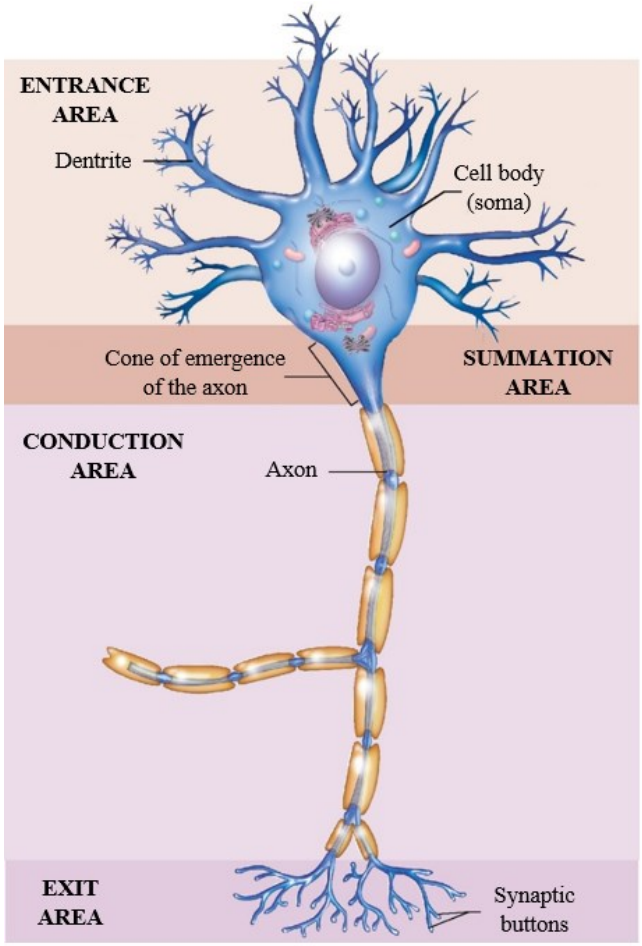


Figure 5. *Functional regions of the neuronal plasma membrane* [2].

1.4 The central nervous system: the brain

The CNS or neuroaxis, defined as the set of nerve formations contained within the cranial cavity and the vertebral canal, consists of the brain and spinal cord [3]. Hence, the CNS is the main integrator of sensory input and motor output, therefore it is able to analyse the incoming information and activate responses to changes that threaten the homeostatic balance of the organism [2].

The brain is one of the largest organs in an adult [2] and is also the most voluminous part of the CNS, the one contained in the cranial cavity [4]. To describe it with rounded numbers, it is estimated that the human brain contains about 100 billion neurons (about 10% of the total number of nervous system cells present in the brain), and 900 billion glia cells [2]. Furthermore, the average brain weight of an adult is about 1250g (1308g for males and 1171g for females - Chiarugi 1917) [4].

Figure 6 shows the six major divisions of the brain, named from bottom to top, are: medulla oblongata, pons, midbrain, cerebellum, diencephalon, and cerebrum [2]. Very often the term brain stem is used to indicate medulla oblongata, pons and midbrain [4].

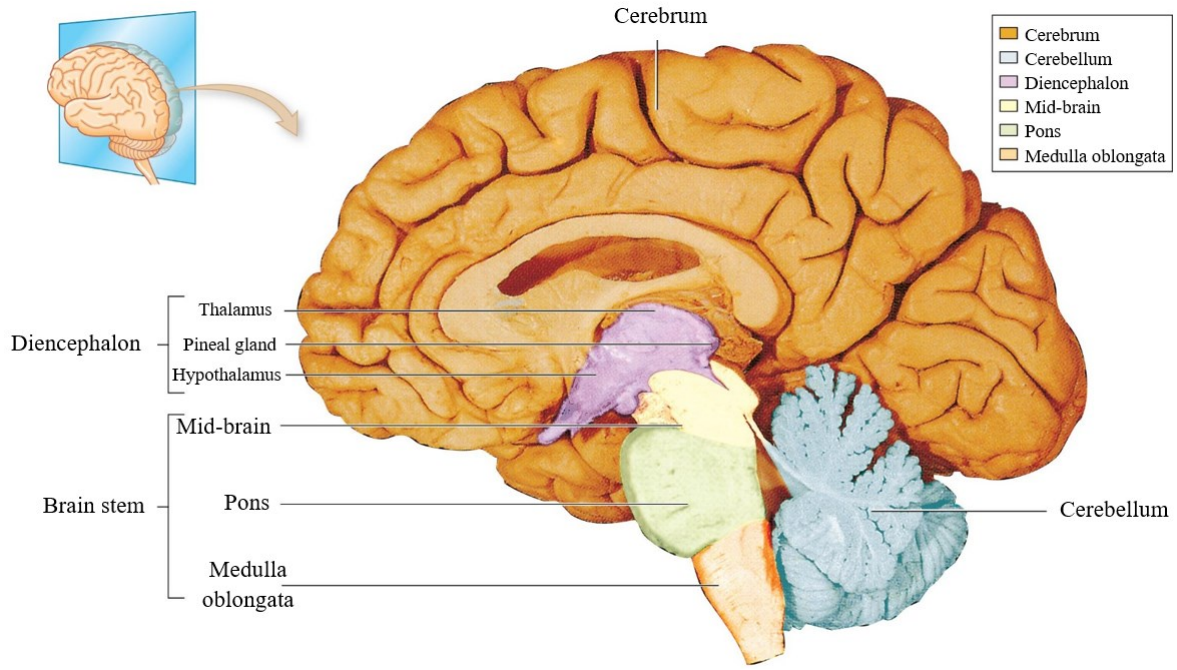


Figure 6. ***Division of the brain*** *- Midsagittal section of the brain showing the characteristics of its main compartments* [2].

1.4.1 The cerebrum

The iconic gray mantle of the human brain, which appears to make up most of the mass of the brain, is the cerebrum (Figure 7) [1]. The cerebrum, the largest and highest portion of the brain, is constituted of two halves: the left and right cerebral hemispheres [2], which are separated by a deep hemispherical longitudinal fissure, called the cerebral sickle [4]. The surface of the hemispheres, called cerebral cortex [2, 4], is wrinkled and the rest of the structure is beneath that outer covering [1]. Deep within the cerebrum, the white matter of the corpus callosum provides the major pathway for communication between the two hemispheres of the cerebral cortex [1].

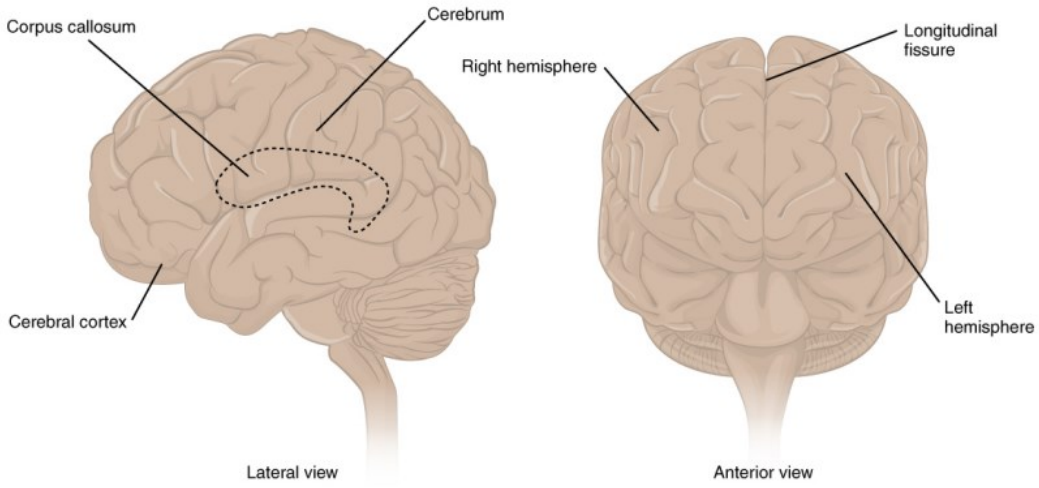


Figure 7. ***The cerebrum*** *- The cerebrum is a large component of the CNS in humans, and the most obvious aspect of it is the folded surface called the cerebral cortex* [1].

In mammals, the cerebrum comprises the outer gray matter that is the cortex and several deep nuclei that belong to three important functional groups: the basal nuclei (responsible for cognitive processing, the most important function being that associated with planning movements), the basal forebrain (it contains nuclei important in learning and memory) and the limbic cortex (the region of the cerebral cortex that is part of the limbic system, a collection of structures involved in emotion, memory, and behaviour) [1].

The cerebrum is covered by a continuous layer of gray matter that wraps around either side of the forebrain - the cerebral cortex [1]. This thin, extensive region of wrinkled gray matter is responsible for the higher functions of the nervous system [1]. Moreover, the folding of the cortex maximises the amount of gray matter in the cranial cavity. During embryonic development, as the telencephalon expands within the skull, the brain goes through a regular course of growth that results in everyone's brain having a similar pattern of folds. The surface of the brain can be mapped on the basis of the locations of large gyri (the ridge of one of those wrinkles) and sulci (the groove between two gyri) [1]. In particular, the cerebral cortex is furrowed by numerous fissures which delimit gyri [4], such as the precentral gyrus, the postcentral gyrus, the cingulate gyrus and the hippocampal gyrus [2]. Using these landmarks, the cortex can be separated into five regions, or lobes: four of them are named after the bones which cover them: frontal lobe, parietal lobe, temporal lobe and occipital lobe (Figure 8) [2]. A fifth lobe, the insula (Reil's island), is hidden from view in the lateral fissure.

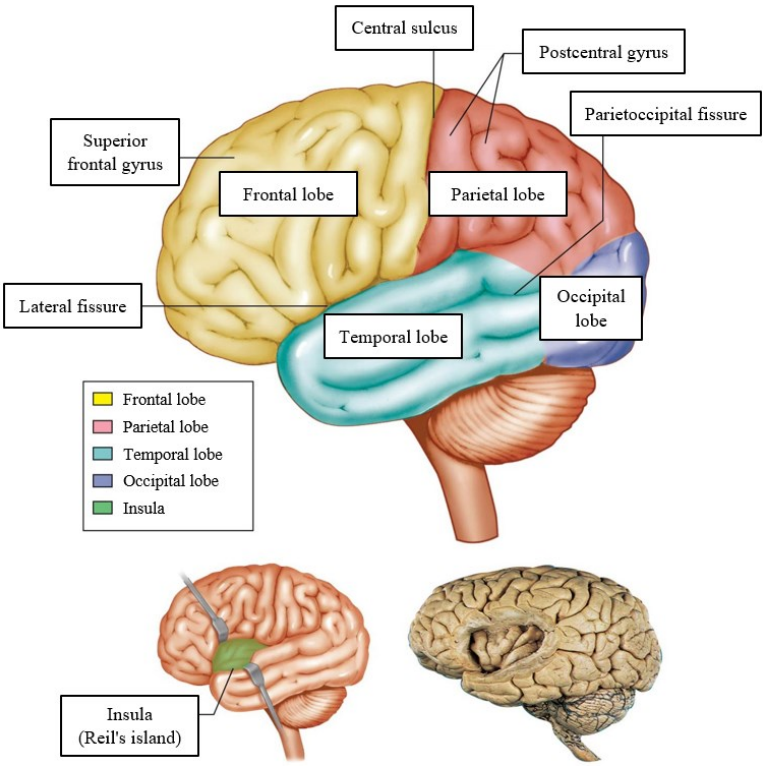


Figure 8. ***Left brain hemisphere, lateral surface*** - *Note the highlighted lobes of the brain* [2].

Figure 8 also shows the two most important fissures, which are the central fissure (Rolandic fissure) and the lateral fissure (Sylvian fissure): the former delimits the frontal lobe from the parietal lobe, while the latter the temporal lobe from the lower portion of the frontal lobe and the parental lobe [4]. Moreover, also the longitudinal fissure and the parietooccipital fissure are reported.

In literature, it is commonly known that certain areas of the cortex in each hemisphere mainly perform a certain function, therefore we speak of cerebral localization. Instead, brain plasticity means that the localization of functions varies from person to person and, when the brain is damaged, even according to the different stages of an individual's life [2].

In the early 1900s, a German neuroscientist named Korbinian Brodmann subdivided the cerebral cortex into numerous areas based on regional differences in the distribution, density, shape, and size of cell bodies, i.e., the cytoarchitecture [11] and divided the cortex into 52 separate regions on the basis of the histology of the cortex [1]. His work resulted on a system of classification known as “Brodmann’s areas”, which is still used today to describe the anatomical distinctions within the cortex. Generally speaking, it is possible to distinguish sensory and motor functions of the cerebral cortex and also about integrative functions of the cerebral cortex, which include consciousness and mental activities of all kinds [2]. Among all the integrative functions, particular attention will be devoted to memory.

1.4.1.1 The memory

Memory is one of the main activities carried out by the human mind and it is defined as the ability to fix, preserve and recall states of consciousness [4]. The cortex is capable of storing and retrieving both short-term and long-term memories [2] and thus the memory trace is formed through these two stages. As it is clear from the name itself, the short-term memory is the ability to store information for a few seconds or minutes [2]. In fact, in the process of short-term memory, which is of limited capacity and of limited duration, after a few minutes, if the process of fixation or consolidation of the memory (learning) has not intervened, it disappears forever [4]. On the other hand, the second stage (long-term memory) is what remains after the first stage of short-term memory has completed. Moreover, short-term and long-term memories are both functions which involve many parts of the cerebral cortex, especially the temporal, parietal, and occipital lobes [2]. In particular, the temporal lobe is associated with primary auditory sensation, known as Brodmann’s areas 41 and 42 in the superior temporal lobe but, because regions of the temporal lobe are part of the limbic system, memory is an important function associated with that lobe [1]. Thus, memory is essentially a

sensory function; memories are recalled sensations such as the smell of Mom's baking or the sound of a barking dog, even memories of movement are really the memory of sensory feedback from those movements, such as stretching muscles or the movement of the skin around a joint [1]. Structures in the temporal lobe are responsible for establishing long-term memory, but the ultimate location of those memories is usually in the region in which the sensory perception was processed [1].

Another important aspect is the evocation of the memory, which occurs through the reactivation of the Papez circuit, whose repeated reactivations improve the consolidation of the memory over time [4]. This explains, whatever the cause of a possible decline in memory, the better preservation of the most ancient memories (Ribot's law) [4].

1.4.1.2 Subcortical structures: the hippocampus

Beneath the cerebral cortex are sets of nuclei known as subcortical nuclei that augment cortical processes. The nuclei of the basal forebrain serve as the primary location for acetylcholine production, which modulates the overall activity of the cortex, possibly leading to greater attention to sensory stimuli [1]. Alzheimer's disease is associated with a loss of neurons in the basal forebrain [1]. The hippocampus (via Latin from Greek ἱππόκαμπος, *seahorse*) and amygdala are medial-lobe structures that, along with the adjacent cortex, are involved in long-term memory formation and emotional responses [1]. Before giving more in-depth information, the hippocampi of the temporal lobes are so named because of their curled shape, which early anatomists thought resembled a seahorse (Figure 9). In general terms, the hippocampal region takes new experiences and turns them into memories that can be stored and accessed later. Individuals who have damage to the hippocampal regions, in both temporal lobes display severe to profound deficits in short-term memory because they cannot crease new memories, and long-term memory abilities might be destroyed as well [12].

The hippocampus develops in the foetal brain by a process of continuing expansion of the medial edge of the temporal lobe in such a way that the hippocampus comes to occupy the floor of the temporal horn of the lateral ventricle [13]. In the mature brain, therefore, the parahippocampal gyrus on the external surface is continuous with the concealed hippocampus [13]. The hippocampus is C-shaped in coronal section and since its outline bears some resemblance to a ram's horn, the hippocampus is also called the *cornu ammonis* [13]. The ventricular surface of the hippocampus is a thin layer of white matter called the alveus, which consists of axons that enter and leave the hippocampal formation [13]. These fibres form

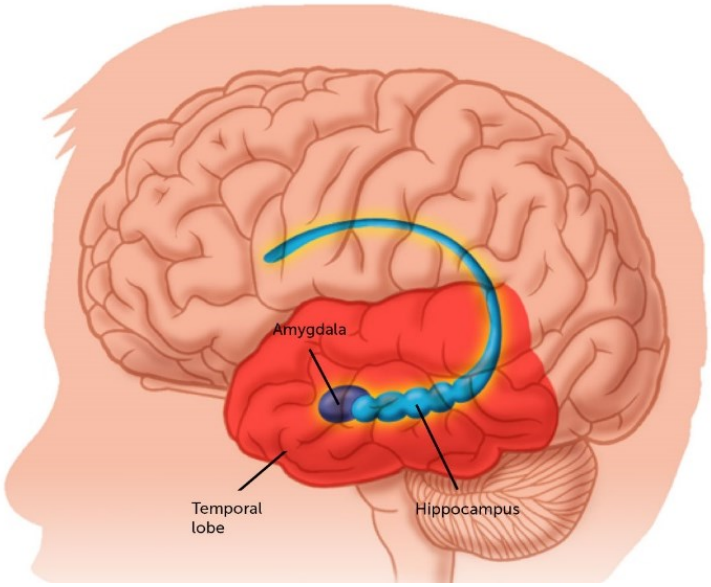
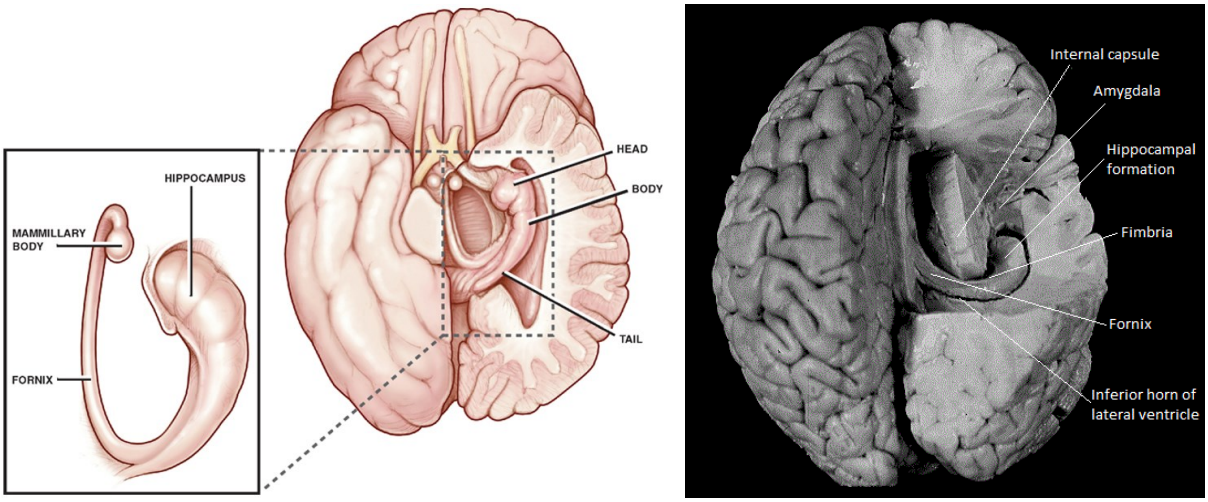


Figure 9. *The temporal lobe, with hippocampus and amygdala.*

the fimbria of the hippocampus along its medial border and then continue as the crus of the fornix after the hippocampus ends beneath the splenium of the corpus callosum (Figure 10). Continued growth of the cortical tissue composing the hippocampus is responsible for the dentate gyrus, which occupies the interval between the fimbria of the hippocampus and the parahippocampal gyrus [13]; its surface is toothed or beaded, hence the name. Although the parahippocampal gyrus is included in the limbic lobe as defined anatomically, most of its cortex is of the six-layered type or nearly so [13]. In the region of the gyrus known as the subiculum, there is a transition between neocortex and the three-layered archicortex of the hippocampus and the anterior end of the parahippocampal gyrus, medial to the rhinal sulcus, is the entorhinal area [13].



(A) (B)

Figure 10. ***The hippocampus*** – (**A**) *Magnification of the position of the hippocampus in the brain*. (**B**) *Opening of the brain with clear view of the hippocampus*

Psychologists and neuroscientists agree that the hippocampus is of relevant importance in the formation of new memories. In particular, there have been recognized different types of long-term memory that are processed differently in the brain. Starting from the declarative (or explicit) memory, it is the knowledge and recall of facts or events that can be recalled to consciousness and the acquisition of an item into declarative memory typically occurs on a single occasion [13]. As previously mentioned, any fact or event is initially held in short-term memory, but it may be forgotten during the course of the next hour or so, otherwise it is moved into long-term storage. If declarative memories are not recalled from time to time, the process of recall will require mental effort, or the memories may be forgotten. Procedural (or implicit) memory is for learned skills, including regularly performed motor tasks and mental activities such as using the common vocabulary and grammatical rules of a language [13]. The learning occurs gradually, and recall is improved with repetition and practice. The best understood functions of the hippocampal formation are the retention of information in short-term memory and its transfer into long-term declarative memory [13].

2. Alzheimer's disease

Dementia is a syndrome – usually of a chronic or progressive nature – which leads to deterioration in cognitive function (i.e. the ability to process thought) beyond what might be expected from the usual consequences of biological aging [14]. This degeneration can progress to worsening memory, attention span, intellectual capacity, personality, and motor control [2], but consciousness is not affected [4]. The impairment in cognitive function is commonly accompanied, and occasionally preceded, by changes in mood, emotional control, behaviour, or motivation [14]. Moreover, progressive loss of cognitive functions can be caused by cerebral disorders like Alzheimer's disease (AD) or other factors such as intoxications, infections, abnormality in the pulmonary and circulatory systems, which causes a reduction in the oxygen supply to the brain, nutritional deficiency, vitamin B12 deficiency, tumours, and others [15].

Pathological cognitive impairment is a highly disabling and continuously increasing condition all over the world, both in industrialized and developing countries, by virtue of the aging of populations [4]. According to the World Health Organization (WHO), worldwide, around 55 million people have dementia, with over 60% living in low- and middle-income countries [14]. Moreover, as the proportion of older people in the population is increasing in nearly every country, this number is expected to rise to 78 million in 2030 and 139 million in 2050 [14]. According to a report by WHO, dementia is currently the seventh leading cause of death among all diseases and one of the major causes of disability and dependency among older people worldwide [14]. In fact, dementia has physical, psychological, social and economic impacts, not only for people living with dementia, but also for their carers, families and society at large [14].

It is possible to distinguish many different types of dementia but, in this work, the attention will be devoted to the Alzheimer-Perusini disease (Alzheimer Disease), named after the German psychiatric Alois Alzheimer [15]. In particular, AD is the most frequent form of dementia (60%) [4] and it is considered one of the progressively more frequent diseases in Western countries due to the dramatic increase in the elderly population [4] and can be defined as a slowly progressive neurodegenerative disease characterized by neuritic plaques and neurofibrillary tangles (Figure 11) as a result of amyloid-beta peptide's (Aβ) accumulation in the most affected area of the brain, the medial temporal lobe and neocortical structures [15]. In fact, the prevalence of AD increases progressively with age (5% and 20% over 65 and 80 years, respectively) [4]. At present, there are around 50 million AD patients

worldwide and this number is projected to double every 5 years and will increase to reach 152 million by 2050 [15].

AD has been considered a multifactorial disease associated with several risk factors such as increasing age, genetic factors, head injuries, vascular diseases, infections, and environmental factors [15]. Nevertheless, it is not yet known what exactly causes this disease, however there are strong signs that it has a genetic basis, at least in some families [2]. Moreover, at present, there is no cure for AD, although there are available treatments that just improve the symptoms [15].

2.1 Neuropathology of Alzheimer's disease

The brain of people affected by AD appears to be reduced in volume, with a marked and widespread atrophy in the frontal and temporal cortex (Figure 12), with loss of large neurons, numerically returning in the order of 30-40%, neurons of the nucleus basalis of Meynert show a particularly high numerical reduction, in the order of 70-80% [4]. The primary cortical areas, the cerebellum, the basal ganglia and several thalamic nuclei, on the other hand, have less impairment [4].

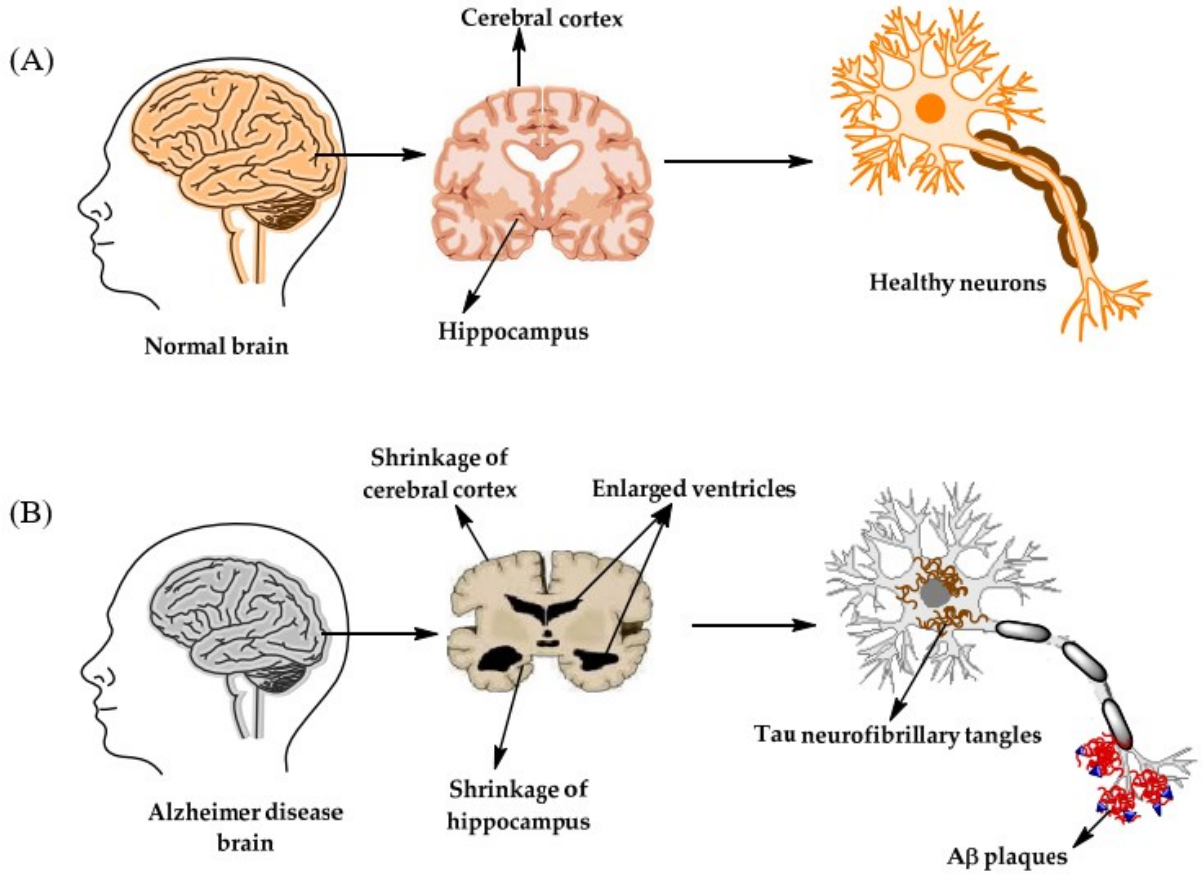


Figure 11. *The physiological structure of the brain and neurons in* (**A**) *healthy brain and* (**B**) *Alzheimer's disease brain* [15].

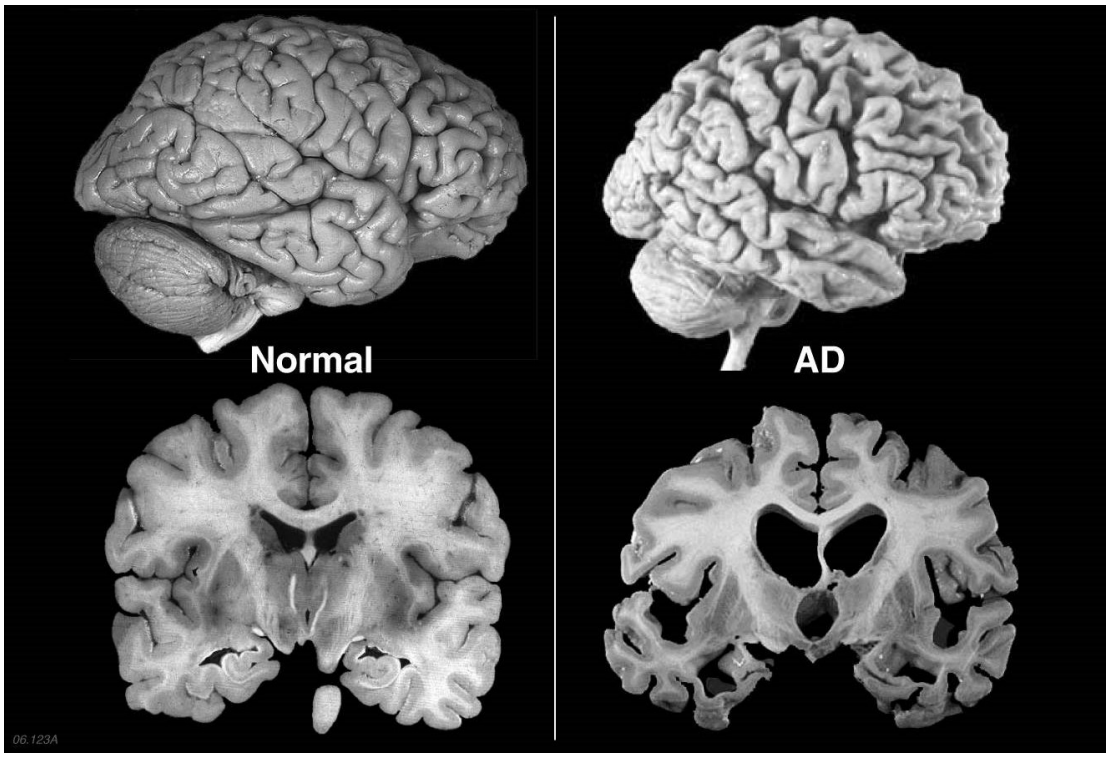


Figure 12. *Brain atrophy in advanced Alzheimer's disease: a healthy brain (normal) and an Alzheimer's disease brain (AD).*

The pathological process initially manifests itself at the level of the medial temporal regions and then spreads almost symmetrically in the two hemispheres, towards the neocortex and towards the subcortical and catecholaminergic cholinergic nuclei of the trunk [4]. Moreover, the involvement of the hippocampus and amygdala in the early stages of the disease causes a particular pathological condition, known as limbic deafferentation, characterized by deficiencies, not only in terms of memory, but also of motivation and affect, because the sensory information transmitted from the primary cortical areas to the associative ones can no longer be integrated through the passage in the limbic circuit [4].

The key pathological hallmarks – extracellular plaques and intracellular neurofibrillary tangles (NFTs) – described by Alois Alzheimer in his seminal 1907 article are still central to the post-mortem diagnosis of Alzheimer's disease (AD), but major advances in the understanding of the underlying pathophysiology as well as significant progress in clinical diagnosis and therapy have changed the perspective and importance of neuropathologic evaluation of the brain [16]. The notion that the pathological processes underlying AD already start decades before symptoms are apparent in patients has brought a major change reflected in the current neuropathological classification of AD neuropathological changes [16].

There are two types of neuropathological changes in AD which provide evidence about disease progress and symptoms and include: positive lesions (due to accumulation), which are characterized by the accumulation of neurofibrillary tangles, amyloid plaques, dystrophic

neurites, neuropil threads, and other deposits found in the brains of AD patients, in addition to negative lesions (due to losses), that are characterized by large atrophy due to a neural, neuropil, and synaptic loss (Figure 13) [15]. Besides, other factors can cause neurodegeneration such as neuroinflammation, oxidative stress, and injury of cholinergic neurons [15].

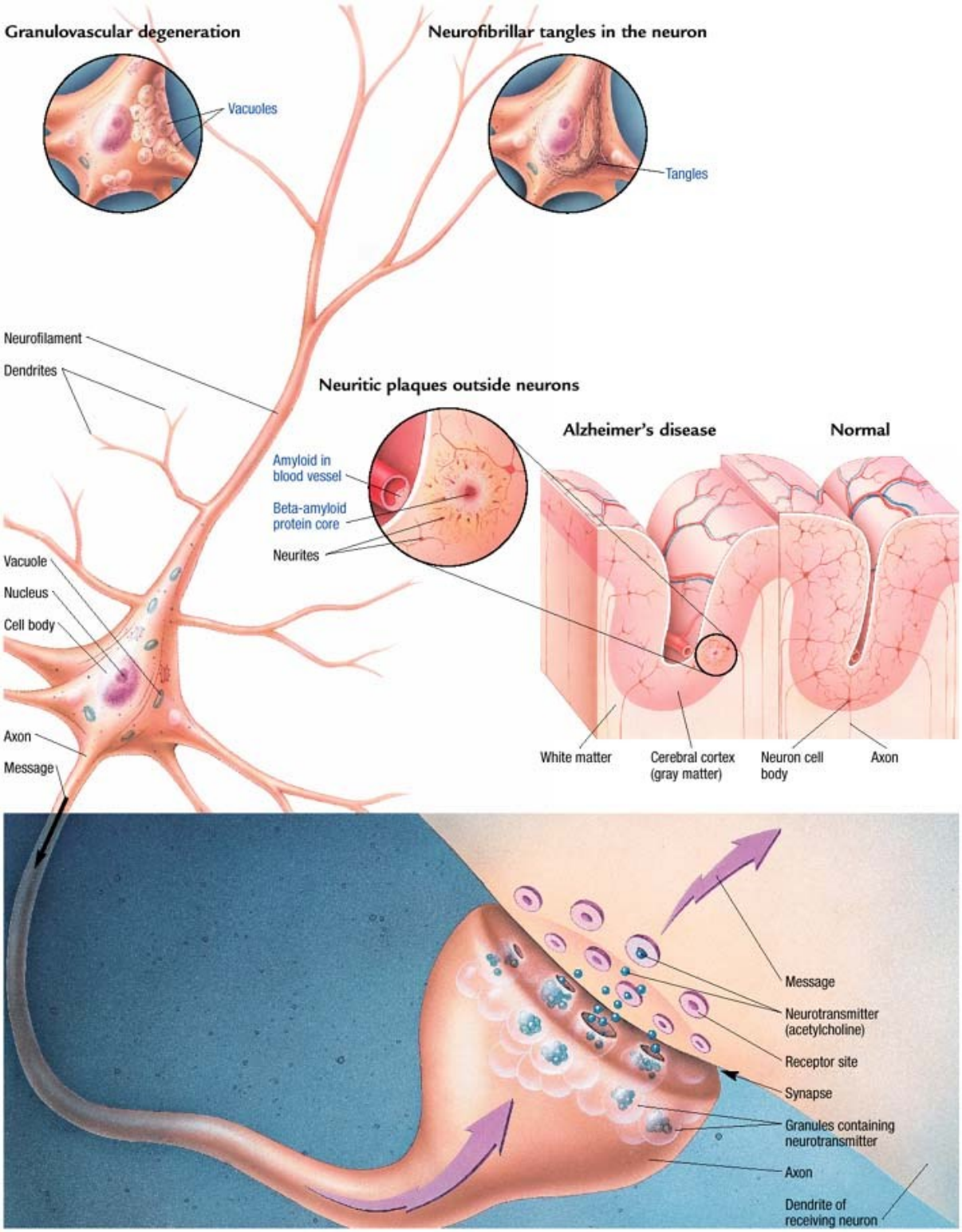


Figure 13. *Tissue changes in Alzheimer's Disease.*

The characteristic elements of AD are senile plaques (SP), neurofibrillary tangles (NFTs) and synaptic loss. However, some of these elements are also found in the normal senescent brain, varying in the pathological condition in terms of density and distribution [4].

First of all, the SP are extracellular deposits of beta-amyloid protein (Aβ) [15]. The identification of Aβ as a central component of extracellular plaques, as well as genetic evidence linking the amyloid precursor protein (APP) and its processing by beta- and gamma-secretase to autosomal-dominant forms of AD, has led to the formulation of the amyloid cascade hypothesis which endures as the favoured pathophysiological framework to understand AD [16]. Multiple different forms of Aβ deposits can be identified in the AD brain, ranging from diffuse, or "lake-like" amyloid over compact, coarse grained, cotton-wool to cored- or senile plaques [16]. The importance of each of these types of Aβ deposits has been studied extensively, and it seems to be more and more evident that diffuse Aβ plaques are probably more benign in nature, as they can be seen in cognitively normal subjects with minimal to no co-existing tau pathology (termed "pathological aging"), while cored plaques, which are often identical to neuritic plaques, are associated with cognitive decline [16]. These Aβ plaques are found predominantly in the cortex of the associative areas [4] and spread through the brain in a predictable fashion, which is summarized in five distinct phases [16]:

- Phase 1: early deposits can be seen in the neocortex;
- Phase 2: Aβ plaques appear in limbic regions including entorhinal cortex, subiculum, amygdala, and cingulate gyrus;
- Phase 3: Aβ deposits in subcortical areas including basal ganglia and thalamus;
- Phase 4: structures of the brainstem including midbrain, pons, and medulla oblongata are affected;
- Phase 5: Aβ plaques can also be found in the cerebellar cortex.

Phases 4 and 5 were associated with the presence of dementia, while phases 1 and 2 were mostly seen in asymptomatic individuals [16]. Aβ plays a major role in neurotoxicity and neural function, therefore, accumulation of denser plaques in the hippocampus, amygdala, and cerebral cortex can cause stimulation of astrocytes and microglia, damage to axons, dendrites, and loss of synapses, in addition to cognitive impairments [15].

NFTs, the second major pathological finding in AD, are formed by aggregates of the microtubule associated protein tau [16]. NFTs are particularly present in the deeper layers of the cortex [4] and they are abnormal filaments of the hyperphosphorylated tau protein that in some stages can be twisted around each other to form paired helical filament [15]. This protein is the major constituent of NFTs in the brains of AD patients, and its evolution can

reflect NFTs morphological stages, which include pre-tangle phase, mature NFTs and extracellular tangles (or the ghost NFTs stage) which results from a neuronal loss due to large amounts of filamentous tau protein [15]. Also in this case, different stages can be distinguished. The very first NFTs were noted in the transentorhinal region of the hippocampal formation (stage I) from which the density of aggregates progresses and also involves the subiculum region of the hippocampal pyramidal cell layer (stage II) [16]. This early presentation of NFT pathology is referred to as "transentorhinal stages". As the disease progresses, NFTs start to impact the entorhinal cortex and the hippocampal pyramid cell layer (stage III), with a further spread of NFT pathology into the adjacent inferior temporal cortex and other neocortical areas such as superior temporal cortex and frontal cortex (stage IV) [16]. These intermediate stages are often referred to as "limbic stages" since the hippocampal formation is most severely affected. In later phases of the disease, the changes intensify in the hippocampal formation but also affect other areas of the neocortex, including secondary association areas and ultimately primary cortical areas and these late disease stages (V and VI) are therefore called "isocortical stages", where pathology in the peristriate area defines stage V, while intraneuronal aggregates in striate area define stage VI [16]. There is also a correlation between these NFTs stages and observed clinical symptoms: the Braak stages V and VI show the strongest association with clinically observed dementia, while stages I and II are encountered not unfrequently in clinically asymptomatic individuals [16].

Lastly, a synaptic damage in the neocortex and limbic system causes memory impairment and, generally, is observed at the early stages of AD [15]. Synaptic loss mechanisms involve defects in axonal transport, mitochondrial damage, oxidative stress, and other processes that can contribute to small fractions, like the accumulation of Aβ and tau at the synaptic sites [15]. These processes eventually lead to a loss of dendritic spines, pre-synaptic terminals, and axonal dystrophy [15].

Nevertheless, the underlying cause of pathological changes in Alzheimer's disease (Aβ, NFTs, and synaptic loss) is still unknown [15].

2.2 The stages of Alzheimer's disease

The onset of clinical symptoms of AD is usually insidious and the slowly progressive course can be divided into several phases. First of all, there is the pre-clinical or the pre-symptomatic stage (may last several years or more), which is characterized by mild memory loss and early pathological changes in cortex and hippocampus, with no functional impairment in the daily activities and absence of clinical signs and symptoms of AD [15]. Subsequently, in the mild

or early stage of AD, several symptoms start to appear, such as a decline in interest, indifference and short-medium term memory deficit [4] or a trouble in the daily life of the patient with a loss of concentration and memory, disorientation of place and time and a change in the mood [15]. In particular, awareness of the memory deficit and the difficulties it entails can lead to a depressive reaction [4]. Then, in the moderate AD stage, the disease spreads to cerebral cortex areas [15] resulting in the fact that the cognitive deficit becomes more and more evident and changes in the personality appear, the memory function is further compromised, the attention, criticism and judgment are reduced, there is an evident decline in work performance and participation in life familiar [4], and difficulties arise in reading, writing, and speaking [15]. Lastly, the severe AD or late-stage involves the spread of the disease to the entire cortex area with a severe accumulation of neuritic plaques and NFTs [15]. In this last stage, the patient needs constant assistance in carrying out daily activities [4] because the patients cannot recognize their family at all and may become bedridden with difficulties in swallowing and urination, and eventually leading to the patient's death due to these complications [15].

2.3 Diagnosis

In 1984, a study group was established under the aegis of the National Institute of Neurological and Communicative Disorders (NINCDS) and the Alzheimer Disease and Related Disorder Association (ADRDA) to formulate diagnostic criteria for AD [4]. Building upon the original 1984 diagnostic criteria, the National Institute on Aging–Alzheimer's Association (NIA–AA) revised the clinical criteria for the diagnosis of mild cognitive impairment (MCI) and the different stages of dementia due to AD in 2011 [17]. The newer criteria allow for the use of current and future biomarkers in the diagnosis of degenerative brain disease [17].

The diagnosis of AD is clinical. Except for brain biopsy, there are currently no laboratory tests for a definite diagnosis of the disease [4]. However, some instrumental investigations can be used, according to the NINCDC-ADRDA criteria, to support the clinical diagnosis [4]. In particular, the development of non-invasive diagnostic imaging recently resulted in a test which increases the diagnostic accuracy in AD [17]. In fact, neuroimaging techniques, in particular computerized axial tomography (CAT) and brain magnetic resonance imaging (MRI) constitute a fundamental step in the diagnostic iter of AD [4]. With these methods, it is possible to evaluate the degree of cerebral atrophy as there is a significant correlation between this parameter and the severity of dementia [4]. In particular, neuroimaging studies (CAT and

MRI) do not show a single specific pattern with AD and may be normal early in the course of the disease. As AD progresses, more distributed but usually posterior-predominant cortical atrophy becomes apparent, along with atrophy of the medial temporal memory structures (Figure 14 A, B) [18]. The main purpose of imaging is to exclude other disorders, such as primary and secondary neoplasms, vascular dementia, diffuse white matter disease, and normal-pressure hydrocephalus; it also helps to distinguish AD from other degenerative disorders with distinctive imaging patterns such as frontotemporal dementia or Creutzfeldt-Jakob disease [18]. Functional imaging studies in AD reveal hypoperfusion or hypometabolism in the posterior temporal-parietal cortex (Figure 14 C, D) [18].

Both positron emission tomography (PET) and single-photon emission computerized tomography (SPECT) provide information on metabolism and regional blood flow. In patients with probable AD these techniques often show a picture of biparietal and temporal reduction of metabolism and cerebral blood flow, even if these findings are to be considered non-specific [4].

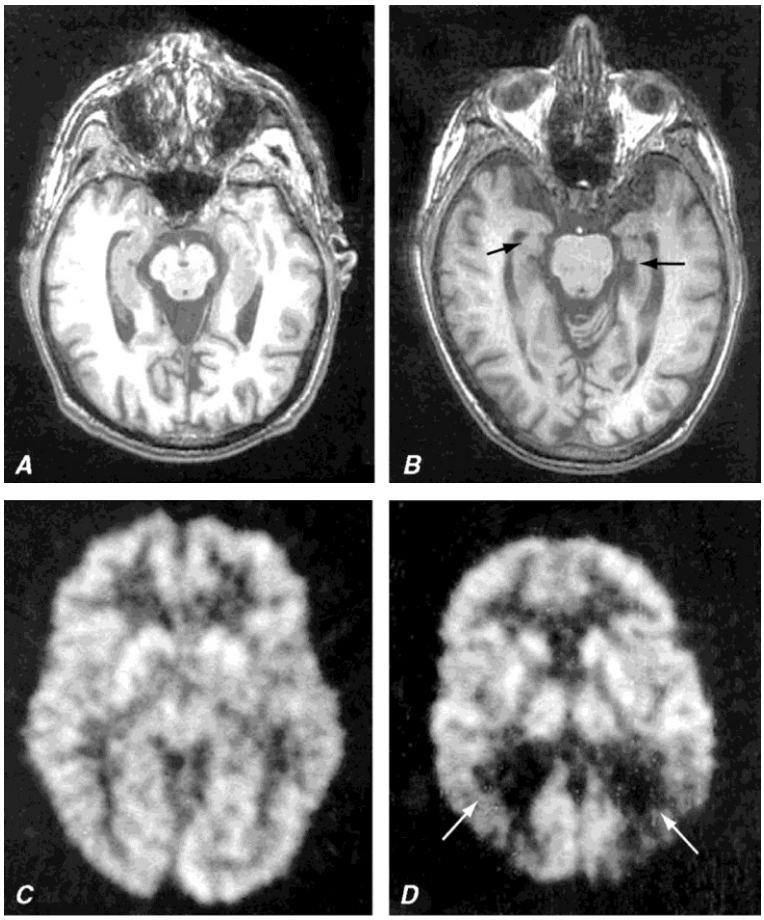


Figure 14. ***Alzheimer's disease*** *- Axial T1-weighted MR images through the midbrain of a normal 86-year-old athlete* (**A**) *and a 77-year-old man with AD* (**B**)*. Note that both individuals have mild sulcal widening and slight dilation of the temporal horns of the lateral ventricles. However, there is a reduction in hippocampal volume in the patient with AD (arrows) compared with the volume of the normal-for-age hippocampus* (**A**)*. Fluorodeoxyglucose positron emission tomography (PET) scans of a normal control* (**C**) *and a patient with AD* (**D**)*. Note that the patient with AD shows decreased glucose metabolism in the posterior temporoparietal regions bilaterally (arrows), a typical finding in this condition.*

Electroencephalographic examination (EEG) in AD shows a significant decrease in the frequency of the alpha band with an increase in power of the theta and delta bands [4]. However, these alterations are present only in the full-blown stages of the disease, while in the initial stages the pattern may still be normal [4]. Quantitative EEG studies have highlighted the existence of a correlation between the severity of the disease and the slowdown of the underlying activity.

Finally, it is essential that patients with suspected AD are subjected to an adequate battery of neuropsychological tests but, as already mentioned, among the most significant tests there are CAT and MRI, where the latter is of particular interest for this work.

3. Image basics and bioimages

Digital image signals are typically represented as two-dimensional (2D) arrays of discrete signal samples [19]. In mathematical terms, an image is a 2D light-intensity function denoted by $f(x, y)$, where the value or amplitude of f at spatial coordinates (x, y) gives the intensity (brightness) of the image at that point. In other words, digital images are composed of individual pixels (this acronym is formed from the words "picture" and "element") (Figure 14), to which discrete brightness or colour values are assigned [20]. In general terms, an image is an array, or a matrix, of square pixels arranged in columns and rows:

$$\text{total number of pixels} = \text{number of rows} \cdot \text{number of columns}$$

For example, by considering 8-bit greyscale image, each picture element has an intensity ranging from 0 to 255 ($2^8 = 256$ values), where values close to 0 indicate darker regions, while values near 255 represent brighter regions, with many shades of grey in the middle.

Instead, the voxel (volumetric pixel or volumetric picture element) is the volume element representing a value on a regular grid in 3D space and so it can be seen as the 3D counterpart of the 2D pixel (Figure 15). For imaging techniques like computed tomography (CT) and MRI, a volume is acquired slice by slice, and each slice is reconstructed from several measures in different angulation allowing voxel assignment [20].

There are several types of images, such as binary images, greyscale images, colour images and videos. The latter is a collection of images in a proper sequence at a certain frame frequency. Contextualising to the case of medical imaging, CT and MRI scans are relatively short videos composed by slices, collected in the right spatial sequence.

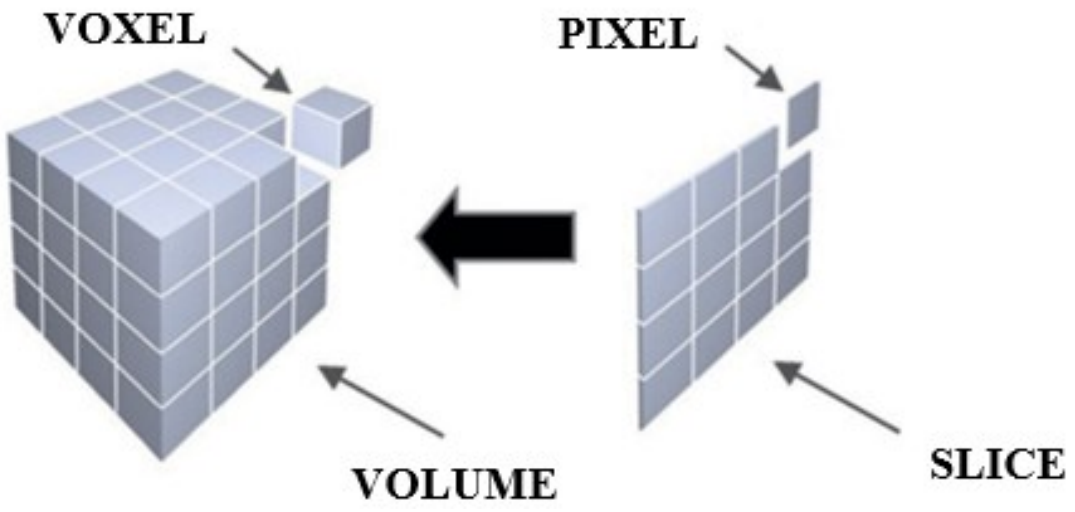


Figure 15. *Pixel and voxel.*

3.1 Biomedical imaging

Medical imaging plays a key role in modern medicine as it allows for the non-invasive visualization of internal structures and metabolic processes of the human body in detail. This aids in disease diagnostics, treatment planning, and treatment follow-up by adding potentially informative data in the form of patient-specific disease characteristics [21]. The science and engineering behind the sensors, instrumentation and software used to obtain biomedical imaging has been evolving continuously since the x-ray was first invented in 1895 [22] and so the amount of healthcare imaging data is rapidly increasing [21]. In fact, biomedical imaging has developed from early, simple uses of X-rays for diagnosis of fractures and detection of foreign bodies into a compendium of powerful techniques, not only for patient care but also for the study of biological structure and function, and for addressing fundamental questions in biomedicine [23]. Technological developments in digital radiography, X-ray computed tomography (CT), nuclear (including positron emission tomography (PET)), ultrasound, optical and magnetic resonance imaging (MRI) have produced a spectrum of methods for interrogating intact 3-dimensional bodies non-invasively [23].

Imaging can provide uniquely valuable information about tissue composition, morphology and function, as well as quantitative descriptions of many fundamental biological processes. In recent years, biomedical imaging science has matured into a distinct and coherent set of ideas and concepts, and it has attained a position of central importance in much medical research [23]. Continuing developments in imaging technology have expanded the application of imaging to new areas, such as the study of gene expression or the functional organization of the brain, and it is important to highlight that imaging science develops applications that use information derived from images for both research and clinical use (e.g., using fMRI to differentiate sub-classes of neuropsychiatric disorders, or to guide surgical procedures) [23].

Nevertheless, major technical advances continue to be made in all modalities, while the development of faster, more powerful computers has led to advanced methods of image analysis and processing algorithms that can be used to extract valuable, quantitative information from images [23]. In fact, the ability to detect, diagnose and monitor pathological, physiological and molecular changes by imaging is of fundamental importance for the management of disease, for personalized interventions, and in basic biological research. However, it has been experienced an increasing difficulty for radiologists and clinicians to cope with the mounting burden of analysing the large amounts of available data from disparate data sources, and studies have highlighted sometimes considerable inter-observer

variability when performing various clinical imaging tasks, thus it follows that there is an evolving need for tools that can aid in diagnosis and decision-making [21].

As it is commonly known, there are many different types of imaging techniques in medicine, among which attention will be devoted to MRI for the scope of this work. However, it is worth to mention that the basic principles of all imaging techniques are the same: a beam of wave passes through the body/area under diagnosis, transmits or reflects back the radiation which will be captured by a detector and processed to get an image pattern, but the type of wave differs depending on the imaging modality (e.g., radio frequency waves are used for MRI) [24].

3.1.1 Magnetic resonance imaging

Magnetic Resonance Imaging (MRI), now widely known for its usefulness as a medical diagnosis tool (it is a powerful diagnostic technique for soft tissues [24]) and for the variety of clear pictures of the body's interior obtained in a harmless and non-invasive manner, had its foundations laid more than 60 years ago in physics experiments designed to measure properties of the nuclear spins of hydrogen atoms [25].

MRI system implies strong and uniform magnetic field together with radiofrequency waves. Suitable resonant radiofrequency is applied to the patient from the scanner, the waves pass through the tissues or any region that hold hydrogen atoms in the body, viz., water molecules, which get excited and return back to the equilibrium state using the energy from oscillating magnetic field which will be captured by the scanner and digitally processed [24]. Hence MRI is best suited for visualization of soft tissues, tendons and ligaments, but it is also applicable in detection of some lesions in brain as shown in Figure 16. The major advantage of using MRI is to vary the contrast of the image. In fact, minute alteration in the radio wave frequency and the magnetic field can alter the contrast of the image which highlights different types of tissues [24]. Another advantage of MRI is that it can construct images in any plane (axial/horizontal) which is unfeasible in CT. More generally, the MRI scans constitute 3D data (thus, volumetric data) given by the spatial sequence of the slices.

There are different types of MRI. MRI exploited in the measurement of diffusion of water molecules inside the body is known as Diffusion MRI, which is valuable in diagnosis of neurological disorders like multiple sclerosis and in stroke [24]. Instead, the change in neural activity can be diagnosed using functional MRI (fMRI), hence widely applied is neurological disorders, while other application would be the real time MRI, which monitors the moving

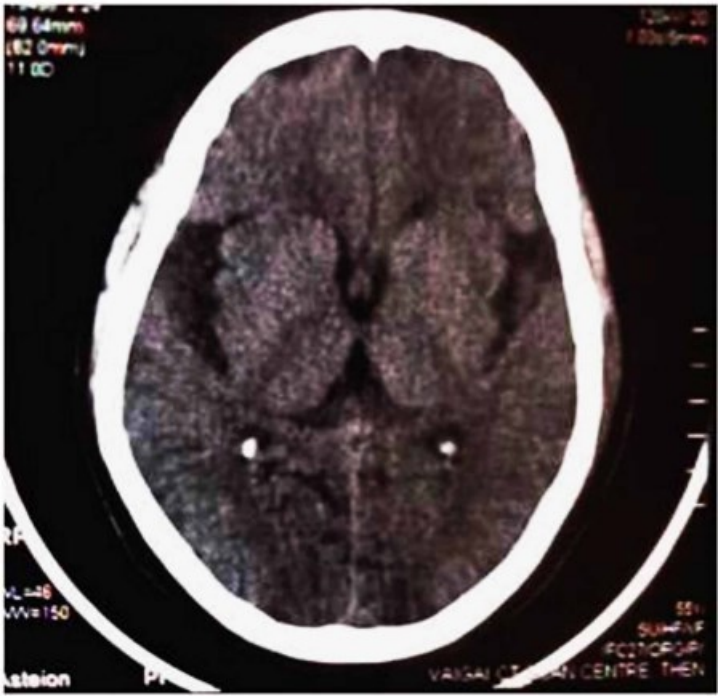


Figure 16. *MRI Image of a brain showing lesions* [24].

objects in real time, as indicated by the name itself [24]. At this point, it is worth to remind that MRI technique is also one of the mostly used for the AD diagnosis.

4. The importance of interpretable Alzheimer's disease diagnosis

Image interpretation may be understood in the sense of analysing an abstract scene that corresponds to the ambiguous goal of developing a 'visual sense for machines', which is as universal and powerful as that of humans [20]. Recent advances in artificial intelligence (AI) have started permitting into healthcare, among those the so-called Deep Learning (DL) methods, which consist of non-linear modules that can learn multiple levels of representations automatically from high dimensional data without any need of explicit feature engineering by humans [21].

Focusing on the topic of this work and so devoting the attention to MRI images for the identification of AD, it is worth to say that the cerebral changes (i.e., hippocampal and parietal lobe atrophy) visible from MRI images are considered to lack specificity for imaging-based AD diagnosis, but advanced machine learning paradigms such as DL offer ways to derive high accuracy predictions from MRI data [26].

There are DL approaches such as convolutional neural networks (CNNs), which may be implemented for MRI and multimodal data-based classification of cognitive status. However, despite the promising results, these models have yet to achieve full integration into clinical practice for several reasons. Briefly, the main reasons are the lack of external validation of DL algorithms, since most models are trained and tested on a single cohort, and the growing notion in the biomedical community that DL models are 'black-box' algorithms, which means that they neither elucidate the underlying diagnostic decision nor indicate the input features associated with the output predictions [26]. So, considering all these aspects, it is easy to realise the clinical potential of DL whose diffusion is slowed down by all these drawbacks, the overcoming of which would be crucial to harness the potential of DL algorithms to improve patient care and to pave the way for explainable evidence-based machine learning in the medical imagining community [26].

The primary purpose of AI tools for medical imaging is to aid (not replace) clinicians in their decision-making by combining multiple factors into a model that returns an actionable output but, without any explanation of this output, the utility of the model is limited as it does not unveil the reasoning process, limitations, and biases [21]. Interpretability of DL systems cannot only unravel any faulty processes within the algorithms, but also enables the discovery of other important information in the imaging data that otherwise might go unnoticed [21].

Nevertheless, unravelling the black-box nature of the DL systems is not only a legal and ethical requirement but is also essential for fostering clinical trust and for troubleshooting systems. Moreover, interpretability methods can also reveal new imaging biomarkers to understand the specifics of the DL model [21].

Overall, interpretability of DL networks can be defined as an attempt to explain the decision-making process of the model in a way that is understandable for the end-uses, and it refers to any technique that attempts to answer the question why the model is making a certain prediction for the medical image analysis tasks.

4.1 Most used interpretability techniques applied to deep learning algorithms

So far it has been said that AI solutions have the purpose to aid clinicians in performing their work more efficiently and accurately, and not to replace them, but this requires understanding on the side of the clinical experts and so interpretability can be incorporated during the design process of the deep neural network [21].

In this chapter there is the description of interpretation methods for neural networks and these methods, in general terms, visualize features and concepts learned by a neural network, explain individual predictions and simplify neural networks [27].

DL has been very successful, especially in tasks that involve images and texts such as image classification and language translation. The success story of deep neural networks began in 2012 and, since then, we have witnessed a Cambrian explosion of deep neural network architectures, with a trend towards deeper networks with more and more weight parameters [27].

To make predictions with a neural network, the data input is passed through many layers of multiplication with the learned weights and through non-linear transformations [27]. A single prediction can involve millions of mathematical operations depending on the architecture of the neural network, thus there is no chance that we humans can follow the exact mapping from data input to prediction because we would have to consider millions of weights that interact in a complex way [27]. To interpret the behaviour and predictions of neural networks, we need specific interpretation methods, which are described in the current chapter. Especially, there are two main reasons why it makes sense to consider interpretation methods developed specifically for neural networks instead of using model-agnostic methods (i.e., local models or partial dependence plots). First of all, neural networks learn features and

concepts in their hidden layers, and we need special tools to uncover them [27]. Secondly, the gradient can be utilized to implement interpretation methods that are more computationally efficient than model-agnostic methods that look at the model "from the outside" [27].

Post-hoc attribution-based techniques provide interpretation of model output and are most relevant during model training and at model deployment [28]. Since these post-hoc interpretability methods provide explanations for the predictions after the DL modes has been trained, they can offer local or global explanations. The former case identifies the attributes and features of a particular image that the DL model considers important for prediction, while the latter aim at identifying the common characteristics that the DL model considers when associating images with that particular class [21].

Figure 17 shows an overview of interpretability methods for DL solutions in medical image analysis. The black-box DL solution can be made more desirable for clinical use by incorporating interpretability during the design or execution phase.

In the following subparagraphs, there is the description of some of the most common interpretability methods found in literature.

4.1.1 Concept learning model

Concept representation learning can be carried out, amongst others, at concept level in which each feature representation is labelled with the concept that owns the feature. In particular, nowadays, DL models are trained to infer the label y directly from the input image x. However, it is usually not possible for radiologists to understand the reason behind the prediction of the DL models using the same high-level concepts c used to arrive at a diagnosis, hence, it is advantageous to explain the outputs of the model in terms of human-interpretable concepts [21]. This problem can be solved by first predicting these high-level concepts (such as semantic features) from the input image and then using these concepts to predict the label [21]. In other words, we approach this problem by revisiting the simple idea of first predicting an intermediate set of human-specified concepts c, then using c to predict the target y [29]. These models are trained on data points (x, c, y), where the input x is annotated with both concepts c and target y. At test time, they take in an input x, predict concepts $\hat{c}$ and then use those concepts to predict the target $\hat{y}$ (Figure 18) [29].

These models require concepts generated by experts as input during training time along with the image and label. Conceptual alignment deep neural networks (CADNNs) utilize some hidden neurons to learn human-interpretable concepts while other neurons are trained freely, and they achieve performance comparable to deep neural networks trained without

interpretability constraints [21]. There has been published many studies about this and it has been seen that the clinicians can intervene at test time to change the predicted value of the clinical concept to observe the effect on the final prediction and this intervention resulted in a performance improvement (Figure 19) [21].

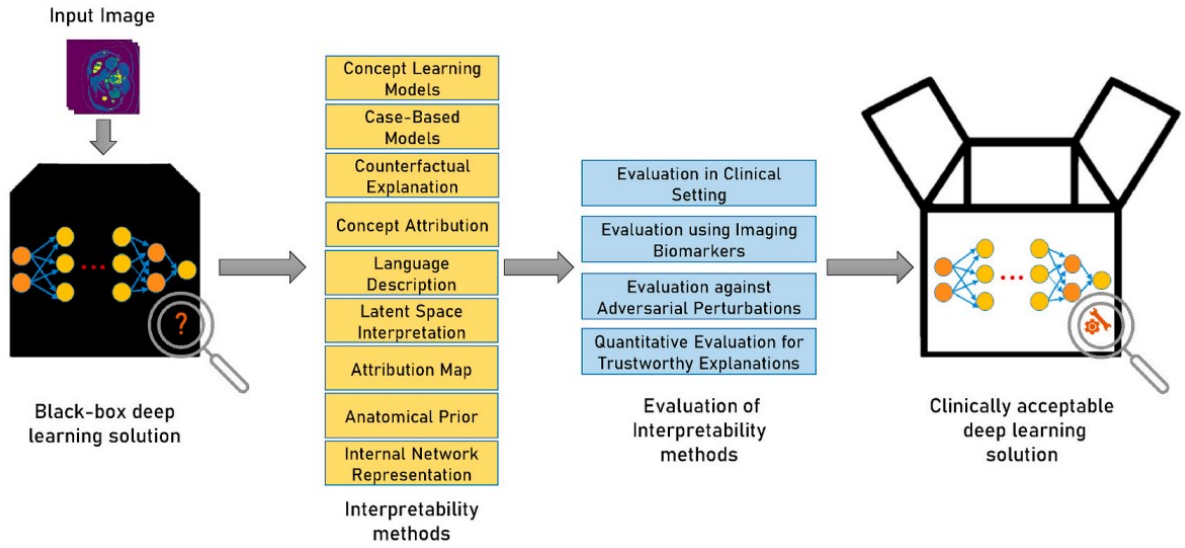


Figure 17. *An overview of interpretability methods for DL solutions in medical image analysis. The black-box DL solution can be made more desirable for clinical use by incorporating interpretability during the design or execution phase* [21].

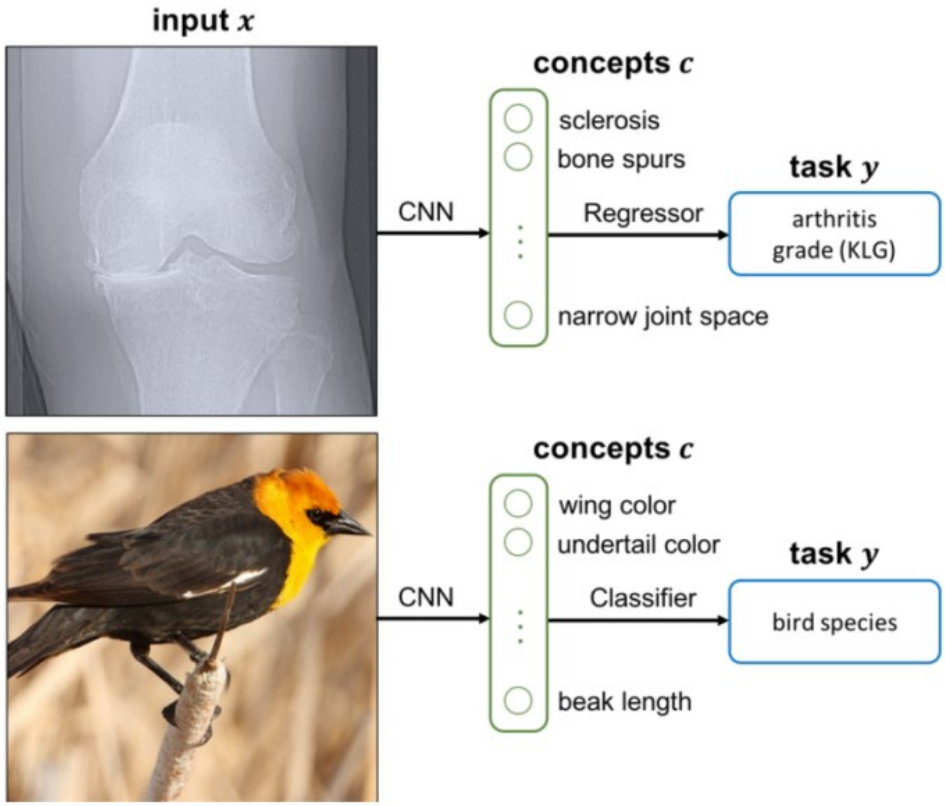


Figure 18. *Concept models first predict an intermediate set of human-specified concepts* $\mathbf{c}$*, then use* $\mathbf{c}$ *to predict the final output* $\mathbf{y}$*. Two applications are here illustrated: knee x-ray grading and bird identification* [29].

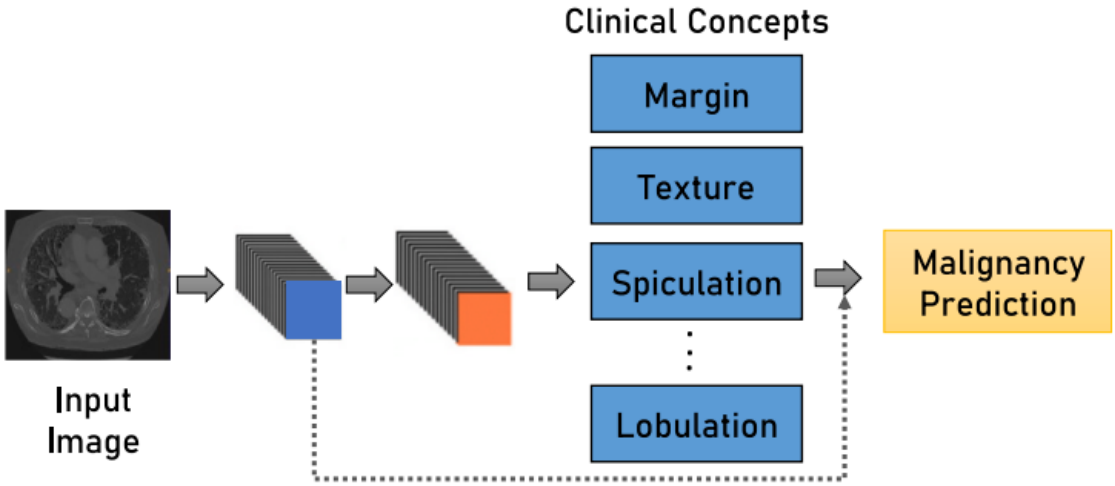


Figure 19. *Concept Learning Models first predict clinical concepts, and the final prediction is made using these human interpretable concepts. The final prediction is based either only on the clinical concepts or a combination of clinical concepts and other deep features (dashed line)* [21].

Capsule networks encode information (e.g. pose, scale, etc.) about each feature using vectorized representation in contrast to scalar features maps used in CNN [21]. In this case, low-level features extracted from CNN layers can be combined with predicted clinical concepts for diagnosis and the Hierarchical Semantic Convolutional Neural Network (HSCNN) consists of three modules for the extraction of generalised low-level features which are fed to the second module, which in turn classifies the presence or absence of five nodule semantic characteristics that reflect diagnostic features relevant for radiologists (e.g., texture, margin) and then third module predicts nodule malignancy based the low-level features from the first module and high-level visual attributes from the second module [21]. It has been shown that 2D explainable capsule network outperforms HSCNN [21].

4.1.2 Case-based model

Case-based models are another kind of network architecture for deep learning that naturally explains its own reasoning for each prediction, by comparing the features extracted from an input image against class discriminative prototypes [21]. In this case, the network architecture is chosen first, and afterwards one aims to interpret the trained model or the learned high-level features [30]. Prototype classification is a type of case-based reasoning that is inherently interpretable because the final predictions are made by taking a weighted sum of similarity scores between features extracted from input and prototypes [21].

The word “prototypes” is overloaded and has various meanings. In fact, in some cases, a prototype is very close or identical to an observation in the training set, and the set of prototypes is representative of the whole data set, while in other contexts, a prototype is not required to be close to any one of the training examples, and could be just a convex

combination of several observations [30]. In addition, in few-shot and zero-shot learning, prototypes are points in the feature space used to represent a single class, and distance to the prototype determines how an observation is classified [30]. For example, Prototypical Part Network (ProtoPNet) consists of a convolutional layer, followed by a prototype layer and then a fully connected layer (Figure 20) [21] and utilizes the mean of several embedded "support" examples as the prototype for each class in few-shot learning [30]. The convolutional layer consists of a trimmed standard CNN pipeline that acts as a feature extractor, the prototype layer takes patches from the convolution layer as input and learns class discriminative prototypes during training [21]. A similarity score is computed after comparison against each prototype and a fixed-size feature map is used for comparison with prototypes and then the fully connected layer then makes predictions based on these similarity scores [21]. These convolutional and prototype layers are trained first and the loss function comprises misclassification loss, cluster cost, and separation cost and then, in the second step, the fully connected layer is trained [21]. It is worth to mention the fact that Mohammadjafary et al. utilized ProtoPNet with DenseNet121 architecture for Alzheimer's disease classification [31].

4.1.3 Counterfactual explanation

A counterfactual explanation describes a causal situation in the form: "If X had not occurred, Y would not have occurred" [27]. For example: "If I hadn't taken a sip of this hot coffee, I would not have burned my tongue". Event Y is that I burned my tongue; cause X is that I had a hot coffee. Thinking in counterfactuals requires imagining a hypothetical reality that contradicts the observed facts (for example, a world in which I have not drunk the hot coffee), hence the name "counterfactual" [27]. More specifically, in deep learning, a counterfactual explanation is an image that is produced by applying minimal perturbations to the original image to bring a maximum change in the classifier's prediction and switch the predicted class of the original image [21, 27]. Moreover, counterfactual explanation not only helps in identifying the diseased area but also aids in understanding the changes that need to be made to switch the classifier's prediction (Figure 21) [21].

A simple and naive approach to generate counterfactual explanations is searching by trial and error and this approach involves randomly changing feature values of the instance of interest and stopping when the desired output is predicted [27]. However, there are better approaches than trial and error. First, we define a loss function, which takes as input the instance of interest, a counterfactual and the desired (counterfactual) outcome, then we can find the counterfactual explanation that minimizes this loss using an optimization algorithm [27].

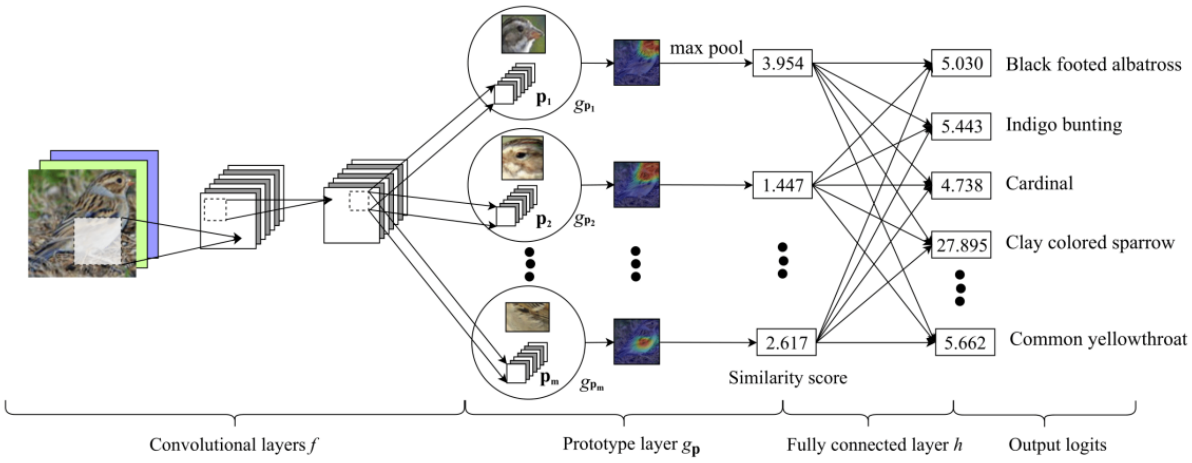


Figure 20. *ProtoPNet architecture* [32].

Many methods proceed in this way but differ in their definition of the loss function and optimization method.

Generative Adversarial Networks, consisting of a generator and a discriminator working in an opposing manner, are widely used for counterfactual images synthesis because of their ability to model target data distribution, but they can be difficult to train due to loss function instability and high sensitivity to hyper-parameters [21]. On the other hand, counterfactual images can by synthetized by perturbing the latent space of an autoencoder, but the resolution of the generated images is limited [21].

4.1.4 Concept attribution

Concept attribution provides global explanations for the deep neural network in terms of high-level image concepts [21]. In particular, Testing with Concept Activation Vectors (TCAVs) was proposed to generate global explanations for neural networks but, in theory, it should also work for any model where taking directional derivative is possible [27]. TCAVs method quantifies the influence of a high-level image feature on the decision of the model and a linear classifier is trained to differentiate between examples containing the concept of interest and random examples [21] and so, for any given concept, TCAV measures the extent of that concept's influence on the model's prediction for a certain class [27]. Since TCAV describes the relationship between a concept and a class, instead of explaining a single prediction, it provides useful global interpretation for a model's overall behaviour [27].

Concept Activation Vector (CAV) is simply the numerical representation that generalizes a concept in the activation space of a neural network layer [27] and is orthogonal to the classification boundary of the linear classifier [21]. TCAV method utilizes CAV and

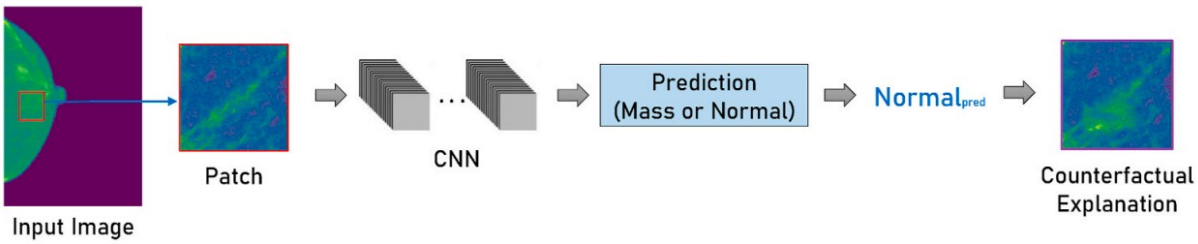


Figure 21. *Counterfactual explanation for breast mass prediction in an image patch from a mammogram. The DL model predicts the label of the input image patch as normal, and the counterfactual explanation is obtained by applying minimal perturbation to the input image to change the model's prediction* [21].

directional derivative to determine the importance of a particular concept for classification in terms of TCAV score [21].

After all, a CAV is trained by user-selected concept and random datasets. If datasets used to train the CAV are bad, the explanation can be misleading and useless and thus, a simple statistical significance test is performed to help TCAV become more reliable [27]. That is, instead of training only one CAV, multiple CAVs can be trained using different random datasets while keeping the concept dataset the same [27]. A meaningful concept should generate CAVs with consistent TCAV scores. Moreover, it is also suggested to apply a multiple testing correction method in case of multiple hypotheses [27].

It can be challenging to create a labelled dataset for different concepts to obtain the CAVs especially in the field of medical imaging. In particular, Automated Concept-based Explanation can be seen as the automated version of TCAV because it goes through a set of images of a class and automatically generates concepts based on the clustering of image segments [27].

4.1.5 Attribution map

Some approaches to model interpretation try to attribute the model output to different parts of the image input. In general, they produce heatmaps that describe the importance of different parts of an image to the model decision on a pixel-by-pixel basis [28], but these heatmap-based explanations do not offer any information on how these salient regions contribute to the prediction [21]. The available approaches generally fall into three groups: perturbation-based approaches, backpropagation- or gradient-based approaches and decomposition-based approaches [28].

Starting from the perturbation-based approaches to model interpretation, they involve altering different parts of an image and seeing how those perturbations change the output of the model [28]. The commonality of the approaches is the underlying idea that when important parts of

an image are perturbed, the output of the model is strongly affected, and when unimportant parts of an image are perturbed, the output of the model is unaffected [28].

The occlusion method as a means of performing model interpretation was first introduced in 2014 by Zeiler and Fergusfor the interpretability of deep neural networks [21] and this technique consists of systematically occluding parts of an image and monitoring how strongly the perturbation influences model output [28]. Image parts that, when occluded, strongly affect the output of the model are assigned high importance, while image parts that have little effect on model output when occluded are of low importance [28]. In general, multiple inferences need to be performed for the same input image, it can be computationally expensive to generate occlusion attribution maps if small portions of the image are perturbed and the resolution of the attribution map is constrained by the choice of the patch size that is altered at a time [21]. For the purpose of this work, it is of high relevance to mention that Tang et al. utilized occlusion maps for the interpretability of a model developed for the diagnosis of AD [21].

Local Interpretable Model-agnostic Explanations (LIME) was introduced by Ribeiro et al. in 2016 and it can be used to explain the prediction of any classifier, but for this review we only consider LIME in the context of image models [28]. LIME for images works by first identifying groups of contiguous pixels with similar intensities called superpixels and the image is then perturbed by turning subsets of superpixels "off" by replacing the value of all pixels in the superpixel with the mean intensity value of that superpixel [28]. In particular, each superpixel corresponds to an input feature and the model's predictions for these perturbed images are the target values [21]. Each perturbed image is weighted by its similarity with the input image using the cosine distance metric and a weighted regression model is trained to estimate the feature importance [21]. Nevertheless, like occlusion, changes in the model output due to the perturbation are used to identify how important each superpixel is to model output, and a heatmap highlighting the important superpixels is produced [28].

An advantage of LIME over occlusion is that LIME uses superpixels that are more likely to correspond to semantically different parts of an image, while occlusion perturbs image patches in a systematic, uniform way, ignoring possible semantic similarity between adjacent pixels [28]. LIME also uses fewer extreme perturbations than occlusion, as the intensities in the perturbed image region are replaced by the mean intensity instead of zeroes, however, there is nothing to prevent modification of either method to remove this difference [28].

Perturbation-based methods can alter parts of the image that are not understandable in clinical term, therefore, there is a need for meaningful perturbations [21]. In literature, there has been

proposed Guideline-based Additive eXplanation that mitigates this problem by first generating features using rule-based segmentation and anatomical irregularities according to the set guidelines and then a perturbation-based analysis is performed to obtain understandable explanations in terms of feature importance [21].

The second approach mentioned above is the backpropagation- or gradient-based method, which is the method by which weights in a neural network are updated during the model training process [28]. Model interpretation methods do not actually update model weights as occurs during training, yet they rely on backpropagation to compute gradients, and these gradients are combined in different ways to visualize salient parts of an image [28]. In fact, gradient-based methods generate post-hoc attribution maps by utilizing gradients of backpropagation to identify the important parts of the input image for the prediction [21].

Saliency maps were introduced in 2013 by Simonyan et al. and they use gradients to visualize the classification of an image evaluated by a deep convolutional network [28]. In particular, these saliency maps highlight regions of the input image that the deep neural network considers important for prediction by computing the gradient of the output concerning the input pixels using backpropagation [21]. In the introductory paper, the authors offer two uses for saliency maps: class maximization visualization and image-specific class saliency maps.

Class maximization uses gradient ascent to produce an image that maximizes the activation of that class, and therefore can be interpreted as being most representative of that class [28]. Formally, class maximization finds an image I of class c for which a class score S_c is maximised:

$$argmax\ S_c(I) - \lambda \| I \|_2^2$$

where λ is a regularization parameter.

Image-specific class saliency maps are image- and class-specific heatmaps that represent the importance of individual pixels to the assignment of the image to a class, providing an assessment of which parts of an image are most important to the model [28]. Saliency mapping is sometimes also referred to as “sensitivity analysis”, but it should be noted that it is a separate technique from the perturbation-based methods. Here, the heatmap $Sal_c(x)$ for a class c is computed directly as the derivative of the model output score $F_c(x)$ with respect to each pixel in the input image x through backpropagation [28]:

$$Sal_c(x) = \frac{dF_c(x)}{dx}$$

Because of its simplicity, saliency mapping is one of the most widely implemented methods for model interpretation in medical imaging to date [28].

In 2018, an expansion of class saliency maps was proposed with the introduction of iterative saliency maps, whose objective is to identify less discriminative image regions that may have been ignored in the initial saliency map [28]. The method works by iteratively computing a saliency map, inpainting the most salient image regions identified, and computing the saliency map again, and this process repeats until the perturbed image is no longer classified as containing an abnormality, or a maximum number of iterations is reached [28]. Then, the final iterative saliency map is computed as a weighted sum of the saliency maps computed at each step.

Despite their popularity, saliency mapping has the drawback that it provides no indication as to whether a pixel provides evidence for or against a class, only that the classification is sensitive to that pixel [28]. Several authors have also noted that in binary classification settings, saliency maps lose their class specificity, because if a feature is important for distinguishing between two classes, it may be highlighted by a saliency map for both classes [28].

Guided Backpropagation (GBP) is an extension to saliency maps and the DeConvolution approach (DeConvNet) and it works by computing the gradient and setting the negative gradient to zero during backpropagation [21]. The difference between these approaches lies in how backpropagation through Rectified Linear Unit (ReLU) – an activation function commonly used in CNNs – activation layers of the network is handled. In general, during the forward pass, neurons with negative output are clamped to zero by ReLU by definition ($\mathrm{ReLU(x)} = \max(0, \mathrm{x})$). This idea was then extended to computing gradients in the backward pass by clamping to zero negative gradients [28]. Finally, GBP combines these two ideas, zeroing out signal through neurons that have either negative output during the forward pass or negative gradient during the backward pass and this produces a heatmap that highlights only pixels that provide positive evidence for a classification [28].

GBP was evaluated as a method for visualizing AD diagnosis on brain MRI, but it was found that the visualizations produced by guided backpropagation are less discriminative than those produced by other methods [28].

Class Activation Maps (CAMs) were first introduced in 2016 and they localize class-specific image regions that the model considers important for classification [21]. To generate CAMs, a global average pooling layer is added after the last convolutional layer and the output of the global average pooling layer is then linearly combined to produce class predictions [21]. CAM for each class is then obtained by taking a weighted sum of the last convolutional layer activations [21]. The class activation map $\mathrm{CAM_c}(x)$ for a class c and image x is defined as:

$$\mathrm{CAM_c}(x) = \sum_k w_k^c \, f_k(x)$$

where w_k^c are the weights for class c in the final network layer, and $f_k(x)$ is the corresponding feature map prior to global average pooling [28]. Thus, $\mathrm{CAM_c}(x)$ is a class-specific heatmap that indicates discriminative image segments [28]. Multi-Layer Class Activation Maps (MLCAM) is an extension of CAM that can be incorporated at different CNN layers [21].

A drawback of class activation mapping is that it places some restrictions on network architecture as it requires a global pooling layer, followed by a fully connected layer as the last layers before the output layer [28]. However, in order to address this limitation of CAM, gradient-weighted class activation maps (grad-CAM) were introduced such that they produce visual explanations that do not require re-training or changes to the architecture like CAMs and they allow to explain activations in any layer of the network [21]. In grad-CAM, the weights are the gradients of the class score with respect to each feature map, instead of requiring that the weights be taken from a fully connected layer [28].

The Integrated Gradient mitigates the saturation problem of gradients [21]. In this method, it is critical to select a baseline that corresponds to a near-zero score, thus a complete black image is a suitable choice for a baseline [21]. The gradients are aggregated for all the points occurring in small steps between the input and baseline [21].

The third approach previously mentioned is the decomposition-based method for model interpretation, which seeks to decompose the prediction of the model to a heatmap that describes how much each pixel contributes to the prediction [28]. Whereas perturbation- and gradient-based methods for interpretation highlight parts of the image that, if altered, affect the prediction of the model, decomposition-based methods identify parts of the image that directly provide evidence for the model decision [28].

Layer-wise relevance propagation (LRP) was introduced in 2015 and it does not rely on gradients to generate the heatmap, but work by computing relevance scores that distribute the output of the final layer amongst nodes in the previous layer [28]. This process continues recursively until the input layer of the network is reached, producing a relevancy score heatmap that can be overlaid over the input image [28]. LRP is a method based on pixelwise decomposition of non-linear classifiers that generates a heatmap by evaluating a relevance score [21]. LRP ensures that the total relevance for all the layers starting from the classification output $f(x)$ to the input layer is the same [21]. Under these constraints, a neuron is highly relevant if it has a high activation and a high contribution for a neuron of the next layer that have a high relevance score [21].

LRP has been used in some medical imaging applications, among which it is important to mention its use to interpret CNN-based AD diagnosis on MRI. The authors find that LRP heatmaps from their trained model highlight the hippocampal volume, which has been used to diagnose AD and predict disease progression (Figure 22) [28]. They also compared LRP to guided backpropagation and concluded that LRP may be more valuable than guided backpropagation for their task because the difference in heatmap scores between Alzheimer's disease and healthy controls was more evident for LRP [28].

4.2 Most used interpretability techniques applied to Alzheimer's disease

The purpose of this paragraph is to describe the interpretability techniques mainly used in the application field of AD starting from medical images and, mainly, from MRI images. However, since this research topic is quite innovative, it is difficult to find a considerable number of studies about it. Nevertheless, according the recent and general review of interpretability methods applied to medical images by Salahuddin et al., it appears that Mohammadjafari et al. utilized ProtoPNet with DenseNet121 architecture (a case-based model) for AD classification [21]. Instead, Baumgartner et al. proposed an attribution method based on Wasserstein Generative Adversarial Networks (WGAN – counterfactual explanation)

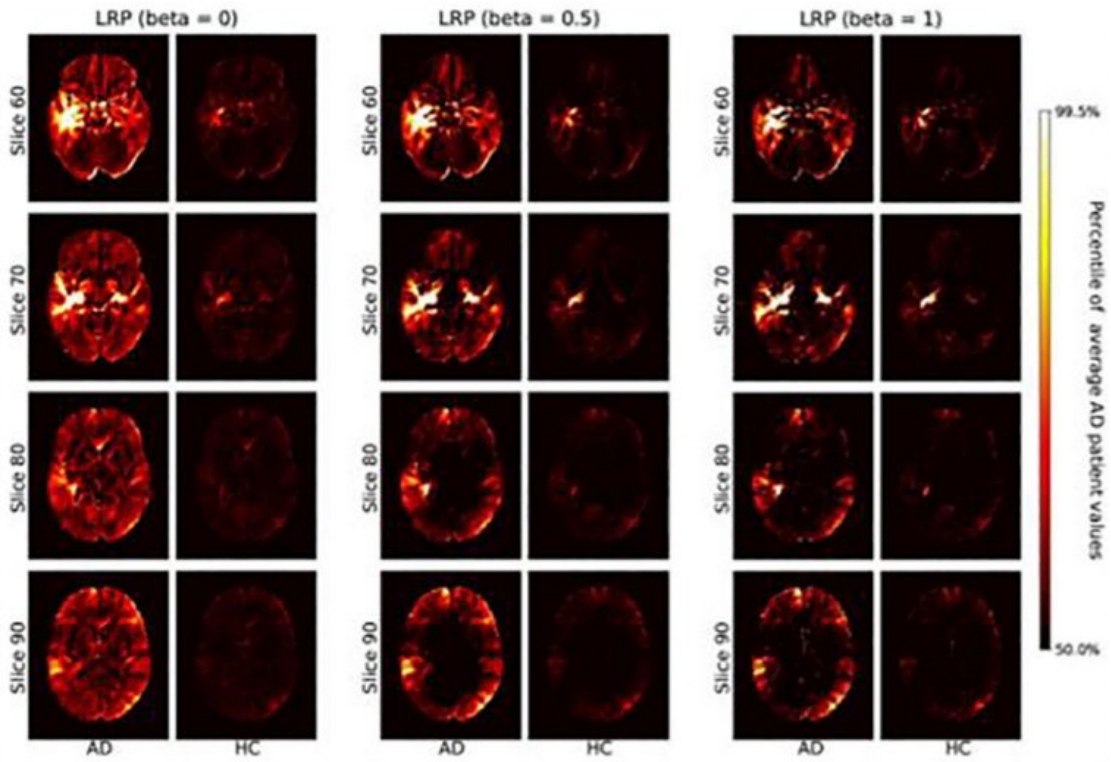


Figure 22. *Layerwise relevance propagation for model interpretation in visualizing evidence for AD on brain MRI* [28].

that requires a baseline class y and a class of interest x. WGAN estimates a function $M(x_i)$ such that when it is added to the image from class x_i, it becomes indistinguishable from class y. This attribution technique was validated on a synthetic dataset and MRIs from patients with AD and mild cognitive impairment [21]. In addition, Tang et al. utilized occlusion maps for interpretability of a model develop for the diagnosis of AD, while Boehle et al. demonstrated that LRP can be used to explain AD classification in 3D CNNs with high inter-patient variability and quantitatively showed that LRP relevance maps correlate with clinical knowledge [21]. Moreover, it is important to mention that Eitel and Ritter carried out a quantitative comparison of four attribution methods for Alzheimer's disease classification and showed that LRP and GBP produce the most coherent explanations.

4.2.1 Literature review

In this section, there is a review and summary of the different studies based on DL, anatomical MRI for AD classification and interpretability techniques.
First of all, it has been searched for articles on PubMed and Scopus published up to date and the flowchart of the whole literature research is reported in Figure 23. In particular, the query contains words linked to the following concepts: Alzheimer's disease, deep learning, interpretability / explainability, MRI and, in a second moment, also the concept of 3D was added. The words matching these concepts were identified in the abstract and titles of the documents and so the queries were the following:

PubMed query:
("Alzheimer's disease" [Title/Abstract] OR "Alzheimer's" [Title/Abstract] OR "Alzheimer" [Title/Abstract]) AND (“magnetic resonance” [Title/Abstract] OR “MR” [Title/Abstract] OR “magnetic resonance imaging” [Title/Abstract] OR “MRI” [Title/Abstract]) AND ("Deep Learning" [Title/Abstract] OR "Convolutional Network" [Title/Abstract] OR “CNN” [Title/Abstract] OR "Neural Network" [Title/Abstract]) AND ("interpretability” [Title/Abstract] OR “explainability" [Title/Abstract])

Scopus query:
TITLE-ABS("Alzheimer's disease" OR "Alzheimer's" OR "Alzheimer") AND TITLE-ABS(“magnetic resonance” OR “MR” OR “magnetic resonance imaging” OR “MRI”) AND TITLE-ABS("Deep Learning" OR "Convolutional Network" OR "CNN" OR "Neural Network") AND TITLE-ABS("interpretability” OR “explainability")

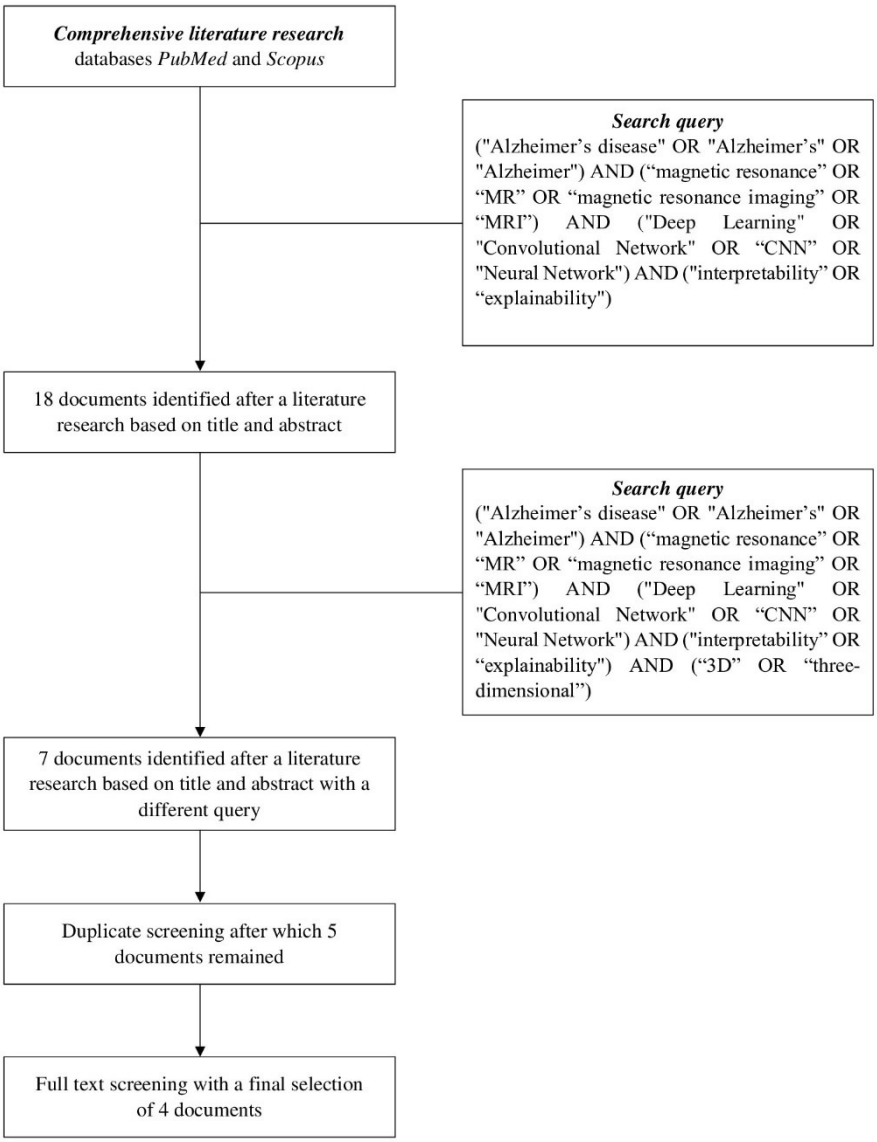


Figure 23. *Flowchart of the literature research.*

Typing just the first four concepts, 7 results were obtained from PubMed, while 11 results from Scopus, thus a total of 18 documents were found.

Instead, also adding the last concept, the queries were the following:

PubMed query:

("Alzheimer's disease" [Title/Abstract] OR "Alzheimer's" [Title/Abstract] OR "Alzheimer" [Title/Abstract]) AND ("magnetic resonance" [Title/Abstract] OR "MR" [Title/Abstract] OR "magnetic resonance imaging" [Title/Abstract] OR "MRI" [Title/Abstract]) AND ("Deep Learning" [Title/Abstract] OR "Convolutional Network" [Title/Abstract] OR CNN [Title/Abstract] OR "Neural Network" [Title/Abstract]) AND ("interpretability" [Title/Abstract] OR "explainability" [Title/Abstract]) AND ("3D" [Title/Abstract] OR "three-dimensional" [Title/Abstract])

Scopus query:

TITLE-ABS("Alzheimer's disease" OR "Alzheimer's" OR "Alzheimer") AND TITLE-ABS("magnetic resonance" OR "MR" OR "magnetic resonance imaging" OR "MRI") AND TITLE-ABS("Deep Learning" OR "Convolutional Network" OR "CNN" OR "Neural Network") AND TITLE-ABS("interpretability" OR "explainability") AND TITLE-ABS("3D" OR "three-dimensional")

In this case, only 3 results from PubMed and 4 results from Scopus were obtained, thus a total of 7 documents were found. In order to make the search more focused on this work, as a first attempt, only the results including the 3D concepts were used. However, it has been noticed that, among the 7 articles, there were some duplicates, which were removed and so 5 records were identified. However, one more document was discarded because, although the abstract and title included all the concepts of the query, that method was a "black box", lacking sufficient interpretability to explain the exact reason for better or worse results in a particular case (Figure 23). Tables 1 and 2 show an overview of the 4 records with their characteristics.

Study	***Dataset***	***Modality***	***Task***	***Section***
Shahamat et al., 2020	ADNI, ABIDE	MRI	AD and autism classification	4.2.1.1
Dyrba et al., 2020	ADNI	T1-weighted volumetric MRI	Comparison of algorithms for generating heatmaps to visually explain the learned patterns of AD classification	4.2.1.2
Guan et al., 2021	ADNI-1, ADNI-2	MRI	Early MRI-based diagnosis of AD	4.2.1.3
Turkan et al., 2021	ADNI	MRI, PET	AD identification	4.2.1.4

Table 1. *Overview of the 4 studies found in PubMed and Scopus. ADNI stands for Alzheimer's Disease Neuroimaging Initiative. ABIDE is Autism Brain Imaging Data Exchange. GAMB method is the Genetic Algorithm based Brain Masking. sMRI is the structural MRI. SHAP stands for Shapley Additive exPlanations.*

Study	*Neural Network*	*Interpretability method*	*Accuracy*	*Area under the curve (AUC)*	*Section*
Shahamat et al., 2020	3D CNN	GABM method	85%	N.A.	4.2.1.1
Dyrba et al., 2020	3D CNN	Deep Taylor decomposition, LRP, Grad-CAM and guided back-propagation	75.2%	0.93	4.2.1.2
Guan et al., 2021	3D CNN (ResNet)	score-CAM	87.18%	0.94	4.2.1.3
Turkan et al., 2021	3D CNN (VoxCNN8, VoxCNN16, VoxATT)	Occlusion, 3D Ultrametric Contour Map, 3D Gradient-Weighted CAM, SHAP	83%, 87%, 92% (VoxCNN8, VoxCNN16, VoxATT, respectively)	0.87, 0.91, 0.94 (VoxCNN8, VoxCNN16, VoxATT, respectively)	4.2.1.4

Table 2. *Overview of the 4 studies found in PubMed and Scopus. CNN stands for Convolutional Neural Network. GAMB method is the Genetic Algorithm based Brain Masking. MRI is the structural MRI. SHAP stands for Shapley Additive exPlanations.*

4.2.1.1 Shahamat et al., 2020

As it is clear from the Table 1, the task of this paper was the AD classification and the autism classification. However, for the purpose of his work, only the parts connected to the AD are summarised. First of all, an overview of the whole proposed framework is proposed in Figure 24, and it consists of four major steps: (1) pre-processing, (2) classification, (3) genetic algorithm based brain masking (identification of knowledgeable brain regions), and (4) experimental results.

Regarding the AD, in this paper, a set of 140 MRI scans (70 normal control subjects and 70 AD) has been downloaded from AD Neuroimaging Initiative (ADNI) site and used for the experiments. All these MRI scans were pre-processed, registered and normalised and then cropped to an $80 \times 80 \times 80$ voxels sub-volume with the brain centred. Then, a 3D-CNN model (Figure 25) was designed and trained from scratch for classification and the input layer

has size $80 \times 80 \times 80$ to accept pre-processed MRI scans, which is then followed by a dropout layer with keep probability of 50% to reduce over fitting. The first convolutional layer consists of 8 filters with size $5 \times 5 \times 5$. After applying a ReLU activation function to the convolution's results, a max-pooling operator is used with window size $2 \times 2 \times 2$ and it reduces the input scan size to $40 \times 40 \times 40$. The second convolutional layer has 16 filters with size $3 \times 3 \times 3$ and, after applying ReLU function and max-pooling operator, the data size is reduced to $20 \times 20 \times 20$. The third convolutional layer has 32 filters with size $3 \times 3 \times 3$ and, after applying ReLU and pooling on the results, the data size is reduced to $10 \times 10 \times 10$. Subsequently, two fully connected layers are used for data classification: the first fully connected layer has 32000 input and 1024 output neurons, and it is followed by a dropout layer with keep probability 50%; The second fully connected layer has 1024 input and 2 output neurons (same as the number of classes). Finally, a softmax layer and a classification layer are used to provide labels for the input MRI scans. Regarding the ADNI dataset, in a 5-fold cross validation mode, the accuracy was 0.85, thus acceptable.

For identification of knowledgeable brain regions, in this paper, the Genetic Algorithm based Brain Masking (GABM) method was proposed and all the brain regions were involved in model training, but only a number of them were selected by the GABM method. By changing the GABM parameters, 4 different experiments were performed to each dataset, but just the ones regarding ADNI are of interest for the purpose of this work. Nevertheless, any change in the parameters will affect the number of involved brain regions in the final mask. In fact, the proposed GABM method is applied to discover most important brain regions and discarding the redundant part of the brain MRI scans to the disease under study. Precisely, the test accuracy of 3D-CNN + GABM method on the ADNI dataset was 0.85 when the parameters of the GABM method were $\alpha = 0.03$ and $\beta = 0.97$ and this accuracy was obtained using only 41 brain regions, which is equal to the obtained accuracy using all 96 brain regions. Finally, the proposed GABM method has found 6 to 65 brain regions in ADNI dataset with respect to the model parameters. This experiment proved that some brain regions may be redundant and the proposed GABM can find them properly.

Overall, the results shown that besides the model interpretability, the proposed GABM method has increased the final performance of the classifier in some cases.

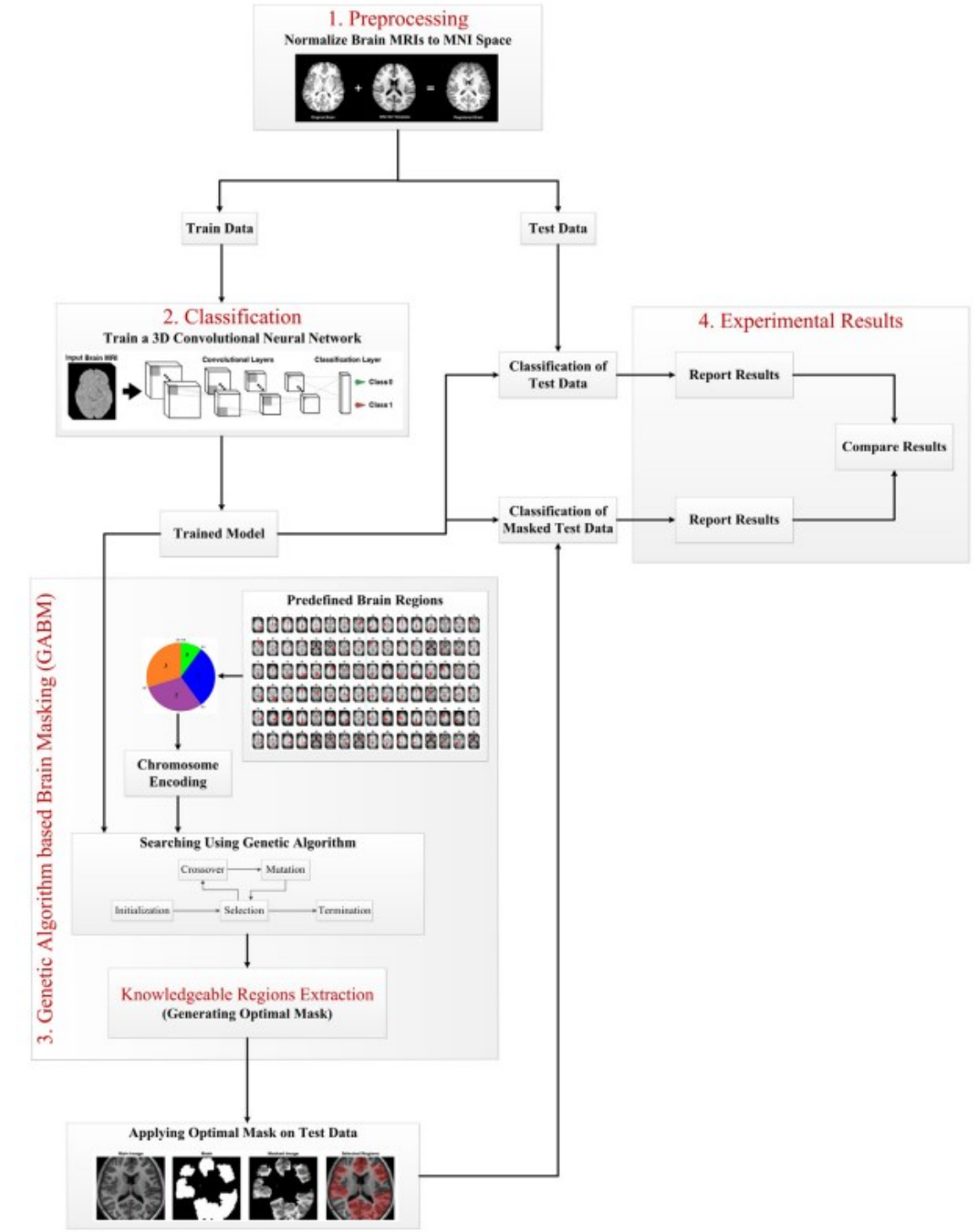


Figure 24. *Overview of the proposed framework* [33].

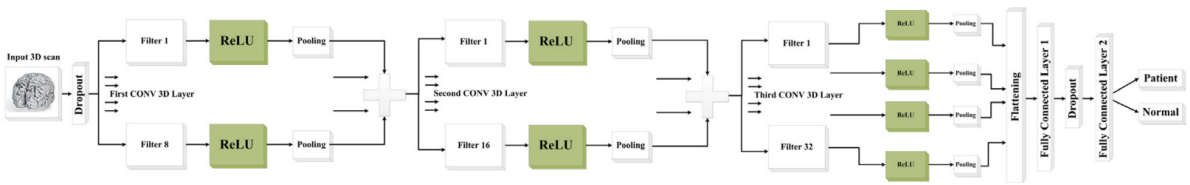


Figure 25. *The proposed 3D-CNN architecture for MRI classification* [33].

4.2.1.2 Dyrba et al., 2020

The purpose of this study was to make a comparison of algorithms for generating heatmaps to visually explain the learned patterns of AD classification. Also in this case, the MRI data were obtained from the ADNI and, in total, the sample included 662 cases consisting of 198 patients with AD dementia, 219 patients with amnestic MCI, and 254 cognitively normal controls. After having segmented the MRI scans into grey and white matter, having normalised and modulated them, the images were taken as input for the CNN (Figure 26), which has become the state-of-the-art technique for various image classification tasks. In this case, it is important to highlight that CNN model achieved excellent diagnostic accuracy for separating AD dementia from controls, comparable to other approaches from the literature, as well as for separating MCI patients from controls. Moreover, as computational complexity is considerably higher for 3D CNN models compared to 2D CNN models used for general purpose image detection tasks, there is a high potential of model overfitting, in contrast to a very limited number of MRI scans available for training. This problem was addressed by applying image pre-processing, by reducing the number of layers resulting in a shallower network, including three convolutional layers, with in total approximately 6400 parameters, and by using data augmentation to multiply the data available for training and to improve the stability and robustness of the model.

Various visualization methods were tested and approximately the same image regions were highlighted across them and, as expected, the hippocampus are showed the highest relevance for the AD and MCI patients. In conclusion, for clinically oriented research, deep Taylor decomposition and LRP with parameters $\alpha = 1$ and $\beta = 0$ rule showed the most promising relevance maps (network activation patterns) with strongest focus and less scatter and these approaches mainly showed positive relevance scores for the AD class and suppressed the negative relevance against AD.

Figure 26. *Convolutional neural network model layout proposed by Dyrba et al. (2020).*

4.2.1.3 Guan et al., 2021

In this paper, a deep learning framework for sMRI-based AD diagnosis was proposed and this framework accepts 3D sMRI scans as input and outputs diagnostic labels (Figure 27). The framework was evaluated on two independent datasets from ADNI (ADNI-1, ADNI-2) for AD classification and MCI conversion prediction. As a primitive feature extractor, a light-weight 3D CNN (based on ResNet) is used and, in order to tackle the trade-off between better representation learning and increased risk of overfitting, a parallel attention-augmented bilinear network is devised. Specifically, the parallel attention-augmented blocks model long-range interdependencies and asymmetrically project the learned features to lower dimensions. Finally, the compressed features of the parallel branches are combined using bilinear pooling to model localized feature interactions.

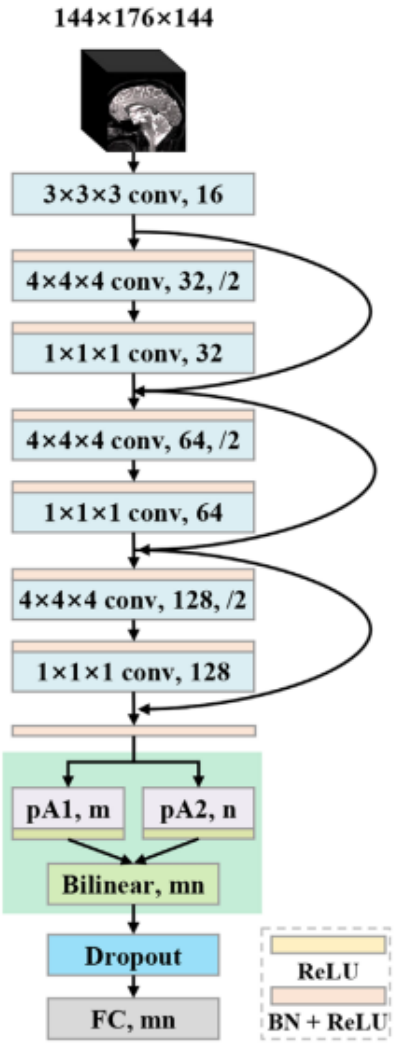


Figure 27. *Architecture of the proposed parallel attention-augmented bilinear network, where 'pA' refers to the parallel attention-augmented blocks* [34].

Taking a 3D brain image as input, a backbone network first extracts primitive features, which are then passed to parallel attention-augmented blocks for extracting long-range interactions. The network architecture consists of a root convolutional layer and three residual units. The root convolutional layer accepts the 3D images, with 3D kernel of size $3 \times 3 \times 3$ and 16 output channels. The next three residual units have the same structure except for the output channels. To cover regions of interest with a sufficiently large receptive field, we choose to use a large kernel size for convolutions. In addition, we use strided convolutions to down-sample the features. Feature down-sampling is achieved via strided convolutions. No average-pooling or max-pooling layer is used. Specifically, each residual unit has two convolutional layers: the first layer has kernels of size $4 \times 4 \times 4$ with stride 2; the next layer has kernels of size $1 \times 1 \times 1$ with stride 1. The number of output channels is doubled for the first layer while unchanged for the second layer. The layer with kernel size $4 \times 4 \times 4$ symmetrically applies zero paddings to ensure that the output feature map is exactly half the size of the input feature map. Before each convolutional layer, a batchnorm (BN) layer and a ReLU are cascared. After three residual units, the feature map is down-sampled eight times to $18 \times 22 \times 18$. Then, another combination of BN and ReLU is inserted before the next parallel attention-augmented blocks.

Score class activation mapping (score-CAM) was used to visualize the discriminative areas where the network focused on. The heat maps are obtained by a linear combination of activation maps and weights, which are forward passing scores on target class. In practice, an image is fed into the fully trained network and use the feature maps output by the pA-blocks to generate two different 3D heat maps. Then, the heat maps are upscaled to the same size as the input image.

Although the proposed framework achieved competitive diagnostic results, there is still room for improvement.

4.2.1.4 Turkan et al., 2021

In this paper, the purpose is the identification of AD by using ADNI dataset. In this study, the Koorolev's 3D VGG model (VoxCNN8) was taken as base model and it was extended by using more filters, resulting in VoxCNN16, in which the feature filter channels were doubled, and in VoxATT, in which is Dot Attention block was added to the VoxCNN16 after the convolution blocks and so a 3D dot attention mechanism was applied, inspired by the 2D dot attention. These architectures are represented in Figure 28.

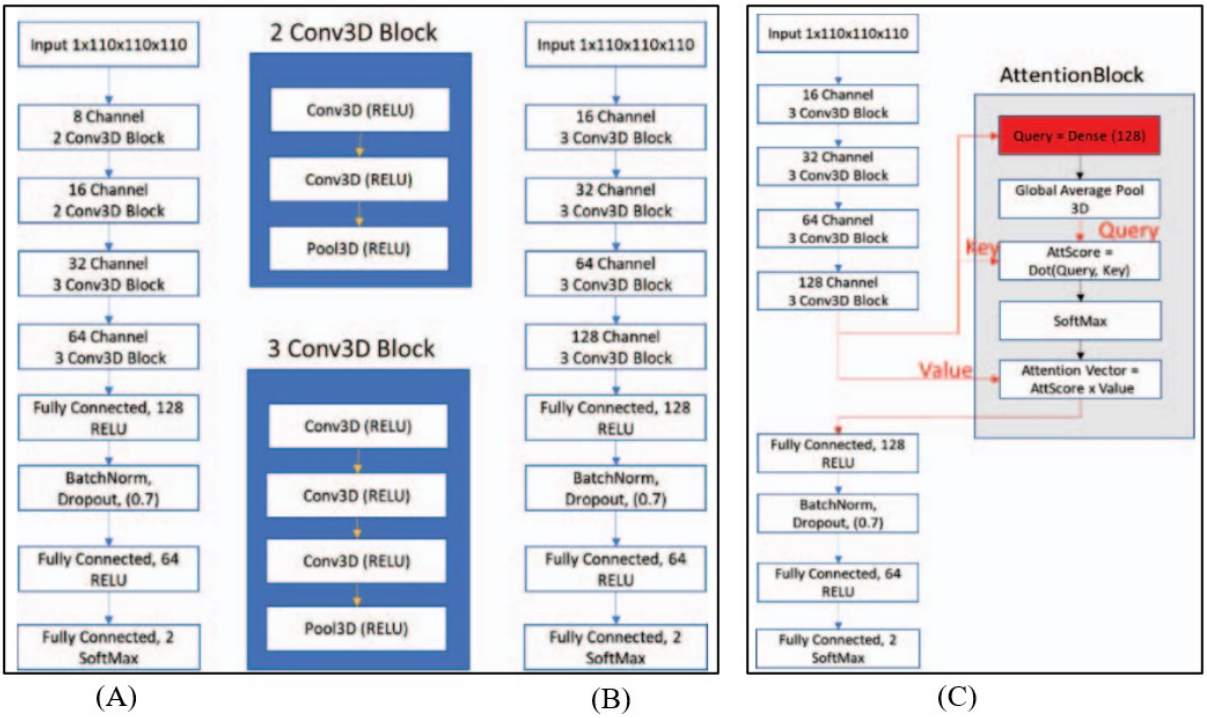


Figure 28. (**A**) *VoxCNN8 is a standard VGG architecture that starts with 8 Channels. It uses two 2Conv3D Blocks, followed by two 3Conv3D Blocks.* (**B**) *VoxCNN16 architecture starts with 16 channels and uses four 3Conv3D Blocks.* (**C**) *VoxATT architecture is designed by adding Attention Block to the VoxCNN16 model* [35].

Regarding the interpretation aspect, four different interpretability methods are compared in order to explain the predictions of the proposed models: occlusion, 3D Ultrametric Contour Map (3D-UCM), 3D Gradient-Weighted Class Activation Mapping (3D-Grad-CAM) and SHAP. The occlusion method failed to express the difference between target classes, 3D-UCM gave more detailed regions, while the results of the 3D-Grad-CAM method were not interpretable in the deep network models. Lastly, SHAP results showed more distinctive regions compare to the other methods. With attention, visual interpretability results became sharper and distinctive.

4.2.2 Discussion

In the previous chapters, an overview of the most useful concepts related to this work has been given and it has been understood that AD is one of the main leading causes of death worldwide and it is going to grow due to the increased life expectancy. Moreover, because of a lack of understanding of AD by patients and their family members, most patients suffer from moderate and severe stages of AD at the time of diagnosis and have missed the optimal intervention stage [36]. Hence, early diagnosis of AD might be fundamental in order to slow down the development of AD. In fact, while there is no cure for AD, early diagnosis and accurate prognosis may enable or encourage lifestyle changes, neurocognitive enrichment,

and therapeutic interventions that strive to improve symptoms, or at least slow down mental deterioration, thereby improving the quality of life [37].

Clinically, there are different forms of neuroimaging techniques available for AD diagnosis, including MRI. As a matter of fact, the structural MRI (sMRI) measurement is considered as a marker of AD progression, which can help to detect the structural abnormalities and track the evolution of brain atrophy, typical of AD [36, 38]. However, nowadays, to my knowledge, the AD identification process is still performed manually by specialists, which is expensive and time-consuming. To solve this issue, recently, with the rapid development and wide application of AI in the medical field, computer-aided diagnosis (CAD) of AD using neuroimaging may be an auxiliary method to assist physicians [36]. More precisely, with the continuous development of DL, several attempts based on DL have been employed to analyse the MRI data by constructing models avoiding manually extracting features. In fact, DL, in particular CNNs, has proved to be an effective method of feature extraction from images and has provided state-of-the-art solutions in different image understanding and recognition tasks [36]. Thus, it is well known that DL helps to solve such a complex diagnostic problem by leveraging hierarchical extraction of input data features to improve classification [39]. Although DL-based models have achieved great classification performance for AD diagnosis, it is still an undetermined since subjects' MRIs have relatively small differences in anatomic abnormalities, and it is necessary to dig out moderately subtle changes in disease progression from high denominational of MRI sequences data [39].

CNNs (DL method) are widely used for image analysis and analysis of complex data and so, among the available machine learning methods, CNNs have been increasingly used in the Alzheimer's biomarker identification task, given its power to learn discriminative representations hierarchically in an automated fashion [37, 40]. According to the literature, most of the proposed studies related to AD used 2D inputs, while studies that used 3D inputs focused basically on binary classifications and, with their work, Folego et al. in 2020 released one of the first models ready to use, encouraging open science and reproducible research and also setting a starting point for researchers working with 3D MRIs [40]. This is noticeable also from Table 2, in which is it clear that all the found studies concerning AD and 3D inputs used CNNs, even though of different types. In particular, for MRI classification, Shahamat et al. (2020) and Dyrba et al. (2020) proposed the 3D CNNs reported in Figure 25 and Figure 26, with an accuracy of about 85% (AUC not available) and 75.2% (AUC of 0.93), respectively. Furthermore, Guan et al. (2021) used another type of 3D CNN based on ResNet to first extract primitive features, which are then passed to a parallel attention-augmented bilinear

network for extracting fine-grained representations, giving an accuracy of 87.18% and an AUC of 0.94. Finally, Turkan et al. (2021) proposed different improvements of the 3D VGG model (VoxCNN8), which are the VoxCNN16 and the VoxATT. With an accuracy of 92% and AUC of 0.94, VoxATT model focused more on distinctive regions than the other models.

In recent years, several approaches have been introduced exploiting MRI data for distinguishing AD and its prodromal dementia stage and they can be categorized in four main categories: voxel-based methods, methods based on Regions-of-Interest (ROI), patch-based methods, and approaches that leverage features from whole-image-levels [41]. In literature, a few machine learning studies used extracted brain structures or cortical thicknesses, and some used 3D patches from predetermined locations across the brain, but not whole-brain MRI data, to predict mild cognitive impairment MCI to AD conversion and it seems that there a no published studies on DL to this prediction using longitudinal and whole-brain 3D MRI, with the only exception for the one of Ocasio et al., 2021 [37].

DL approaches CNNs, which may be implemented for MRI and multimodal data-based classification of cognitive status, despite the promising results, have yet to achieve full integration into clinical practice mainly because DL models are 'black-box' algorithms, which means that they still lack interpretability. Nowadays, much attention has been given in order to try to solve this issue and so different interpretability algorithms have been proposed in literature. However, since it is a novelty, only a few studies have been published, which are the ones reported in the previous subchapters. By analysing them, it is possible highlight that they are quite different one another since different interpretability techniques are described and discussed, thus direct comparison between them it is hard to be made. Because of that, it can be said that still there appears to be confusion about which the most accurate and reliable interpretability technique is. In fact, as it is summarised in Tables 1 and 2, many different methods have been proposed, providing different result. So far, according to what reported in literature (Tables 1 and 2), it seems that one of the best methods applicable to AD is the LRP technique, which can be used to explain AD classification in 3D CNNs with high inter-patient variability and these LRP relevance maps correlate with clinical knowledge. In fact, according to Dyrba et al. (2020), Grad-CAM and guided backpropagation methods showed more scattered activations or random areas, which is something confirmed by Turkan et al. (2021) in which it is reported that the results of the 3D Grad-CAM method were not interpretable in the deep network models, while SHAP results showed more distinctive regions compared to the other proposed techniques. Instead, Guan et al. (2021) used score-CAM to generate the heatmaps, which only highlighted discriminative regions, but this interpretation is still too

coarse to make DL techniques enter into clinical practice to support decision-making procedures. Thus, as mentioned above, according to Dyrba et al. (2020) it seems that the best technique applicable to AD is the LRP because the clinical maps correlate with clinical knowledge.

Obviously, an interpretable AD diagnosis using DL techniques is an extremely new research topic and this is highlighted also by the fact that, during the literature research, just a few documents were found.

5. Convolutional neural networks applied on volumetric magnetic resonance scans

Before going into the details of what has been performed in the experimental part of the thesis, after having introduced all the theoretical concepts and having understood the current situation in literature, it is also good to make a brief overview of the technologies and tools used to carry out this work.

5.1 Data and methodology

5.1.1 Data selection

OASIS-3 dataset is the most recent subset of the OASIS database and it is the one which has been used in this work. Since only one raw T1w 3D sMRI scan was considered for each subject, this dataset is characterized by 275 scans: 145 AD and 130 CN. Each scan contains 256 stacked slices, having an original resolution of 176 pixels × 256 pixels, a thickness of 1 mm and a pixel size of 1 mm.

AD scans belong to 74 anonymized women and 71 anonymized men ranging from 52 to 95 years in age, whereas CN scans belong to 81 anonymized women and 49 anonymized men ranging from 45 to 86 years in age. In particular, all the scans were acquired with 1.5 T and 3.0 T Siemens scanners and stored as Digital Imaging and Communications in Medicine (DICOM) files, then converted to compressed Neuroimaging Informatics Technology Initiative (NIFTI) files and finally to NumPy arrays.

5.1.2 Environmental setup

Google Colab, or "Colaboratory", allows writing and running Python in your browser making available to anyone (with certain limitations) free access to GPUs for a maximum of 12 consecutive hours [42]. Colab notebooks are saved on your Google Drive account and are accessible and executable by anyone with permission to access them, facilitating code sharing. In the free version, a Colab notebook provides a CPU with 12 GB of RAM and 97 GB of disk space on the instance of the virtual machine to which you connect (of which 31 GB already occupied by the entire Colab environment and libraries already pre-installed) and also a GPU with 12 GB of RAM and 60 GB of disk space (of which always 31 GB already occupied). Using a Colab Pro account you can keep an instance active for up to 48 consecutive hours and

you have access to more resources: 225 GB of disk space for the CPU with 12 GB of RAM, 147 GB of disk space for the GPU and always 12 GB of RAM and 225 GB of disk space for the TPU with 33 GB of RAM, in addition to a higher "priority" in accessing the required resources and a longer time required before being disconnected due to inactivity (about 2-3 hours from the experiments carried out).

As already understood, Goodle Colab allows writing and executing code in a multitude of languages, including mainly Python code. Python is a high-level object-oriented programming language widely used in numerical computing, data analysis, distributed applications and scripting and it is continuously updated and has many libraries that make it suitable for the most varied uses [43]. In particular, the focus is on this language since the whole code of the current thesis was implemented in Python.

NumPy is an open-source library written for the Python programming language (hence the name of the library) and is one of the fundamental and most used libraries for scientific computation of data in Python [44]. In the field of ML and above all image analysis, NumPy is practically universally used for the representation of images through n×m matrices, where n×m are the dimensions of the image, and many other types of data.

Keras is another open-source library written for Python specifically for ML and neural networks. The aim of Keras is to provide a clear and easily usable interface by a human operator for the creation and development of deep neural networks, providing clear and concise APIs and trying to minimize the number of operations required to develop a neural network, from its creation to its training up to the final phase of testing and fine-tuning of the parameters [45].

The ones illustrated are the two main libraries used for the implementation of the code for this work and they are already pre-installed and ready for use in the Colab environment as well as being automatically and constantly updated to their most recent version.

5.1.3 3D convolutional neural network

5.1.3.1 Pre-trained C3DKeras

The 3D convolutional neural network (3D CNN) model used is called C3D for Keras which, as the name suggests, is an adaptation of a convolutional 3D network model originally developed for Caffe in the paper by Du et al. [46] in order to make it compatible with Keras. It is in turn a modification of the BVLC caffe model which was trained on the Sports-1M dataset which contains video clips of various sports in order to recognize the type of sport

contained in each video clip. In this work, the same pre-trained C3DKeras network provided was adapted to the case of the AD in order to then be able to apply interpretability techniques. Figure 28 shows the model summary.

The C3D has been trained to take 16 slices as input, where each slice consists of a 112×112 pixel RGB image, thus the dimensions required by this network are 16×112×112×3. It was therefore necessary to adapt the original dimensions of the dataset to these dimensions.

The feature extractor of the network is composed of 5 layers of 3D convolution to which as many levels of MaxPooling3D are alternated, in order to reduce the amount of data to be processed while trying to lose as little information as possible. At that point, through a layer called Flatten, all the features obtained are flattened into a vector of dimensions 8192×1.

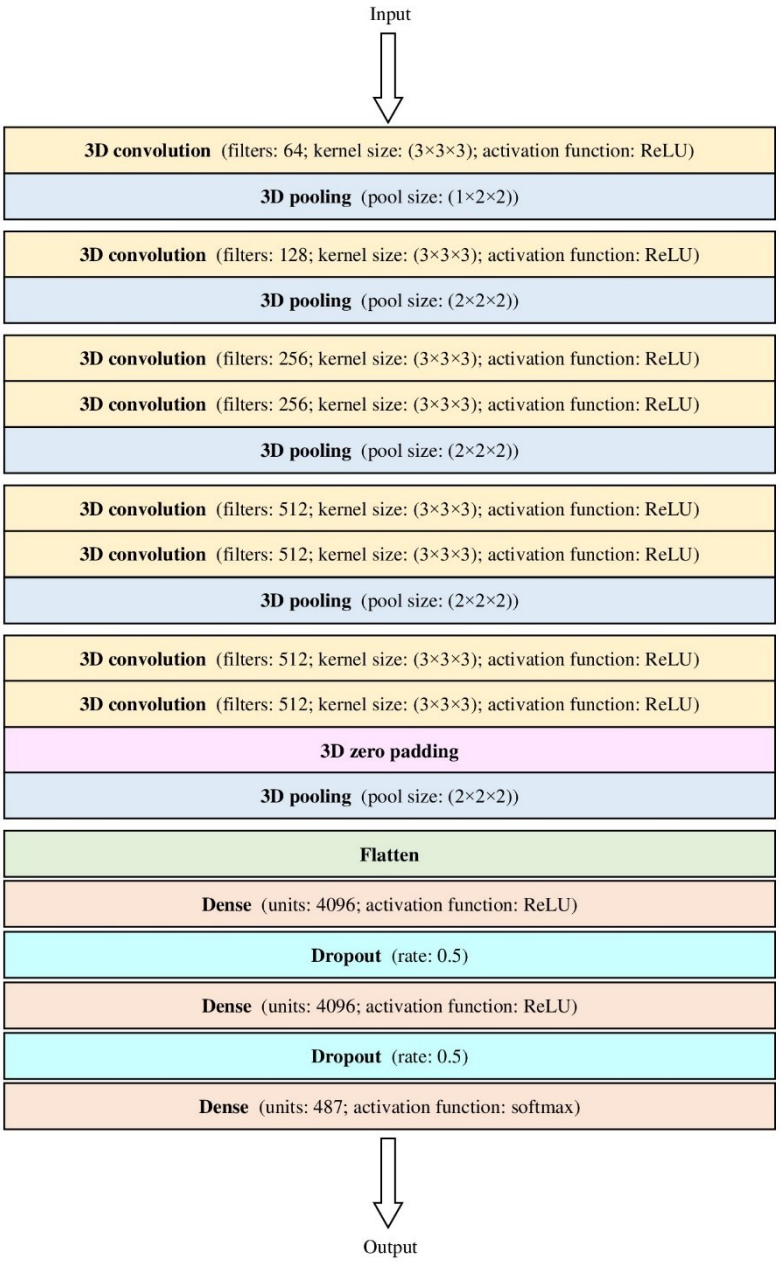


Figure 28. *Model summary of the C3DKeras adapted to the case of Alzheimer's disease.*

At this point, since it is a binary and not multiclass classification, the number of nodes of the fully connected layers has been reduced and, above all, the number of output layers has been reduced to 2 in order to obtain, as output from the network, the probability of a sample belonging to the two classes to be evaluated.
Furthermore, the weights of the pre-trained network have not been changed, otherwise the training of the network is carried out from scratch and the previously learned features are lost. To do this it is possible to "freeze" certain layers, so that their weights are not updated during the training phase. In this way, it is possible to load into the model the weights obtained from a previous training and be sure that those weights will never be modified, while during the training the weights of the new layers to be trained for the new task to be performed will be updated.

5.1.3.2 Data preparation

The first step to be performed was build the dataset according to the shape required by C3DKeras. In fact, first of all the slices were resized to 112×112, which are the x and y needed by C3DKeras, leaving an intra-slice distance of 1 mm. Then, a zero padding was executed to add black slices in order to match the dimensions (16×112×112×3) required by the C3DKeras. So, in order to extract blocks of 16 slices, the zero padding was performed. In particular, if the number of slices in the array is 16 or a multiple of 16, the algorithm takes blocks of 16. Otherwise, if the number of slices is not 16 or its multiple, but it is lower, the algorithm adds the remaining slices through the zero padding. In this way, blocks of 16 are obtained. Finally, in order to match the dimension relative to the channels, the conversion from 1 channel (gray scale) to 3 channels RGB. In fact, since the C3DKeras is pre-trained on sport's video, it works on coloured videos, thus it was necessary to perform also this kind of conversion. In this way, the original dataset has been adapted to the dimensions of 16×112×112×3 required by the C3DKeras. In fact, it was obtained an array divided in 4 groups of 16 slices each (4×16×112×112×3).
Moreover, while preparing the dataset, only the 50 slices centred on the hippocampus were used mainly for two reasons. First of all, according to the neurologists, the right and left hippocampal areas are the ones mainly hit by the AD especially in the beginning stages of the disease. Thus, the idea was that also our algorithm should mainly focus on that part of the brain. On the other hand, while selecting the 50 most significant slices, the load of the dataset was lowered in order to better manage the memory.
All this procedure was performed both on AD and CN data.

The whole dataset was split into training set, validation set and test set. In particular, for the training set 176 scans were obtained, each of them with dimensions of 4×16×112×112×3. Instead, for the validation set, there are 44 scans because the 20% of the training set was used. Finally, the test set is composed of 55 scans because it is the 20% of the whole dataset.
At this point, it was necessary to fit the input shape to the C3DKeras from 4×16×112×112×3 to 16×112×112×3, for the training, validation and test sets.

5.1.3.3 Neural network classification

The hyper-parameter selection has high influence on the model performance and so they need to be investigated and tuned. In particular, the following hyper-parameters were used:

- Loss function: binary crossentropy;
- Adam optimizer;
- Dropout rate: 0.5;
- 50 epochs;
- Patience: 5;
- Batch size: 1;

Usually, these hyper-parameters are changed in case of over-fitting or under-fitting. In the former case, the validation performance is significantly higher than the training performance, while the latter case is the opposite. Nevertheless, in both cases, the model does not correctly work, and it does not make a good generalization.
In the case of this work, the values are quite acceptable.

5.1.4 Time-distributed convolutional neural network

5.1.4.1 End-to-end VGG16 + ConvLSTM

The second convolutional network model used is the VGG16 + ConvLSTM (Figure 29). First of all, this is not a pre-trained neural network like the C3DKeras, thus it does not require a fixed input shape of the data. Moreover, in its architecture, the VGG16 2D CNN passes the image through a stack of convolutional layers and spatial pooling is performed by five max-pooling layers, which follow some of the convolutional layers. On top of the features extracted by VGG16, Convolutional Long Short-Term Memory (ConvLSTM) is used for classification. At this point, a ConvLSTM layer is characterised by 8 convolutional filters with a kernel of 3×3 and it was chosen because both spatial and temporal AD features are involved in the classification. In addition, there is a Dropout layer, thanks to which a few

unites are randomly removed from the model during the training phase, reducing the overall complexity of the neural network. Then, a Flatten layer is present in order to flatten all the extracted features into a big mono-dimensional tensor. This layer is followed by a Dense layer with 256 neurons and ReLU as activation function, which helps the model consider non-linear effects. The next layer is another Dropout followed by another Dense layer with 2 neurons and Softmax as activation function.

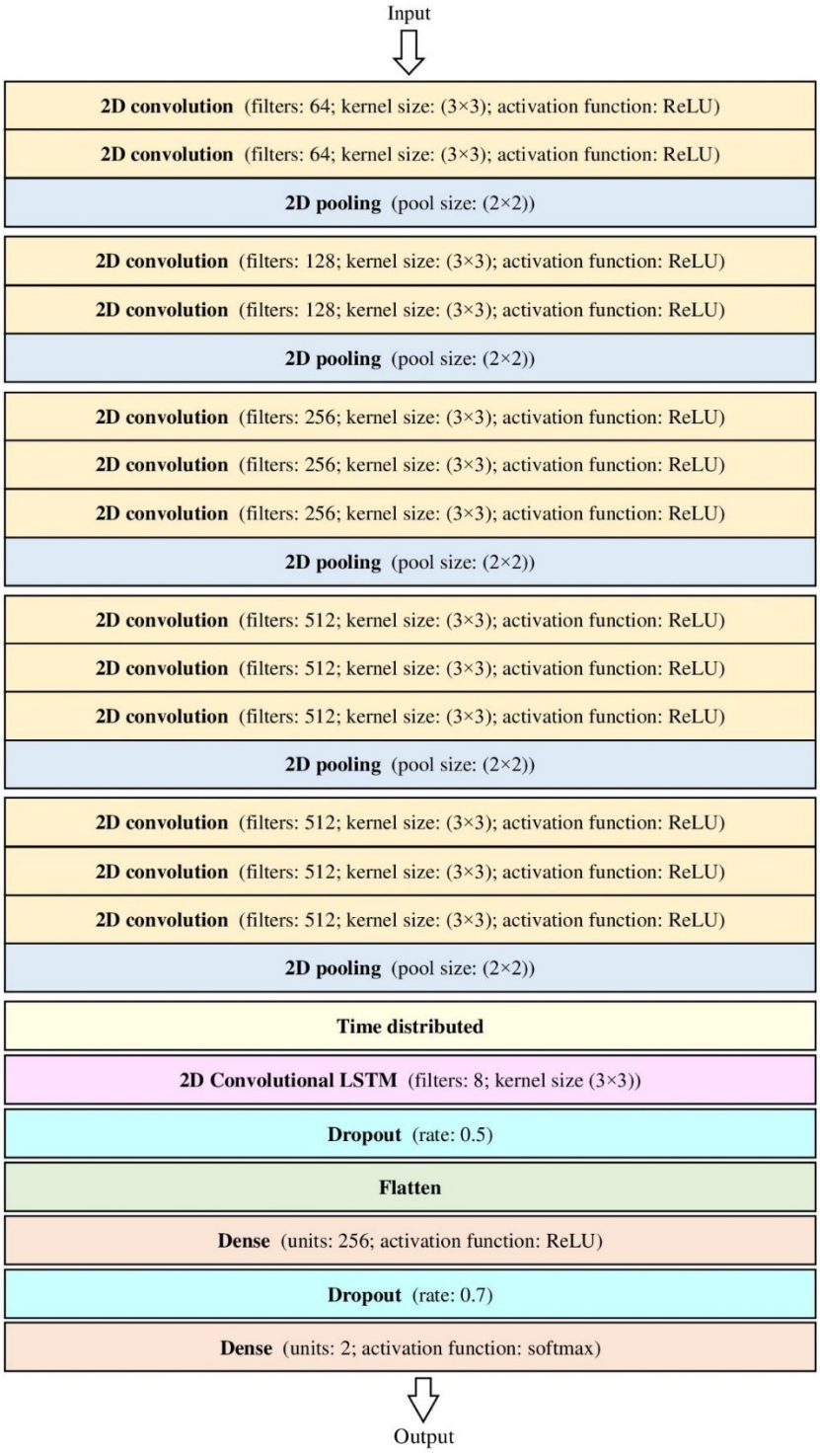


Figure 29. *Model summary of the end-to-end VGG16 + ConvLSTM.*

5.1.4.2 Data preparation

The first step to be performed was the loading of the data and their preparation. Like in the C3DKeras case, the 50 most significant slices per scan centred in the hippocampal region were selected. Moreover, in order to have a dimensionally uniform dataset, all the scans were reshaped to 147 pixels × 192 pixels per slice. Furthermore, there was no need to change the dimension of the channel, which is 1 (not 3 as the RGB case), because this neural network in not pre-trained, thus it does not require any fixed shape of the input data. Since we are not dealing with coloured images, the channel dimension was left to 1. So, the resulting shape was (275×50×147×192×1).

The whole dataset was split into training set, validation set and test set. In particular, the 80% of the whole set is used for the training set, while the remaining 20% constitutes the test set. Instead, the 20% of the training set is reserved as validation set.

5.1.4.3 Neural network classification

Since it is known that hyper-parameter selection has high influence on the model performance, the following hyper-parameters were used:

- Loss function: binary crossentropy;
- SGD optimizer;
- Dropout rate (first Dropout layer): 0.5;
- Dropout rate (second Dropout layer): 0.7;
- 50 epochs;
- Patience: 5;
- Batch size: 10;

Usually, these hyper-parameters are changed in case of over-fitting or under-fitting. However, in this case, it is not necessary a further tuning of the parameters.

5.1.5 Neural network evaluation

In order to evaluate the model performance, the following classification metrics were taken into account, both for the C3DKeras and the VGG16 + ConvLSTM: precision (PR), sensitivity (SE), F1-Score (F1-S), accuracy (ACC), Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). Specifically, a 0.5 discrimination threshold was chosen to compute the following metrics:

- Precision: $$PR = \frac{TP}{TP + FP} \quad (1)$$
- Sensitivity: $$SE = \frac{TP}{TP + FN} \quad (2)$$
- F1-Score: $$F1 - S = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (3)$$
- Accuracy: $$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where TP stands for True Positive and it is the case in which AD subjects are correctly classified as affected by AD; TN stands for True Negative and it is when CN subjects are correctly classified as healthy subjects; FP stands for False Positive and it is when CN subjects are wrongly classified as subjects affected by AD; FN stands for False Negative and it is when AD subjects are wrongly classified as healthy subjects.

5.2 Results

In this part of the thesis, the results obtained from the two different CNNs analysed are reported. In particular, in the following, it will be possible to read the numerical results related to AD because they are those of interest for the purpose of this work.

5.2.1 3D convolutional neural network

Table 3 reports the performance in classifying AD of the C3DKeras neural network. Instead, Figure 30 focuses on the ROC curve and AUC values, which are of great interest in the biomedical engineering field.

***Precision* (%)**	***Sensitivity* (%)**	***F1-Score* (%)**	***Accuracy* (%)**
59	100	74	64

Table 3. *C3DKeras neural network evaluation results.*

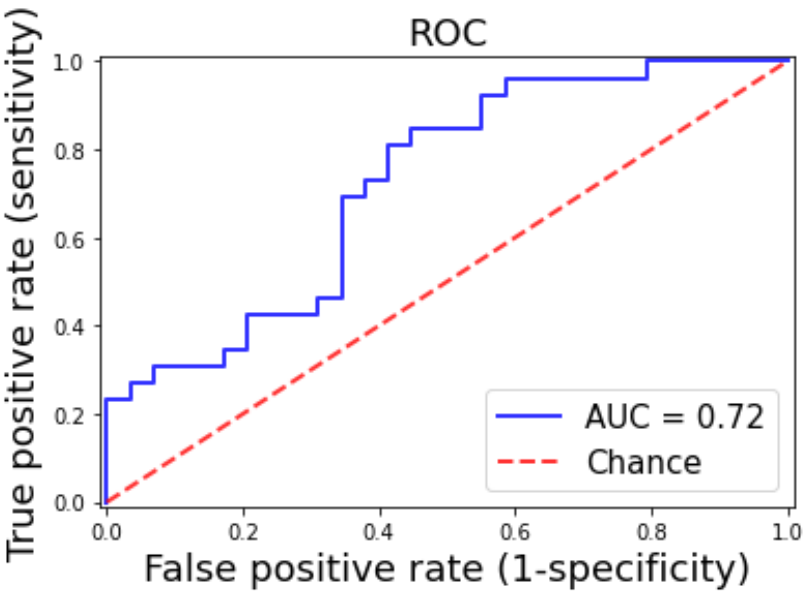


Figure 30. *Receiver Operating Characteristic (ROC) curve of the C3DKeras neural network.*

5.2.2 Time-distributed convolutional neural network

Table 4 shows the performance in classifying AD of the second neural network analysed in this work and so of the VGG16 + ConvLSTM neural network. Figure 31 is the ROC curve and AUC values.

***Precision* (%)**	***Sensitivity* (%)**	***F1-Score* (%)**	***Accuracy* (%)**
82	90	86	84

Table 4. *VGG16 + ConvLSTM neural network evaluation results.*

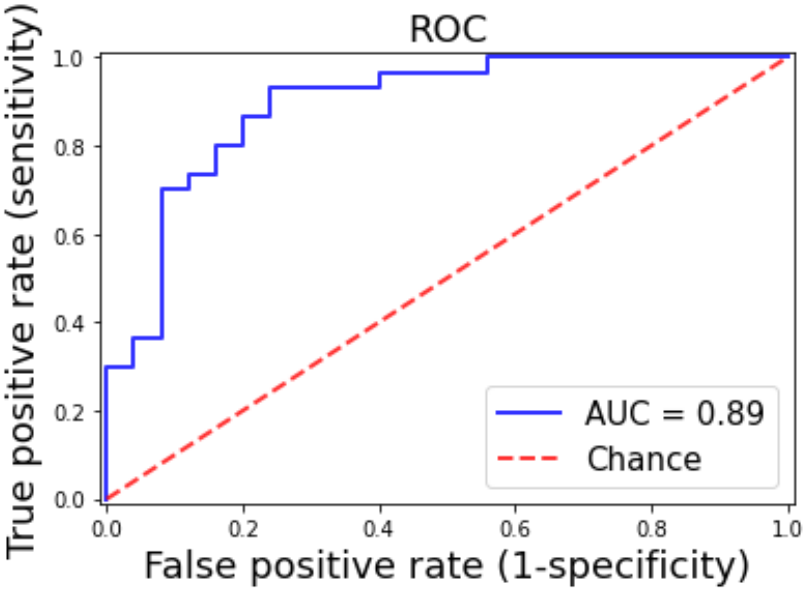


Figure 31. *Receiver Operating Characteristic (ROC) curve of the VGG16 + ConvLSTM neural network.*

5.3 Discussion

This study proposed a comparison between two different neural networks. In particular, the first one was a 3D CNN (C3DKeras) pre-trained on Sports-1M dataset containing video clips of various sports in order to recognize the type of sport contained in each video clip. Instead, the second neural network was a time-distributed CNN without any pre-training and, thus, trained from scratch. Due to the fact that the C3DKeras is a pre-trained neural network, it requires a fixed shape of the input data, which is not needed in the time-distributed case. This implies also that the input to the time-distributed CNN has a higher resolution if compared to the resolution of the input data of the 3D CNN under consideration. In fact, the input data of the time-distributed CNN has a resolution of (147×192), while the C3DKeras needs an input resolution of (112×112).

Moreover, from Table 3 and Table 4, it is possible to notice that the accuracy is much higher for the time-distributed CNN (84%) with respect to the 3D CNN (64%). However, the fact that the accuracy value of the C3DKeras would not have been so good was something expected, since this neural network was pre-trained on sport video clips. Furthermore, it can be highlighted that also the precision of the time-distributed CNN (82%) is higher than that of the 3D CNN (59%) and the same can be said for the F1-Score (86% and 74%, respectively).

Since in biomedical field sensitivity and AUC are of great importance, it is of great importance to discuss about them. Firstly, the sensitivity should be as high as possible because it means that there are a few false positive and so an AD case is diagnosed as such. In particular, from the results, it is noticeable that the sensitivity of the 3D CNN is 100%, which is an optimal value, while the one of the time-distributed CNN is 90%. The latter has a lower sensitivity than the 3D CNN, but it is still a very good value. Thus, in both cases, it can be said that the neural network is sensitive to correctly classify the AD scans. Instead, regarding the AUC, it is important to be discussed because it is independent of the threshold used to evaluate the metrics, which is at 0.5 in this case. From Figure 30, it is possible to see that the AUC value is at 72%, which means that the classifier in the 3D CNN can correctly discriminate between AD and CN at 72%, which can be considered an acceptable value. On the other hand, Figure 31 highlights that the time-distributed approach can better discriminate between AD and CN because the AUC is 89%.

A further remark is about the computational cost, which is lower for the time-distributed CNN because it manages streams of slices and not whole 3D blocks like in the 3D CNN.

Table 5 reports a summary of the overall comparison between the two different neural networks under consideration, with their characteristics.

Considering everything has been discussed so far, it can be stated that the time-distributed approach is preferable to the 3D CNN because it is trained from scratch, the input data have not a fixed shape to be respected and thus the resolution of the input data can be enhanced. Moreover, the computational cost is lower because the time-distributed CNN manages streams of slices, it is able to correctly classify the AD scans with high sensitivity and it can also better discriminate between AD and CN.

Type of CNN	***Type of training***	***Input type***	***Input shape***	***Computational cost***	***Sensitivity (%)***	***AUC (%)***
3D CNN	Pre-trained	Volumetric scans	16×112×112×3	High	100	72
Time-distributed CNN	From scratch	Volumetric scans	50×147×192×1	Low	90	89

Table 5. *Overall characteristics of the 3D CNN and time-distributed CNN.*

Conclusion

This thesis had a dual purpose. On one hand, it has been firstly performed a state-of-the-art analysis of the studies which applied interpretability techniques to AD diagnosis in order to have a clear idea about the current trends. On the other hand, two CNNs have been compared in order to realise which is the best in analysing volumetric MRI scans in terms of both classification performance and required computational effort. In particular, for this latter task, a pre-trained 3D CNN (C3DKeras) and an end-to-end time-distributed CNN were exploited.

To conclude, it can be said that both purposes have been pursued. On one hand, it has been realised that in literature there is still uncertainty concerning the best interpretability technique to be applied to the AD case, even though it is clear that attribution map approaches, such as LRP and GBP, seem to produce the most coherent interpretations. On the other hand, an end-to-end time-distributed CNN resulted to be the best approach between the exploited CNNs because of its higher classification outcomes and also lower computational cost.

Once realised that a time-distributed CNN is more suitable for the computerized AD diagnosis with respect to a 3D CNN, a future development of this thesis could lead to the application of an interpretability module to the time-distributed CNN in order to make a step forward in the direction of an interpretable AD diagnosis, which is crucial in the medical field.

Ringraziamenti

Sono giunta al termine di questo elaborato, ma sono anche giunta al termine del mio percorso universitario e, se sono arrivata fin qui, non è solo grazie alla mia determinazione e al mio essere testarda. In cuor mio, sento quindi il dovere di porre alcuni sentiti ringraziamenti alle persone che mi hanno accompagnato in questo fondamentale periodo della mia vita, che mi hanno visto crescere, mi hanno aiutato a crescere, ma anche aiutato a rialzarmi nei momenti di difficoltà.

Un primo ringraziamento va alla Prof.ssa Laura Burattini per avermi offerto l'opportunità di lavorare a questa tesi, a Dr. Selene Tomassini e Dr. Agnese Sbrollini per la guida costante e competente perché mi hanno accompagnato in questo percorso di tesi da dicembre fino ad oggi. È stato un percorso tutt'altro che semplice, pieno di ostacoli ma anche ricco di nuove conoscenze.

A mia madre, mio padre e mio fratello, che costituiscono il pilastro della mia vita. Se oggi sono la persona che sono e se sono arrivata fino a qui è anche grazie a voi e a tutti i sacrifici che avete sempre fatto per me. Mi avete insegnato cosa significa la parola *famiglia* ed è l'insegnamento più importate per la vita, non si impara ad amare dai libri di testo ma lo si impara con l'amore che si riceve e voi me ne avete dato tanto.

A voi, nonni miei, che oggi mi state guardando da lassù e sono certa che vi sentiate fieri della "nipotina" che sono diventata e del traguardo di vita che ho raggiunto. Mi avete sempre insegnato che per ottenere ciò che si vuole, bisogna lottare perché niente ci viene regalato. Eccomi qui oggi con il sorriso in viso, un sorriso che nessuno può levarmi perché, anche nei momenti di sconforto, ho pensato a voi e non ho mollato. Eravate, siete e sarete sempre la mia forza.

A Luna che in punta di piedi è entrata a far parte dei miei giorni, diventando una compagna di studio formidabile e una spalla sicura. Abbiamo condiviso gioie, sudore, lacrime, sorrisi, paure, paranoie, abbiamo affrontato ostacoli che sembrava fossero più grandi di noi, abbiamo studiato nei luoghi più strani e alle ore più improbabili. L'abbraccio che ci siamo date dopo l'ultimo esame resterà sempre dentro di me, perché solo noi sappiamo quanti sacrifici ci fossero dietro quel momento di pura gioia.

Ad Alessandra, la mia socia da ormai una ventina d'anni. Non so se c'è qualcuno al mondo che mi conosca meglio di te. Lo sport ci ha unite dal primo giorno che ci siamo viste in palestra, io piccina con i capelli corti corti e tu ancora più piccina ma con un cuore immenso. Le nostre allenatrici, Giorgia, Simona e Cristiana ci hanno sempre insegnato cosa significa

lottare e farlo insieme, non ci hanno trasmesso solo i loro insegnamenti per eseguire correttamente un lancio, un giro o un equilibrio… ma ci hanno insegnato cosa significa vivere, organizzarsi le giornate per fare sempre ciò che si ama. Ve ne sarò grata per tutta la vita.

A tutti i miei compagni di atletica, perché è grazie a voi se ogni sera torno a casa meno nervosa, stanca ma felice. Tra tutti, un grazie doveroso va a Virgi e Dado, che ancora mi chiedo dove trovino la pazienza per sopportarmi e supportarmi, ma i veri amici si distinguono anche per questo. Infine, e non certo per ordine di importanza, un grazie alla mia allenatrice Annalisa, capace di trasmettere una dose infinita di positività. Sempre solare anche in quei momenti in cui ti stai allacciando le scarpe chiodate e ti espone il programma del giorno, ma se sta sorridendo più del solito capisci di doverti preoccupare perché non sarà un allenamento facile… ma con lei tutto diventa più leggero.

Purtroppo ringraziarvi tutti uno per uno mi risulta impossibile perché ogni giorno mi rendo conto di quante siano le persone che dimostrano di volermi bene. Ricordatevi che, anche se il vostro nome non è scritto esplicitamente su questo foglio di carta, vi ho impressi nel mio cuore e vi sarò per sempre grata per tutto l'amore che mi date continuamente.

Grazie di tutto.

Ancona, 18 luglio 2022

Giada Bernardi

References

[1] OpenStax College, “Anatomy & Physiology”, 2013, OpenStax, http://cnx.org/content/col11496/latest/

[2] Kevin T. Patton, Gary A. Thibodeau, “Anatomia e fisiologia”, Edra Masson, 2011, 7th ed.

[3] Glauco Ambrosi, Dario Cantino, Paolo Castano, Silvia Correr, Loredana D’Este, Rosario F. Donato, Giuseppe Familiari, Francesco Fornai, Massimo Gulisano, Annalisa Iannello, Ludovico Magaudda, Maria F. Marcello, Alberto M. Martelli, Paolo Pacini, Mario Rende, Pellegrino Rossi, Chiarella Sforza, Carlo Tacchetti, Roberto Toni, Giovanni Zummo, “Anatomia dell’uomo”, Edi.Ermes, 2006, 2nd ed.

[4] Bruno Bergamasco, Roberto Mutani, “La neurologia di Bergamini”, 2007, Edizioni Libreria Cortina Torino.

[5] Walter F. Boron, Emile L. Boulpaep, “Medical Physiology”, 2017, 3rd ed.

[6] Lovinger D.M., “Communication networks in the brain: neurons, receptors, neurotransmitters, and alcohol”, Alcohol Res Health, 2008;31(3):196-214. PMID: 23584863; PMCID: PMC3860493.

[7] Ludwig P.E., Reddy V., Varacallo M., “Neuroanatomy, Central Nervous System (CNS)”, 2021 Oct 14, In: StatPearls [Internet], Treasure Island (FL): StatPearls Publishing, 2022 Jan., PMID: 28723039.

[8] Azarfar A., Calcini N., Huang C., Zeldenrust F., Celikel T., “Neural coding: A single neuron's perspective”, Neurosci Biobehav Rev, 2018 Nov;94:238-247. doi: 10.1016/j.neubiorev.2018.09.007, Epub 2018 Sep 15, PMID: 30227142.

[9] Walter F. Boron, Emile L. Boulpaep, “Medical physiology : a cellular and molecular approach”, 2009, Updated 2nd ed.

[10] Colón-Ramos D.A., “Synapse formation in developing neural circuits”, Curr Top Dev Biol, 2009;87:53-79. doi: 10.1016/S0070-2153(09)01202-2. PMID: 19427516; PMCID: PMC7649972.

[11] Amunts K., Zilles K., “Architectonic Mapping of the Human Brain beyond Brodmann”, Neuron, 2015 Dec 16;88(6):1086-1107. doi: 10.1016/j.neuron.2015.12.001. PMID: 26687219.

[12] M. Hunter Manasco, “Introduction to Neurogenic Communication Disorders”, 2014, Pap/Psc Edition

[13] John A. Kiernan, “Barr's The Human Nervous System: An Anatomical Viewpoint”, 2008, 9th ed

[14] World Health Organization, World Health Organization (2021), “Dementia”, Available online: https://www.who.int/news-room/fact-sheets/detail/dementia (accessed on 2 September 2021)

[15] Breijyeh Z., Karaman R., “Comprehensive Review on Alzheimer's Disease: Causes and Treatment”, Molecules, 2020 Dec 8;25(24):5789, doi: 10.3390/molecules25245789, PMID: 33302541; PMCID: PMC7764106.

[16] Trejo-Lopez J.A., Yachnis A.T., Prokop S., “Neuropathology of Alzheimer's Disease”, Neurotherapeutics, 2021 Nov 2, doi: 10.1007/s13311-021-01146-y, Epub ahead of print, PMID: 34729690.

[17] Weller J., Budson A., “Current understanding of Alzheimer's disease diagnosis and treatment”, F1000Res, 2018 Jul 31;7:F1000 Faculty Rev-1161, doi: 10.12688/f1000research.14506.1, PMID: 30135715; PMCID: PMC6073093.

[18] Hauser S., Josephson S., “Harrison's Neurology in Clinical Medicine”, 2013, 3rd ed.

[19] Wang Z., Bovik A.C., Simoncelli E.P., “Structural approaches to image quality assessment”, in Handbook of Image and Video Processing, 2005, Academic Press, 2nd ed.

[20] Thomas M. Deserno, “Fundamentals of Biomedical Image Processing” In: Thomas M. Deserno (eds) “Biomedical Image Processing. Biological and Medical Physics, Biomedical Engineering”, Springer, 2010, pp. 1-51, doi: 10.1007/978-3-642-15816-2_1.

[21] Salahuddin Z., Woodruff H.C., Chatterjee A., Lambin P., “Transparency of deep neural networks for medical image analysis: A review of interpretability methods”, Comput Biol Med, 2021 Dec 4;140:105111, doi: 10.1016/j.compbiomed.2021.105111, Epub ahead of print, PMID: 34891095.

[22] IEEE Engineering in Medicine & Biology Society, “Biomedical Imaging & Image Processing”, https://www.embs.org/about-biomedical-engineering/our-areas-of-research/biomedical-imaging-image-processing/

[23] John C. Gore, Foreword, “Biomedical Imaging: Applications and Advances”, pages xix-xxi. DOI: 10.1016/B978-0-85709-127-7.50015-6

[24] Santhi V., Acharjya D.P., Ezhilarasan M., “Emerging Technologies in Intelligent Applications for Image and Video Processing”, 2016, in the Advances in Computational Intelligence and Robotics (ACIR) Book Series

[25] Pierre-Jean Nacher, “Magnetic Resonance Imaging: From Spin Physics to Medical Diagnosis”, 2007

[26] Qiu S., Joshi P.S., Miller M.I., Xue C., Zhou X., Karjadi C., Chang G.H., Joshi A.S., Dwyer B., Zhu S., Kaku M., Zhou Y., Alderazi Y.J., Swaminathan A., Kedar S., Saint-Hilaire M.H., Auerbach S.H., Yuan J., Sartor E.A., Au R., Kolachalama V.B., “Development and validation of an interpretable deep learning framework for Alzheimer's disease classification”, Brain, 2020 Jun 1;143(6):1920-1933, doi: 10.1093/brain/awaa137, PMID: 32357201; PMCID: PMC7296847.

[27] Molnar C., “Interpretable Machine Learning: A Guide for Making Black Box Models Explainable”, 2022, 2nd ed., https://christophm.github.io/interpretable-ml-book/

[28] Huff D.T., Weisman A.J., Jeraj R., “Interpretation and visualization techniques for deep learning models in medical imaging”, Phys Med Biol, 2021 Feb 2;66(4):04TR01, doi: 10.1088/1361-6560/abcd17, PMID: 33227719; PMCID: PMC8236074.

[29] Koh P.W., Nguyen T., Tang Y.S., Mussmann S., Pierson E., Kim B., Liang P., “Concept Bottleneck Models”, 2020, Proceedings of the 37 th International Conference on Machine Learning, Online, PMLR 119, 2020.

[30] Li O., Liu H., Chen C., Rudin C., “Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions”, 2017, The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)

[31] Mohammadjafari S., Cevik M., Thanabalasingam M., Basar A., “Using ProtoPNet for Interpretable Alzheimer’s Disease Classification”, 2021, The 34th Canadian Conference on Artificial Intelligence.

[32] Chen C., Li O., Tao C., Barnett A.J., Su J., Rudin C., “This Looks Like That: Deep Learning for Interpretable Image Recognition”, 2019, 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[33] Shahamat H., Saniee Abadeh M., “Brain MRI analysis using a deep learning based evolutionary approach”, Neural Netw, 2020 Jun;126:218-234, doi: 10.1016/j.neunet.2020.03.017, Epub 2020 Mar 28. PMID: 32259762.

[34] Guan H., Wang C., Cheng J., Jing J., Liu T., “A parallel attention-augmented bilinear network for early magnetic resonance imaging-based diagnosis of Alzheimer's disease”, Hum Brain Mapp, 2022 Feb 1;43(2):760-772, doi: 10.1002/hbm.25685, Epub 2021 Oct 22, PMID: 34676625; PMCID: PMC8720194.

[35] Turkan Y., Tek F.B., "Convolutional Attention Network for MRI-based Alzheimer’s Disease Classification and its Interpretability Analysis", 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 1-6, doi: 10.1109/UBMK52708.2021.9558882.

[36] Liang S., Gu Y., “Computer-Aided Diagnosis of Alzheimer's Disease through Weak Supervision Deep Learning Framework with Attention Mechanism”, Sensors (Basel), 2020 Dec 31;21(1):220, doi: 10.3390/s21010220, PMID: 33396415; PMCID: PMC7795039.

[37] Ocasio E., Duong T.Q., “Deep learning prediction of mild cognitive impairment conversion to Alzheimer's disease at 3 years after diagnosis using longitudinal and whole-brain 3D MRI”, PeerJ Comput Sci, 2021 May 25;7:e560, doi: 10.7717/peerj-cs.560, PMID: 34141888; PMCID: PMC8176545.

[38] Zhang X., Han L., Zhu W., Sun L., Zhang D., “An Explainable 3D Residual Self-Attention Deep Neural Network For Joint Atrophy Localization and Alzheimer's Disease Diagnosis using Structural MRI”, IEEE J Biomed Health Inform, 2021 Mar 18;PP, doi: 10.1109/JBHI.2021.3066832, Epub ahead of print, PMID: 33735087.

[39] Hosseini-Asl E., Ghazal M., Mahmoud A., Aslantas A., Shalaby A.M., Casanova M.F., Barnes G.N., Gimel'farb G., Keynton R., El-Baz A., “Alzheimer's disease diagnostics by a 3D deeply supervised adaptable convolutional network”, Front Biosci (Landmark Ed), 2018 Jan 1;23(3):584-596, doi: 10.2741/4606, PMID: 28930562.

[40] Folego G., Weiler M., Casseb R.F., Pires R., Rocha A., “Alzheimer's Disease Detection Through Whole-Brain 3D-CNN MRI”, Front Bioeng Biotechnol, 2020 Oct 30;8:534592, doi: 10.3389/fbioe.2020.534592, PMID: 33195111; PMCID: PMC7661929.

[41] Esmaeilzadeh S., Belivanis D.I., Pohl K.M., Adeli E., “End-To-End Alzheimer's Disease Diagnosis and Biomarker Identification”, Mach Learn Med Imaging, 2018 Sep;11046:337-345, doi: 10.1007/978-3-030-00919-9_39, Epub 2018 Sep 15, PMID: 32832936; PMCID: PMC7440044.

[42] Google Colab. Google colab, https://colab.research.google.com/notebooks/intro.ipynb.

[43] Python. Python, https://www.python.org/.

[44] NumPy. Numppy,https://numpy.org/.

[45] Keras. Keras, https://keras.io/.

[46] Tran D., Bourdev L., Fergus R., Torresani L., Paluri M., "Learning Spatiotemporal Features with 3D Convolutional Networks", 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497, doi: 10.1109/ICCV.2015.510.