VISUALIZATION OF DEEP NETWORKS TRAINED FOR BIPOLAR DISORDER CLASSIFICATION BY USING FNIRS MEASUREMENTS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

OĞUZHAN BABACAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

FEBRUARY 2021

Approval of the thesis:

VISUALIZATION OF DEEP NETWORKS TRAINED FOR BIPOLAR DISORDER CLASSIFICATION BY USING FNIRS MEASUREMENTS

submitted by **OĞUZHAN BABACAN** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. İlkay Ulusoy
Head of Department, **Electrical and Electronics Engineering** _____

Prof. Dr. İlkay Ulusoy
Supervisor, **Electrical and Electronics Engineering, METU** _____

Examining Committee Members:

Prof. Dr. Uğur Halıcı
Electrical and Electronics Engineering, METU _____

Prof. Dr. İlkay Ulusoy
Electrical and Electronics Engineering, METU _____

Prof. Dr. Nevzat Güneri Gençer
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. Yeşim Serinağaoğlu Doğrusöz
Electrical and Electronics Engineering, METU _____

Assoc. Prof. Dr. Bora Başkak
School of Medicine, Ankara University _____

Date: 03.02.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Oğuzhan Babacan

Signature :

ABSTRACT

VISUALIZATION OF DEEP NETWORKS TRAINED FOR BIPOLAR DISORDER CLASSIFICATION BY USING FNIRS MEASUREMENTS

Babacan, Oğuzhan
M.S., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. İlkay Ulusoy

February 2021, 71 pages

Deep learning applications have achieved impressive performances on many medical problems such as classification of disorders, effects of a treatment or unspotted symptoms of a disease, etc. While modern deep learning progress is impressive in such areas, genuine understandings of its working principles are not clear. For that matter, the term black box has often been associated with deep learning algorithms. The majority of previous studies have concentrated on networks' successes and have computed their performances in terms of accuracy levels. However, this thesis focuses on disintegrating the internal working mechanisms of neural networks into intuitive and understandable components. It makes them easy to understand and to interpret from medical experts' perspectives. With this purpose in mind, pre-trained Convolutional Neural Networks and Residual Neural Networks are utilized by using time-series neuroimaging data, i.e. Functional Near-Infrared Spectroscopy (fNIRS) measurements, belonging to two classes, namely healthy and bipolar, and their visualization outputs are attained. Since these outputs are complex time-series data, they are analyzed by statistical methods such as chi-square and t-tests so that the intrinsic features of healthy and bipolar subjects specific to their classes are obtained. Results

are compared with previous medical studies and are analyzed so that potential reasons behind the classification results are provided. The contribution of this thesis is providing an inference about visualization outcomes of different neural networks, which are trained for the bipolar disorder classification using fNIRS data. Therefore, this study tries to fill the void between medical researchers and deep learning experts.

Keywords: fNIRS, Deep Learning, Bipolar Disorder, Classification, Visualization, Class Activation Map

ÖZ

BİPOLAR HASTALIĞI SINIFLANDIRMASI İÇİN FNIRS ÖLÇÜMLERİ KULLANILARAK EĞİTİLMİŞ DERİN AĞLARIN GÖRSELLEŞTİRİLMESİ

Babacan, Oğuzhan
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. İlkay Ulusoy

Şubat 2021 , 71 sayfa

Derin öğrenme uygulamaları, hastalıkların sınıflandırılması, bir tedavinin etkileri veya bir hastalığın belirlenmemiş semptomları gibi birçok tıbbi problem üzerinde etkileyici performanslar elde etmiştir. Modern derin öğrenme gelişimi bu alanlarda etkileyici olsa da çalışma prensiplerinin gerçek kavrayışı net değildir. Bu nedenle, derin öğrenme algoritmaları genellikle kara kutu terimiyle ilişkilendirilmiştir. Önceki çalışmaların çoğu, ağların başarılarına odaklanmış ve performanslarını doğruluk seviyeleri açısından hesaplamışlardır. Bununla birlikte, bu tez, sinirsel ağların iç çalışma mekanizmalarının sezgisel ve anlaşılır bileşenlere bölünmesine odaklanmaktadır ve bunların anlaşılmasını ve tıp uzmanlarının bakış açısından yorumlanmasını kolaylaştırmaktadır. Bu amaçla, önceden eğitilmiş Evrişimsel Sinir Ağları ve Artık Değerli Sinir Ağlarına sağlıklı ve bipolar olmak üzere iki sınıfın zaman serisi verileri olan Fonksiyonel Yakın Kızılötesi Spektroskopi (fNIRS) ölçümleri uygulanmış ve bunlara ait görselleştirme çıktıları elde edilmiştir. Bu çıktılar karmaşık zaman serisi verileri oldukları için ki-kare ve t-testleri gibi istatistiksel yöntemlerle analiz edilerek sağlıklı

ve bipolar deneklerin kendi sınıflarına özgü iç özellikleri elde edilmiştir. Sonuçlar önceki tıbbi çalışmalarla karşılaştırılıp ve analiz edilerek sınıflandırma sonuçlarının ardındaki olası nedenler verilmiştir. Bu tezin katkısı, fNIRS verileri kullanılarak bipolar bozukluk sınıflandırması amacıyla eğitilmiş farklı sinir ağlarının görselleştirme sonuçları hakkında bir çıkarım sağlamaktır. Bu nedenle, bu çalışma tıp araştırmacıları ile derin öğrenme uzmanları arasındaki boşluğu doldurmaya çalışmaktadır.

Anahtar Kelimeler: fNIRS, Derin Öğrenme, Bipolar Bozukluk Hastalığı, Sınıflandırma, Görselleştirme, Sınıf Aktivasyon Haritası

To my family

ACKNOWLEDGMENTS

I would like to express my deepest appreciation and gratitude to my advisor Prof. Dr. İlkay Ulusoy for her help and supervision. This work would not have been possible without her continuous guidance and advice.

I would like to thank the rest of my thesis committee members Prof. Dr. Uğur Halıcı, Prof. Dr. Nevzat Güneri Gençer, Assoc. Prof. Dr. Yeşim SerinAğaoğlu Doğrusöz and Assoc. Prof. Dr. Bora Başkak for being in my jury and sharing their insightful comments and questions.

I am grateful to my family for their endless support, understanding and love. My mother Yadigâr Babacan, my father Yaşar Babacan and my sweet sister Tuğçe Babacan have stood by me along all the happy and difficult times of my life as I have deeply experienced after the recent unfortunate accident. I know that we will be there for each other whatever may come. I am also thankful to my beloved Afra Nazlı for always showing her pure love and encouragement.

I would like to express my gratitude to Neşet Ertaş for himself and the existence of his music.

Last but not least, I would like to thank Barış Can Çam and my study partner Haluk Barkın Evgin for their help and contributions, and Alptekin Meriç for his motivation and unforgettable friendship.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Networks
DNN	Deep Neural Networks
CNN	Convolutional Neural Networks
FCN	Fully Convolutional Network
RESNET	Residual Neural Network
CAM	Class Activation Map
GAP	Global Average Pooling
fNIRS	Functional Near Infrared Spectroscopy
fMRI	Functional Magnetic Resonance Imaging
MRI	Magnetic Resonance Image
EEG	Electroencephalography
VF-Task	Verbal Fluency Task

CHAPTER 1

INTRODUCTION

1.1 Motivation and Problem Definition

Nowadays, the applications of deep learning have achieved impressive performance on many computer vision-related tasks, such as object detection, speech recognition, image retrieval, etc. On the other hand, as mentioned in [1], it appears that the discriminatory power of a feature is heavily dependent on the difficulty of the task. A more straightforward classification task results in simpler features even when the data is the same. As a result of that the CNN automatically generates features of just enough complexity to perform the task at hand. In other words, training on a more challenging task (e.g. a larger number of classes) yields features that are more informative and special to the representative class members.

Nonetheless, while these deep neural networks provide superior performance, the lack of disintegration of their internal working mechanism into intuitive and understandable components makes them hard to identify and interpret from a human vision perspective [2]. Consequently, when today's intelligent systems fail, they fail mysteriously without any warning or explanation, leaving an inconsistent output behind. To establish trust in intelligent systems and make progress in their integration into our everyday lives, we are supposed to develop 'transparent' models that explain what, how, and why they predict. Typically, this transparency is beneficial on three sides of Artificial Intelligence (AI) development. First, when AI is significantly weaker than humans and not reliably deployable, transparency gives users the ability to identify failure cases and to spend their efforts on improving the system in the required directions. Second, in the case where AI's power is equivalent to that of humans,

the aim is to preserve trust and confidence in the system. Finally, when AI is excessively stronger than humans, the purpose of developing transparent models is machine teaching [3], which is a concept that a machine lends assistance to humans about how to make better decisions [2].

While modern deep learning progress is impressive, model developers do not fully understand its working principles. For that matter, the term "black box" has often been associated with deep learning algorithms. How could a model's results be trusted if there is no convincing answer about how it works? In order to find a satisfying answer for these kinds of questions, the lack of interpretability and transparency of neural networks from the learned features to the underlying decision processes is needed to be handled.

Making sense of why a particular model misclassifies data or behaves poorly can be challenging for model developers. Similarly, end-users interacting with an application that relies on deep learning to make decisions may question its reliability if no explanation is given by the model or become confused if the explanation is inexplicable. For example, if a self-driving car makes an awful decision and harms a person, the reason behind it could not be quantified. Then, there would be no way to fix it, which could lead to even more disasters. These challenges are often worse due to the requirement of having a large dataset to train a great majority of deep learning models. As threatening as these problems are, they will likely become even more widespread as more AI-powered systems are deployed in the world.

Therefore, a general sense of model understanding is beneficial and often required to address the aforementioned issues. Without taking a look at what is happening inside the neural network, model developers remain unaware of all this. In a nutshell, it is not possible for developers to know where the network is looking and what it is concentrating on, or what features it is relying on while making a decision. Thus, visualizing neural networks is essential in making them work robustly in practical, real-world use cases.

Researchers compete to reach the most accurate and stable models, thanks to a great variety of medical data. However, as mentioned at the beginning of the chapter, these models cannot extract biological insights lied behind the decision mechanism. Visu-

alization of deep networks has bridged this gap as it provides an opportunity to look at the inside of the network. This is crucial for medical researchers to receive feedback about the questions they investigate, such as which symptoms are more distinctive for diagnosing a disease or which part of treatment is more effective.

1.2 Proposed Methods and Models

Motivated by the aforementioned problems, this study's main objective is to provide interpretable feedback that highlights the reason taken by the pre-trained classifier. In other words, we aim to visualize the activation maps of networks trained for the diagnosis of bipolar disease by using neuroimaging data to understand the networks' operation. Training and tuning hyperparameters of neural networks are out of this thesis's scope, and our study covers the process after getting well-trained models from Evgin's work [4].

Hence state-of-the-art visualization method, heatmap of class activation, is applied as a base method during the thesis, and heatmaps of CNN and ResNet models are generated. As the input is time series data collected from healthy and bipolar subjects while completing a psychological test, the visualization method outputs are time-series data. They indicate heatmaps of activation in the time scale for each of the subjects. However, directly observing these heatmaps and making a remark according to them is unfavorable since the data comes from different subjects. Their interactions with the environment, which might trigger their activation maps for a brief time, are unpredictable. For this reason, in order for psychiatrists to obtain reliable and meaningful results, the traditional visualization approach is needed to be updated.

In this study, after getting activation heatmaps of networks, we further calculate average activations for each of 6 unequal time partitions, which are parts of the overall test, explained in detail in Section 3.2. Then, by using these average activations of each partition, the differentiation of healthy and bipolar groups, as well as the comparison of different networks, are analyzed with statistical tools in the scope of this thesis. Thanks to this work, essential parts of the test for classification and the distribution of activations belong to healthy and bipolar subjects can be introduced into

medical experts' concerns.

In accordance with this purpose, the chi-square test of independence is applied to reveal whether bipolar disorder is related to a specific partition number that shows the maximum activation in the neural network. On the other side, the Welch's t-test is performed so that healthy and bipolar populations' distributions in terms of the neural network's activation can be examined for each partition of the test. The difference between these populations' means is investigated. In order for this study to achieve its goal, both CNN and ResNet models are operated with various tuning options.

1.3 Contributions and Novelties

While neural networks can deal with large amounts of monotonous data and learn to classify data based on training examples, researchers still need to comprehend the cause and effect relation, supposing that they are interested in explanation instead of absolute classification. In her thesis [5], Hoşgören, who is a psychiatrist, investigated whether healthy and bipolar subjects have distinguishable patterns in terms of the verbal fluency test score and asserted that there is no remarkable difference between those of both subject groups. However, thanks to Evgin's work [4], with the power of deep learning, bipolar and healthy subjects have been classified successfully over 75 percent by using their medical recordings obtained while performing the verbal fluency task.

Nevertheless, it is not enough for medical experts to benefit from this study effectively. They need to know how the classification process has been done, which part of the data with which reason is more important, etc. This thesis contributes to a better understanding of the working mechanism of neural networks used for the classification of bipolar disorder.

In [6], our first study was concluded as the classification of bipolar disorder by using time-series neuroimaging data with 1D CNN. It was one of few publications announced in the field of bipolar disorder classification with a neural network, furthermore using fNIRS data for this purpose made it unique among others. In his study [4], Evgin took a step forward and implemented state-of-the-art networks with higher

accuracies for the same problem and data. Our thesis's novelty comes from being the leading publication that makes an inference about outcomes of visualizations of different neural networks, which are trained for the classification of bipolar disorder with Functional Near-Infrared Spectroscopy (fNIRS) data. Therefore, by this study, we tried to fill the void between medical researchers and deep learning experts and put a milestone for fresh research.

1.4 The Outline of the Thesis

Chapter 1 introduces the scope of the thesis by stating the problem and the motivation behind the study. The approach followed and the contribution to the problem are also provided.

Main methods of visualization of neural networks are described, and the terms and concepts to be known are introduced in Chapter 2. The related works regarding visualization methods are examined in detail. Moreover, the literature survey of recent approaches and methods are given at the end of this chapter.

In Chapter 3, the working principle of a neuroimaging method, namely fNIRS, is explained. The attainment and properties of the input dataset are provided in detail.

Chapter 4 explains the detailed theoretical background that is needed to understand the thesis. It introduces state-of-the-art deep neural networks used for generating models, which are inputs of visualization methods. Furthermore, in this chapter, the detailed background information of inferential statistics used while drawing conclusions is described.

Chapter 5 describes the neural network architectures specific to the classification of bipolar disease by using the fNIRS dataset. Moreover, the particular visualization method, i.e. class activation map, is examined, and the modifications made on it to obtain more useful inferences are provided. At the end of the chapter, the results attained by statistical methods are summarized, and medical inferences that may lie behind results are discussed.

Chapter 6 concludes the thesis by summarizing the work and results. Finally, the

future work arising from this study is mentioned.

CHAPTER 2

VISUALIZATION OF NEURAL NETWORKS AND RELATED WORKS

2.1 Neural Network Visualization Methods

Visualizing the output of a machine learning model is a great way to see how it progresses, regardless of the neural network's complexity. Mainly training or validation errors, as well as their accuracies, are concerned for the sake of the success criterion while training deep networks. These two components provide information about how the network performance is at each epoch. However, the working principle of the Neural Network (NN), i.e. the way it succeeds at the classification, can be observed and learned by way of the visualization of the neural network. Therefore, to clarify the decision-making processes about why and how NN models generate their outputs, a visualization method is needed to be employed by transforming the complex internal features of the network into visually observable patterns.

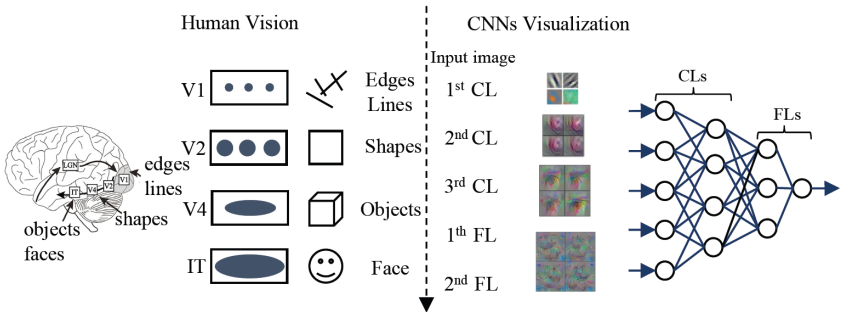

Figure 2.1: The analogy between Human vision and CNN visualization [7]. Left: the human brain processes the target features through multiple visual neuron blocks is on the left. Right: the main steps of the CNNs visualization are shown

Neural network visualization takes the human visual cortex system as a reference [2] and Figure 2.1 shows the similarities between them. On the left part of the figure, the human brain processes the target features through multiple visual neuron blocks [8], each of which may be specific to different features like colors, edges, and shapes [9]. While humans recognize a face, the neuron blocks with small receptive fields, i.e. V1, are used for basic features such as edges and corners. In deeper blocks, i.e. V2 and V4, complex features such as particular shapes and objects are detected. The final decision-maker block, i.e. IT, consists of visual neurons having the most extensive receptive fields, resulting in sensitivity to the entire face. On the other side of the figure, the main steps of the CNNs visualization are shown. Characteristically, CNNs extract basic features such as edges and colored spots in the first layer. Next, through deeper layers, more comprehensive features such as some parts of objects and shapes are detected. Finally, in the fully-connected layers, the final decision is made.

[10] and [11] state that a pattern to which a unit's response is maximum is a representation of what that unit is doing. In other saying, the maximum response of an internal unit to input provides a fair characterization of what the unit does. That is the reason some researchers study on different approaches and try to show the connection between the feature space of a network and the input. They feed a vast dataset to the network and investigate the inputs that activate neurons most [12], [13], [8].

Considering the rapid developments of NNs, the visualization has been extended to interpret their working mechanism. In the rest of the chapter, this study divides visualization methods into three main sections according to which part of the network to be visualized, namely *Visualization of Intermediate Layer Activations, Visualization of Convolutional Filters, Visualization of Heatmaps of Class Activation*.

2.1.1 Visualization of Intermediate Layer Activations

One way to investigate how a NN classifies the input is to look at the output of its intermediate layers. By doing so, outputs of each filter of each layer are derived so that they can be used to determine the way each filter in the network serves for classification. Looking at the different outputs from convolutional layer filters, it is expected to see how different filters in different layers are trying to highlight or

activate different input parts.

According to Chollet [14], visualizing intermediate activations is to display the feature maps of various convolution layers of a network when a particular input provided. Each channel of the input encodes relatively independent features. Therefore, it is required to plot every channel's contents independently to visualize the overall feature maps.

Since the first layer is used as a collection of various edge detectors, its activations generally contain the information present in the input. As getting deeper layers, the activations become increasingly abstract and less interpretable. By looking at the example of cat image classification, a cat-likely face or body that existed in the image might be recognized in the very first layer. On the other hand, higher-level concepts, such as cat ear and cat feather, are encoded by the deeper layers.

Briefly, higher layers of the network extract less information about the image's visual contents and more information about the image class. Deep neural network effectively acts as an information distillation pipeline. It starts working with the raw data going in (in the cat image classification, the RGB pictures) and being repeatedly transformed. By this means, the irrelevant information is filtered out, and useful information is magnified and refined. The activations' sparsity increases with the depth of the layer. While all filters are activated in the first layer, the filters of the following layers become deactivated. Through deeper layers, more and more filters get blank, meaning that the pattern encoded by the filter is not found in the input image.

A widely used activation maximization algorithm [10], [11] which stands behind the visualization of layer activations is provided, in detail. Since the aim is to observe the representation of what the layer is doing, a pattern in which the layer responds maximally is needed to be found. For this purpose, given a network of N layers, starting from an input image, the input image pixels are modified to maximize the output of a target layer $l < N$. Thanks to Equation 2.1 and Equation 2.2, the visualization of layer activations could be handled as an optimization problem. Let θ represent the parameters of a network (weights and biases), x be the input sample, $x^{(i)}$ be the feature maps of convolutional layer i of the network, i.e. the activation of layer i, and l_i be the layer's activation function; the aim is to look for x^* such that maximize

the $f\left(x^{(i)}\right)$, where f denotes the objective function. The maximization activation algorithm assumes a fixed θ.

$$x^{(i)} = l_i(\theta, x) \tag{2.1}$$

$$x^* = \arg_x \max f\left(x^{(i)}\right) \tag{2.2}$$

Although this is a non-convex problem, a local maximum could be found by simple gradient ascent in the input space. After calculating $x^{(i)}$ by Equation 2.1, the gradient of the objective function with respect to that, $\frac{df}{dx^{(i)}}$, is calculated. Then, the backpropagation is performed to get the gradient of the objective function with respect to the input, x. Finally, the input image is renewed by gradient ascent using Equation 2.3, where λ is the learning rate of the model.

$$x := x + \lambda \frac{df}{dx} \tag{2.3}$$

In Figure 2.2, activations of all filters of *block1_conv2* layer of VGG16 model [15] are plotted when the Lena image is used as input. As it can be observed, each of the outputs contains different features, almost half of which are more apparent than others. When we take a closer look at the activations, thanks to Figure 2.3, it can be stated that the left image, i.e. the activation of filter 46 of the *block1_conv2* layer, focus on the feather on the hat. In contrast, the right one, the activation of filter 52 of the *block1_conv2* layer, is interested in the background of the image and some parts of Lena's face.

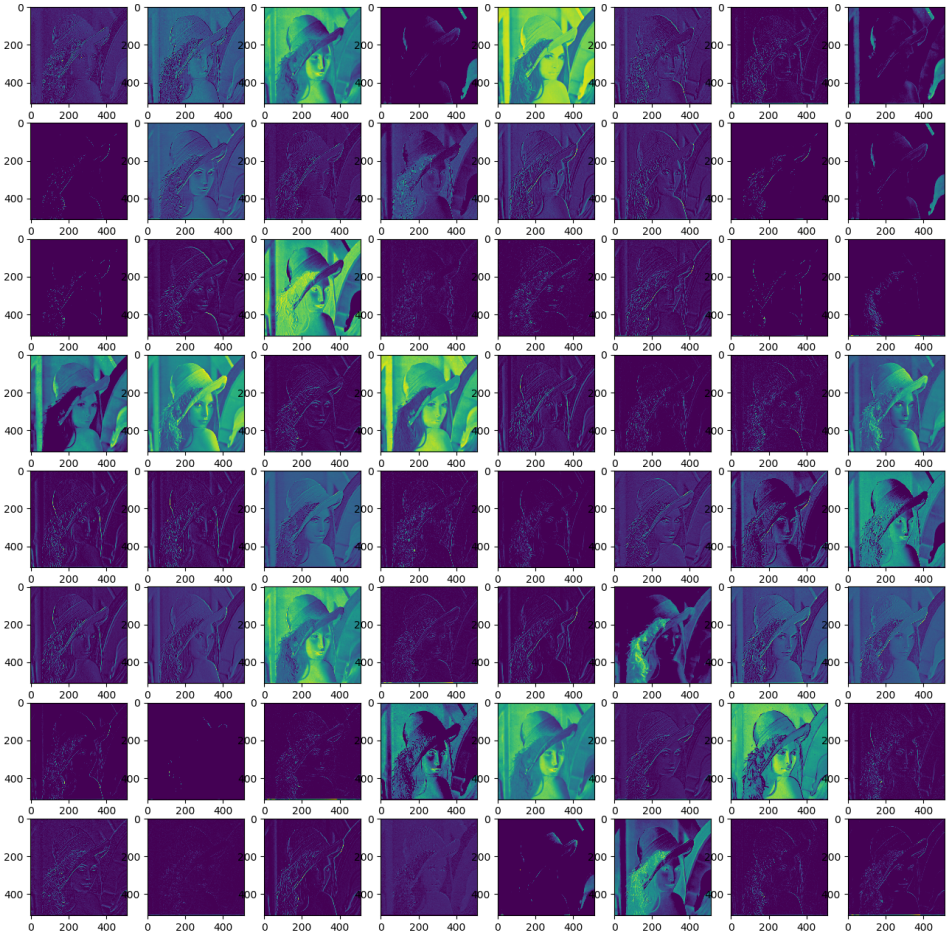

Figure 2.2: Visualization of activations of all filters of block1_conv2 layer of the VGG16 model

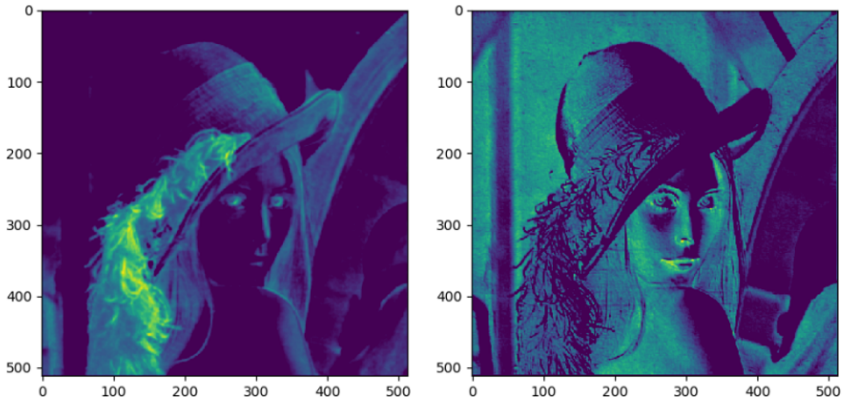

Figure 2.3: Visualization of activations of Filter 46 (left) and Filter 52 (right) of block1_conv2 layer of the VGG16 model

Unlikely Figure 2.2, most of the activations in Figure 2.4 has more abstract features since it contains activations of all filters of *block2_conv2* layer of the model because of the aforementioned reason that going through deeper layers results in more complex features extracted by the model.

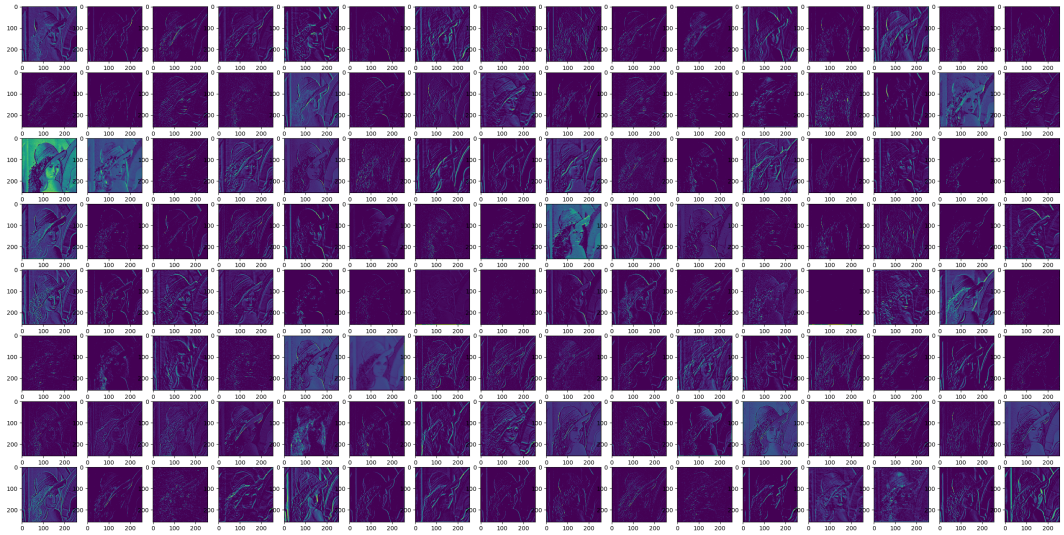

Figure 2.4: Visualization of activations of all filters of block2_conv2 layer of the VGG16 model

2.1.2 Visualization of Convolutional Filters

The alternative approach inspecting the model's working principle is to represent the visual pattern to which each filter of the network responds. Thanks to his book [14], Chollet broadly explained the process, which mainly relies on the gradient ascent in input space. In order to maximize the response of a specific filter, the gradient descent is applied to the value of the input image. Starting with a blank input image, this process generates an input image that the target filter gives the maximum response.

After taking a closer look at these images of filters from different convolution layers, it becomes apparent what different layers are trying to learn from the image data provided to them. The patterns encoded in filters of starting layers seem to be very basic, composed of lines and other basic shapes, which tells us that the earlier layers learn about basic features in images like edges, colors, etc. However, as going deeper

into the network, the patterns become more complex. Hence, it can be stated that the deeper layers are learning about much more abstract information. These layers begin to resemble textures that existed in natural images such as eyes, leaves, posture, and so on to generalize the classes and not the specific image. This is why some empty filters in deeper layers can be encountered, as stated in the previous section. Those particular filters are not activated for that image; put it differently, the image does not have the information that the filter was interested in.

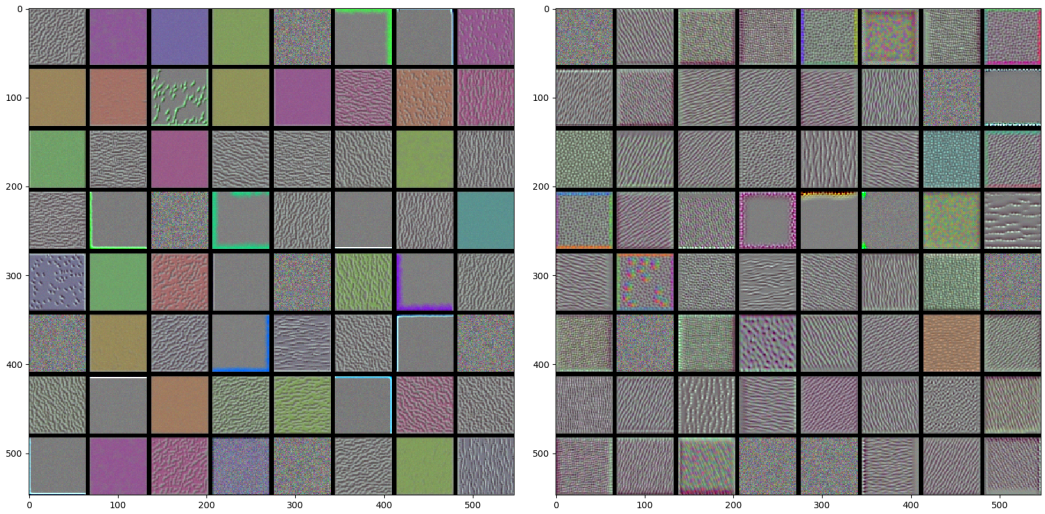

Figure 2.5: Visualization of convolutional filters of block1_conv1 and block2_conv1 layers of the VGG16 model

All 64 convolutional filters from *block1_conv1* (left) and *block2_conv1* (right) layers of VGG16 model [15] are given in Figure 2.5; whereas, *block3_conv1* (left) and *block4_conv1* (right) layers are in Figure 2.6. When these filters are analyzed, the abovementioned deductions can be verified such that the very first layer's filters are representatives of simple shapes while the last layer's filters are experts for particular patterns.

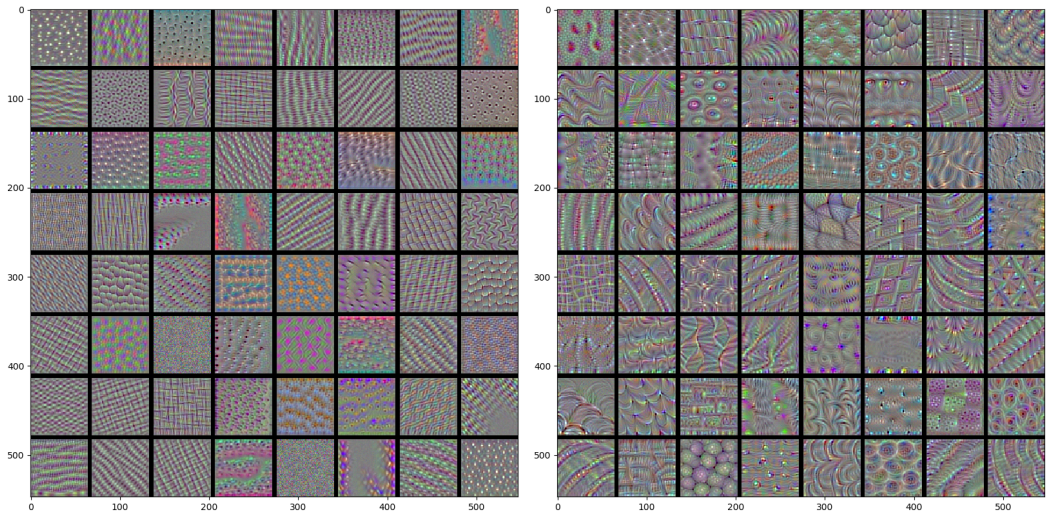

Figure 2.6: Visualization of convolutional filters of block3_conv1 and block4_conv1 layers of the VGG16 model

The process is as follows: after building a loss function that maximizes the value of a filter in a layer, the stochastic gradient descent is used to maximize the activation value by adjusting the values of the input.

2.1.3 Visualization of Heatmaps of Class Activation

2.1.3.1 Class Activation Map

The other visualization method for debugging the decision process in classification networks is the class activation map (CAM). It is a technique to identify the discriminative regions in the input specific to a class. In other words, CAM produces heatmaps of class activation over input data so that regions in the data relevant to the class can be observed. As stated by [14], a class activation heatmap of an image is a 2D grid of scores associated with a specific class. Every location in the input is used for the generation of the heatmap, indicating how important each location is concerning the class. For instance, when a cat image is fed into a cat vs. dog convolutional network, its CAM visualization provides a heatmap for the class ‘cat’, indicating which parts of the image are cat-like.

As indicated by [16], to be able to create a CAM, the network architecture is restricted to contain a global average pooling (GAP) layer after the final convolutional layer. The loss for average pooling benefits from the network to identify all discriminative regions of an object. As a result of this, the image regions' importance for a specific class can be identified by projecting back the output layer's weights on the feature maps. The activation heatmaps might differ from layer to layer in the network since all layers view the input image differently and create a unique abstraction of the image based on their filters. Because the class prediction result depends majorly on the final convolutional layer, it is suggested to focus on it as the target layer.

The algorithm of CAM, provided in [16], assumes that the bias term is ignored as it has almost no impact on the classification performance. For a given image, let $f_k(x, y)$ denote the activation of unit k in the last convolutional layer at (x, y). Then F^k, the output of the GAP, becomes $\sum_{x,y} f_k(x, y)$. Therefore, the class score which is the input to softmax, S_c, is $\sum_k w_k^c F^k$ for a given class c, where w_k^c is the weight corresponding the class and unit indicating the importance of F^k for class c. After putting F^k into S_c, the Equation 2.4 is obtained:

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y) \tag{2.4}$$

The class activation map for class c, M_c, is defined as a weighted linear sum of the visual patterns at different locations as given in Equation 2.5. Figure 2.7 clearly shows the relationship between weights and activations while generating the class activation map.

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{2.5}$$

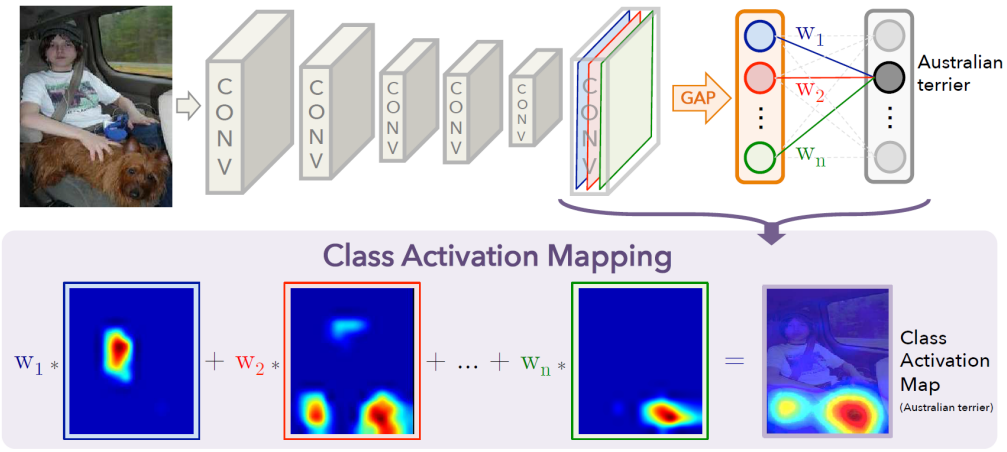

Figure 2.7: The relation between weights and activations while generating the class activation map [16]

Thus, the class score can be expressed as $S_c = \sum_{x,y} M_c(x, y)$ which proves the fact that the class activation map calculated from the spatial grid (x, y) results in the classification of class c. By upsampling the class activation map to the input image's size, the most relevant regions for the particular category are identified on the image. Some examples of highlighted image regions are provided in Figure 2.8.

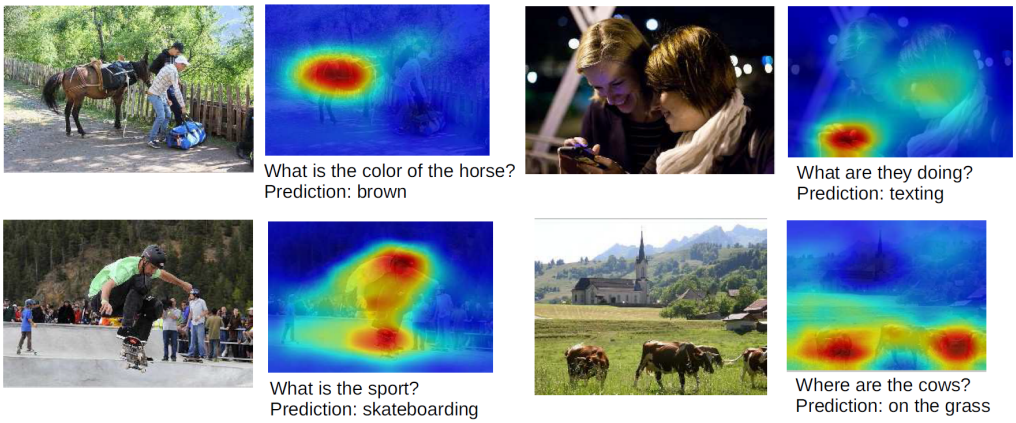

Figure 2.8: A couple of examples of highlighted image regions for the predicted answer class in the visual question answering [16]

2.1.3.2 Gradient-weighted Class Activation Mapping

The original CAM method described in the previous section requires feature maps to be directly connected to the softmax layer. In other words, it is used for classification purposes in CNNs having a GAP layer after the final convolutional layer and then a dense layer without any fully-connected layers between them. Therefore, since CAM is not applicable for most of the cases, in order to use it, the network is needed to be modified and to be retrained. Thanks to the study [2], there is an alternative and more general deep visualization method: Gradient-weighted Class Activation Mapping (Grad-CAM), which computes the gradients of the target function with respect to the layer outputs efficiently with backpropagation in order to produce a localization map highlighting the important regions on the input. This algorithm is simply shown in Figure 2.9. Grad-CAM is applicable to a wide range of CNN models without architectural changes or retraining. On the other side, given that the network is a CAM-computable structure, Grad-CAM converges to CAM.

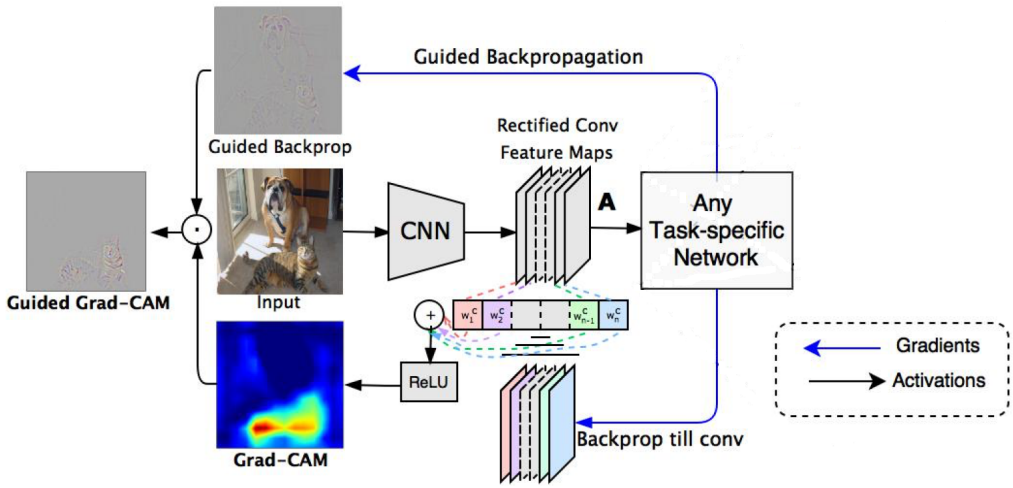

Figure 2.9: The algorithm behind Grad-CAM [2]

Chollet explains the logic behind the Grad-CAM implementation as weighting a spatial map of "how intensely the input activates different channels" by "how important each channel is with regard to the class," resulting in a spatial map of "how intensely the input activates the class" [14]. Formally, in order to obtain the Grad-

CAM, $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$, of width u and height v for any class c; first, the gradient of y^c, the score for class c, with respect to feature maps A^k of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k}$ is needed to be computed, where k is a unit of the layer. Next, these back-propagation gradients are global-average-pooled to calculate the neuron importance weights, α^c_k, as given in Equation 2.6:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i,j}} \tag{2.6}$$

α^c_k captures the interest of feature map k for a target class c. It is combined with forward activation maps and the result is followed by a ReLU operation to get the Grad-CAM output given in Equation 2.7.

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{2.7}$$

Note that the output is in the same dimension as the last convolutional layer which necessitates a resizing process in order to apply to on the original input. On the other hand, the ReLU process is used so that features with positive effects on the class of interest could be observable. Without the ReLU, i.e. not neglecting negative valued pixels, the Grad-CAM localization map might emphasize more than the target class and the classification performance might decreases.

[14] puts forward the following assumption: Let the second before the last layer has K feature maps, A^k, which are pooled by using average pooling method and linearly transformed to obtain the score of class c as shown in Equation 2.8.

$$S^c = \sum_k w^c_k \frac{1}{Z} \sum_{i,j} A^k_{ij} = \frac{1}{Z} \sum_{i,j} \sum_k w^c_k A^k_{ij} \tag{2.8}$$

Note that when Grad-CAM is applied to a specific architecture that has weights w^c_k between feature maps and outputs $\alpha^c_k = w^c_k$. This result proves that Grad-CAM is a strict generalization of CAM. Figure 2.10 illustrates the dog-classification Grad-CAM output when a photo that has a dog and a cat in it, given as input. On the left, the heatmap is seen while the right side is the original photo with the heatmap applied to it.

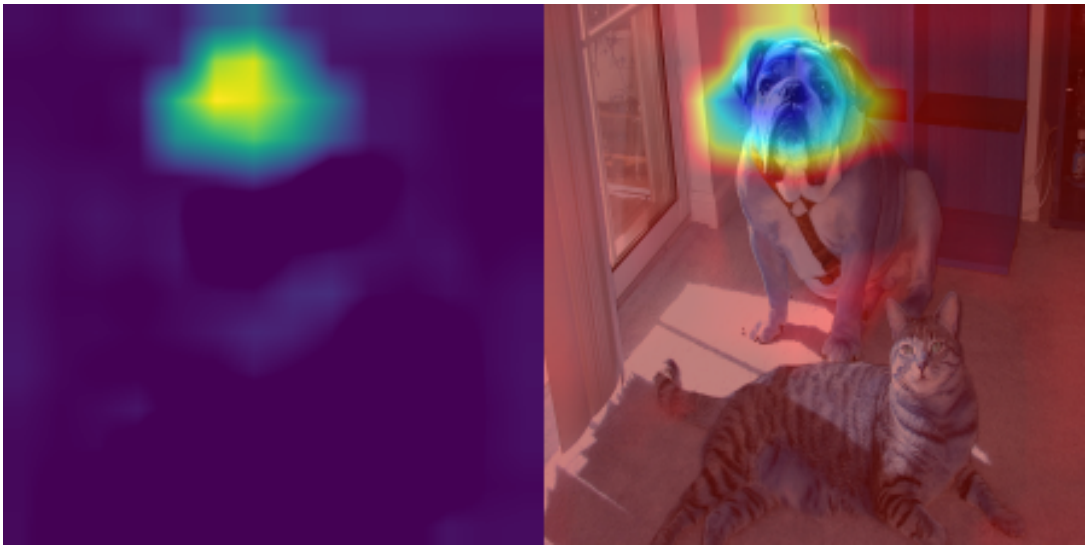

Figure 2.10: The Grad-CAM output when a photo that has a dog and a cat in it is given into a network trained to detect dogs.

2.2 Related Works

Neuroimaging methods that are handy tools for medical experts to attain images of the brain have been popular for a long time to observe the brain's functioning, identify mental illnesses, or even explain the relationship between a disorder and its symptoms. With the development of a promising deep learning field, several methods have recently been published to diagnose mental disorders by using neuroimaging data.

For the purpose of mental disorder classification, there are some state-of-the-art studies have been published. In [17], Lehmann et al. benefited from an electroencephalogram (EEG) to obtain the record of electrical activities of the brain and use them in their classification algorithm with the purpose of early recognition of Alzheimer's. In [18], magnetic resonance images (MRI) of the brain was used to reveal the incidence of schizophrenia and to estimate the effectiveness of pharmacological treatment by the help of an artificial vision algorithm. Besides, in [19], Morillo et al. announced a web platform, namely the Psycho Web, to diagnose patients' mental disorders.

Other than absolute classifications of psychological diseases by using neuroimaging data for training neural networks, the visualization methods applied into these net-

works have been getting more popular day by day since it is an effective way to give medical experts feedback on biological insights that are lied behind the decision mechanism. In [20], Riaz et al. studied with fMRI time-series signals of the Attention Deficit Hyperactivity Disorder patients and control group and constructed a CNN-based deep learning model for classification. They also analyzed feature importance maps for both classes learned by their method and stated that the importance value assigned by the network to particular functional connectivity was different for both classes. Another study was conducted with an fMRI dataset practiced by Sarraf et al. [21], to classify Alzheimer's Disease via deep convolutional neural networks. They generated heat maps of weights, filters, and activations belong to different layers of different networks. Their findings reveal significant differences between Alzheimer's and healthy subjects in various regions of the brain. Similarly, in their work, Feng et al. [22] classified Alzheimer's disease by using cortical morphometric measures derived from structural MRI dataset in spherical CNN. In order to analyze the human cortex behaviours, they generate class activation maps in the spherical form defined in SO(3) space.

Oviedo et al. [23] used an alternative data type, X-ray diffraction (XRD) data, in order to classify XRD patterns of subjects. By using CAM, they visualize the main discriminative regions of an XRD pattern, such as peak and series of peaks, that were used to classify all the training data belonging to a certain class. Consequently, they could identify the root cause of misclassification and could design a more robust experiment. Based on the class activation mapping method, Shi et al. [24] used real-time data from inertial measurement unit (IMU) in a CNN-based model and obtained the heatmap of the original time-series data, which highlights the contributing region. By analyzing this region in detail, they manually extract effective characteristics such as which interval of time-series data can best predict the occurrence of a fall. Ghosh et al. [25], on the other hand, proposed a novel method, i.e. cue-combination for Class Activation Map (ccCAM), that can be used for the networks in which inserting the Global Average Pooling (GAP) layer is not available. To identify the frequency bands that involved important information, they generated the ccCAMs for all time points and group-averaged (averaging across all time points and subjects in a group) the maps. After having two 2D maps for the control and exercise groups, Ghosh et al.

found significantly different behaviours between them.

Gao et al. [26] proposed a novel method named Dense-CAM, which is the combination of DenseNet and CAM in order to visualize the whole network and to generate more accurate and more robust deep model visualization. Thanks to the skip connections between the last layer and any of the front layers in DenseNets, they attained more information for visualization than the original CAM, whose resolution is restricted to the last convolutional layer's size. They asserted that unlike to visualizing the network's final process, obtaining the combined features of the whole network is much more beneficial while interpreting visualization outputs.

In order to compare the resolutions of the CAM from a different viewpoint, the study of Fawaz et al. [27] can be analyzed. They generated visualizations of Fully Convolutional Network (FCN) and Residual Neural Network (ResNet) trained for time series classification (TSC). They explored that thanks to the skip connections in it, the ResNet could filter out discriminative regions with higher confidence than FCN, resulting in the lower accuracy of FCN. Moreover, they observed that CNN is able to localize a given discriminative shape regardless of where it appears in the time series. It proves CNN's capability of learning time-invariant warped features. As a conclusion, they emphasized that interpretable analysis of TSC with a DNN is a significant research area since it enables to identify which regions of a time series data constitute the reason for the classification.

To the best of our knowledge, Wang et al. [28] is the first comprehensive study that introduced one-dimensional CAM with an application to TSC. Their proposed FCN and ResNet were able to classify time series data from scratch. They observed that the discriminative regions of time series data for the right classes are highlighted when visualizing these networks. The visual demonstrations of both networks' filters were observed similarly, even though ResNet tends to overfit the noncomplex data much easier than FCN. However, they expressed that after making an effort to regularize the model, the gradients could flow directly through the bottom layers by using shortcut connections in the ResNet, which vastly improved the model's interpretability. They suggested benefiting from the visualization of ResNet rather than CNN when the data is large and complex.

CHAPTER 3

FUNCTIONAL NEAR-INFRARED SPECTROSCOPY AND DATASET

3.1 Functional Near-Infrared Spectroscopy

Neuronal activity can be determined considering the changes in the brain's oxygenation level since cerebral hemodynamics' variation is directly related to functional brain activity. These changes can be detected with the help of an optical apparatus and light in the near-infrared range. Typically, an optical apparatus consists of a light emitter and a light detector. Photons that interact with tissue are either absorbed or scattered. Light in the near-infrared range with the wavelengths of 700 to 900 nm (the optical window) can penetrate most biological tissues by means of the low absorbance of their main components, as reported by Ayaz et al. [29].

Fortunately, in the optical window, the absorption spectra of oxygenated and deoxygenated hemoglobin (Hb) remain separate. Hence the spectroscopic separation of these compounds is possible when the wavelengths of 695 and 830 nm are used. Once the photons are sent into the human head, they are either scattered by different layers of the head (skin, skull, brain, etc.) or absorbed mainly by oxy- and deoxy-Hb [30]. A photo-detector placed nearly 3 cm away from the light source collects the photons that travel along the "banana-shaped path" between the source and detector due to scattering. The illustration of emitter-detector pairs and the banana-shaped path is given in Figure 3.1.

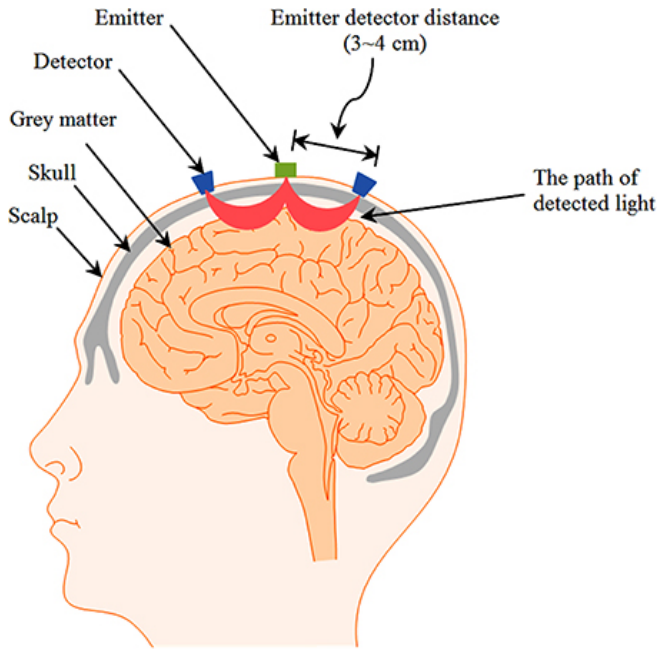

Figure 3.1: The illustration of emitter-detector pairs as well as the banana-shaped path used in the fNIRS measurement [31]

According to Cui et al. [32], functional near-infrared spectroscopy (fNIRS) is an optical non-invasive neuroimaging modality to monitor the concentration of both oxy- and deoxy-Hb particles. It allows making a set of observations of the neural activity [33], while the subject is able to perform a particular task because of the portability and compactness of its apparatus. This ability provides an opportunity to take measurements under natural conditions with subjects sitting on a chair as seen in Figure 3.2. In other words, the fNIRS system's interest is taking a snapshot of the cortical activity across brain regions by acquiring data at each channel with separately located probes.

Unlike to functional magnetic resonance imaging (fMRI); fNIRS does not require the participant to stand still on a bed. Thus, fNIRS can be used for more naturalistic experiments, including face-to-face communication or natural body movements, and is well suited for real-time applications. However, it is challenging to improve signal quality and reduce noise induced by external contributions such as scalp blood flow, blood pressure, heart rate, or head motion.

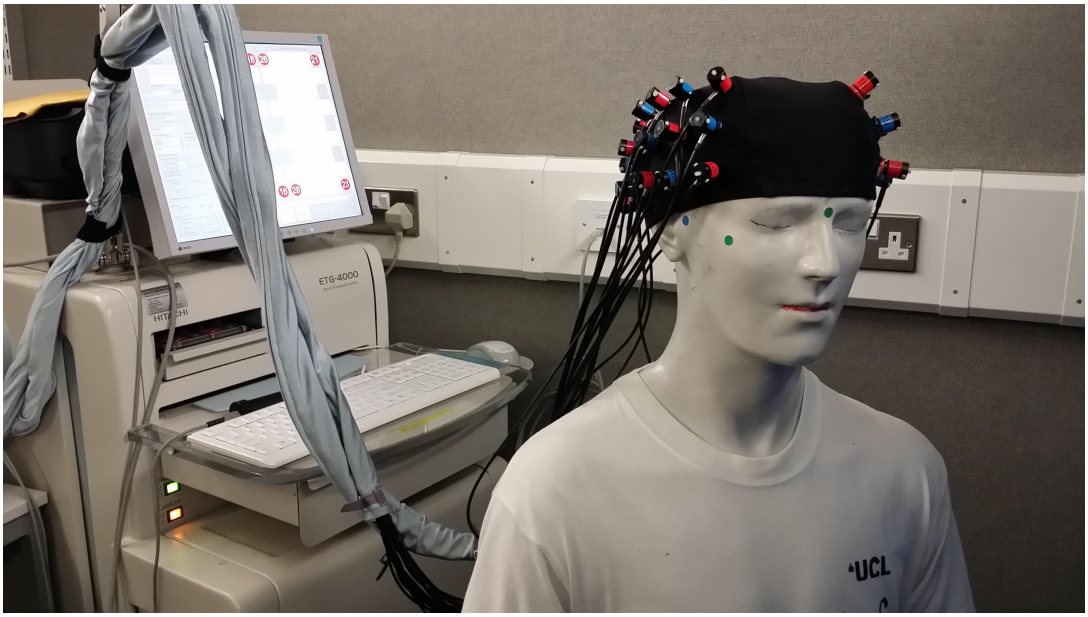

Figure 3.2: Hitachi ETG-4000 optical imaging system to be used for fNIRS measurements [30]

3.2 Dataset

The neuroimaging data used in this thesis is the fNIRS data obtained from Ankara University Brain Research and Application Center laboratory during the period between 1 April 2014 and 30 July 2014, and it is firstly used by Hoşgören [5] for her medical study. Besides the dataset, the information provided in this section is mainly taken from her study [5].

The Hitachi ETG-4000 optical imaging system (HitachiMedicalCo., Tokyo, Japan) is used to collect 24-channel fNIRS data with the resolution of 100 milliseconds, i.e. with the frequency of 10 Hz, enabling a detailed clarification of temporal changes in relative cerebral blood volume. 2 channels from 24 probes of the device are available indeed; however, the only difference is the wavelengths operated in them. Moreover, fNIRS signals are affected by physiological activities such as oscillations of systemic arterial circulation (0.1 Hz), and breathing (0.2-0.3 Hz) as stated by [34]. This problem is overcome by the high-pass (5 Hz), low-pass (0.001 Hz) and motion average filters. During measurements, participants' body movements are automatically de-

tected by the ETG-4000 device, which makes it possible to produce motion average filters. A researcher blind to the study group manually analyzes channels responsible for these artifacts, and outlier data is omitted from the study. According to Allin et al. [35], fNIRS studies with the verbal fluency test prove that subjects with the bipolar disease tend to have lower prefrontal activation. Whereas, Hoşgören [5] does not detect any differences between healthy and the bipolar subjects in terms of brain activation during the test.

All processes concerning the data measurement and experiment are approved by the Ethics Committee of Ankara University Medical Faculty. Before the study, these processes are also announced to all participants, and each of them gives approval both verbally and in written. Bipolar subjects in this dataset are the patients who are diagnosed bipolar disorder according to the DSM-IV-TR and whose IQs are above at least 80. Furthermore, they have all been in the remission from the disease, which is defined as absence or minimal symptoms of both mania and depression, for at least one month. On the other side, healthy subjects do not have any psychological disorder according to the SCID-I. Moreover, there is no difference between healthy and bipolar subjects in terms of age, gender, IQ, and educational background.

Verbal fluency test is one of the tests that can be performed while fNIRS data being collected. In this test, a subject is asked to generate as many words starting with the given letter as she/he can in a certain time. In the original form of the test /f/, /a/, /s/ letters are used to generate words; however, thanks to Tumaç [36], in the adapted form to Turkish language, the most commonly used 3 letters in the Turkish words, i.e. /k/, /a/, /s/, are chosen. The participant's overall test point is calculated by summing up the numbers of words generated by each of these letters in a minute. In the study that the fNIRS data collected, the verbal fluency task (VF-task) is designated by adopting the verbal fluency test. It comprises two blocks in a single session with a total duration of 225 seconds.

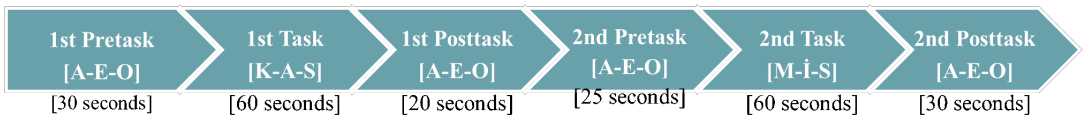

Figure 3.3: The time table of the verbal fluency test with its 6 partitions shown

In the first 30 seconds, namely the first pre-task, the participant is asked to repeat /a/, /e/, /o/ letters sequentially without a break so that the brain's basic activity while talking can be determined. After that, the first task is carried out. The participant is supposed to generate as many words as she/he can, starting with the letter /s/ for 20 seconds, then with the letter /a/ for another 20 seconds, and finally, with the letter /k/ for once again 20 seconds. Next, during 20 seconds, the participant again repeats /a/, /e/, /o/ letters after and after. The second block is then repeated with the sequence of the second pre-task, second task, and second post-task; 25, 60, and 30 seconds respectively. In the second task, the letters being used for word generation are /m/, /i/, and /s/, sequentially. All 6 partitions belong to the VF-test are given in the timetable in Figure 3.3.

Since a single subject's data is collected from 24 channel devices with 100 milliseconds resolution, from each channel 2251 samples, 54024 samples in total, are taken. As there are 2 different wavelengths for each channel, the data consists of 108048 samples. Even though 82 subjects participating in the test, only 71 of them are ready to use. This is because 5 of them are left out due to the lack of the noise cancellation process, and 6 of them have missing parts in their fNIRS data.

Thanks to Evgin's study [4], working with basic architectures, overfitting occurs owing to outliers in the data, regardless of any modification made. To find them out, the subjects are trained and tested using random shuffles with relatively small test sets. After each subject's success, some subjects have insufficient classification successes even with a great number of trials. Consequently, the worst 10 subjects are removed resulting in that 61 subjects, 33 control and 28 bipolar, are left.

CHAPTER 4

BACKGROUND INFORMATION

4.1 Deep Learning

In recent times, deep learning has impressed daily life in various areas such as intelligent web search, online advertising, and pattern recognition. Besides, by improving deep learning methods, the human-level AI has been developed as well, see [37], [38] for more discussions. With deep learning techniques, computers are allowed to operate without being explicitly programmed. They are constructed with algorithms that they can learn from data and make data-driven decisions. Even though it is quite complex to understand how neural networks work, the logic behind them is straightforward.

First of all, a neural network consists of neurons connected to each other, while each connection of the NN has a weight that determines how important the relationship of the connected neurons is. Besides, each neuron has an activation function that defines the neuron's output. The decision of how the network being updated according to the loss function is given by an optimizer. After training the network, each neuron's weights and bias are learned, which is the unique part of deep learning. Until this learning procedure is finished, the neural network process iteratively in the forward and backward directions.

In the forward propagation, the input data is passed through the network. All neurons in a layer receive the information from the previous layer's neurons, process it according to the activation function, and transmit it to the next layer's neurons. After the data crosses all the layers with this procedure, the final layer is reached, and the label prediction is made. Once the forward propagation process is done, a loss func-

tion is used to calculate the error and to measure how good/bad the predicted result is compared to the correct result.

Next, the loss score is calculated, and this information is propagated backward, starting from the final layer. It means that the loss information is sent back to all the neurons in the hidden layers contributing to the prediction. This process is repeated, layer by layer, until all the neurons are informed about their individual contributions to the total loss, and they are updated to lessen this loss. Finally, when forward propagation and back propagation operations are repeated numerously, the weights of connections between neurons are adjusted so that the loss can be as close as possible to a certain extent. When the number of layers and units in a single layer increases, the NN is called Deep neural networks (DNN), representing functions with higher complexity. The whole process is visualized in Figure 4.1.

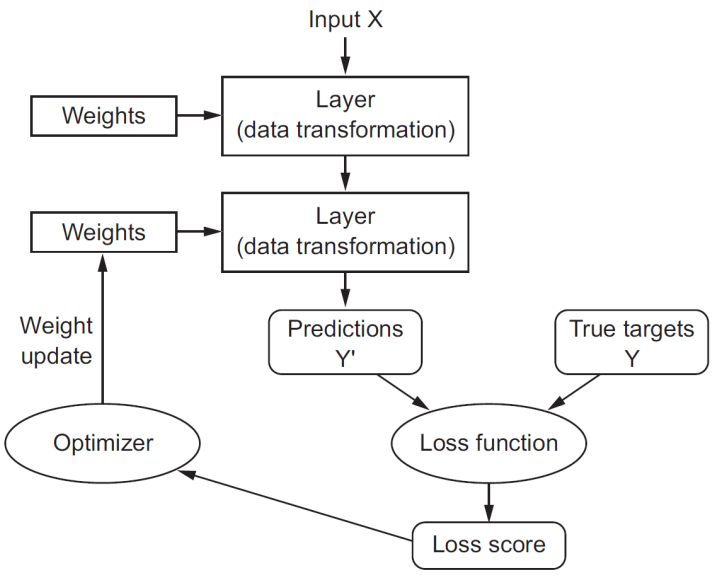

Figure 4.1: The general algorithm behind neural networks [14]

In this study, the labeled multivariate time series dataset belongs to healthy and bipolar subjects is used to detect the disease. Therefore, a supervised multi-channel deep neural network is needed to be constructed as a classifier. As reported in [39], Convolutional Neural Network (CNN) and Residual Neural Network (ResNet) are able to classify time-series datasets with premium performance. Considering that, Convolutional Neural Network and Residual Neural Network are chosen to be built for this

classification problem during the experimental procedure.

4.1.1 Convolutional Neural Network

Having been used for numerous applications, CNNs have been constructed with many structures containing complexly interconnected layers. According to Qin et al. [7], the convolution filters convolve with their inputs resulting in learned features. The neurons of deeper layers are expected to extract more complicated features. Finally, after the whole feature map is extracted, the network converges to the classification output. One of the most common CNN architectures can be seen in Figure 4.2.

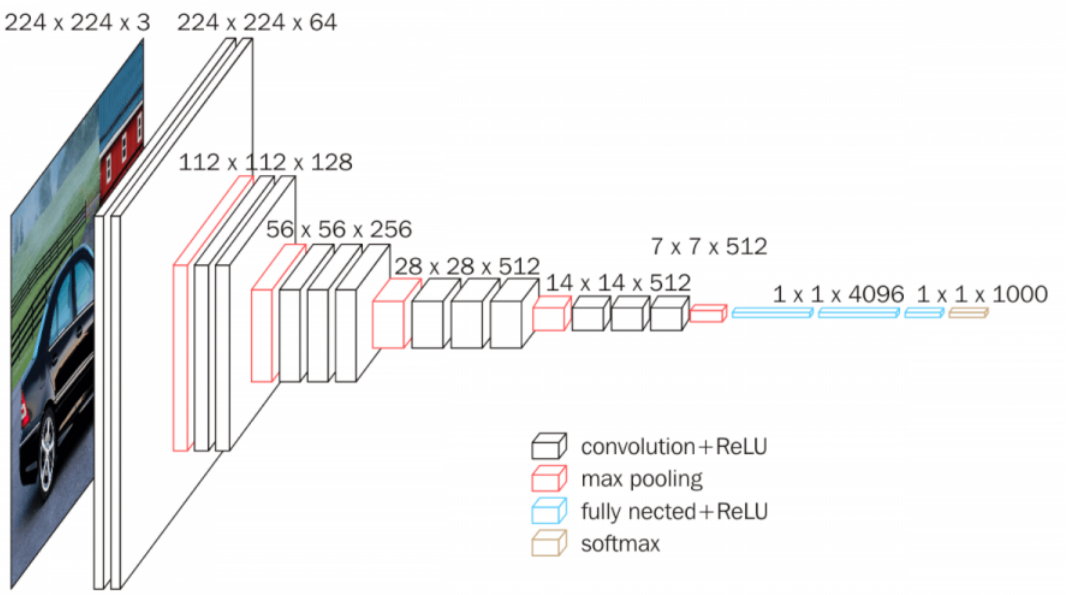

Figure 4.2: The VGG16 network, which is the most common CNN architecture

Therefore, with the increase in complexities of problems, networks become more complex, and the number of layers, the layer depth, increases as stated in [7]. Moreover, for training and testing procedures, the network is supported by sophisticated and well-defined algorithms and back propagation methods. Consequently, a vast amount of labeled data is required to enhance network ability by iteratively training the extensive neurons, as well as the interconnection between them.

4.1.2 Residual Neural Network

Deep learning algorithms are based on the idea that the deeper the hierarchy of layers, the higher the representations of patterns. Therefore, in order to get a higher accuracy level, researchers tend to increase hierarchical compositions through deep networks. However, by adding new layers one after another, a network could encounter the degradation problem. It arises when the network's accuracy goes up until to a level and becomes saturated, which is followed by rapid degradation, see Figure 4.3. Unexpectedly, such degradation is not caused by overfitting but by adding more layers to a suitably deep model as verified in [39].

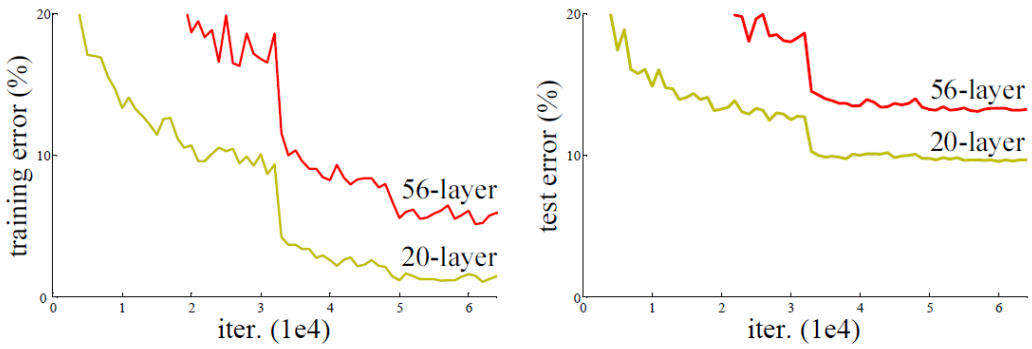

Figure 4.3: Training error (left) and test error (right) on the 20-layer and 56-layer networks [39]

Considering the presence of the degradation problem, optimizing the solution by constructing a deeper network is nothing but adding shortcut connections between nonsuccessive layers as He et al. introduces [39]. They suggested a residual block architecture with identity mapping as shortcut connections.

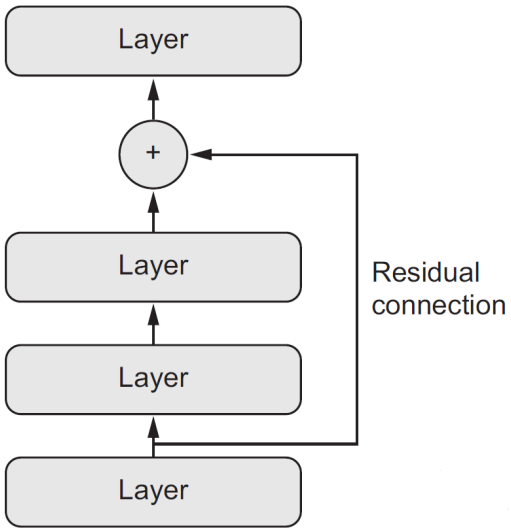

Figure 4.4: The illustration of a residual layer with a shortcut connection

As shown in Figure 4.4, shortcut connections do not add extra parameters and do not increase computational complexity. Moreover, lower layers' inputs are forwarded to deeper layers in the form of shaped information instead of abstract information. Thus, the operation of the network is favorable to increase accuracy.

4.2 Statistics

Under some circumstances, it may become too difficult to obtain a population's true data. At this stage, inferential statistics come out as a solution to use a sample from the population to make reasonable guesses about the population's features. However, it is important to use unbiased samples in the statistical tests. If the sample is not a fair representation of the population, then valid statistical inferences cannot be made. Inferential statistics have two main uses on populations:

- making predictions about them
- testing hypotheses to speculate about them

Provided that collected data on income and job of a randomly selected sample of people in a country, inferential statistics can be applied to estimate the mean income

or test the hypothesis of the relationship between income and job about the whole employed population of the country thanks to the sample data.

In this study, no estimation about populations is made; instead, it is analyzed whether the data of 2 groups, healthy and bipolar subjects, is used to obtain information about their corresponding populations' insights. Hence, we are interested in testing the hypotheses side of statistics. We can use it for the purpose of populations' comparison or inquisition of the relationship between variables.

Inferential tests are either parametric or non-parametric. As parametric tests usually have stricter requirements than non-parametric tests, they can make more statistically robust inferences for a population. The preconditions that the sample data is supposed to meet are as follows:

1. **Normality of data**: the data must be normally-distributed.
2. **Independence of observations** (i.e. no auto-correlation): The observations/-variables included in the test must not be related to each other.
3. **Homogeneity of variance**: the variance of a population being compared must be similar to that of other populations. If one of them has a significantly different variance, it will limit the test's effectiveness.

If data does not meet the assumptions of normality or homogeneity of variance, a non-parametric statistical test can be performed. Non-parametric statistical tests are used with the data that can be labeled without providing any quantitative. Each observation corresponds to no more than a single category. On the other side, parametric tests assume specific characteristics of a data set. Statistical tests are divided into three categories: *tests of comparison, correlation, and regression*.

4.2.1 Statistical Test Types

4.2.1.1 Comparison Tests

Comparison tests are proper tools to analyze if the means of populations differ from each other in order to observe the effect of a categorical variable on some other char-

acteristics of the population. If three or more groups are compared, or multiple pairwise comparisons are made, using an ANOVA or a post-hoc test is a reliable solution. On the other hand, the comparison of two groups' means, such as the average electricity use of two cities, can be made by using the t-test.

In this study, there are two groups in the population, and the categorical variable, namely bipolar or healthy, are used as the predictor. On the other hand, the outcome can be measured as a quantitive variable, such as the average activation value of visualization output. Hence, the t-test is chosen as one of the statistical tests used in Section 5.3 to observe whether subjects with bipolar disease have different distributions of visualization outputs compared to healthy subjects.

4.2.1.2 Correlation Tests

If two variables associate with each other can be investigated by the correlation test. There are three main correlation tests. While the most powerful one is Pearson's r test, the Spearman's r test is suitable for interval variables when the data is not normally distributed. Other than these, the chi-square test uses nominal variables. Therefore, it is used in the experiments of this study to check whether there is a relation between the predictor of the categorical variable, i.e. bipolar or healthy, and its outcome, i.e. the time partition that maximum visualization output occurs.

4.2.1.3 Regression Tests

Regression tests inquire if changing a predictor variable affects an outcome variable. Most of the regression tests are parametric, meaning that the data is supposed to be normally distributed. However, since this study's fNIRS data does not meet the regression test conditions, this type of statistical test is neglected in the experiments.

4.2.2 Definitions of Statistical Parameters

4.2.2.1 P-value

Calculating a test statistic is a necessity for all statistical tests. It refers to a value providing the difference between the test's variables' relationship and the null hypothesis of no relationship. As an alternative for the test statistic, a probability value, i.e. p-value, can be used. It defines a probability that the results from the sample data of a population are occurred by chance. To illustrate, a p-value of .01 means that the probability of the results happening by chance and the null hypothesis being true is 1%. In a general manner, the p-value is decided as .05 (5%).

4.2.2.2 Confidence Interval

To overcome the statistical test's variability, it is more convenient to provide an interval estimate instead of a single value for a parameter. It is called the confidence interval, and it is related to a confidence level, which defines the probability of the interval comprising the parameter estimate. A 95% confidence interval means that if a research is repeated 100 times by using the same method but different sample data, it is expected to observe the parameter estimate residing in the specified range for 95 times. However, it cannot be deduced that the actual parameter of the population lies within that range. In order to be sure about that, it is required to collect the data of the full population. However, provided that random sampling and suitable sample size are ensured, the parameter is supposed to be in the confidence interval a certain percentage of the time.

4.2.3 Statistical Tests Used in Experiments

4.2.3.1 Chi-Square Test

Chi-square tests are non-parametric statistical tests for categorical variables. Types of chi-square test, as well as the interpretation of the test, are given in the following sections.

The Goodness of Fit Chi-Square Test The goodness of fit chi-square test investigates whether the distribution calculated from the theoretical frequencies coincides with that of observed frequencies of the sample. In other words, it is a good way to check if the sample's frequency distribution matches what is expected from the broader population. Given that the actual frequencies are close to the theoretical ones, the chi-square statistic will be small, which concludes in there will be consistency between theoretical and actual distributions. The mathematical deduction for the corresponding hypothesis is given in Equation (1.9).

Null Hypothesis (H_0): The data follows a specified distribution. χ^2 Equation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E} \tag{4.1}$$

where, E_i is the expected value and O_i is the observed value for bin i when the data follows the specified distribution. To calculate the expected value for bin i:

$$E_i = N * p_i \tag{4.2}$$

Here, p_i is the hypothesized proportion of observations for bin i and N is the total sample size.

The Chi-Square Test of Independence On the other side, the chi-square test for independence compares two nominal variables from a single sample by using a contingency table to check whether they are independent. In other words, it can be used to reveal the significance of the relationship between two categorical variables. For example, after collecting data on hobbies and job for each participant, one can benefit from this test to investigate whether hobbies are related to the occupation.

Null Hypothesis (H_0): The two categorical variables are independent, i.e. there is no relationship between the variables.

To calculate the expected value for a cell:

$$E_{ij} = \frac{R_i C_j}{N} \tag{4.3}$$

where R = row, C = column, N = total, for ith row and jth column.

χ^2 Equation:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E} \tag{4.4}$$

Performing Chi-Square Test Theoretically, when the actual and expected values were the same, the chi-square test statistic would be zero. Therefore, it can be stated that the smaller the test statistic, the more similar the observed and expected data.

A chi-square test statistic is a single value that defines the significance of the difference between observed values and the expected values, considering that there is no relationship in the population. The decision of if a test statistic is sufficient to indicate a significant difference is given by comparing the calculated chi-square value and a critical value from a chi-square table. On the condition that the chi-square value is greater than the critical value, then the difference is significant enough concerning the specified criteria. As an alternative, the p-value can also be used for the same reason. After the null and alternate hypotheses are set, a test statistic, as well as a p-value, are calculated and interpreted.

4.2.3.2 T-Test

The t-test is used to show the significance of the difference of two groups' means. It benefits from the hypothesis testing to determine if two groups are different or if a process affects the population. The t-test also investigates if the differences could have happened by chance; in other words, it reveals how significant the differences are. The t-test is a parametric test; therefore, the same assumptions that the data is supposed to meet provided in the previous subsection are valid as other parametric tests.

Types of T-Test The decision of which t-test needed to be used is given based on two things. Firstly, the t-test is divided into three types concerning if the comparison is made between the groups from a single population or two populations:

- **Paired t-test:** Suitable, if the groups of the same population at different times

are used (e.g. the distance covered by a team in the first and second halves of a match).

- **Two-sample t-test (i.e. the Welch's t-test):** Valid, if the groups come from two separate populations (e.g. comparison of the average test scores of males and females).

- **One-sample t-test:** Performed, if one group is compared against a known value (e.g. comparing the income of a coal miner to the minimum wage of 2825 Turkish Liras).

The other criteria on variations of the t-test is which direction of the difference being tested:

- **Two-tailed t-test:** Performed, if the only criteria is two populations being different from one another.

- **One-tailed t-test:** Performed, in order to check whether the mean of one population is greater or less than the other.

In this thesis, we benefit from the t-test to compare the means of two populations, i.e. bipolar and healthy subjects. By this means, we investigate whether two groups, bipolar and healthy, are distinguishable from each other regarding their visualization outputs. Hence, the two-tailed, two-sample t-test is preferred.

Performing T-Test The t-test evaluates the difference between two group means by comparing the difference in the means of each group, considering both groups' standard errors. In other words, the t-value gives the ratio of the two groups' difference versus the difference inside the groups. A t-value of 5 means that a sample from a group is five times different from the other group as it is within the original group.

The formula for the two-sample t-test, i.e. the Welch's t-test, is given in Equation 4.5. In this formula, n_1 and n_2 are the numbers of observations in each group being compared; while x_1 and x_2 are their means, respectively. s_2 is the standard error of

these two groups and t is the t-value.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)}} \tag{4.5}$$

As in other inferential statistical tests, in the t-test, a larger test statistic value indicates a more significant difference between the groups. The calculated t-value can be compared with the corresponding critical value from the t-value chart according to the degrees of freedom and the specified confidence interval. Suppose the calculated t-value is greater than the true t-value from the confidence table. In that case, the null hypothesis that the two groups are equal can be rejected, and it is concluded that the two groups are different.

CHAPTER 5

IMPLEMENTATION

This section introduces the architectures and methods used to classify bipolar disease on the subjects chosen. The features are extracted using the proposed architectures, and the classification is applied using these methods. Due to the limited number of participants, only 41 subjects, 21 from the control group and 20 from bipolar patients, are chosen as the training set. For all the networks being evaluated, the same training sets are used for a fair comparison.

NVIDIA GTX 1650 GPU is used for the training process, and the code is implemented by using Tensorflow 2.1.0 and Keras API 2.3.1. Adam Optimizer, an algorithm for first-order gradient-based optimization of stochastic objective function, is chosen as an optimizer in all networks. It is based on adaptive estimates of lower-order moments. It is computationally efficient and is well suited for large problems in terms of data or parameters. Most importantly, it is also appropriate for non-stationary objectives and problems with very noisy or sparse gradients, making it suitable for fNIRS data. Decay rates of the first and second moments of the optimizer are set to 0.8 and 0.9. The networks are trained with a batch size of 5 samples for all networks. Furthermore, the learning rates of all networks are set to $3e^{-5}$. Other than these settings, special tuning processes related to the CNN and the ResNet are covered in the following sections.

5.1 Deep Neural Networks

5.1.1 Convolutional Neural Network

The models constructed and tuned by Evgin [4] are taken as a baseline for both the CNN and ResNet architectures used in this study. In the fully convolutional neural network, the sigmoid activation function is preferred over the ReLU activation function since it gives better performance when considering that fNIRS data has negative feature points. The network has three convolutional layers directly connected. After the first layer, there is a batch normalization process in order to overcome difficulties caused by the nature of fNIRS data. Each layer has 128 filters with the kernel size of 16, and they are convolved by a stride size of 3. After the last convolutional layer, global average pooling (GAP) and dense layer are inserted to minimize overfitting by reducing the total number of parameters in the model. The classification's key layer is the GAP layer, which establishes the mapping relationship between feature maps and classes. This layer is crucial for this study because the Class Activation Map is chosen as a visualization method during the experiment. Hence, to generate CAM, the network architecture must have a GAP layer after the final convolutional layer, and then a linear (dense) layer, as declared by [16]. The layer structure is visualized in Figure 5.1.

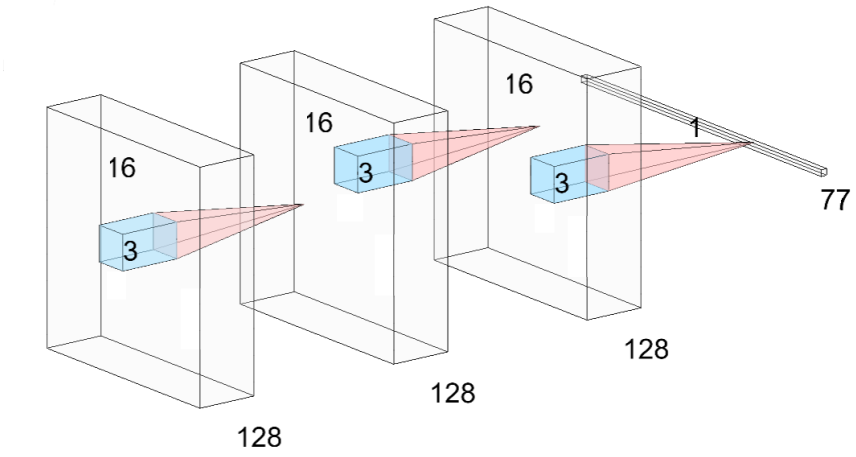

Figure 5.1: The diagram of the 3-layer Convolutional Neural Network that is used in the study

5.1.2 Residual Neural Network

In the CNN model, when the number of layers getting greater than 3, the accuracy sharply diminishes, and the model is not trained as expected. Therefore, to evaluate deeper models while preserving the capability of training, Residual Neural Networks are chosen as an alternative model for this research. Thanks to Evgin study [4], a tuned ResNet is developed with appropriate parameters. After different trial and error cycles, the number of layers and between which layers shortcuts tie are decided. The best performance model includes six layers and shortcut connections between input-3^{rd} layer and 3^{rd}-6^{th} layers. The first 3 layers constitute the first block, while the second block consists of the remaining 3 layers. Each layer of these blocks and the shortcut layers include 64 filters. Besides, the kernel sizes of the first, second, and third layers of each block are 16, 8, and 4, respectively, while the kernel sizes of both shortcut layers are 1. The diagram of the model is provided in Figure 5.2.

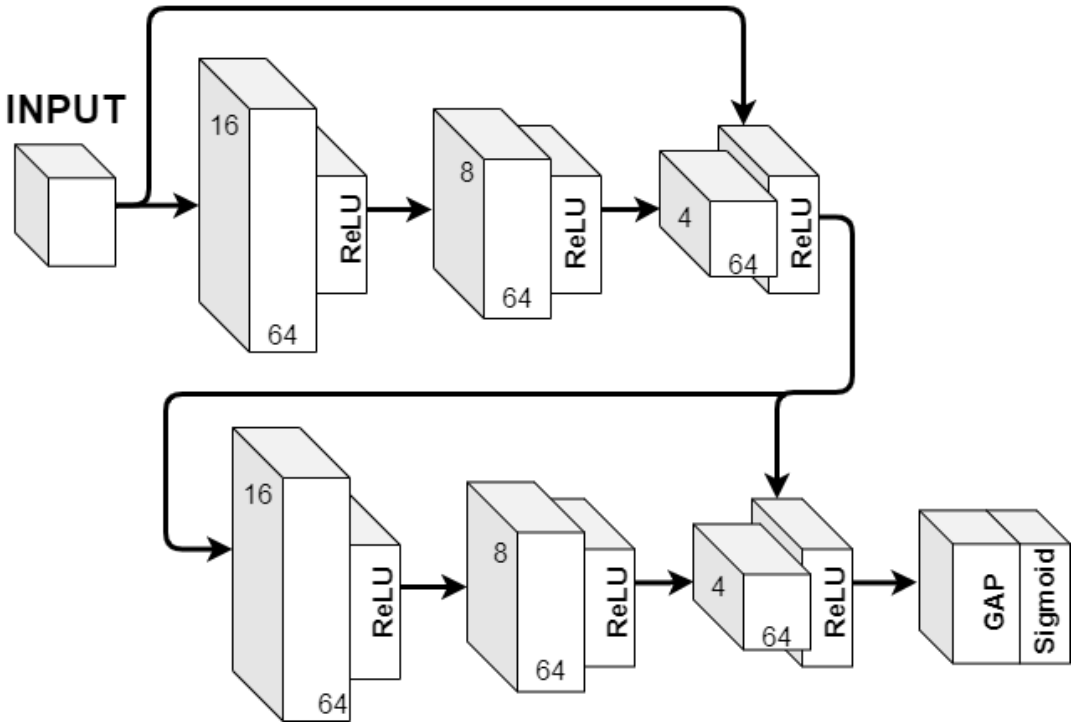

Figure 5.2: The diagram of the Residual Neural Network that is used in the study

Like the CNN model, the ResNet model has batch normalization after the first convolutional layer. It ends with a dense layer after the GAP layer, which is the condition

for CAM.

5.2 Class Activation Maps

This study aims to visualize activation maps of networks that are trained for the diagnosis of bipolar disease by using time-series multivariate fNIRS data to understand the networks' decision mechanism. Experiments are conducted with the CNN and ResNet models detailed in the previous section to evaluate class activation maps according to the purpose. From the visualization methods mentioned in Section 2.1, visualization of intermediate layer activations and visualization of convolutional filters are unsuitable for this case. This is because it is too complex to interpret intermediate layer activations and filters when time-series input is used because the human perception of its visualization is not so understandable as an image visualization. The Grad-CAM method, on the other hand, converges to CAM in case the network already has a CAM-computable structure [2]. Additionally, in deeper layers, the features become more sparse and localized, and visualization helps to explore any potential dead filters, i.e. all zero features for many inputs. All in all, the visualization of the last convolutional layer is the most useful method, and it is chosen among other alternatives for this study.

Consequently, 1-D CAM makes it possible to identify which regions of time series input constitute the bipolar disorder classification; hence, it is preferred as an application to time-series data throughout the study. 1-D CAM is a time-series output whose content coincides with that of input data explained in Section 3.2, see an example of it in Figure 5.3.

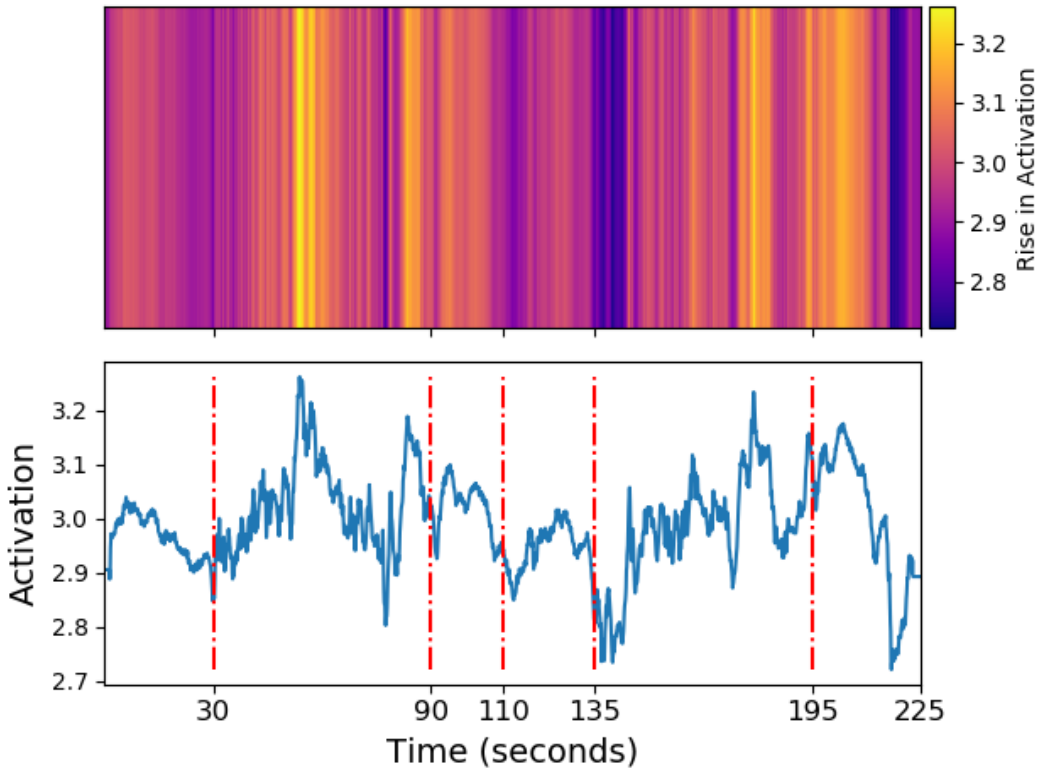

Figure 5.3: The Class Activation Map generated by the ResNet model for the 'Subject 35', with the healthy label. Partitions of the VF-test lie between red dashed lines. Their names from left to right: *Pre-task 1, Task 1, Post-task 1, Pre-task 2, Task 2, Post-task 2*

The above figure is the class activation map belonging to the 'Subject 35' generated from a ResNet model. In this example and others, the 24-channel fNIRS data is used unless otherwise specified. There is a heatmap changing with time samples on the upper side of the figure, making it a time-dependent heatmap. The change in heatmap from yellow to dark blue corresponds to changing from highly discriminative region to region with no contribution to the classifier's decision. The color bar for the heatmap is provided on the right of the above figure, and it can be used for other figures given below as a reference. Below the heatmap, the activation map's actual output can be seen with the time in seconds on the x-axis and activation magnitude on the y-axis. The x-axis values from 0 to 225, which corresponds to 225 seconds verbal fluency test period in which there are 6 unequal partitions explained in Section 3.2

in detail. On the other hand, the activation distribution varies from subject to a subject according to the value of the subject's raw fNIRS data collected during the test. Therefore, it can become meaningless while comparing few subjects because their raw data magnitudes can vary due to the environmental factors to which the subjects are exposed. However, considering all healthy and bipolar subjects as groups, their distributions gain meaning and make it possible to acquire significant interpretations. Furthermore, the activation stays consistent in the same subject's heatmap, enabling us to compare activations belonging to different time samples of the same subject.

In Figure 5.4, CAM outputs of the CNN and ResNet models belong to the 'Subject 32', who is labeled as healthy, are given. It shows that both networks' activation maps follow similar distribution; however, the ResNet's output has a sharp rise and fall, whereas CNN contains a smooth pattern. The reason for the ResNet models having noisy shapes is that the size of the last convolutional layer of the ResNet model is the same as the input data, i.e. 2250, while that of CNN is much smaller, i.e. 77, which results in smoother distribution and lower resolution compared to the ResNet.

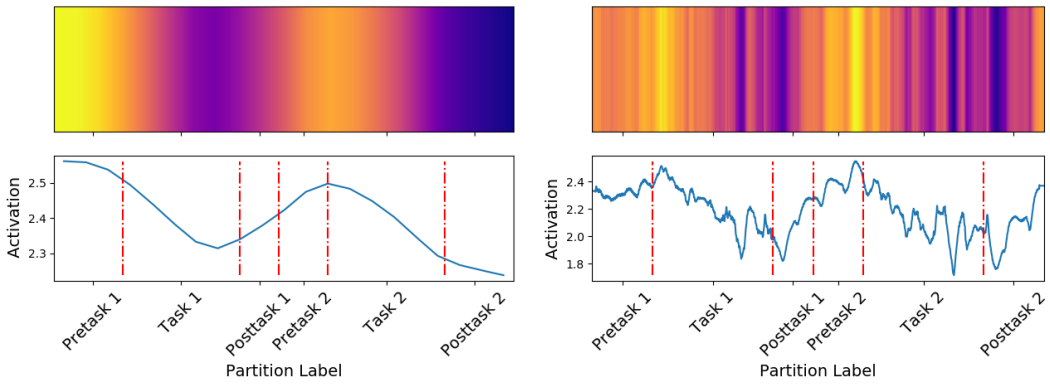

Figure 5.4: The Class Activation Map generated by the CNN (on left), and ResNet (on right) networks for the 'Subject 32' with the healthy label.

In the above graph, there are two peaks where this subject's classification decision mostly comes from. The first peak starts with the beginning of the test, i.e. the first pre-task of the test, and consistently drops till the end of the first VF task. Then, after the Post-task 1 and the Pre-task 2 with relatively low activation, the second peak occurs with the beginning of the second VF task and activation decreases over time

until the end of Post-task 2.

5.2.1 Modifications on CAM

Other examples of heatmaps of activation in time for different subjects, namely the 'Subject 64' and 'Subject 15', with the bipolar label are given in Figure 5.5 and Figure 5.6, respectively.

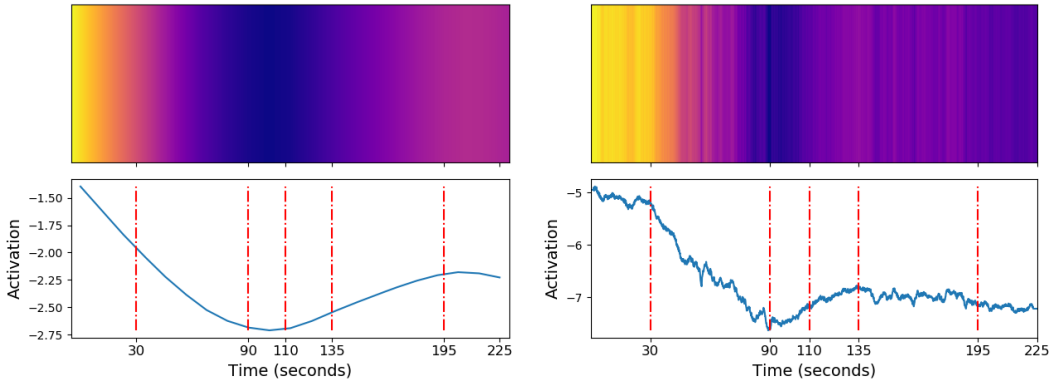

Figure 5.5: The Class Activation Map generated by the CNN (on left), and ResNet (on right) networks for the 'Subject 64' with the bipolar label

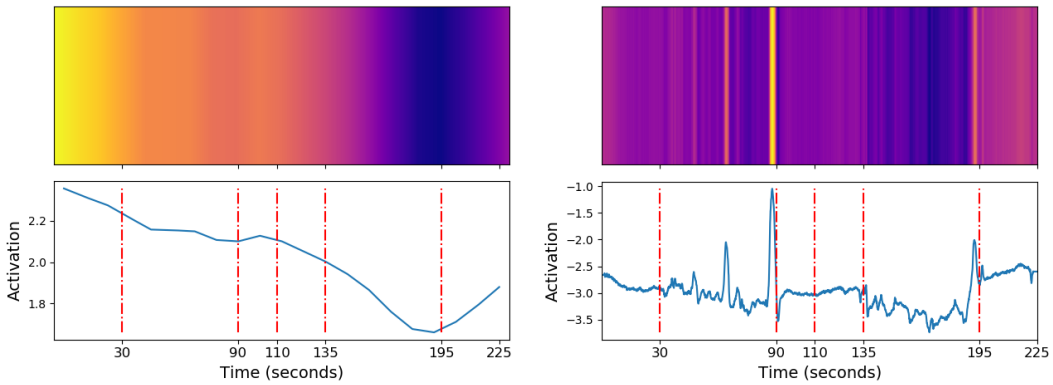

Figure 5.6: The Class Activation Map generated by the CNN (on left), and ResNet (on right) networks for the 'Subject 15' with bipolar label

As it can be deduced from these figures, the CAM outputs of subjects even with the same label, can have different distributions in the time scale. Consequently, it is

impractical to directly observe these heatmaps and make a consistent remark from them. Therefore, the traditional visualization approach is needed to be updated.

The CAM outputs acquired from our networks have the same dimension as the time-series input data. They involve 2250 time points corresponding to the 225 seconds VF-test period in which there are 6 unequal partitions explained in Section 3.2 in detail. Psychiatrists generally focus on these partitions instead of single time points. They make inferences about partitions, more specifically the whole task durations, not a single time sample. Therefore, our study gains meaning if the VF-test is divided into these partitions and if each partition's characteristic is revealed and compared. For this purpose, we average activations of each partition so that the statistical tests can be conducted in a single partition and their results can make sense from the perspective of psychiatrists. Besides, averaging of activations enables us to generalize the distribution in the partition while the characteristic feature of the partition is preserved. As an alternative modification, we sort 6 partitions of the test to find the partition with the maximum activation distribution. This method gives us the flexibility to discover the partitions' importances in the whole test duration while differentiating bipolar and healthy subjects.

5.2.1.1 Averaging of Activations

Between boundaries of a partition, which is explained in Section 3.2, the average of all activations is calculated and is assigned as the new activation value of that partition. This process is applied for all partitions, and the heatmap is updated accordingly. Thanks to the application of the averaging method, we have 6 activation values, one for each partition of fNIRS data of a subject. It enables us to compare those of all subjects with respect to their classes. One example of heatmaps with averaged activation is shown in Figure 5.7. After this phase, the t-test is one of the most suitable statistical methods to check whether there is a significant difference between healthy and bipolar subject populations, which will be performed in Section 5.3.2.

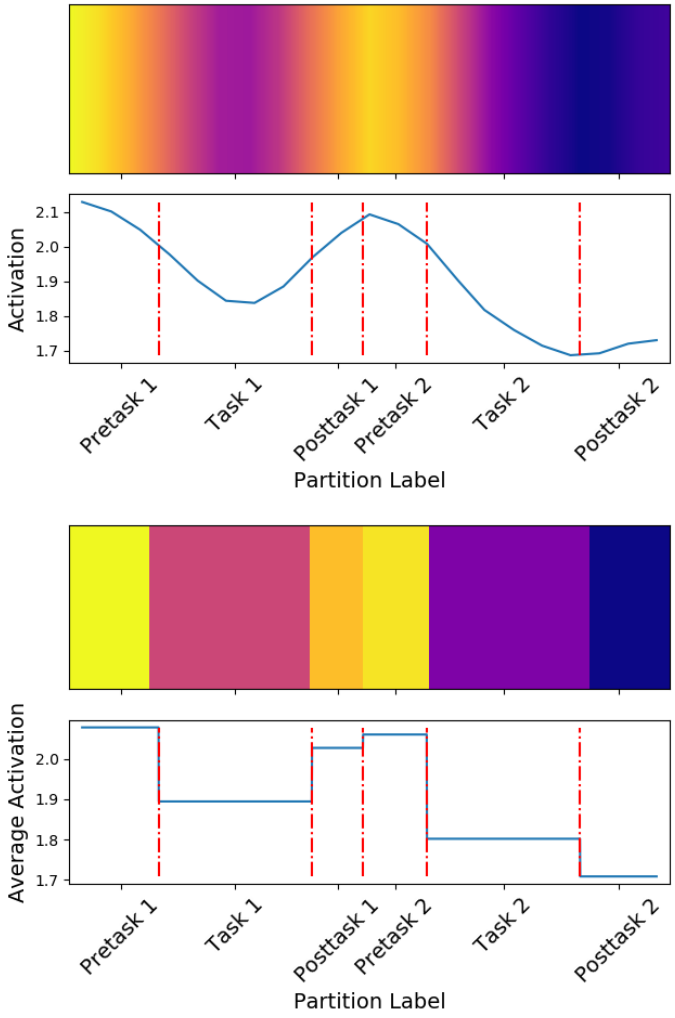

Figure 5.7: The CAM (above), and the average CAM (below) generated by the CNN for the 'Subject 60' with bipolar label

5.2.1.2 Sorting of Activations

Other than averaging activations of each partition of a subject, we sort them in order starting from the most active partition to the least active one. Therefore, we are able to examine if bipolar disorder is related to a specific partition of class activation maps and if it is detected by only looking at that partition of fNIRS data of a subject. To simplify what is done with the sorting of activations method, Figure 5.8 can be examined.

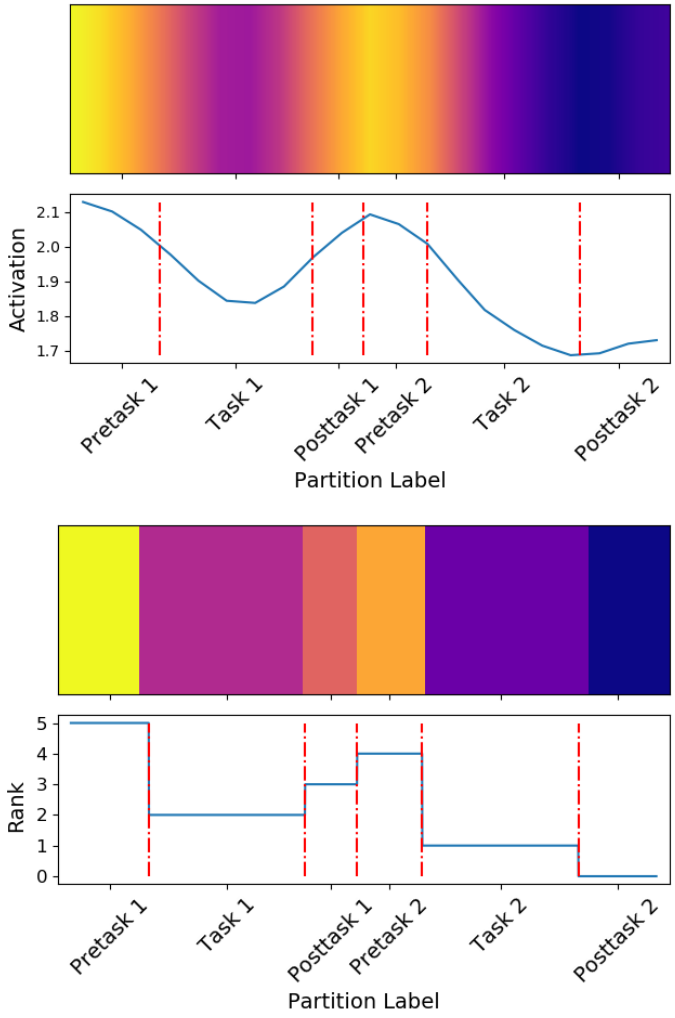

Figure 5.8: The CAM (above), and the sorted average CAM (below) generated by the CNN for the 'Subject 60' with bipolar label

In this figure, the CAM modified with the sorting method for the CNN model is given. As shown from the above graph, the maximum activation is in *Pre-task 1* which results in assigning the biggest rank, i.e. 5, for this partition in the below graph of the figure. Then, the second maximum activation occurs in the middle of the above graph corresponding to *Pre-task 2* having the second biggest rank, i.e. 4, in the below graph. This sorting process continues until the least active partition , i.e. Rank 0, is determined.

5.3 Results

Applying modifications on heatmaps, we can benefit from inferential statistical tests as the independence of observations and homogeneity of variance properties hold. It means medical experts can understand and interpret how fNIRS data of bipolar and healthy subjects behave. Since we investigate the pre-trained networks' insights and try to understand how they classify bipolar disorder, we are obliged to study with the training set because it is the only dataset that the networks benefit from while learning the classification features and updating themselves accordingly. In other words, we use the training set for the statistical tests as this study covers the analysis of the trained networks and the extraction of the biological insights lied behind the classifiers' decision mechanism.

5.3.1 Chi-Square Test

For the goodness of chi-square test, we benefit from the sorting of activations method explained in previous section. After that, we have all training subjects with their 6 partitions sorted. Finally, partitions with maximum activation of healthy and bipolar subjects can be used in the goodness of chi-square test. However, it is not useful in this case since the expected distribution of partitions with maximum activations for healthy and bipolar subjects can not be estimated. The only inference obtained with this test is whether the distribution follows the uniform distribution, and this can not provide any meaningful information.

The chi-square test of independence, on the other hand, can give information about the relationship between bipolar disorder and the partition number that maximum activation occurs. While investigating this, two different networks, namely the CNN and ResNet, are used with two different input combinations. These combinations are generated by changing the channel number of fNIRS data. As given in Section 3.1, the number of probes of the fNIRS device is 24, and there are two different wavelengths used in each probe to collect samples. Even the fNIRS data related to both wavelengths resemble each other; it can be used as it is the 24-channel or 48-channel data. Despite there is no significant difference between both input combinations, the

features that networks extract might differ. For this reason, we use 24-channel input data, as well as the 48-channel one, to obtain more information about the two populations and the network.

The results belong to CNN models with the combinations mentioned above are given in Table 5.1. When the alpha value is specified as .05, the table indicates we cannot reject the null hypothesis such that having bipolar disorder is independent of the partition number that maximum activation occurs. In other words, we can not state whether a person is bipolar by looking at his/her partition number.

Table 5.1: Results of the chi-square test of independence for CNN models

	χ^2-value	p-value
The CNN model with 24-channel data	8.92	.112
The CNN model with 48-channel data	7.55	.183

After getting no effective results from the chi-square test of independence with CNN models, the alternative way is checking the same null hypothesis for the ResNet models with the same input combinations. Table 5.2 shows that reliable results exist for the ResNet with 24 and 48 channel data (bold, in the table). The former's p-value is .035, while that of the latter is .039. They are both smaller than the alpha value of 0.05. Stated in other words, the likelihood of the occurrence of the relation between the partition number and bipolar disorder by chance is below %5 in these models. Hence, we are able to reject the null hypothesis and state that the most activated partition number is dependent on whether a subject is bipolar or not.

Table 5.2: Results of the chi-square test of independence for the ResNet models

	χ^2-value	p-value
The ResNet model with 24-channel data	**12.01**	**.035**
The ResNet model with 48-channel data	**11.74**	**.039**

By considering these results, we investigate deeply which part of the verbal fluency test activates most the visualization output, i.e. which part constitutes the basis of networks' decisions on the bipolar disorder classification. For this purpose, we independently conduct the chi-square test on each partition such that we divide the timeline as 1 partition and the remaining. In other words, the effect of each partition is checked while assuming the remaining partitions as a single partition. Accordingly, Table 5.3 is constructed and it is shown that only *Pre-task 1* has a meaningful p-value (bold, in the table). On the other hand, there is no bipolar subject whose maximum activation resides in *Post-task 1*, *Pre-task 2*, and *Task 2*; therefore, it is not applicable to conduct the test for these partitions. In conclusion, it is most likely that the partition number with the maximum activation of bipolar patients is *Pre-task 1*.

Table 5.3: Singular partition results of the chi-square test of independence for the ResNet model with 24-channel data

	χ^2-value	p-value
Pre-task 1 vs remaining parts	**7.64**	**.006**
Task 1 vs remaining parts	0.06	.802
Post-task 1 vs remaining parts	N/A	N/A
Pre-task 2 vs remaining parts	N/A	N/A
Task 2 vs remaining parts	N/A	N/A
Post-task 2 vs remaining parts	1.19	.276

Table 5.4 gives the chi-square test of independence result for the ResNet model with 48-channel input data. The p-value of *Pretask 1* is significant, which supports the aforementioned deduction such that bipolar subjects have their maximum activation very likely in *Pretask 1*. Besides, in this model, none of all subjects has the maximum activation at *Posttask 2*, instead of *Pretask 2* from the previous model.

Table 5.4: Singular partition results of the chi-square test of independence for the ResNet model with 48-channel data

	χ^2-value	p-value
Pretask 1 vs remaining parts	**7.25**	**.007**
Task 1 vs remaining parts	0.43	.51
Posttask 1 vs remaining parts	N/A	N/A
Pretask 2 vs remaining parts	0.36	.549
Task 2 vs remaining parts	N/A	N/A
Posttask 2 vs remaining parts	N/A	N/A

5.3.2 T-Test

In the previous section, the chi-square test is carried out so that the partition number's dependency with maximum activation on the bipolar disorder can be checked. However, in order to dive into deeper characteristics of healthy and bipolar subjects, we need to understand the distribution of these two populations. At this point, the t-test comes out as a solution such that whether there is a particular pattern in average activations of bipolar subjects distinguished from that of healthy subjects and their means, as well as standard deviations, are investigated for that purpose. For the use of the t-test, we need to average activations of each partition in itself, as we did in the chi-square test. Then, since we have average activations of each partition for all bipolar and healthy subjects, we are able to analyze the distribution of both populations for each partition independently. However, before applying the t-test, it is required to check if the dataset meets three main assumptions of the parametric statistical test. The first two are the independence of observations and the homogeneity of population variance, and they are both valid for our dataset. The normality of the dataset, which is the last assumption, is checked with an additional process. For this purpose, after collecting average activations of both populations, we perform the Kolmogorov-Smirnov test for goodness of fit with the null hypothesis that populations follow normal distribution patterns.

P-values of the test for the ResNet models are much greater than .05, which implies that we cannot reject the null hypothesis. Hence, there is strong evidence of the average activations of healthy and bipolar subjects being normally distributed. To illustrate, the distribution of average activations belong to populations of healthy and bipolar subjects in the *Task 2* is given in Figure 5.9. These results are obtained by the ResNet model with 24-channel data; however, results from the ResNet model trained with 48 channel input data have similar distributions of average activations of healthy and bipolar subjects. They both fulfill the condition of normal distribution. On the other hand, the Kolmogorov-Smirnov test results of CNN models imply that average activations' distributions belong to healthy and bipolar subjects are not normally distributed.

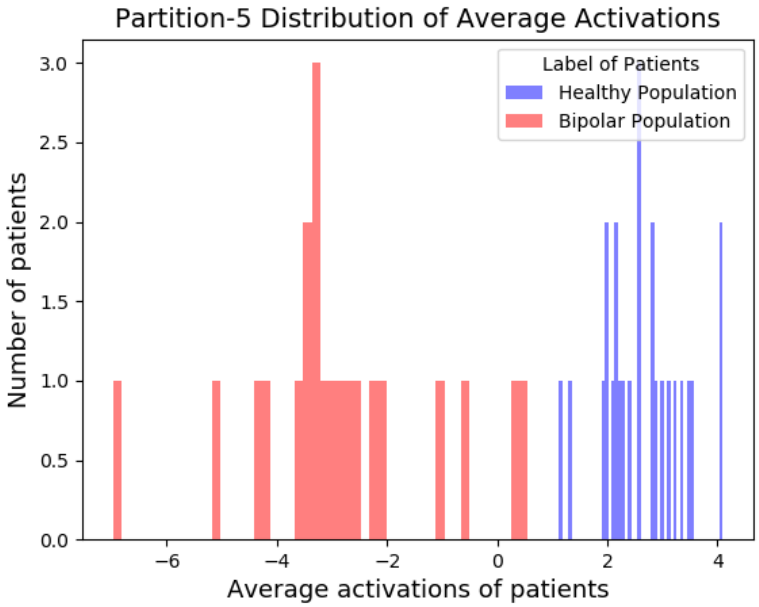

Figure 5.9: The histogram plot of average activations of *Task 2* belonging to the ResNet model with 24-channel data

Average activations of bipolar subjects shown in Figure 5.9 with red lines are located close to each other; whereas, those of healthy subjects reside in their neighborhood, seen with blue lines. It is possible to make an inference that both populations' variances differ from each other. To support this inference, means and standard deviations of healthy and bipolar populations for each partition are given in Table 5.5 and Table

5.6, respectively. In the former, the means of partitions stay between 2.63 and 2.94, while those of the latter are between -3.16 and -1.64. On the other hand, standard deviations are less than approximately 0.75 for the healthy population; whereas, they range from 2.665 to 3.704 for the bipolar population.

Table 5.5: Means and standard deviations of average activations of 25 healthy subjects belonging to the ResNet model with 24-channel data

	Mean	Standard Deviation
Pretask 1	2.94	0.735
Task 1	2.85	0.767
Posttask 1	2.86	0.596
Pretask 2	2.88	0.596
Task 2	2.63	0.566
Posttask 2	2.65	0.698

Table 5.6: Means and standard deviations of average activations of 21 bipolar subjects belonging to the ResNet model with 24-channel data

	Mean	Standard Deviation
Pretask 1	-1.64	2.665
Task 1	-2.12	2.974
Posttask 1	-2.50	3.429
Pretask 2	-2.50	3.288
Task 2	-2.87	2.948
Posttask 2	-3.16	3.704

After the normality of the dataset is proven, since we have two populations with different sizes and variances, the Welch's t-test is applied for each of the partitions independently. The null hypothesis for the test is that two populations come from the same origin; however, they can be distinguishable from each other in proportion

to the difference between the calculated t-value and the value of the test statistic. T-values and p-values are provided in Table 5.7. When the degrees of freedom of the population is calculated as 30, and the confidence interval is 95%, the value of the test statistic is found as 1.697. By comparing this value with t-values from Table 5.7, we find a significant difference between healthy and bipolar populations regarding their visualization outputs since calculated t-values are much greater than 1.697. Table 5.8 confirms this result with better performance while using the ResNet model with 48-channel data.

Table 5.7: T-values and p-values of all partitions for the ResNet model with 24-channel data

	t-value	p-value
Pretask 1	11.59	$2.08e^{-12}$
Task 1	12.97	$1.23e^{-12}$
Posttask 1	12.38	$2.34e^{-12}$
Pretask 2	12.65	$1.24e^{-12}$
Task 2	13.61	$1.92e^{-13}$
Posttask 2	12.86	$7.57e^{-13}$

Other than comparing healthy and bipolar subjects partition by partition, we investigate whether the difference in average activations between two populations by taking into consideration the entire test duration, in other words putting all 6 partitions as a whole. For that purpose, we average the activations of the entire test for all subjects. Then we apply the Kolmogorov-Smirnov test on both populations to check if their distributions are normal. Nevertheless, test results belong to CNN models showing that average activations' distributions belong to healthy and bipolar subjects are not normally distributed. However, distributions of average activations from the ResNet models are normally distributed, enabling us to study on them and evaluate the results.

Table 5.8: T-values and p-values of all partitions for the ResNet model with 48-channel data

	t-value	p-value
Pretask 1	16.73	$7.00e^{-18}$
Task 1	18.82	$7.14e^{-18}$
Posttask 1	17.97	$1.03e^{-16}$
Pretask 2	17.97	$4.21e^{-17}$
Task 2	19.20	$5.93e^{-18}$
Posttask 2	17.51	$5.47e^{-17}$

Table 5.9 demonstrates that means of healthy and bipolar subjects for the ResNet model with 24-channel data are 2.78 and -2.47, respectively; whereas variances of those are 0.61 and 2.93. For the ResNet model with 48-channel data, these values are 2.79 and -6.05 for mean and 0.98 and 3.63 for the variance. These results indicate that differences between means and variances of both populations are much greater in the ResNet model with 48-channel data. After obtaining average activations of healthy and bipolar subjects, we apply the Welch's t-test for both the ResNet models. According to Table 5.9 T-values and p-values of the ResNet model with 24-channel input data are 12.97 and $.43e^{-12}$, respectively; whereas those of the ResNet model with 48-channel data are 19.20 and $.57e^{-17}$. When the degrees of freedom is calculated as 27, and the confidence interval is decided as 95%, the corresponding value of the test statistic is 1.703. Comparing this value with t-values from Table 5.9, we find a significant difference between healthy and bipolar populations with regards to their visualization outputs, since the t-values from Table 5.9 are much greater than 1.703.

Table 5.9: Means and standard deviations of average activations of 21 bipolar and 25 healthy subjects when the performance of the whole test duration is investigated

	Healthy Population		Bipolar Population			
	Mean	Standard deviations	Mean	Standard deviations	t-value	p-value
The ResNet with 24-channel input	2.78	0.61	-2.47	2.93	12.97	$4.3e^{-13}$
The ResNet with 48-channel input	2.79	0.98	-6.05	3.63	19.20	$5.7e^{-18}$

5.4 Discussions

It is important to investigate the specificity of certain neurophysiological abnormalities to bipolar disorder in order to determine both the similarities and differences between bipolar and healthy people. Even within the same task, i.e. VF-task, the reported changes in the frontal lobe function between bipolar and healthy subjects are so far inconsistent since outcomes of research differ from each other. Therefore, it is open for improvement, and new inferences are needed to be revealed. In [40], Onitsuka et al. review fNIRS studies of bipolar disorder and state that increases of oxy-Hb and total-Hb in bipolar disorder patients are significantly smaller than those in healthy subjects during the VF-task. In addition, their findings suggest that during the VF-task, remitted subjects with bipolar disorder exhibit bilateral hypofrontality, which means a state of decreased cerebral blood flow in the prefrontal cortex of the brain. In [41], Kameyama et al. monitor changes in oxy-Hb concentration during the cognitive task using frontal and temporal probes of two sets of 24-channel fNIRS devices. They find that oxy-Hb increases in the bipolar disorder group are smaller than those in the healthy control group during the early period of a VF-task, larger than those in the healthy control group during this task's late period. Since the activations of healthy subjects are higher than those of bipolar ones, it is expected to produce distinguishable patterns in their visualization outputs as well.

Although previous researchers generally focus only on the VF-task duration, in our study, our neural networks are trained by using the whole time period of the verbal fluency test in order to investigate patterns that may lie behind the resting time partitions, i.e. pre-task and post-task partitions, as well as the actual task partitions. The results given in Section 5.3 are obtained for each time partitions comparing healthy and bipolar subjects. Thanks to Table 5.2, we are able to state that the most activated partition of the verbal fluency test is dependent on whether a subject is bipolar or not. Besides, Tables 5.3 and 5.4 reveal that it is most likely that the partition with the maximum activation of bipolar patients is the *Pretask 1* of the test, i.e. the resting state. This finding is crucial since it tells us that the differentiation between bipolar and healthy subjects can occur even outside the test's task partitions.

Furthermore, from Tables 5.7 and 5.8 given in the previous section, it can be deduced that the significance of the difference between populations increases comparatively in *Task 1* and *Task 2* parts of the verbal fluency test. Hence, it can be said that healthy and bipolar subjects' performances differentiate through the experiment but especially during actual task parts. It is unexpected since these results state that even in the resting time, healthy and bipolar subjects produce brain signals that comprise different patterns specific to their labels. Around *Task-2*, t-values are greater than those of *Task-1*, which implies that differences between distributions of healthy and bipolar subjects' average activations are slightly more apparent in *Task-2* compared to *Task-1*, which coincides with the aforementioned findings of [40] and [41]. On the other side, the results of the case where the entire VF-test duration investigated as a whole are given in Table 5.9. These outcomes prove the aforementioned statement that distributions of healthy and bipolar subjects' average activations are significantly different with distinctive characteristics.

Standard deviations of partitions for healthy subjects given in Table 5.5, are small since the model is able to grab a common pattern for healthy subjects; as a consequence, visualization outputs of healthy subjects have similar distributions. It is expected as a result of which brain activities of healthy subjects act in similar patterns. Bipolar subjects, on the other hand, spread on a large area in Figure 5.9, with greater standard deviations for all partitions. As it is known that bipolar subjects are in the same phase of the disease while conducting the verbal fluency test [5], these differ-

ences of average activations between patients are not due to the variation of clinical stage. Instead, they are most likely to be originated from differences in the severity of disease for each individual.

The difference in the pattern in bipolar patients from healthy controls may be a sign of neurodegeneration that occurs during the illness. Moreover, differences in activations along the time course cause inconsistent results in cognitive activation regarding bipolar disorder. Tests mentioned in the previous section are intense supporters of these claims. This study may open a new window into psychiatrists so that they can work through the aforementioned outcomes.

On the other hand, our test results show that the ResNet models outperform those of CNN, which are in agreement with the deep learning literature. The study of Fawaz et al. reveals the dominance of ResNet as the best performing deep network across different domains, including the time domain [27]. The ResNet has higher success in training the network due to its deep flexible architecture. The ResNet is able to filter out discriminative regions with higher confidence, and its output has a higher resolution, which enables us to visualize by using more samples. As mentioned in 4.1.2, this ability comes from the shortcut connection between convolutional blocks providing the network to learn to skip unnecessary convolutions since the gradients can flow directly through the bottom layers in the network. Therefore, the statistical tests conducted during this study give more stable results for the ResNet models.

CHAPTER 6

CONCLUSIONS

6.1 Conclusion and Future Works

In this study, we analyze the decision mechanism of neural networks trained for the classification of the diagnosis of bipolar disorder by using the time-series neuroimaging data, i.e. fNIRS measurements, taken during the verbal fluency test. The main aim is to provide assistance to psychiatrists for understanding the characteristic features of fNIRS data belong to bipolar patients.

In order to accomplish this goal, we benefit from CAM to visualize the Convolutional Neural Network and Residual Neural Network. Since our dataset is time-series, the visualization process differs from the traditional methods which use the image data. Therefore, we modify the class activation maps so that the heatmap of time-series data can be generated. After the modification, we are able to observe the importance of every time interval of the verbal fluency test for any subject by visualizing the heatmaps of activation in the time scale.

Directly observing these heatmaps and making an inference according to them is unfavorable since the data comes from different subjects. Also, their interactions with the environment, which might trigger their activation maps for a brief time, are unpredictable. For this reason, in order for psychiatrists to obtain more trustworthy and understandable results, we divide the heatmaps into partitions in accordance with those of the verbal fluency test. Then, we calculate the average activations for each partition. Using these average activations, we analyzed the differentiation of healthy and bipolar groups and the comparison of CNN and ResNet models with statistical tools. We apply the chi-square test of independence to reveal whether bipolar disorder

is related to a specific partition of the test that shows the maximum activation in the neural network. When we use the ResNet models, our findings reveal that it is possible to state that the most activated partition of the verbal fluency test is dependent on whether a subject is bipolar or not. Moreover, the maximum activations belonging to bipolar subjects occur most likely in the test's first resting time. Hence, given that the maximum activation of a subject is in the first resting time, most likely she/he has the bipolar disease.

On the other side, we perform the independent t-test so that the distributions of the average activations of healthy and bipolar populations can be examined for each partition of the test. According to our results, when both CNN and ResNet models are used, the significance of the difference between populations increases comparatively in the verbal fluency test's first and second tasks. Hence, it can be said that healthy and bipolar subjects' performances differ through the experiment but especially during actual task parts. In other words, healthy and bipolar subjects produce brain signals which comprise different patterns specific to their labels. Furthermore, thanks to our findings, the differences between distributions of healthy and bipolar subjects' average activations are slightly more apparent in the second task compared to the first one.

Moreover, we find that standard deviations of average activations of the healthy population for all partitions of the test are small compared to those of the bipolar population since neural networks are able to grab a common pattern for healthy subjects. It is expected as a result of which brain activities of healthy subjects act in similar patterns. In other respects, the bipolar population has high standard deviations due to the variability of the disease severity for each individual, such as the neurodegeneration that occurs during the illness.

After getting more subjects' fNIRS data, more trustworthy and stable networks with higher accuracy can be obtained. Then, for future work, the experiment and the statistical tests that have been formerly failed due to the lack of a sufficient number of subjects can be repeated, and the results can be evaluated by comparing them with the current findings. Besides, the fNIRS data consists of time and channel information, and in this study, we focus on showing the importance of the test's time intervals.

By this means, we provide psychiatrists with information that which part of the test is more suitable to distinguish the bipolar and healthy subjects. However, the comparison of the channels can be studied as another future work. The most significant channels that correspond to specific brain regions can be decided and used by psychiatrists to distinguish bipolar subjects.

REFERENCES

[1] A. Punjabi and A. K. Katsaggelos, "Visualization of feature evolution during convolutional neural network training," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 311–315, 2017.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, p. 336–359, Oct 2019.

[3] E. Johns, O. M. Aodha, and G. J. Brostow, "Becoming the expert - interactive multi-class machine teaching," 2015.

[4] H. B. Evgin, "Deep learning for the classification of bipolar disorder by using fnirs measurements," Master's thesis, The Graduate School of Natural and Applied Sciences of Middle East Technical University, 2 2021.

[5] Y. H. ALICI, *Bipolar Bozukluk Hastalarında Sozel Akıcılık Ve Dusunce Akıcılığı Performansları Sırasında Prefrontal Korteks Aktivitesinin Psikososyal Uyum, Islevsellik, Yasam Kalitesi Ve Iç Goru Ile Iliskisi*. PhD thesis, Ankara Universitesi Tip Fakültesi, 7 2015.

[6] H. B. Evgin, O. Babacan, I. Ulusoy, Y. Hosgoren, A. Kusman, D. Sayar, B. Baskak, and H. D. Ozguven, "Classification of fnirs data using deep learning for bipolar disorder detection," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, 2019.

[7] Z. Qin, F. Yu, C. Liu, and X. Chen, "How convolutional neural network see the world - a survey of convolutional neural network visualization methods," 2018.

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013.

[9] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[10] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Technical Report, Univeristé de Montréal*, 01 2009.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014.

[13] I. Goodfellow, Q. Le, A. Saxe, H. Lee, and A. Ng, "Measuring invariances in deep networks.," pp. 646–654, 01 2009.

[14] F. Chollet, *Deep Learning with Python*. Manning, Nov. 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2015.

[17] C. Lehmann, T. Koenig, V. Jelic, L. Prichep, R. John, L.-O. Wahlund, Y. Dodge, and T. Dierks, "Application and comparison of classification of alzheimer's disease in electrical brain activity (eeg)," *Journal of neuroscience methods*, vol. 161, pp. 342–50, 05 2007.

[18] B. Cao, R. Y. Cho, D. Chen, M. Xiu, L. Wang, J. C. Soares, and X. Y. Zhang, "Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity," *Molecular Psychiatry*, vol. 25, pp. 906–913, June 2018.

[19] P. Morillo, H. Ortega, D. Chauca, J. Proaño, D. Vallejo, and M. Cazares, *Psycho Web: A Machine Learning Platform for the Diagnosis and Classification of Mental Disorders*, pp. 399–410. 01 2020.

[20] A. Riaz, M. Asad, E. Alonso, and G. Slabaugh, "Deepfmri: End-to-end deep learning for functional connectivity and classification of adhd using fmri," *Journal of Neuroscience Methods*, vol. 335, p. 108506, 01 2020.

[21] S. Sarraf, D. DeSouza, J. Anderson, and G. Tofighi, "Deepad: Alzheimer's disease classification via deep convolutional neural networks using mri and fmri," *bioRxiv*, 08 2016.

[22] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, "Discriminative analysis of the human cortex using spherical cnns - a study on alzheimer's disease diagnosis," 2018.

[23] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, R. Savitha, B. L. DeCost, S. I. P. Tian, G. Romano, A. G. Kusne, and T. Buonassisi, "Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks," 2019.

[24] J. Shi, D. Chen, and M. Wang, "Pre-impact fall detection with cnn-based class activation mapping method," *Sensors*, vol. 20, p. 4750, 08 2020.

[25] A. Ghosh, F. dal Maso, M. Roig, G. D. Mitsis, and M.-H. Boudrias, "Deep semantic architecture with discriminative feature visualization for neuroimage analysis," 2018.

[26] K. Gao, H. Shen, Y. Liu, and D. Hu, "Dense-cam: Visualize the gender of brains with mri images," pp. 1–7, 07 2019.

[27] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Mining and Knowledge Discovery*, vol. 33, p. 917–963, Mar 2019.

[28] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 1578–1585, 2017.

[29] H. Ayaz, M. Izzetoglu, S. Bunce, T. Heiman-Patterson, and B. Onaral, "Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy," in *2007 3rd International IEEE/EMBS Conference on Neural Engineering*, pp. 342–345, 2007.

[30] D. Boas, A. Dale, and M. A. Franceschini, "Diffuse optical imaging of brain activation: Approaches to optimizing image sensitivity, resolution, and accuracy," *NeuroImage*, vol. 23 Suppl 1, pp. S275–88, 02 2004.

[31] N. Naseer and K.-S. Hong, "Fnirs-based brain-computer interfaces: A review," *Frontiers in Human Neuroscience*, 01 2015.

[32] X. Cui and S. Bray, "Functional near infrared spectroscopy (nirs) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics," *NeuroImage*, vol. 49, pp. 3039–46, 11 2009.

[33] D. Heeger and D. Ress, "What does fmri tell us about neuronal activity?," *Nature reviews. Neuroscience*, vol. 3, pp. 142–51, 03 2002.

[34] Y. Hoshi, "Functional near-infrared optical imaging: Utility and limitations in human brain mapping," *Psychophysiology*, vol. 40, pp. 511–20, 07 2003.

[35] M. P. G. Allin, N. Marshall, K. Schulze, M. Walshe, M.-H. Hall, M. Picchioni, R. M. Murray, and C. McDonald, "A functional mri study of verbal fluency in adults with bipolar disorder and their unaffected relatives," *Psychological Medicine*, vol. 40, no. 12, p. 2025–2035, 2010.

[36] A. Tumaç, "Normal deneklerde frontal hasarlara duyarli bazi testlerde performansa yas ve egitimin etkisi," Master's thesis, Istanbul Universitesi Sosyal Bilimler Enstitusu Psikoloji Bolumu, 7 1997.

[37] D. Yu and L. Deng, "Deep learning and its applications to signal and information processing [exploratory dsp]," *IEEE Signal Processing Magazine*, vol. 28, no. 1, pp. 145–154, 2011.

[38] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, p. 85–117, Jan 2015.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[40] T. Onitsuka, N. Oribe, and S. Kanba, *Neurophysiological findings in patients with bipolar disorder*, pp. 197–206. Supplements to Clinical Neurophysiology, Elsevier B.V., 2013.

[41] M. Kameyama, M. Fukuda, Y. Yamagishi, T. Sato, T. Uehara, M. Ito, T. Suto, and M. Mikuni, “Frontal lobe function in bipolar disorder: A multichannel near-infrared spectroscopy study,” *NeuroImage*, vol. 29, pp. 172–84, 02 2006.