Handling Overfitting in Deep Learning Models and Their Applications

Isaac F. Bensaid BSc

February 23, 2022

Contents

1 Introduction

Given the uncertainty of the future, decision-making is a difficult task. If we can accurately predict the future, decision-making becomes easier. While making predictions might seem easy, making accurate ones is hard. When Data Scientists attempt to predict the future, many obstacles can arise. One obstacle is Overfitting, a common problem in Deep Learning (DL). Data Scientists often rely on historical data to predict the future. For instance, assume we would like to predict the outcome of the next presidential election. There is a tremendous amount of historical data about presidential elections. The question becomes, how do we select Hyperparameters – the data that could produce the most accurate prediction results – from a large pool of historical elections data. When Data Scientists do not choose their data wisely, Overfitting could undermine their predictions' accuracy. In some cases, a high number of Hyperparameters could increase the complexity of the DL Model and increase the likelihood of Overfitting [48]. Historical data helps Data Scientists train DL Models. Data Scientists can feed DL Models raw data — data that is not preprocessed (e.g., Magnetic Resonance Imaging) — and expect them to make accurate predictions without being explicitly programmed. Data Scientists train DL Models using a set of chosen training data: a collection of situations for which the desired output is known. For example, a collection of election data in which Data Scientists know the election outcome beforehand. The goal is that the DL Model will correctly predict the output when fed validation data – data that the DL Model did not encounter during its training and its output was unknown. The ability to perform correct predictions on validation datasets is called Generalization [25]. Overfitting occurs when a DL model performs correct predictions on training data and performs incorrect predictions on validation data [48]. We can explain Overfitting intuitively because all data – related to the mentioned election outcome prediction example – can be divided into two groups. The first group consists of data helpful for accurate election outcome prediction (also known as Hyperparameters). The second group consists of useless data for accurate election outcome prediction (also known as noise [48]). DL Models tend to overfit by memorizing properties of noise that do not serve them well during the prediction phases. The higher the uncertainty and complexity for any prediction task, the more noise in its historical data. We may reduce Overfitting if we successfully determine which historical data to ignore [25]. One of the most prominent qualities in DL is how well a model performs on unseen data. Making accurate predictions on unseen data is a fundamental element of any DL Model [9]. Many Data Scientists have been attempting to reduce Overfitting to improve prediction accuracy [47] [13]. However, up until today, there is no standard solution for Overfitting. No single technique will work perfectly for every problem. Each problem required a unique solution which may include more than one technique [54]. For example, Biologists have relied on DL to prevent or cure diseases [66]. In gene therapy — the practice of inserting corrective genetic material into defective cells — Biologists have used DL Models to develop or improve non-pathogenic viruses. Physicians use Non-pathogenic viruses to deliver corrective genetic materials to defective cells [10]. These viruses are called vectors. Biologists believe they do not

cause any human diseases because they have been disabled of any pathogenic effects [46]. During the delivery process, many hurdles can arise. For example, the immune system can destroy the delivery virus vector, preventing the delivery of the corrective genetic material. Overfitting could arise when Data Scientists predict which potential vectors are capable of evading the immune system — to deliver the corrective genes to the targeted cells. However, this thesis will focus on Overfitting in the medical diagnosis of neurodegenerative diseases. Before we shift our focus to the medical diagnosis of neurodegenerative diseases, we will elaborate on Overfitting from a broader scope. Humans can overgeneralize sometimes. For example, assume we are conducting business with an American businessman for the first time. If we find the American businessman unethical, we might inevitably claim that all American businessmen are unethical. In DL, a similar issue called Overfitting can occur if the Model does not generalize well. Overfitting could occur because the Model learns the noise in the training set, impacting its performance on unseen data. The noise picked up in the training set may not apply to unseen data, leading to inaccurate predictions. Linear Modules tend to be less likely to overfit. The more complex the DL Model, the more it will be prone to Overfitting. Convolutional Neural Networks (CNN) Models — a DL model — contain multiple non-linear hidden layers, making them complex models to learn complicated relationships. Unfortunately, Overfitting is a common problem that prevents CNNs from making accurate predictions. Overfitting occurs when CNNs fails to make accurate predictions on unseen datasets. Overfitting is still considered a standing problem in DL [47]. Regularization was imposed by [47] on the fully connected Layer(FCL) — more on FCL later — of the CNN to reduce Overfitting. In a typical CNN, weights of FCLs make up most of the parameters of the network. So, first, [47] sparsified connections to the FCLs by applying L1 Regularization on the weights of the FCLs, then reduced small weights to zero. Next, [47] also applied L2 Regularization on all weights of the CNN to keep the weights around zero. It was found in [47] that the numerous parameters of FCLs are a contributor to Overfitting. The number of parameters in CNN is vast. The high number of features makes the model complex, which tends to result in **Overfitting**. In biotechnology, therapeutics should meet high safety standards during their development. Meeting safety standards requires an accurate prediction of any adverse side effects. For example, in gene therapy, predicting the safety level of Viral Vectors could be a challenge due to Overfitting. Several individuals developed cancer during gene therapy clinical trials using viral vectors to deliver corrective genes. In a study done by [72], it was demonstrated that Genotoxic Vectors have a unique gene expression signature. Based on this finding, [72] developed a "surrogate assay for genotoxicity assessment" (SAGA). SAGA uses this unique gene expression signature to distinguish genotoxic vectors from safe vectors, using DL to recognize the said gene expression signature. SAGA achieved an accuracy of 90.91% on a dataset of vectors with known genotoxic potential. This level of accuracy is arguably high, but given that gene therapy could lead to cancer, methods to identify safe gene therapy vectors require a higher degree of accuracy.

1.1 Problem Statement

Suppose we would like to create a Model for neurodegenerative disease medical diagnosis. To create the Model, we need to train it with a medical records dataset. The complexity and uncertainty of this Model are likely to be high, increasing the likelihood of Overfitting. If we do not address Overfitting, it can arise and discredits the accuracy of our Model – which could result in unwanted prediction results. For example, Parkinson's disease (PD) is a common progressive neurodegenerative disease — characterized by motor and non-motor symptoms. While some diseases can be diagnosed with a lab test — such as cholesterol level and blood pressure measurements — Parkinson's disease cannot. Physicians rely on medical history and physical examination to diagnose Parkinson's disease. Physicians look for classic motor symptoms: resting tremor, stiffness, and slowness of movement. Additionally, vocal defections are a common symptom in the early stages of Parkinson's disease [14]. A CNN was trained by [39] to recognize patients with such vocal defections using sets of vocal features. However, Overfitting prevented accurate classification between healthy individuals and PD patients. Luckily, [39] used two methods to reduce Overfitting — L2 regulations and Dropout (more on those methods later). It was stated in [32] and [51] that the pen pressure and speed of PD patients are reduced compared to healthy patients. It was also mentioned in [51] that a difference in handwriting and sketching abilities is noticeable among PD patients. It was also stated by [51] that the impairment in handwriting is directly proportional to the severity of the disease. In [68] data augmentation techniques such as contrast illumination, thresholding, flipping, and rotation were applied to reduce Overfitting. As a result, the Model achieved 99.22% accuracy in [68]. In [59] two CNN models were used for classification of PD patients. Transfer learning —using a pre-trained network— and data augmentation methods such as jittering, scaling, Time-Warping were used by [59] to mitigate Overfitting. The model in [59] achieved 97.62% accuracy. However, [59] noted that there was not a significant difference in accuracy between using transfer learning and training a model from scratch. In [67] transfer learning, removing features, and data augmentation were used to mitigate Overfitting. It was argued in [67] that the reliability of the data set is compromised because sketches of waves and spirals were drawn in different pencils. By thinning the thickness of the drawings, [67] unified the drawings and removed noise from the data. The model in [67] achieved a 93.33% precision. It was stated in [67] that the main limitation in the classification of PD using CNN is the lack of data. In [51] data augmentation was used to perform horizontal and vertical shifts. The early stop method was used by [51] in such a way that if the validation loss does not decrease for consecutive 18 epochs by reducing the learning rate twice by the fraction of 0.8, then the Model needs to be stopped. The model in [51] achieved an accuracy of 93.3%. The use of a fine-tuned VGG-19 for screening Parkinson's Disease based on a handwriting dataset was investigated and experimented by [58]. The last fully-connected layer of the original pre-trained VGG-19 with a two-neurons layer was replaced by [58]. To minimize Overfitting, [58] adopted data augmentation based on image rotation. Fifty percent of the outputs of the first two fully-connected layers were dropped by [58] to reduce Overfitting

further. The Model in [58] achieved an accuracy of 88%. It was concluded in [50] that image input layers, activation function, loss function, and weights initialization are the main factors that Data Scientists need to consider when mitigating Overfitting. As such, we will do an extensive review of those factors. On the other hand, Alzheimer's disease (AD) is a disorder that slowly destroys memory and thinking skills, making it difficult for patients to perform daily functions. Early detection of AD is essential for providing effective medical treatment. Deep learning-based Convolutional Neural Networks (CNN) are used to improve the detection of Alzheimer's disease. The earlier the detection of Alzheimer the more effective the treatment [73]. Two prominent datasets in this field are Alzheimer's Disease Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies OASIS. The CNN Model used in [52] used 12 layers, which consists of convolutional, max pooling, dense, and flatten layers. In addition, the activation functions Sigmoid, ReLU, and Leaky ReLU were used by [52]. The Model in [52] was trained and tested on the Open Access Series of Imaging Studies (OASIS) dataset. The Model in [52] achieved an accuracy of 97.7%, which is higher than any other existing CNN model published on the dataset at the time of publication. To address overfitting [52] used image resizing (image resizing reduced the time taken to train the Model) and denoising (suppressing noise from the images). However, the use of different activation functions was not explored in [52]. In addition, the multi-class classification of AD patients was not explored in [52], as they merely performed binary classification. In [41] three pre-trained models were used on the OASIS dataset via transfer learning as well as a model trained from scratch by [41]. The accuracy of the four models was compared by [41]. AlexNet, GoogLeNet, ResNet50 —the most prominent pre-trained CNN models [73] — were trained by [41]. A CNN model similar in architecture to AlexNet was designed by [41], instead of using cross-channel normalization, [41] implemented batch normalization and reduced the size of the filters, making the Model a 22-layer CNN. It was shown in [41] that scratch training could be more accurate than pre-trained models by achieving high accuracy of 98%, which was higher than all pre-trained models. However, [41] merely performed binary classification. Multi-class classification of AD patinas — which is mush highly prone to Overfitting— was not explored by [41]. The dependency on a high number of training images and the demand for deep network architecture optimization were emphasized in [28] as frequent limitations for deep learning algorithms. Transfer learning was enchained by [28] to resolve these issues. In [28] VGG and Inception — pre-trained models — were utilized to perform binary classification of PD patients. The fully-connected layer was re-trained by [28] with only a small number of MRI images. Image entropy was employed by [28] to select the most informative slices of MRI images for re-training the fully-connected layers. Via experimentation on the OASIS MRI dataset, [28] show that with a small amount of training data, the Model in [28] was able to achieve a 96.25% accuracy. It was argued by [28] that the informed selection of the training data was the reason behind the high accuracy despite the small amount of data used and minimal parameter optimization. Albeit, multi-class classification of AD patients was not explored by [28]. In [69] transfer learning was used to classify between Mild Cognitive Impairment and healthy patients. The Alexnet CNN was fine-tuned by [69]. The Model in

[69] achieved an accuracy of 98.35%. The early diagnosis of Alzheimer's disease is essential for the effective treatment of AD. In [37] built a CNN model to predict the individual diagnosis of Alzheimer's disease (AD) and mild cognitive impairment who will convert to AD based on MRI scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Transfer learning and data augmentation were applied by [37]. Deformation, cropping, rotation, flipping, and scaling were applied by [37]. The model in [37] achieved a 98% accuracy. To the best of our knowledge, a comprehensive study on several overfitting methods to increase the accuracy of Multi-class classification of AD patients has not been conducted yet.

1.2 Artificial Neural Network

Deep Learning is based on Artificial Neural Networks(ANN). Every ANN contains many perceptrons. Perceptrons are a function that can take any form of data as input and return a boolean value as output [27]. A perceptron has weights, a bias, a weighted sum, input values, and an activation function [27]. To understand how ANN works, we need to understand how perceptions work.

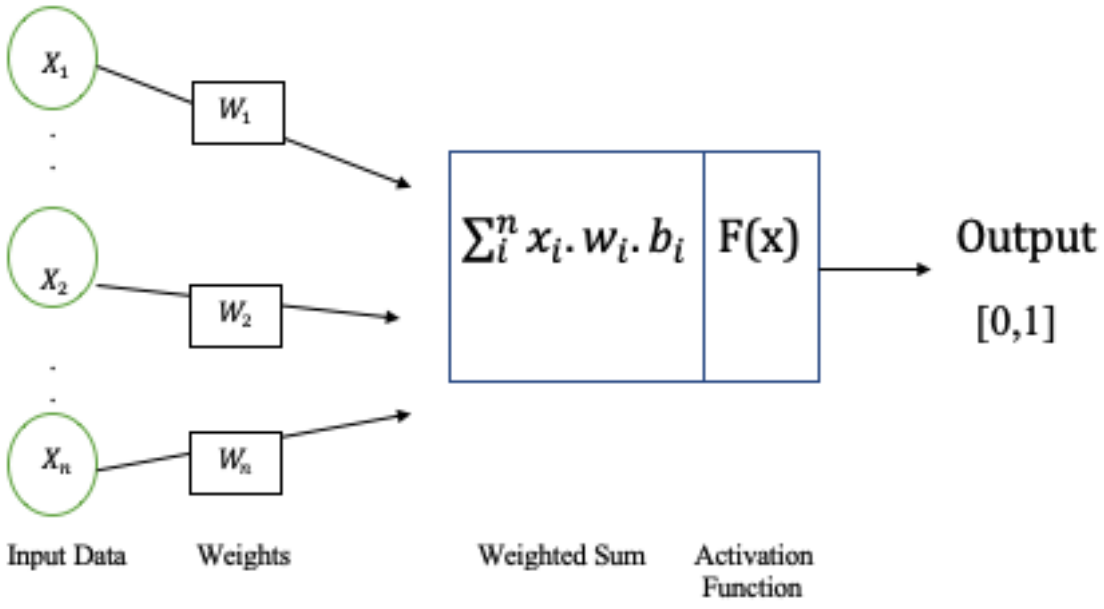


Figure 1: An image of a single perceptron. Adapted from [27]

As shown in Figure 1, all the inputs x are multiplied with their weights w, then we add all of the multiplied values and call the result the Weighted Sum. Finally, the weighted sum gets passed to the activation function.

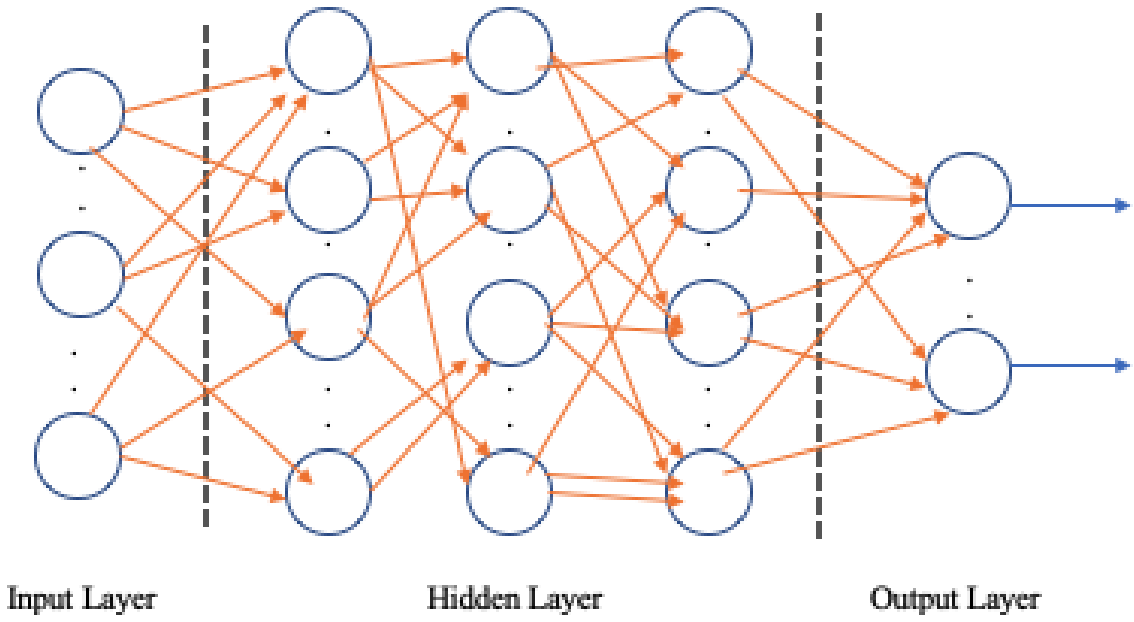


Figure 2 shows the entire Neural Network Architecture. Adapted from [27]

Every node in the Neural Network architecture shown in Figure 2 contains a perceptron. If we were to zoom closer at each node in the Neural Network Architecture, we would see the perceptron shown in Figure 1. As we can see, there are many layers to the Neural Network; the first layer is called the input layer, the last layer is called the output layer, and the layers in between are called the hidden layer. Each perceptron layer can receive input from the previous layer and produce output to the next layer. Therefore, the Neural Network architecture can vary based on the type of Neural Network used and the problem we solve. For instance, CNN will have a slightly different architecture than what we just saw in Figure 2.

1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are based on convolutional layers that extract local information (e.g., apples in a fruit collection image) from an input picture. Each node in a convolutional layer is linked to a subset of neurons that are spatially connected. A max-pooling layer is added after convolutional layers to minimize computational complexity. This layer compresses feature maps by picking the maximum response in a small region. Following pairs of convolutional and pooling layers, there are fully connected layers in which each neuron has connections to all activations in the preceding layer. Fully connected layers aid in the discovery of non-linear relationships between convolutional layer-extracted local information. Following the fully connected layers, a soft-max layer normalizes the outputs to the appropriate values. Data Scientists train CNNs using the back-propagation algorithm. At each iteration, the weights associated with the neurons in the convolutional layers are changed to minimize the loss function. This function calculates the difference between the actual and predicted values. CNNs focus on data that has a gird-like representation. For

example, images that we represent as a 2-D grid of pixels. CNNs employ a linear mathematical operation known as convolution. The integral of the product of two real-valued functions after one has been inverted and shifted is known as convolution [7]. Let f and g be two real-valued functions, the convolution of the two functions is written as $f * g$, denoting the convolution operator with the symbol "$*$" [24][5].

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau)\, d\tau$$

Depending on the convolution operation, the CNN model will transfer the input into some form. Therefore, convolution can be detrimental to input data. For example, if we use a CNN to detect traffic lights, the bright color location is crucial. Changes to the location of the bright color can be an adverse impact caused by convolution. Pooling can help preserve the image's bright color after the convolution layer's transformations. CNN models use pooling to preserve computational resources by extracting the most vital features from the preceding convolutional layer. CNN architectures can vary based on their specific application. Typically, CNNs consist of an input and output layer and several hidden layers in between. There is no one conventional CNN architecture. The type of architecture used in a CNN depends on the problem and its data. For example, it could be the case that a particular benchmark has a specific CNN architecture that is known to be the best. As Data Scientists conduct further research, this architecture will continue to evolve [25]. CNN can be used in the medical field to help doctors diagnose patients. For example, in [63], CNN was used to assess neurological damage in Parkinson's disease patients using 3D brain images. Two CNN architectures differing on the output layer shape, the activation function, and the loss function were used by [63] used. When using the same dataset of 508 3D brain images, it was evident that one of the different architectures outperformed the other by achieving higher classification accuracy of Parkinson's disease patients. The difference in prediction accuracy shows how modification in CNN architecture can improve its performance.

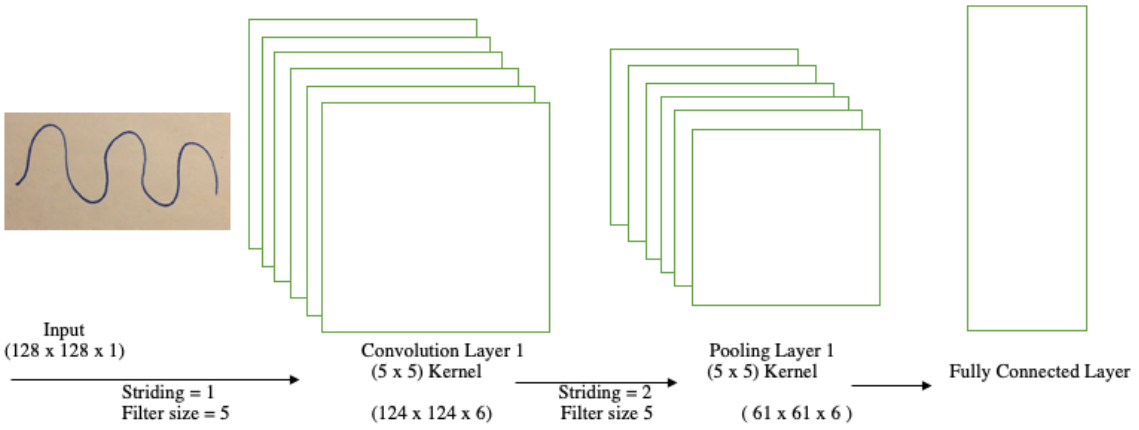


Figure 3 shows a sample of CNN Architectures. We adapted this architecture from our Parkinson's Disease experiment in section 1.4. The author of this thesis drew the wave above.

Figure 3 predicted whether a patient has Parkinson's disease from their hand-written wave. First, the image has dimensions height = 128 pixels, width = 128 pixels, and a depth of 1. Next, the image moves to the convolution layer shown in figure 3. Applying six filters of size 5×5 yields an image of height = 124 pixels, width = 124 pixels, and a depth of 6. Then it gets passed to the pooling layer, applying the same amount of filters and resulting in the same depth with an image of height = 61 pixels width = 61 pixels. We repeat this process until we flatten the image to a single layer. Then we feed it to a network similar to the one shown in Figure 2 to classify it into the appropriate category [24].

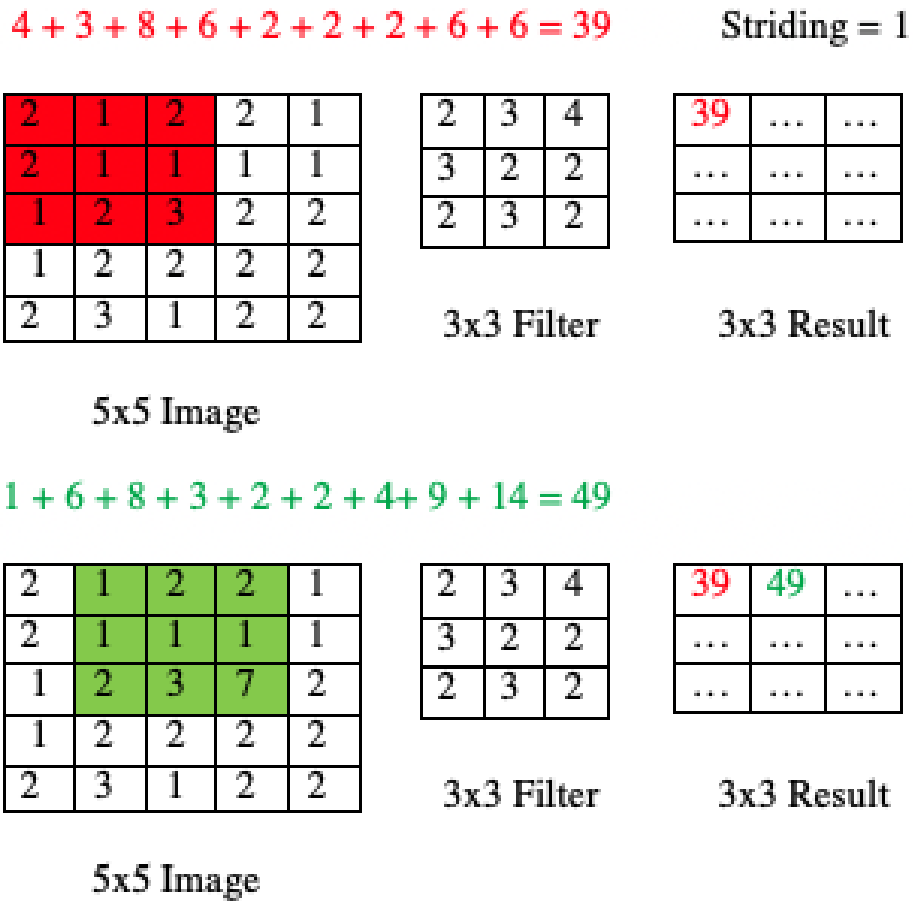


Figure 4 shows a breakdown of the convolution layer.

For simplicity, we will assume that the wave image is in 5×5 format. As shown in Figure 4, applying a 3×3 filter will result in a 3×3 image. Every value in the filter is multiplied by its corresponding value in the original image. With Striding —the amount of movement over the image — equals one, the filter moves one step to the right and repeats the same process [24] [44].

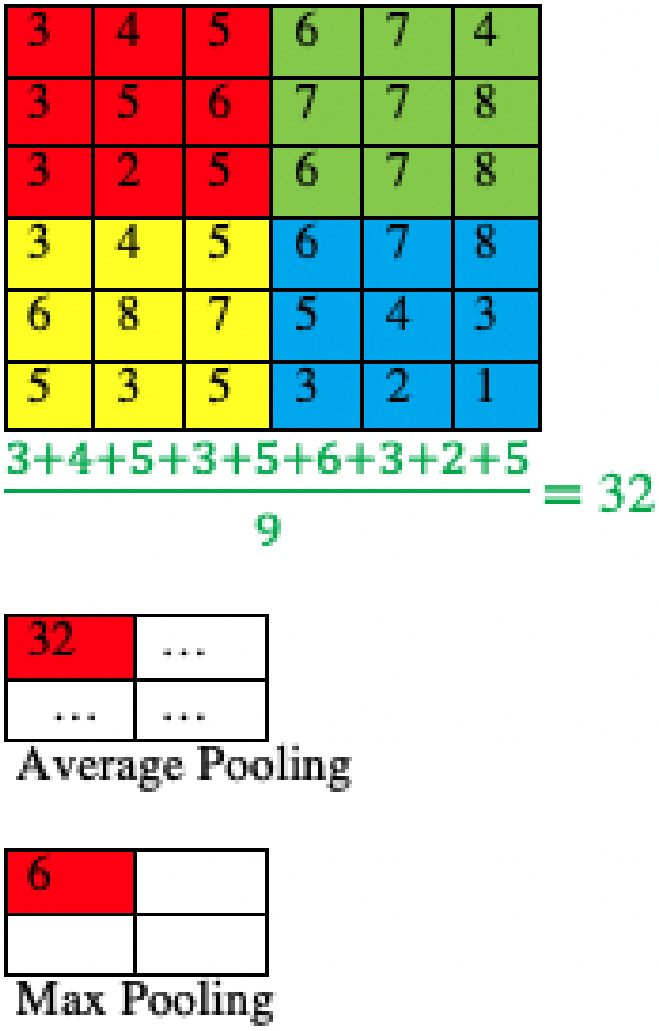


Figure 5 shows a breakdown of the pooling layer.

In Figure 5, we zoom in into the pooling layer. Simply put, we take a small portion of the image, and we apply either max-pooling — taking the maximum number of the portion we select – or average pooling, taking the average of the portion we select and rounding it up [24] [44].

1.4 Practical Example of Overfitting

To demonstrate Overfitting, we used a dataset that contains medical records of Parkinson's disease patients. We programmed this experiment in Keras (Python interface for artificial neural networks [29]). The dataset we used contains two sub-directories. The first sub-directory has a collection of wave drawings of healthy patients, while the second sub-directory has a collection of wave drawings of patients with Parkinson's disease. The dataset is available [64] here.. The data was collected by [32] from participants who provided their written and verbal consent to participate in the formation of this dataset. As we mentioned earlier, Motor symptoms such as

speed of drawing and pen pressure make detecting Parkinson's disease possible [32] [51]. The dataset consists of 102 images. We used 72 images to train a CNN model and 30 images to test it. Due to the small size of the dataset, the CNN model is highly prone to Overfitting. Therefore, data Augmentation — creating more data — is needed in this situation to reduce Overfitting. We built a CNN model with four convolutional layers, followed by another four max-pooling layers to classify the Parkinson's Disease data into the sub-directories mentioned above: (Health patients, patients with Parkinson's disease). We used a flatten and two dense layers after the convolutional and pooling layers. Figure 6 shows the specification of the CNN Model.

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1 (Conv2D)              (None, 128, 128, 128)     3328

 max_pooling2d_12 (MaxPoolin  (None, 40, 40, 128)      0
 g2D)

 conv2 (Conv2D)              (None, 40, 40, 64)        204864

 max_pooling2d_13 (MaxPoolin  (None, 12, 12, 64)       0
 g2D)

 conv3 (Conv2D)              (None, 12, 12, 32)        18464

 max_pooling2d_14 (MaxPoolin  (None, 4, 4, 32)         0
 g2D)

 conv4 (Conv2D)              (None, 4, 4, 32)          9248

 max_pooling2d_15 (MaxPoolin  (None, 1, 1, 32)         0
 g2D)

 flatten_3 (Flatten)         (None, 32)                0

 fc1 (Dense)                 (None, 64)                2112

 fc3 (Dense)                 (None, 2)                 130

=================================================================
Total params: 238,146
Trainable params: 238,146
Non-trainable params: 0
_________________________________________________________________
```

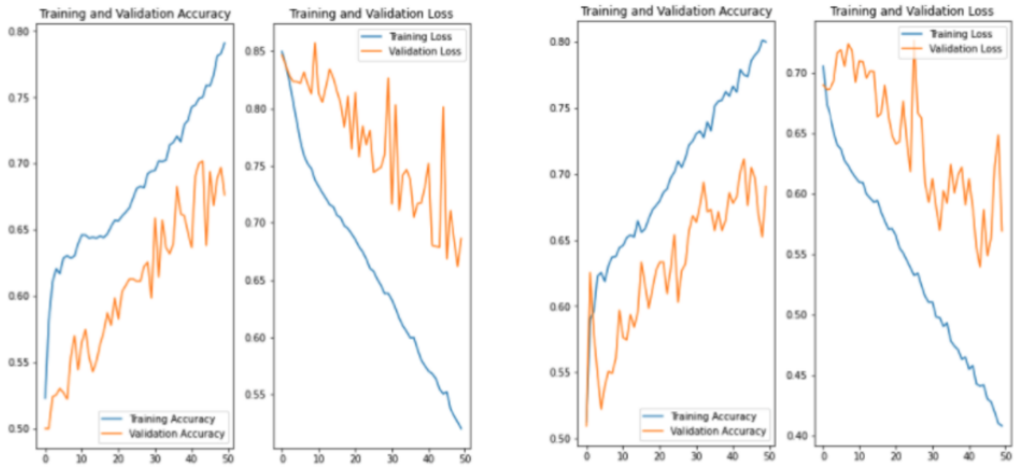
Figure 6 CNN Configuration

As shown in figure 6, images were rescaled to 128 by 128 for better resolution, and the convolutional layers were of size 128, 64, and two of size 32, respectively. To reduce Overfitting and improve the Model's accuracy, we used the following overfitting techniques: Dropout Regularization, Weight Constraints, Weight Initialization, L2 Regularization. Due to the dataset's limited size, we generated an additional 5110 images using data augmentation. We rotated the original image by 360 degrees, and we flipped it virtually and horizontally to produce other images from the original images in the dataset.

(a) Dropout: 57.41 Accuracy (b) Weight Constraints: 60.71 Accuracy

Figure 7

Figure 7 shows the model performance using Weight Constraints and Dropout Regularization. The graph on the left in section (a) shows the training accuracy in blue and the validation accuracy in orange. The graph on the right in section (b) shows the training loss in blue and validation loss in orange. The same is true in section (b).



(a) L2 Regularization: 68.37 Accuracy (b) Weight Initialization: 69.57 Accuracy

Figure 8

Figure 8 shows the model performance using L2 Regularization and Weight Initialization. The graph on the left in section (a) shows the training accuracy in blue and the validation accuracy in orange. The graph on the right in section (b) shows the training loss in blue and validation loss in orange. The same is true in section (b).

Method	Accuracy
Dropout Regularization	57.41 %
L2 regularization	68.37 %
Weight Constraints	60.71 %
Weight Initialization	69.57 %

Table 1

As we can see from table 1, figure 7, and figure 8 above. The module performed better when using Weight Initialization and L2 Regularization than Weight Constraints and Dropout Regularization.

1.5 Discussion

Overfitting takes place when the margin (gap) between the training error and test error is large [25]. The models have overfitted the data, as shown in Figures 2 and 3. Training accuracy and validation accuracy are off by a large margin — the lower the loss, the higher the accuracy. As we can see in some of the graphs above, the training set has a low loss, which yields higher accuracy, but the validation set has a high loss which yields low accuracy. A high difference between training and validation loss is a sign of Overfitting. In other words, the models performed well in the training set but performed poorly in the validation set, resulting in Overfitting. While we tried to reduce Overfitting by using a few methods. Such as Dropout Regularization, Weight Constraints, Weight Initialization, L2 Regularization. Based on Figures 2 and 3, there is still room for improvement. We saw a slight reduction in Overfitting after applying Weight Initialization and L2 Regularization on the Parkinson's Disease dataset because the training and validation accuracy are closely aligned. Due to Random Access Memory and GPU limitations, we could not produce a more significant number of augmented data and train the models for a longer time. If we train the Model with more data and longer time, there may be room for improvement, with a higher RAM and GPU.

2 Methods

Suppose we build a DL Model to detect objects (e.g., apples in a fruit collection image). To build the Model, we need to train it with a dataset that contains images of the object of interest. Depending on the complexity of the Model, Overfitting can arise and deters the Model from correctly detecting the object of interest. There has been a lot of research and development to mitigate Overfitting. Several methods that help to reduce Overfitting have been developed [25][48], such as Regularization, Early Stopping, Dropout, Data Augmentation, Transfer Learning as well as the methods mentioned in this section. Additionally, we can be interested in understanding how learning, cognition, and creative behaviors emerge [61]. To conduct an artificial intelligence-driven study, we may use microscopic brain images, such as (Human brain MRI, Electrocorticography, Gene Atlas, and Tracer Injection). Any model we build using the said data will be complex and depend on several hyperparameters. The proliferation of hyperparameters will most likely lead to Overfitting.

2.1 Regularization

The objective of Regularization is to prevent the DL Model from Overfitting by making it simpler or by adding elements to compensate for data-scarce [54]. For example, L2 Regularization adds a penalty term to the cost function to decrease the values of the Model's weights and simplify it. Regularization inhibits the development of complicated models in order to prevent Overfitting. We can simplify a DL model by altering its capacity: A model's capacity can fit a wide variety of functions. Models with high capacity can learn from complex data at the expense of Overfitting. It was concluded in [31] that increasing a CNN models' capacity helped its units better adapt and specialize to its tasks at the expense of overfitting [25]. For example, in a linear regression problem, our hypothesis space or capacity is the set of all linear functions. We can increase our capacity by adding the set of all polynomials to our hypothesis space. This increase will most certainly lead to Overfitting because our Model will have many functions to fit our data. The right capacity size depends chiefly on the problem's complexity and how much we can tolerate Overfitting. For example, simple regression has a single explanatory variable [3]. It means that we will use a single variable to predict another variable. For example, If we merely used age to predict the likelihood of Parkinson's Disease. In reality, any effort to make predictions using a single variable without attention to the other factors could result in statistical difficulties (termed "omitted variables bias"). For example, we say that age increases the likelihood of Parkinson's Disease without looking into other factors.

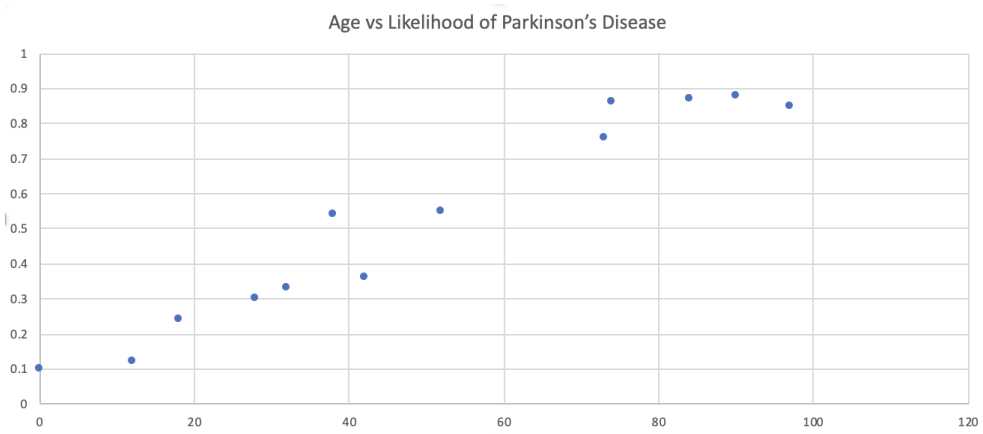


Figure 9: This data generated randomly for illustration.

Take Figure 9, for example. We can see that sometimes older people tend to develop Parkinson's disease. However, the relationship is not perfect. It seems that age does not suffice for an accurate prediction of the likelihood of Parkinson's disease. Therefore, we may deduce that other factors could influence the likelihood of Parkinson's disease; we may refer to them as "noise". Nevertheless, we can write the above relation mathematically as follows:

$$CI = A\beta + \epsilon,$$

Where β is the effect of an additional year of age on the likelihood of Parkinson's disease and ϵ is the "noise". CI is the dependent variable; A is the independent variable. The task of regression is to produce an estimate of β, based on the data provided and taken ϵ into account, meaning that we need to calculate the error ϵ and minimize it. One way to calculate the error is to calculate the difference between the actual and predicted values and average them. However, this might be problematic since there is a chance that the averaged variable will cancel each other out. For this reason, we take the sum of the Square, formally known as Residual Sum of Squares (RSS)[4]:

$$RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2,$$

Since this is a relatively simple example, we can shift our attention to more deep learning methods to minimize error. We will start with Dropout and move to Data augmentation, Ridge Regularization, Bayesian Regularization, and Early Stopping. The main objective of DL models is to calculate the prediction error and minimize it. Calculating prediction error and minimizing it can happen using several strategies including Regularization.

2.1.1 Dropout

Combining different DL models built for the same task and averaging their prediction results can reduce Overfitting. However, such a process requires a lot of computational resources. Dropout helps to finesse this situation by temporarily dropping units from a neural network along with their associated connections. The units to be dropped are chosen at random. Units that survived Dropout will constitute a "thinned" neural network. The left diagram in Figure 5 shows a thinned neural network. Ultimately the predictions of the thinned networks are averaged to mitigate Overfitting. In [15][25] it was shown that classification using Dropout is better than 1000 sub-networks using Monte Carlo approximation. In [15] it was also shown that Dropout is better than L1 and L2 Regularization. The computation and memory complexities of Dropout are ideal at O(n) – which is a function that increases linearly to reflect the time and space complexity of Dropout [54]. A linear complexity – as if the running time of a given algorithm increases linearly with the algorithm input – is highly preferable over higher-order functions/time complexities. For example, some algorithms have a quadratic time complexity typically denoted as $O(n^2)$. While Dropout ameliorates Overfitting, it also helps to combine enormous neural network architectures sufficiently [15].

2.1.2 Training with More Data - Data augmentation

The quantity and quality of the training dataset are a big factor in the Model's overall performance and accuracy. Expanding the training dataset will help improve the performance and accuracy of the Model. If we add noisy data, this technique will not work. However, training with more data can help improve the accuracy of the Model, especially in complicated models [48]. There is, however, a downside to this method. The high amount of data will increase the training time. Training data could be expensive to obtain. In the experiment done in [13] to combat overfitting in the ImageNet dataset – a dataset with 15 million labeled images mapped to 22,000 categories – [13] used data augmentation. First, [13] used "label-preserving transformation," which added more data to the training set. Random 224×224 patches from the 256×256 images were extracted by [13]. The CNN was merely trained on the extracted patches by [13]. The enlarging of the training set allowed for a more complex model without suffering from substantial Overfitting. Second, [13] performed Principal Component Analysis (PCA) — the method used to reduce the dimensionality of large datasets while preserving most of its information [18] — on the set of RGB (stands for Red Green Blue: combining these colors can produce images on screens) pixel values through the training set and alter the intensities of the RGB channels in training images. To each training image, [13] adds multiples of the found principal components, with magnitudes proportional to the corresponding eigenvalues times a random variable drawn from a Gaussian with mean zero and standard deviation 0.1. Therefore to each RGB image pixel. $I_{xy} = [I^R_{xy}, I^G_{xy}, I^B_{xy}]^T$ [13] add the following quantity:

$$[P_1, P_2, P_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T$$

where α_i and λ_i are ith eigenvector and eigenvalue of the 3×3 covariance matrix of RGB pixel values, respectively, and α_i is the aforementioned random variable [13]. Each α_i is drawn only once for all the pixels of a particular training image until that image is used for training again, at which point it is redrawn. This scheme approximately captures an important property of natural images, namely, that object identity is invariant to changes in the intensity and color of the illumination. This scheme reduces the top-1 error rate by over one percent.

2.1.3 L1 Regularization

As [48] summarized it, for neural networks, the objective is to find a perfect set of weights and biases. Finding an ideal set of weights and biases becomes complicated when the number of hyperparameters increases. To find an ideal set of weights and biases, the Model needs sufficient data for learning. The amount of data needs to be proportional to the number of hyperparameters. The Model's complexity increases as the number of hyperparameters increases. Even though some do not affect its accuracy, overfitting models consider all hyperparameters. So, we need to minimize the weights of hyperparameters that do not affect the Model's accuracy. Since we do not know which hyperparameters impact the Model's accuracy, L1 Regularization attempts to minimize all of them by minimizing the Model's cost function. Regularization using L1 introduces a "penalty term" into the cost function, as shown by the following formula:

$$\Omega(\omega) = \|w\|_1 = \textstyle\sum_i \|w_i\| \ ,$$

2.1.4 Weight Decay known as (L2 or ridge regularization)

We only mentioned one way to control a model's capacity. Instead of excluding or including more functions in our hypothesis space, we can also alter our capacity by expressing a preference for a specific function in the hypothesis space. This can be done by adding a parameter norm penalty $\Omega(\theta)$ to the objective function J [25] [34]:

$$\hat{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta) \ ,$$

Where α is a hyperparameter that weights the relative contribution of the norm penalty term Ω relative to the standard objective function $J(x, \theta)$. Setting α to 0 results in no regularization. Larger values of α correspond to more regularization [25] [34]. As was shown in [34] picking different parameter norms Ω leads to different preferred solutions that suits the problem in question. In [34]:

$$\hat{J}(\omega; X, y) = \tfrac{\alpha_1}{2}\omega^T\omega + \alpha_2|\omega|_1 + J(\omega; X, y) \ ,$$

was used as objective function J. With $\frac{\alpha_1}{2}\omega^T\omega + \alpha_2|\omega|_1$ as a parameter norm penalty they were able to reduce overfitting and produce a state of the art results for denoising autoencoders: a special type of neural network which takes corrupted data as input and predict the uncorrupted data. But in ridge regularization to reduce variance and generalization error – the proportion of incorrect output – in a DL model, researchers added:

$$\Omega(\theta) = \frac{1}{2}\|w\|_2^2 \ ,$$

as parameter norm penalty to the objective function. This term is known as weight decay – also referred to as Tikhonov's regularization – that could also be added to the cost function in order to restrict the model's parameters from increasing and/or penalizes the parameters of the DL model which results in a smaller squared L2 norm weight [70] [1].

2.1.5 Bayesian Regularization

Bayesian Regularization imposes prior distributions – distributions that express a prior opinion before taking any facts into account – on a model's parameters and penalizes its weights [20]. However, typical neural network models use the mean square errors (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(t_i - a_i)^2,$$

as an objective function and do not optimize the weights of the model [57]. In typical neural networks models, the optimal configuration of its weights is derived by minimizing MSE. Bayesian regularization neural networks apply a probability distribution to the Model's weights. The objective function in Bayesian Regularization includes network weights. It is written as [57]:

$$F(\omega) = \alpha E_\omega + \beta E_D,$$

Where E_D is MSE. β and α are objective parameters. E_ω is the sum of squared network weights also known as weight decay:

$$E_\omega = \frac{1}{n}\sum_{i=1}^{n}(w_j)^2.$$

A gradient-based based optimization method is used to minimize the function $F(\omega)$ and optimize the hyperparameters β and α. Here is an overview of this method. For more details see [12][2]. Given a M as a neural network model, ω as vector weights and N as total number of network weights. $P(\omega|, \beta, M)$ is the prior density. $P(D|\omega, \beta, M)$ is the likelihood function. $Z_D(\beta)$, $Z_\omega(\beta)$ and $P(D|\alpha, \beta, M)$ are the normalization factors. Based on Bayesian rule we have:

$$P(\omega|D, \alpha, \beta, M) = \frac{P(D|\omega,\beta,M)P(\omega|\alpha,M)}{P(D|\alpha,\beta,M)}$$

Since the prior distribution of network weight is viewed as a distribution of Gaussian, we have [19]:

$$P(D|\omega, \beta, M) = \frac{e^{-\beta E_D}}{Z_D(\beta)}$$

$$P(\omega|\alpha, M) = \frac{e^{-\alpha E_\omega}}{Z_\omega(\alpha)}$$

$$Z_D(\beta) = (\frac{\pi}{\beta})^{\frac{n}{2}}$$

$$Z_\omega(\alpha) = (\frac{\pi}{\alpha})^{\frac{N}{2}}$$

Now the optimal weight should maximize the posterior probability $P(\omega|D, \alpha, \beta, M)$ [12]. Bayesian Regularization was shown by [55] to reduce Overfitting in inverse modeling for filters using Deep Neural Network. Moreover, [55] increased the prediction accuracy of the Neural Network by using Bayesian Regularization when trying to predict coupling Matrix from Simulated S-Parameters.

2.1.6 Model Sparsity

It was stated in [54] that traditional machine-learning techniques were limited in their abilities to take original data in its raw form as input to the model. On the other hand, Data Scientists feed Deep learning models with data, and the Deep learning models automatically extract informative features to help with the task in question. Deep learning models are non-linear; they use their complex dimensionality to extract informative and complex features layer by layer. It was aslo stated by [54] that as the dimensionality of the model grows, it becomes more prone to overfitting. Due to their complexity and high dimensionality, deep learning models can extract irrelative and non-informative features. Data scientists use regularization to promote sparsity and reduce dimensionality by adding a penalty term to the model cost function. The penalty term reduces the number of neuron activation, which reduces the number of irrelevant features extracted by the deep learning model. As [38] puts it, a subset of the model parameters has a value of precisely zero. Weights with a zero value will not affect the model and have negligible computational power.

2.2 Early Stopping

Stopping the training early before the Model overfits the training dataset can reduce Overfitting. By monitoring the generalization error of the Model – a difference between the loss/error of a training dataset and its test dataset [40] – and training the Model multiple times with different values to select the number of epochs that produce the lowest error rate [8]. This method can be used to detect when Overfitting starts during supervised training; training is then stopped before convergence to avoid Overfitting. This strategy iteratively modifies the Model to improve its fit to the training data. However, improving the Model's fit to the training data may increase the generalization error.

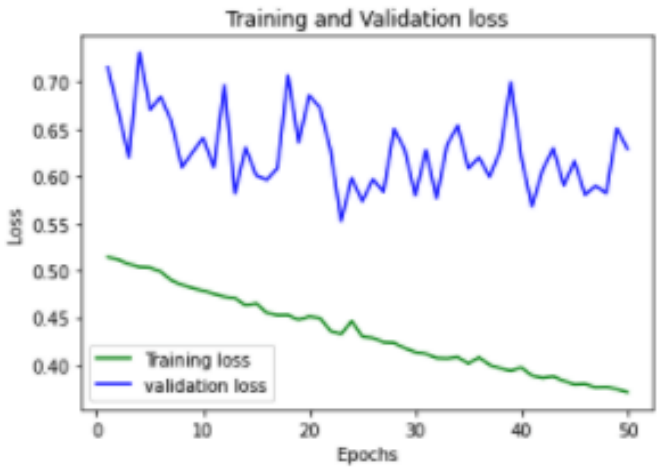


Figure 10 shows how training and validation errors decreased in Parkinson's Disease Dataset experiment from section 1.4.

When we run the DL Model, the training and validation errors decrease in parallel until epoch 7, then the validation error starts to increase (as we can see from figure 10). Early Stopping uses this fact to check at every iteration. If validation error is increasing, we stop the Model. In general terms, we execute the Model until the error on the validation set has plateaued; while doing so, we keep a record of the model configuration and the time the Model improved. Upon resuming training, we return to the stored configuration [25]. In [65] the model configuration was preserved using the so-called Checkpoint function in Keras . This function saves the Model's configuration once the early stopping algorithm halts the training. Using this approach [65] was able to achieve high classification results for diagnosing Parkinson's disease patients using MRI images. Notably, Early Stopping is equivalent to L2 Regularization, when we have a simple linear model with a quadratic error function and simple gradient descent [25]. This was also shown in [53] where early Stopping was resembled by using the quadratic function:

$$L(\Omega) = \sum_{i=1}^{n} (y_j - f(W, x_i))^2$$

as a loss function and applying gradient descent. Gradient descent is an optimization algorithm used in deep learning models to minimize their objective function $J(\theta)$ by finding its local minimum. The algorithm starts by setting θ equal to a random value. A constant *alpha* known as learning rate is used as a small value that determines the length of the algorithm's step towards the local minimum. Then it computes the derivative of the function $J'(\theta)$ at the given values of θ. This method is repeated until the algorithm achieves a value of θ where the derivative is equal to zero. At this value of θ, the function will be at its local minimum, and the function will converge [25][21].

2.3 Transfer Learning

Most deep learning models are trained from scratch. However, training from scratch has limitations. First, it requires a lot of computational resources. Second, it requires many training data, which is not always available. Finally, training a model from scratch requires tuning of many paraments and is highly prone to overfitting [16] [28] [33]. An alternative to training from scratch is transfer learning. In transfer learning, the knowledge gained by a deep learning model from a task with a lot of data available is transferred to a different task keeping the same weights and parameters. Transfer learning is commonly referred to as fine-tuning. In most cases, layers of the pre-trained model are retrained on the new dataset while the remaining layers and their weights are kept the same [67]. For example, [22] fine-tuned a CNN model pre-trained from natural images dataset to medical image tasks. As such, Transfer Learning enables the use of deep learning models on smaller datasets with only fine-tuning the part of the model. In [22] it was proven that transfer learning is effective even in cross-domain application since [22] used a network trained on natural images in a medical domain. Transfer learning can be enhanced by choosing the most informative data to retrain some fine-tuned model layers. For example, [28] extract the most informative slices of the MRI images to train the network by calculating the entropy of each image. The higher the entropy values of the slice, the more informative it would be.

2.4 Using Multiple Loss Functions

[17] used a DNN model with multiple loss functions to reduce Overfitting. [17] noted that researchers have not well explored the optimization of loss functions to prevent the overfitting problem. [17] used four kinds of the loss function in one Model, softmax loss, pairwise loss, LambdaRank top-1 loss, and LambdaRank top-2 loss. [17] explained that other loss functions might have the potential to prevent the algorithm overfitting to one softmax loss function. [17] was able to achieve a state of the art results on the CIFAR-10, MNIST, and SVHN datasets. [50] enhanced the classification and prediction accuracy of white blood cells images while lowering processing time through the use of deep convolutional neural network (DCNN) architecture by using a modified cross-entropy loss function.

2.5 Modifying The Activation Function

This method shows the correlation between overfitting and activation functions. A modified activation function called: modified-sigmoid is based on the well-known sigmoid function. This activation function can effectively improve the accuracy of the Model and inhibit the overfitting problem [43]. Before we dig deep, let us describe activation functions and their role in neural networks.

2.5.1 Activation Functions

Each layer of the neural network needs a non-linear function called the activation function before generating the output signal. When there is no non-linear activation

function, this neural network layer is transformed linearly. No matter how many layers the network contains, the final output can be expressed by the linear transformation of the input. Therefore, it has no difference from the single-layer neural network [43]. There are three commonly used activation functions:

Sigmoid: Sigmoid activation function:

$$Sigmoid = \frac{1}{1 + e^{-x}}$$

Tanh activation function:

$$Tanh(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$$

ReLu activation function:

$$ReLu(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

The proposed method in [43] introduced a new activation function that showed promising results in terms of reducing Overfitting. The sigmoid activation function is a near S-shaped curve. The modified-sigmoid activation function is an improvement of the Sigmoid activation function. By defining a hyperparameter, the Modified-sigmoid activation function is shown as follows:

$$ModifiedSigmoid(x) = \begin{cases} x, & -w \leq x \leq w \\ Sigmoid, & x < -w \text{ or } x > w \end{cases}$$

The four different activation functions were compared by [43] using (MNIST) dataset for for training and testing the model in this proposed method.

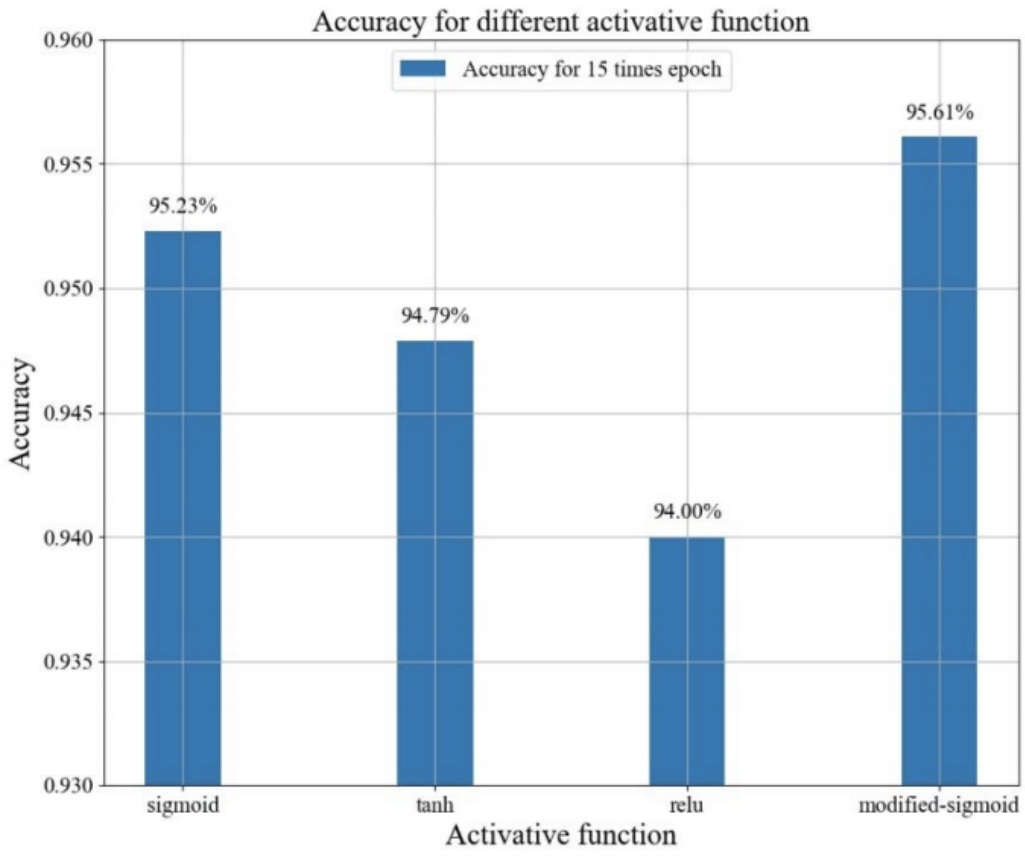


Figure 12: Effect of activation function on model accuracy [43].

As we can see in figure 12, the activation function plays a vital role in preventing the Model from Overfitting. After optimization, it was found in [43] that the neural network possesses the best test accuracy with hyperparameter $w = 2$. Furthermore, the results in [43] show that the Modified-sigmoid function can effectively improve the accuracy of the neural network and prevent the Overfitting of the neural network.

2.6 K-fold Cross-Validation

K refers to the number of groups that a given dataset will be split into. "Fold" refers to the number of resulting subsets. Cross-validation helps estimate how the Model is expected to perform when making predictions on unseen data. The general idea is to shuffle the dataset randomly and split it into k groups. We take each group as a test dataset and the remaining groups as the training set. We fit the Model on the training set and evaluate it on the test set. Once we retain the evaluation score, we discard the model [6]. This method reduces bias and computation time as we repeat the process only ten times when the value of k is 10. Every data point is tested precisely once and is used in training k-1 times. As we increase k, the variance of the resulting estimate is reduced as well.

2.7 Removing Features

Unjustifiable features that do not fit the Model should be removed. Some algorithms remove irrelevant features by default. Removing irrelevant input costs less computational power. This method improves prediction accuracy by the exclusion of irrelevant variables. The Model to be built is simpler and faster when fewer input variables are used [11].[49] experimented if deep learning can extract latent Multiple Sclerosis (MS) ¬– a neurological disease — lesion (damaged areas of the brain) features that when combined with three user-defined MRI measurements (T2w lesion volume, brain volume, and diffusely abnormal white matter DAWM) and eight user-defined clinical measurements (gender, cerebrum, optic nerve, cerebellum, brain stem, spinal cord, EDDS, and cistype) can accurately predict the chance of conversion from clinically isolated syndrome (CIS), an initial stage of MS, to clinically definite MS (CDMS). [49] used a CNN model fed with lesion masks segmented from MRI images. [49] applied the feature selection method to increase the accuracy of the Model and avoid Overfitting. [49] computed the relative importance of each of the 11 user-defined features for classification between converters and non-converters by permuting the features among the training data and computing the average generalization error. They then selected discriminative features by choosing the features whose relative importance is larger than the median importance. The importance of each feature is shown in figure 13.

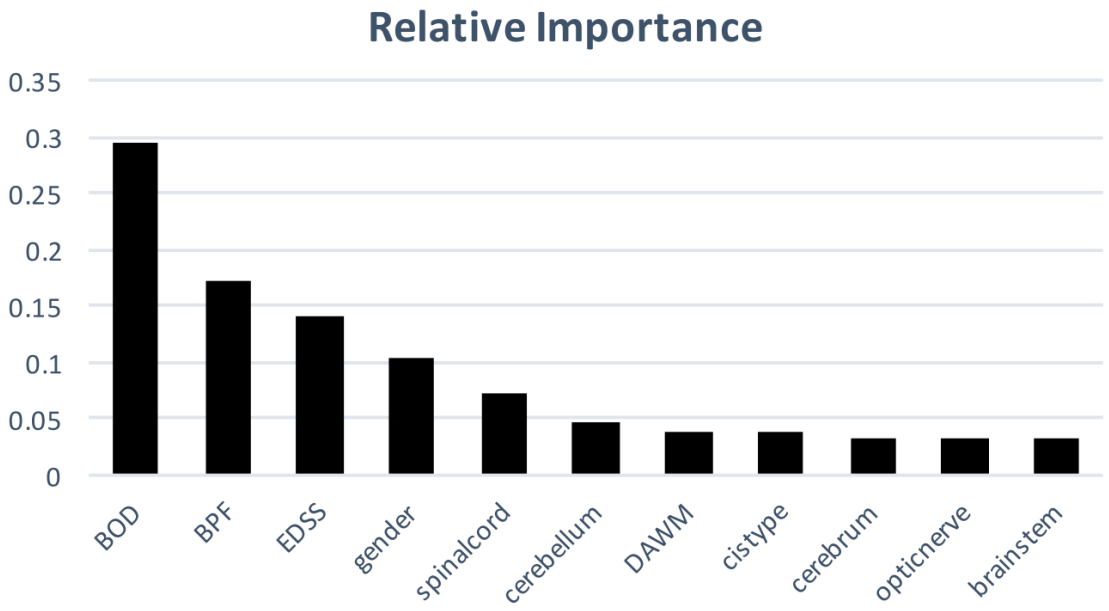


Figure 13: the importance of each feature on the Model's accuracy [49].

The selected features were BOD, BPF, EDSS, gender, spinal cord, and cerebellum. [49] trained the Model using the selected features only and were able to increase the Model's accuracy from 63.6 to 73.8, which amounts to a 10 percent increase. [36] observed that neural networks have many redundant features that are very similar. [36] proposed an algorithm to regularize related features and therefore encourage

features diversity in the model. As a result, [36] eliminated redundant features. This method reduced the computational overhead. [36] was able to achieve a state of the art accuracy and reduce computational overhead as a result of this method

$$J_D(\phi) = \sum_{l=1}^{L} \left(\frac{1}{2} \sum_{i=1}^{n'_l} \sum_{j=1}^{n'_l} \left(\overset{(l)}{\Omega}_{ij} \right)^2 \overset{(l)}{\mathbf{M}}_{ij} \right)$$

To encourage diversity of features, [36] added the J_D term to the loss function $J(\theta; X, y)$. [36] was able to achieve a state of the art accuracy and reduce computational overhead as a result of this method.

2.8 Critical Evaluation of Overfitting Methods using Brain Data

Now that we have discussed some methods to address Overfitting, we will implement them and discuss their advantages and drawbacks. As such, we chose a dataset that contains images of MRI Segmentation. The dataset contains images of healthy brains and brains affected by Alzheimer's — a type of dementia that affects brain functions and interferes with regular daily activities. Accordingly, we split the data into three sets. A training set to train our Model, testing set to test our Model once the training is complete, and a validation set to test our Model during the training phase and help it find the weights that can produce the highest level of accuracy. The dataset has four classes; "Non-Demented," "Mild Demented," "Moderate Demented," and "Very Mild Demented." Each of these classes represents a stage of Alzheimer's. Early detection of Alzheimer's is critical because treating it at an early stage is much easier than at advanced stages.

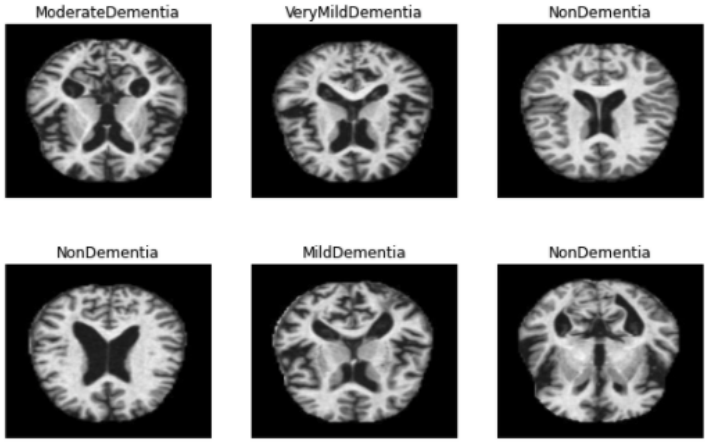


Figure 14 MRI Images

We built a basic CNN model with four convolution layers, four max-pooling layers, and two dense layers. We illustrate the model specification in figure 15.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1 (Conv2D)              (None, 128, 128, 128)     3328

 max_pooling2d (MaxPooling2D  (None, 40, 40, 128)      0
 )

 conv2 (Conv2D)              (None, 40, 40, 64)        204864

 max_pooling2d_1 (MaxPooling  (None, 12, 12, 64)       0
 2D)

 conv3 (Conv2D)              (None, 12, 12, 32)        18464

 max_pooling2d_2 (MaxPooling  (None, 4, 4, 32)         0
 2D)

 conv4 (Conv2D)              (None, 4, 4, 32)          9248

 max_pooling2d_3 (MaxPooling  (None, 1, 1, 32)         0
 2D)

 flatten (Flatten)           (None, 32)                0

 dense (Dense)               (None, 64)                2112

 dense_1 (Dense)             (None, 4)                 260

=================================================================
Total params: 238,276
Trainable params: 238,276
Non-trainable params: 0
```

Figure 15 CNN Configuration

The Model was only able to achieve 55 % accuracy. Additionally, as shown in Figure 16, the training accuracy kept increasing while the testing accuracy pleated around 50%. Therefore, the Model is experiencing Overfitting. Consequently, we will use the methods discussed earlier to improve the Model's accuracy and reduce Overfitting.

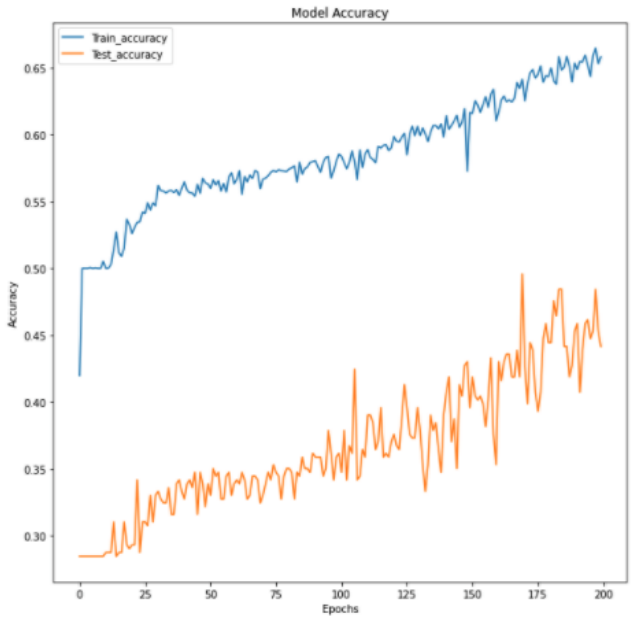


Figure 16 Training vs. Testing accuracy basic CNN model

2.8.1 Data Augmentation and Dropout

Originally the training dataset had a total of 5121 images. 717 MildDemented, 52 ModerateDemented, 2560 NonDemented and 1792 VeryMildDemented. On the other hand, the testing dataset had 1279 images. 179 MildDemented, 12 ModerateDemented, 640 NonDemented and 488 VeryMildDemented. As shown in [26] the left and right regions of the brain are symmetrical. We augmented the data by flipping the image along the horizontal axis. Due to Random Access Memory limitation, we could only produce 77126 images for the training set and 20601 images for the testing set. We tried to balance the dataset by producing more images in the classes with a limited number of images. For example, we only had 12 MildDemented images in the training set. As such, we produced 17925 images of the MildDemented class. Finally, we added two layers of Dropout, each with a Dropout rate of 0.5.

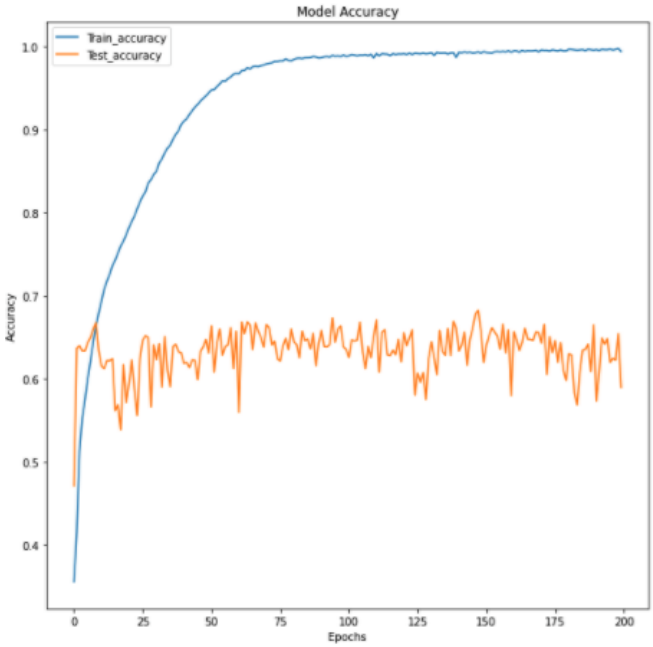


Figure 17 Training vs. Testing accuracy Data Augmentation and Dropout CNN model

We were able to increase the Model's accuracy by 4%. After applying the mentioned Dropout and Data Augmentation techniques, the Model achieved a 59% accuracy. However, as shown in figure 17, the Model is still experiencing Overfitting as the testing accuracy plateaued and the training accuracy increased after reaching roughly Epoch 25.

2.8.2 Data Augmentation and L2 Regularization with and without Dropout

In this section, we trained the Model with Data Augmentation in addition to L2 Regularization. Then, we trained the Model again with Data Augmentation, L2 Regularization, and Dropout regularization.

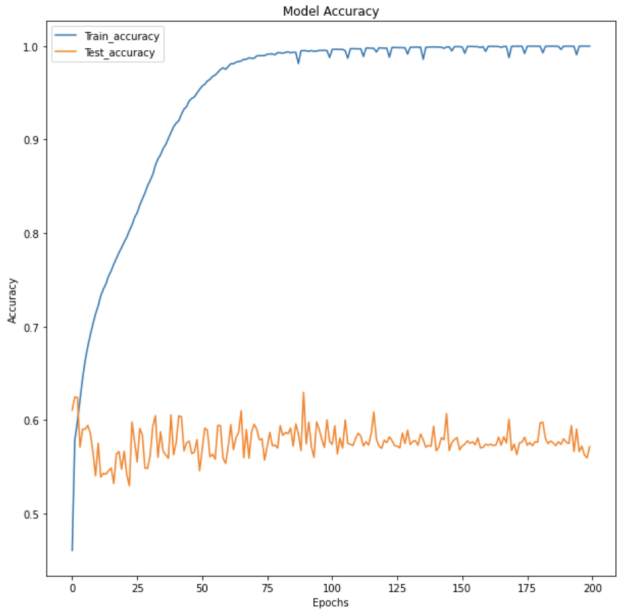


Figure 18 Training vs. Testing accuracy Data Augmentation and L2 Regularization without Dropout CNN model

When training the Model with Data Augmentation and L2 Regularization without Dropout regularization, the Model achieved 57% accuracy. However, compared to training with Data Augmentation and Dropout Regularization, the Model underperformed by 2%. We previously performed 59% accuracy with Data Augmentation and Dropout regularization. As shown in figure 18, the Model is still undergoing Overfitting as the training accuracy is plateauing at around 90%, and the testing accuracy is plateauing around 57%.

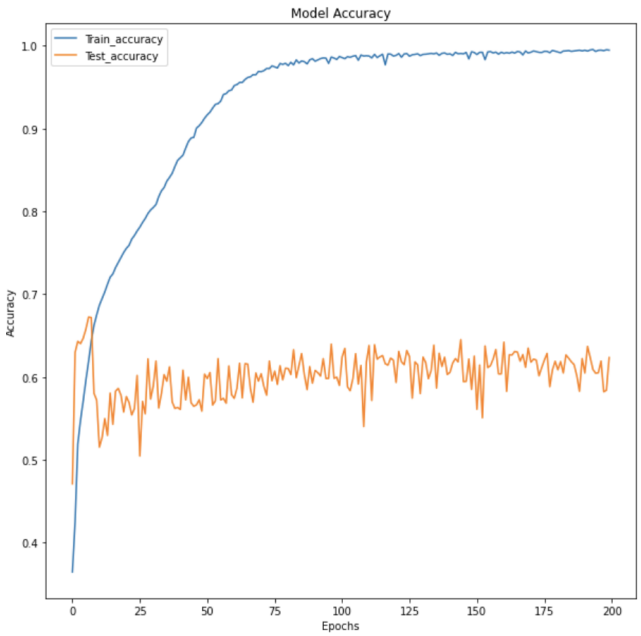


Figure 19 Training vs. Testing accuracy Data Augmentation and L2 Regularization with Dropout CNN model

When training the Model with Data Augmentation and L2 Regularization with Dropout regularization, the Model achieved 62% accuracy. Nevertheless, compared to training with Data Augmentation and Dropout regularization, the Model overperformed by 3%. We previously performed 59% accuracy with Data Augmentation and Dropout regularization. Additionally, the Model overperformed by 5% compared to the Model we trained with merely Data Augmentation and L2 regularization. As shown in figure 19, the Model is still undergoing Overfitting as the training accuracy is plateauing at around 90%, and the testing accuracy is plateauing around 62%.

2.8.3 Early Stopping with Data Augmentation, L2 and Dropout Regularization

In the previous models, we set a specific number of Epoch — the number of passes of the entire training dataset the Model has completed. The Model had to go through the whole training set based on a pre-specified number of epochs regardless of whether its performance improved. Therefore, training without Early Stopping took a long time and did not improve the Model's accuracy. We utilized the Early Stopping method by monitoring the Model's accuracy to save time. Since we are dealing with a classification problem, monitoring the accuracy would be most appropriate instead of monitoring the loss. Once we noticed an improvement in training accuracy, we saved the Model's configuration and returned to it for further training. We halted the training once we noticed that the Model's accuracy had not improved for a long time. We pre-specified how many epochs the Model should go through before we halt the training once we notice no improvement in accuracy. Table 2 shows how the Model improved throughout the training process. Once the Model's accuracy reached 0.63880, it stopped improving for 87 epochs, and that was where we decided to halt the training. The Model may improve if we let it go beyond 87 epochs at the expense of time.

From	To
-inf	0.52968
0.52968	0.57099
0.57099	0.57366
0.57366	0.57832
0.57832	0.58784
0.58784	0.58798
0.58798	0.59177
0.59177	0.60293
0.60293	0.60881
0.60881	0.60973
0.60973	0.63060
0.63060	0.63278
0.63278	0.63613
0.63613	0.63880

Table 2

Overall the Model achieved an accuracy of approximately 64%, Which is an improvement from the previous 62% achieved by the Model trained with Data Augmentation, L2, and Dropout regularization.

2.8.4 Changing the Activation Function and Removing Features

One irrelevant feature that we could have removed to improve the Model's accuracy is the skull from the MRI image. However, as shown in figure 19, the skull was already removed from the dataset. We are currently not aware of any other irrelevant features that we could remove to improve the Model's accuracy.

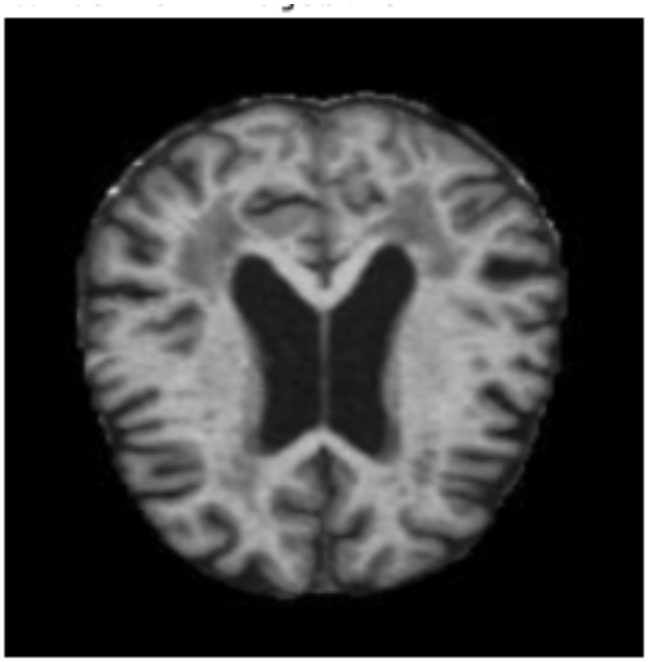


Figure 20 Brain MRI image extracted from the Alzheimer training set.

As shown in Table 3, we trained the Early Stopping with Data Augmentation, L2, and Dropout Regularization model with several different activation functions. The Model improved when using the Hyperbolic Tangent Activation Function, achieving a 66% classification accuracy. However, the accuracy worsened when using other activation functions.

Activation Function	Accuracy
Rectified Linear Unit	63%
Sigmoid	53%
Hyperbolic Tangent	66%
Leaky version of a Rectified Linear Unit	62%
Parametric Rectified Linear Unit	61%
Thresholded Rectified Linear Unit	52%
Softmax	53%

Table 3

2.8.5 Discussion

We reached the highest accuracy when we trained the Model with Early Stopping, Data Augmentation, L2, and Dropout Regularization using the Hyperbolic Tangent Activation Function. One drawback we noticed when using Dropout is a longer training time than training the Model without Dropout. As [35] stated, training a model with Dropout Regularization will significantly increase the training time. It is rational to use dropout in case of enough computational resources [54]. Models trained with Dropout Regularization take a long time than models trained without Dropout Regularization. Dropout creates thinned networks after randomly dropping several units and combines them all together. As stated in [35] training a Dropout neural network with n hidden units can be seen as training a collection of 2n different "thinned" models with extensive weight sharing. [35] improved on the traditional Dropout method by omitting to drop units from the neural network that has a positive impact on the Model's accuracy and opting to dropout units that have no potential to improve the Model's accuracy. Units with high activation values positively impact the Model's accuracy, and units with low activation values have a low impact on the Model's accuracy. Selectively adjusting the amount of Regularization on the Model's weights has increased the accuracy of neural networks by reducing Overfitting. [30] showed that it could also mitigate catastrophic forgetting. Catastrophic forgetting occurs when a network learns task A and forgets it once it starts to learn another task B. Catastrophic forgetting occurs because the weights in the network that are important for task A are changed to meet the objectives of task B. To mitigate catastrophic forgetting [30] reduced Regularization on certain weights based on how important they are to previously seen tasks. Similarly, [42] put more weight on the unit outputs that support correct predictions on the training set when they introduced a novel dropout regularization called Spectral Dropout (SD). SD discarded noisy spectral components during the train and test phases. The spectral Dropout speeds up the convergence rate during the training process. As stated in [45], to reduce computation time, many network weights are redundant, and a regularizer can remove it from the network without much loss of performance. Different from the traditional Dropout method, with the moving of convolution filters on the input feature maps, [56] dropped out different units with fixed or random drop rates, introducing more uncertainty. However, experimental results indicated that [56]'s method could prevent overfitting. [45] used a spare 1 regularizer and applied it to the matrix space of network weight to zero redundant weights and removed their unnecessary connections and neurons. [45] hypothesized that reducing the number of unnecessary connections can reduce computation and memory requirements. [30] tested this method on several datasets and was able to obtain state-of-the-art accuracy. We started with 55% accuracy with a basic CNN model, which did not use any methods to mitigate Overfitting. We ended with a 66% accuracy with the Model that used Early Stopping with Data Augmentation, L2, and Dropout Regularization using the Hyperbolic Tangent Activation Function. Given the high complexity of the task and dataset complexity, an 11% increase in accuracy is considered a good improvement. However, there are still signs of Overfitting, and the Model's accuracy can be improved further. Our

model is very complex compared to the one used in [71] to predict brain age. [71] used a lightweight CNN architecture. [71] kept only one convolutional layer before each MaxPool layer to reduce memory requirements. In addition, [71] removed all the fully connected layers, which not only greatly reduces the number of parameters. [71] reduced the GPU memory consumption by liming the channel numbers of the first layer to 32. We could use this idea in our research to reduce GPU memory consumption. A problem we faced when training our model on the Alzheimer dataset. Using data augmentation and regularization techniques, [71] achieved state-of-the-art results. [71] stated that it was intriguing why the lightweight model performs better than the deeper ones, something we can explore in our feature research. [71] noted that the choice of optimizer might affect the model performance. [71] noticed a change in performance when switching from Adam optimizer to Stochastic gradient descent Optimizer. As [60] concluded, the methods we mentioned and experimented with have produced state-of-the-art results in many fields; they have not gained as much ground over human neuroscience datasets. The high dimensionality and large amounts of noise present in neuroscience datasets, coupled with limited data available that can be reasonably obtained from a sample of human subjects, lead to difficulty training deep learning models. Overfitting can occur despite the presence of regularization techniques. [60] introduced the paired trial classification (PTC) method for classifications. PTC reduces the multiclass classification problem to two classes, potentially simplifying the problems and thus presenting a way of making it easier to achieve robust classification performance from limited training data. Instead of classifying a single example at a time into several classes, [60] instead attempt to classify pairs of examples as belonging to either the same class or different classes. [62] demonstrate that DL methods, if implemented and trained according to the common practices, have the potential to outperform standard machine learning methods with lower computational time and space despite being more complex in their architecture and parameterization, using raw or minimally preprocessed input data. [62] states that DL approaches are just beginning to show successes in diagnostic classification and disease prediction domains.

3 Development of DL Computational Models Addressing Overfitting

4 Applications of DL Handling Overfitting

5 Conclusions and Future Directions

References

[1] Andrey Nikolayevich Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

[2] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[3] Alan O Sykes. An introduction to regression analysis. *Coase Sandor Institute for Law Economics Working Paper*, 1993.

[4] Thomas J Archdeacon. Correlation and regression analysis: a historian's guide. 1994.

[5] Branislav Kisacanin and Dan Schonfeld. A fast thresholded linear convolution representation of morphological operations. *IEEE transactions on image processing*, 3(4):455–457, 1994.

[6] Andrew Y Ng et al. Preventing" overfitting" of cross-validation data. In *ICML*, volume 97, pages 245–253. Citeseer, 1997.

[7] Gregory S Chirikjian and Imme Ebert-Uphoff. Numerical convolution on the euclidean group with applications to workspace generation. *IEEE Transactions on Robotics and Automation*, 14(1):123–136, 1998.

[8] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

[9] Steve Lawrence and C Lee Giles. Overfitting and neural networks: conjugate gradient and backpropagation. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 1, pages 114–119. IEEE, 2000.

[10] Inder M Verma, L Naldini, T Kafri, H Miyoshi, M Takahashi, U Blömer, N Somia, L Wang, and FH Gage. Gene therapy: promises, problems and prospects. In *Genes and Resistance to Disease*, pages 147–157. Springer, 2000.

[11] Juha Reunanen. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3(Mar):1371–1382, 2003.

[12] Zhao Yue, Zhao Songzheng, and Liu Tianshi. Bayesian regularization bp neural network model for predicting oil-gas drilling cost. In *2011 International Conference on Business Management and Electronic Information*, volume 2, pages 483–487. IEEE, 2011.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[14] Janice M Beitz et al. Parkinson's disease: a review. *Front Biosci (Schol Ed)*, 6(6):65–74, 2014.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[16] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[17] Chunyan Xu, Canyi Lu, Xiaodan Liang, Junbin Gao, Wei Zheng, Tianjiang Wang, and Shuicheng Yan. Multi-loss regularized deep neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(12):2273–2283, 2015.

[18] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[19] Murat Kayri. Predictive abilities of bayesian regularization and levenberg–marquardt algorithms in artificial neural networks: a comparative empirical study on social data. *Mathematical and Computational Applications*, 21(2):20, 2016.

[20] Hayrettin Okut. Bayesian regularized neural networks for small n big p data. *Artificial neural networks-models and applications*, pages 28–48, 2016.

[21] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[22] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[23] P Zham, DK Kumar, P Dabnichki, S Raghav, and SM Keloth. Dynamic handwriting analysis for assessing movement disorder. In *13th International Conference of Applied Computing*, 2016.

[24] Ahmed Ali Mohammed Al-Saffar, Hai Tao, and Mohammed Ahmed Talab. Review of deep convolution neural network in image classification. In *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pages 26–31. IEEE, 2017.

[25] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. Deep learning. 2017.

[26] Ammarah Farooq, SyedMuhammad Anwar, Muhammad Awais, and Saad Rehman. A deep cnn based multi-class classification of alzheimer's disease using mri. In *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–6, 2017.

[27] Pranay Reddy Gankidi and Jekan Thangavelautham. Fpga architecture for deep learning and its application to planetary robotics. In *2017 IEEE Aerospace Conference*, pages 1–9. IEEE, 2017.

[28] Marcia Hon and Naimul Mefraz Khan. Towards alzheimer's disease classification through transfer learning. In *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pages 1166–1169. IEEE, 2017.

[29] Nikhil Ketkar. Introduction to keras. In *Deep learning with Python*, pages 97–111. Springer, 2017.

[30] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[31] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Growing a brain: Fine-tuning by increasing model capacity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2471–2480, 2017.

[32] Poonam Zham, Dinesh K Kumar, Peter Dabnichki, Sridhar Poosapadi Arjunan, and Sanjay Raghav. Distinguishing different stages of parkinson's disease using composite index of speed and pen-pressure of sketching a spiral. *Frontiers in neurology*, page 435, 2017.

[33] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z Papadakis, Adrien Depeursinge, Ronald M Summers, et al. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(1):1–6, 2018.

[34] Zong Meng, Xuyang Zhan, Jing Li, and Zuozhou Pan. An enhancement denoising autoencoder for rolling bearing fault diagnosis. *Measurement*, 130:448–454, 2018.

[35] Alvin Poernomo and Dae-Ki Kang. Biased dropout and crossmap dropout: learning towards effective dropout regularization in convolutional neural network. *Neural networks*, 104:60–67, 2018.

[36] Babajide O Ayinde, Tamer Inanc, and Jacek M Zurada. Regularizing deep neural networks by enhancing diversity in feature extraction. *IEEE transactions on neural networks and learning systems*, 30(9):2650–2661, 2019.

[37] Silvia Basaia, Federica Agosta, Luca Wagner, Elisa Canu, Giuseppe Magnani, Roberto Santangelo, Massimo Filippi, Alzheimer's Disease Neuroimaging Initiative, et al. Automated classification of alzheimer's disease and mild cognitive impairment using a single mri and deep neural networks. *NeuroImage: Clinical*, 21:101645, 2019.

[38] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

[39] Hakan Gunduz. Deep learning-based parkinson's disease classification using vocal feature sets. *IEEE Access*, 7:115540–115551, 2019.

[40] Daniel Jakubovitz, Raja Giryes, and Miguel RD Rodrigues. Generalization error in deep learning. In *Compressed Sensing and Its Applications*, pages 153–193. Springer, 2019.

[41] Bijen Khagi, Bumshik Lee, Jae-Young Pyun, and Goo-Rak Kwon. Cnn models performance analysis on mri images of oasis dataset for distinction between healthy and alzheimer's patient. In *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, pages 1–4. IEEE, 2019.

[42] Salman H Khan, Munawar Hayat, and Fatih Porikli. Regularization of deep neural networks with spectral dropout. *Neural Networks*, 110:82–90, 2019.

[43] Haidong Li, Jiongcheng Li, Xiaoming Guan, Binghao Liang, Yuting Lai, and Xinglong Luo. Research on overfitting of deep learning. In *2019 15th International Conference on Computational Intelligence and Security (CIS)*, pages 78–81. IEEE, 2019.

[44] Yu Li, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, 166:4–21, 2019.

[45] Rongrong Ma, Jianyu Miao, Lingfeng Niu, and Peng Zhang. Transformed 1 regularization for learning sparse deep neural networks. *Neural Networks*, 119:286–298, 2019.

[46] Dan Wang, Phillip WL Tai, and Guangping Gao. Adeno-associated virus vector as a platform for gene therapy delivery. *Nature reviews Drug discovery*, 18(5):358–378, 2019.

[47] Qi Xu, M. Zhang, Z. Gu, and Gang Pan. Overfitting remedy by sparsifying regularization on fully-connected layers of cnns. *Neurocomputing*, 328:69–74, 2019.

[48] Xue Ying. An overview of overfitting and its solutions. In *Journal of Physics: Conference Series*, volume 1168, page 022022. IOP Publishing, 2019.

[49] Youngjin Yoo, Lisa YW Tang, David KB Li, Luanne Metz, Shannon Kolind, Anthony L Traboulsee, and Roger C Tam. Deep learning of brain lesion patterns and user-defined clinical and mri features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 7(3):250–259, 2019.

[50] Jaya Basnet, Abeer Alsadoon, PWC Prasad, Sarmad Al Aloussi, and Omar Hisham Alsadoon. A novel solution of using deep learning for white blood cells classification: enhanced loss function with regularization and weighted loss (elfrwl). *Neural Processing Letters*, 52(2):1517–1553, 2020.

[51] Sabyasachi Chakraborty, Satyabrata Aich, Eunyoung Han, Jinse Park, Hee-Cheol Kim, et al. Parkinson's disease detection from spiral and wave drawings using convolutional neural networks: A multistage classifier approach. In *2020 22nd International Conference on Advanced Communication Technology (ICACT)*, pages 298–303. IEEE, 2020.

[52] Emtiaz Hussain, Mahmudul Hasan, Syed Zafrul Hassan, Tanzina Hassan Azmi, Md Anisur Rahman, and Mohammad Zavid Parvez. Deep learning based binary classification for alzheimer's disease detection using brain mri images. In *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1115–1120. IEEE, 2020.

[53] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International conference on artificial intelligence and statistics*, pages 4313–4324. PMLR, 2020.

[54] Reza Moradi, Reza Berangi, and Behrouz Minaei. A survey of regularization strategies for deep models. *Artificial Intelligence Review*, 53(6):3947–3986, 2020.

[55] Guangyuan Pan, Yang Wu, Ming Yu, Liping Fu, and Hongsheng Li. Inverse modeling for filters using a regularized deep neural network approach. *IEEE Microwave and Wireless Components Letters*, 30(5):457–460, 2020.

[56] Hengyue Pan, Xin Niu, Rongchun Li, Siqi Shen, and Yong Dou. Dropfilterr: A novel regularization method for learning convolutional neural networks. *Neural Processing Letters*, 51(2):1285–1298, 2020.

[57] Eduard Sariev and Guido Germano. Bayesian regularized artificial neural networks for the estimation of the probability of default. *Quantitative Finance*, 20(2):311–328, 2020.

[58] Mohamed Shaban. Deep convolutional neural network for parkinson's disease based handwriting screening. In *2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops)*, pages 1–4. IEEE, 2020.

[59] Catherine Taleb, Laurence Likforman-Sulem, Chafic Mokbel, and Maha Khachab. Detection of parkinson's disease from handwriting using deep learning: A comparative study. *Evolutionary Intelligence*, pages 1–12, 2020.

[60] Jacob M Williams, Ashok Samal, Prahalada K Rao, and Matthew R Johnson. Paired trial classification: A novel deep learning technique for mvpa. *Frontiers in Neuroscience*, 14:417, 2020.

[61] Chenzhong Yin, Xiongye Xiao, Valeriu Balaban, Mikhail E Kandel, Young Jae Lee, Gabriel Popescu, and Paul Bogdan. Network science characteristics of brain-derived neuronal cultures deciphered from quantitative phase imaging data. *Scientific Reports*, 10(1):1–13, 2020.

[62] Anees Abrol, Zening Fu, Mustafa Salman, Rogers Silva, Yuhui Du, Sergey Plis, and Vince Calhoun. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):1–17, 2021.

[63] Javier Barbero-Gómez, Pedro-Antonio Gutiérrez, Víctor-Manuel Vargas, Juan-Antonio Vallejo-Casas, and César Hervás-Martínez. An ordinal cnn approach for the assessment of neurological damage in parkinson's disease patients. *Expert Systems with Applications*, page 115271, 2021.

[64] basel99. Parkinson's disease detection, Jul 2021.

[65] Nanziba Basnin, Nazmun Nahar, Fahmida Ahmed Anika, Mohammad Shahadat Hossain, and Karl Andersson. Deep learning approach to classify parkinson's disease from mri samples. In *International Conference on Brain Informatics*, pages 536–547. Springer, 2021.

[66] Kexin Huang, Cao Xiao, Lucas M Glass, Cathy W Critchlow, Greg Gibson, and Jimeng Sun. Machine learning applications for therapeutic tasks with genomics data. *arXiv preprint arXiv:2105.01171*, 2021.

[67] Nusrat Jahan, Arifatun Nesa, and Md Abu Layek. Parkinson's disease detection using cnn architectures withtransfer learning. In *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, pages 1–5. IEEE, 2021.

[68] Iqra Kamran, Saeeda Naz, Imran Razzak, and Muhammad Imran. Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease. *Future Generation Computer Systems*, 117:234–244, 2021.

[69] L Sathish Kumar, S Hariharasitaraman, Kanagaraj Narayanasamy, K Thinakaran, J Mahalakshmi, and V Pandimurugan. Alexnet approach for early stage alzheimer's disease detection from mri brain images. *Materials Today: Proceedings*, 2021.

[70] Ujjwal Kumar and Anamitra Bhar. Studying and analysing the effect of weight norm penalties and dropout as regularizers for small convolutional neural networks. *International Journal of Engineering Research Technology*, 2021.

[71] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871, 2021.

[72] Adrian Schwarzer, Steven R Talbot, Anton Selich, Michael Morgan, Juliane W Schott, Oliver Dittrich-Breiholz, Antonella L Bastone, Bettina Weigel, Teng Cheong Ha, Violetta Dziadek, et al. Predicting genotoxicity of viral vectors for stem cell gene therapy using gene expression-based machine learning. *Molecular Therapy*, 2021.

[73] Xinxing Zhao, Candice Ke En Ang, U Rajendra Acharya, and Kang Hao Cheong. Application of artificial intelligence techniques for the detection of alzheimer's disease using structural mri images. *Biocybernetics and Biomedical Engineering*, 41(2):456–473, 2021.