J. Dawn. Res. Vol. 4 (2022)
www.dawningresearch.org

Journal of Dawning Research

Machine Learning For Alzheimer's Disease Diagnosis: Computer Vision and Recurrent Neural Networking

Research Article

Jack Diskin[1]*, Alison Huenger[1]†

1 Manhasset High School, 200 Memorial Pl, Manhasset, New York 11030, USA

Abstract: In 2013, the Virginia Medical Center found that 12 million Americans are impacted by clinical diagnostic errors each year — mistakes that result in 40,000 to 80,000 annual fatalities. The purpose of this study was to address this issue through the implementation of convolutional neural network (CNN) and long-short term memory (LSTM) machine learning algorithms in python, trained for dementia classification via MRI and demographic data. The CNN is a computer vision model capable of 2-D feature extraction through kernel convolution and pooling operations, in tandem with interconnected vectors (dense layers), containing individual elements referred to as neurons, which provide class predictions. The LSTM is a recurrent neural network composed of independent computational cells designed to perform sequential data analysis. Each of these cells are defined by a cell state, to which information can be appended or removed via regulatory gates that utilize the sigmoid function and pointwise multiplication to make decisions. The final LSTM cell outputs a refined version of its cell state data to prediction-deducing dense layers. These algorithms can be tuned through backpropagation, where an optimization algorithm adjusts the parameters connecting individual neurons based on a given prediction's error. It was hypothesized that both algorithms would achieve greater than 85% classification accuracy, and that the LSTM model would return superior metrics. Prior to utilization, MRI volumes were sliced along the sagittal axis, voided of skeletal features, and noise injected using the gaussian convolution kernel. Data augmentation using rotation, gamma correction, and scaling within a defined set of parameter values (ex. degrees, gamma value, scale factor) yielded a 27,000-fold data quantity increase. The model was then trained for 6 independent trials using cross validation. The demographic data was pre-processed via selective variable removal and null value replacement, and the algorithm was trained based on 9 elements: gender, age, education level, socioeconomic status, MMSE and CDR score, estimated total intracranial volume, atlas scaling factor, and normalized whole brain volume. Both models were also optimized through the adjustment of model hyperparameters. The CNN model achieved a final accuracy of 89.5%, 87.3% precision, and 86.8% recall. The LSTM achieved 82.3% accuracy, 71.7% precision, and 75.8% recall. It can be inferred that the CNN's superior efficacy was a result of superior dataset size, allowing for better abstractualization and fit as the model was able to decipher general delineations between dementia-affected and healthy brains.

Keywords: Alzheimer's Disease • Machine Learning • Diagnosis • CNN • LSTM • MRI

* *E-mail: jackpdiskin@gmail.com*
† *E-mail: alison_huenger@manhassetschools.org*

1. Introduction

1.1. Background & Terminology

According to Greb (2013), roughly 12% to 23% of individuals diagnosed with Alzheimer's disease (AD), the most prevalent neurodegenerative disorder in today's society, are victims of misdiagnosis [1]. These same misguided patients have a 67% chance of being prescribed inappropriate medication, which often leads to unnecessary health complications. Ultimately, the lack of a consistent method for accurate early diagnosis results in $7 trillion to $7.9 trillion in avoidable annual health care expenditures [2]. Evidently, a massive issue lies within the medical community when it comes to the efficient and effective diagnosis of the diseases of the central nervous system (CNS). For these reasons, it is clear that research must focus on developing a more accurate, non-subjective method of dementia diagnosis; one that uses machine learning to limit human error. This breakthrough would benefit American society by decreasing medical expenses, extending the lifespan of dementia patients, and giving individuals a greater chance of participating in cutting-edge clinical trials.

Dementia is a term that refers to the series of diseases that result in the gradual decline of cognitive function and physical viability, which commonly comes in the form of memory loss, altered personality, and a decrease in rational thinking [3]. Some of the most common forms of the ailment include AD, Parkinson's disease (PD), and Lewy Body Dementia. When protein fibrils, such as ß-amyloid (Aß) and alpha-synuclein, deposit to form synaptically obstructive, water-insoluble plaques, the functionality of the CNS is greatly inhibited. Proteotoxicity impedes upon neurotransmission at the synaptic membrane, as well as promotes the development of oxidative stress, which can be defined as the oxidation of vital cellular structures by excessive free radical populations [4]. This cellular damage ultimately results in the expression of the common symptoms that are characteristic of dementia by enhancing neurodegeneration through lipid peroxidation, proteo-oxidation, and the break-down of nucleic acids [5].

In conjunction with an observable decline in neurocognitive function, physicians use a variety of scans and imaging techniques to attempt to pinpoint abnormalities in the brain's of possible dementia patients. Within this study, magnetic resonance imaging (MRI) was used to develop mathematical models capable of identifying AD. MRI scans utilize a magnetic field, radio wave pulses, and computing power that can currently detect abnormalities associated with neurodegenerative diseases, such as AD and mild cognitive impairment (MCI) [6].

In order to eliminate the human error associated with neuro-radiology, computer science can be utilized to develop computer-aided diagnostic (CAD) systems capable of automatically analyzing neuroimagery. CAD systems are founded on a basis of machine learning, a technique used to teach computers how to conduct tasks in a more efficient or successful manner than could be done by humans, without explicit programming [7]. Machine learning significantly outperforms other scientific diagnostic methods due to its rapidity, autonomy, and ability to locate regions of interest (ROI) within input imagery without manual instruction from a software engineer [8]. Within this study, two machine learning algorithms, known as the 2-dimensional convolutional neural network

(2-D CNN) and the Long-Short Term Memory (LSTM) Model, were developed for dementia classification based on MRI brain scans and demographic data.

Within a 2-D CNN, 2 separate neural network sections are conjoined in order to progressively analyze and extract information from image data. The first of which is the convolutional section, which is further subdivided into two mathematical operations: the convolutional unit and the max pooling unit. In a convolutional operation, a 2-D kernel matrix traverses an image pixel value matrix and performs calculations at increments of a predefined size, known as stride. At each of these increments, the kernel matrix multiplies each of its elements by corresponding elements of the image that the kernel currently intersects. These products are then summed together and inputted into a convolved feature [9]. A max pooling operation is similar in that it involves the traversing of input imagery by a kernel, except a max pooling kernel contains no information. Instead, it simply inputs the maximum pixel value within the region it currently intersects into an output matrix, thus, reducing data dimensionality and eliminating noise [10].

Similarly to the 2-D CNN, the LSTM neural network is also utilized for the classification of input data. However, while the CNN is a feed-forward mathematical model, meaning that data flow occurs unidirectionally (from the input layer to the output layer), the LSTM is recurrent. This means that the model possesses feedback connections that allow information to persist within the model in similar fashion to how thoughts and memories persist within the human brain. For this reason, the model is adept at identifying patterns within sequential and general scalar data, making the LSTM a prime candidate for the demographic classification of dementia subjects.

At its core, this specialized RNN is composed of basic units known as cell modules. These blocks of computation are connected within a chain-like structure where data flow occurs with complete freedom of direction. These individual cells are defined by their cell state (Ct), which, in reality, refers to a vector containing the significant information that has thus far been learned by the algorithm. The information defining the cell state is regulated by a series of gates that consist of a sigmoid function, which outputs a value between zero and one that states the extent to which data will be let through the junction.

The gates utilized by the LSTM algorithm include the forget, input, and output gates. The purpose of the forget gate is to examine the previous unit's cell state, and then to determine which data should be omitted or retained through the use of a sigmoid neural network layer. This operation is essential to the model's ability to perform both long-term and short-term memory, as it allows for the network to limit the extent to which it delves into previously encountered information to perform predictions. The input gate and a hyperbolic tangent function (tanh) layer then determine which values should be updated, and how this update should occur. Lastly, the LSTM cell has to determine what values within the current cell state it wants to output to its neighboring unit. Primarily, a sigmoid neural network layer is used to determine which elements with Ct are to be sent to the next cell. Then, the cell state is inputted into the tanh function in order to normalize the vector's values within the range of -1 and 1, and multiplied by the sigmoid-output to create the finalized output vector Ht [11].

Finally, for both networks, the outputs of the recurrent and convolutional operations are fed into dense layers, which consist of interconnected information containing objects, known as neurons [12]. These layers progressively augment inputted data based on calculations with trainable parameters, and ultimately culminate in an output layer containing a number of neurons equal to the number of possible classes. The softmax function is applied to each object, and the output of each calculation represents the probability of the data belonging to each class, thus representing the model's prediction. Initially, the model's predicted values will inevitably be incorrect. Therefore, in order to train the model, the network is fed data, the error, or loss, of its outputs are calculated, and the weights and biases are adjusted accordingly. This process is known as back propagation and is achieved via gradient descent. This algorithm allows the model to compute which connections need to be adjusted, and to what extent the values need to be changed [13].

Ieracitano et al (2018) [14] engineered a 2-dimensional convolutional neural network that was trained to differentiate AD, MCI, and healthy patients based on non-invasive electroencephalography (EEG) readings. A CNN was developed that possessed one convolutional unit, consisting of one kernel convolution and one max-pooling layer. Following the convolutional portion of the network, the final convolved feature was flattened and inputted into a single fully-connected layer, which outputted to the classification, or softmax layer. This set of neurons was able to perform both binary (demented vs non-demented) categorization, as well as ternary (AD vs. MCI vs. Healthy). In terms of results, the average accuracies of binary and ternary classification were 84.62% and 83.33%, respectively. Evidently, this study demonstrates the potential for the use of machine learning to diagnose neurological diseases, as well as suggests the need for research that examines the use of different types of input data for illness-identification. This is because the complexity and time-consuming nature of the preprocessing pipeline used for the analysis of EEG recordings proves problematic. The steps included artefactual pattern removal, epoch segmentation, continuous wavelet transform (CWT) calculation, bispectrum analysis, CWT feature extraction, bispectrum feature extraction, and feature vector preparation. More importantly, pattern removal and epoch segmentation had to be carried out manually by a medical professional, greatly decreasing the clinical practicality and time-efficiency of the model.

Roy et al (2019) [15] attempted to detect Alzheimer's disease in MRI images through the use of a simple CNN architecture. The researchers used the OASIS-1 public dataset in order to train and test their model, and specifically used cross-sectional MRIs. The NIFTI neurological imaging files were converted into the 2-dimensional PNG format and were grayscaled to reduce the dimensionality of the data. Additionally, Roy et al (2019) used each of the imaging viewpoints (coronal, sagittal, transverse, and masked transverse) to create a well-rounded model that could make classifications based on different views of the brain. The neural network model was implemented using python and the keras tensorflow library. Post training, it was found that the neural network was able to accurately classify 80% of the scans. Clearly, this study posits the need for the production of dementia-diagnostic CAD systems that can achieve greater accuracy so that AI solutions to manual diagnosis can be implemented in the clinical sector.

Korolev et al (2017) [8] utilized the previously developed VoxNET and ResNET neural network models in order to classify demented and non-demented individuals using a 3-dimensional CNN model. The VoxNET architecture makes use of volumetric convolutional and deconvolutional operations with batchnorm and dropout modules for overfit-prevention and normalization. The ResNET algorithm was developed in a 2015 machine learning competition, and is defined by special residual blocks, or skip connections, that allow data to flow from initial to final layers without encountering hidden nodes. When performing the binary classification of AD vs Healthy Control, the VoxNET and ResNET models achieved .79 $\pm$ .08 and .80 $\pm$ .07, respectively. Evidently, this lack of performance suggests the need for unique neural network models that can be optimized and implemented for the sole purpose of AD classification. Instead of repurposing algorithms produced for different computer vision tasks, newly-developed models can be better trained to extract information from neurological imagery. Additionally, these results posit the examination of the ability of data augmentation to increase model accuracy by providing a more thorough training dataset to the network. This could be done through the use of techniques such as random rotation, affine transformation, and gamma correction, and would hopefully allow for a vast improvement in performance.

The purpose of this study was to construct and implement two unique AI algorithms, the 2-D CNN and the recurrent LSTM, using autonomous and efficient data augmentation and preprocessing pipelines, as well as to examine their ability to aptly classify demented and non-demented individuals based on MRI scans and demographic data. It was hypothesized that both of the models would achieve greater than or equal to 85% accuracy in accomplishing their respective tasks.. The null hypothesis was that both models would fail to reach the benchmark of 85% accuracy.

2. Methodology

Executive Summary

The following methodology was utilized to develop, implement, and test two machine learning algorithms, namely the CNN and LSTM, for the purpose of dementia classification. The OASIS-3 neuroimaging dataset was used to train the convolutional computer vision algorithm, while the OASIS-2 demographic dataset was used for the LSTM. Preprocessing consisted of creating a data wrangling pipeline for the purpose of selecting viable MRIs, extracting skeletal features, and slicing along the sagittal axis. In order to expand upon the provided MRI volumes, data augmentation was utilized. The pipeline consisted of random rotation, gamma adjustment, gaussian blur, and scaling, allowing for a 10,800-fold increase in data quantity. After implementation in python, the cross-validation technique was used to assess model capability by creating 6 independent accuracy, precision, and recall scores for both networks, which could then be statistically analyzed using various IBM SPSS v26 t-Tests. Model metric values were maximized via the manual optimization of hyperparameters, such as the type of function utilized to record model loss, as well as the rate at which the optimization function applied parameter adjustments.

2.1. Preparing Data

In modern data science, the convolutional neural network is greatly popular due to the model's ability to handle significant variability within its input data, thus requiring minimal preprocessing. However, several augmentations must still be applied to the original dataset in preparation for training. The data preprocessing used during this study followed a similar pipeline to that utilized by Zhang et al (2018) [16], with the two differences being that histogram stretching (HS) and the Functional Magnetic Resonance Imaging of The Brain (FMRIB) Linear Image Registration Tool (FLIRT) were not used. This is due to the fact that HS is a process that allows for the normalization of two MRI scans taken using different machinery, which is a situation that did not arise during this study. FLIRT is a program that aligns images along an atlas within MNI space, and was not used because the program only performs very small adjustments upon the input data and commonly malfunctions if the reference scan is not of high enough quality [17].

Prior to scan preprocessing, 40 healthy and 40 AD-affected patients were selected for use within this study. Healthy and AD-affected individuals have been defined as having clinical dementia ratings (CDR) of 0 and 1-3, respectively. CDR is defined as a professionally-accepted quantification of dementia severity determined via an interview between a physician and a patient or an informant on the patient's behalf (Hughes et al, 1982) [18]. Initially, the brain extraction tool (BET), a CNN machine learning model developed by Smith (2002) [19], was used to remove all features within the MRI except the brain. This eliminated the variability of the unnecessary facial features and other structures that can be seen within the scan. Secondly, a Guassian convolution kernel was used to blur the input data and lessen resolution. This prevented the CNN model from overfitting during training, allowing the network to abstractualize based on the inputted MRI scans. Lastly, 2-dimensional slices of the NIFTI files were taken at Z = -22 millimeters along the sagittal plane. This depth is advantageous for use in Alzheimer's classification studies due to the clear visibility of the hippocampus [20].

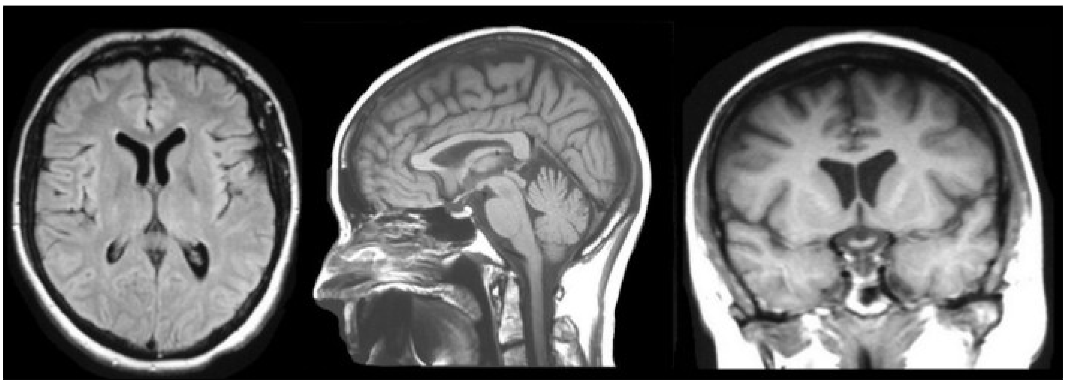

Figure 1. **Transverse (axial), Sagittal, and Coronal MRI axes [left to right] [21].**

For the LSTM neural network, the OASIS-2 clinical demographic dataset was used for model fitting. In order to preprocess the raw set, unnecessary subject classes, such as handedness and MR delay, were voided, and null values were replaced with the mean of a given characteristic. For instance, each null value within the socioeconomic status (SES) column was replaced with 2.5, which represents the mean SES of the subjects examined within the imaging initiative. Ultimately, the classes of gender, age, education level, SES, Mini-Mental State Exam (MMSE) score, CDR score, estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and axis scaling factor (ASF) were utilized for study. The processed data was then imported into python and vectorized by subject to create independent sequences for each individual [22].

2.2. Data Augmentation

In order to sufficiently train a deep learning, convolutional neural network model, an immensely large quantity of input data was necessary. For this reason, the 80 T1-weighted MRI scans selected for use within this study were augmented using various techniques to create a dataset containing more than 100,000 images. The data augmentation (DA) pipeline used to enhance the selected dataset size followed after the methods outlined by Wang et al (2018) [20]. Initially, each of the 80 MRI scans selected were rotated from -15 degrees to 15 degrees, by increments of one unit, in order to yield a 30-fold increase in the size of the training set. Secondly, gamma correction, a mathematical process used to adjust the brightness, or gamma, of an image, was used to further expand the available data. The gamma value of each MRI was set within the range of 0.7 to 1.3, following increments of 0.02, in order to create 30 new scans from each individual image. The third method of augmentation was to inject each image with different levels of noise through the use of a gaussian convolution operation. The second to last DA technique was to scale each of the images by a factor of 0.7 to 1.3, with increments of 0.02 between each value. Lastly, 30 random transformations were applied to each image that maintain parallelism and angle measure. These various methods allowed for the creation of a dataset 10,800-times larger than the set that was initially defined.

2.3. Model Development

2.3.1. CNN Algorithm

The convolutional model developed within this study consisted of a combination of 2-D convolutions, Max-Pooling operations, and dense layers, in addition to rectified linear activation functions. The categorical cross-entropy loss function was utilized to assess model error, while the Adam optimizer was used for backpropagation with a dynamic learning rate as defined by a scheduler callback mechanism. The model began with a series of convolutional operations, followed by a single pooling layer, which then fed into a dense layer region. The first two convolutional units utilized kernels of size 3, while the second pair of operations downsized to a 2x2 convolution. Additionally, both 2-D and 1-D dropout units were used to prevent overfitting during training [23]. The entirety of the model was implemented using the TensorFlow library and was trained through the use of the model.fit() package. The model was trained for 70 epochs, with a batch size of 32 scans. This means that the model was trained on the entire dataset 70 times, and only adjusted parameters after viewing 32 unique pieces of data.

2.3.2. LSTM Algorithm

The developed LSTM recurrent neural network consisted of an initial embedding layer, which had the purpose of converting the demographic sequential data into low-length matrices that could be decoded using an automatically defined embedding dictionary matrix. This technique is able to greatly reduce the required memory for model training, as it is abundantly more efficient than the alternative of one-hot encoding. The transformed input data was then pushed through an LSTM neural network layer consisting of 260 independent cells. The output of this network was then inputted into a series of fully-connected layers in order to transform the value into two numbers representing the probability of the data belonging to either the dementia-affected or healthy control class. Two dropout units (p=0.4) after the LSTM layer and the first dense node. The LSTM model was trained for 30 epochs, using a batch size of 16 subjects [24].

2.4. Model Training & Testing

The OASIS Brains MRI data was divided into six, equally sized and class-balanced segments in order to assist in the validation of the success and accuracy of the CNN model. This division was carried out using stratified cross-validation. K-Fold Cross Validation (KFCV) is a statistical analysis technique that divides the entire dataset into k folds, or sections, of equal size. Then, k-1 folds are used to train the network, while the remaining fold is used to assess the model's accuracy. Finally, this process is repeated over C(k, k-1) iterations with the section used for testing being changed each time, resulting in the production of k accuracy values that can be averaged to determine the final averaged accuracy score of the networks [25]. The LSTM model was trained on six, equally sized and class-balanced data segments as well, using the previously described validation technique. Each model was trained on a unique combination of input data for 6 iterations, with an independent accuracy value being generated each time. Alterations to model parameters were conducted by the Adaptive Moment Estimation (Adam) algorithm. This optimization tool has been chosen due to the fact that it combines the best properties of SGD with those of RMSProp - another popular optimizer that is used in many neural network models [26].

2.5. Optimization

In an attempt to create the most accurate CNN and LSTM models possible for dementia classification, the previously described models were tuned through the adjustment of hyperparameters. These modifications were done in an attempt to decrease the loss of the network and increase accuracy. The parameters that were changed consist of the kernel sizes and stride lengths of each of the convolutional and MaxPooling units, as well as the number of neurons outputted by the first dense layer. Additionally, the activation function used by the neurons within the first dense layer was changed if the alteration increased the model's accuracy. The activation functions that were tested included the basic rectified linear function, the leaky rectified linear function, the sigmoid function, and the binary step function. Lastly, the learning rate used by the Adam optimizer was adjusted during training in order to discover the parameter's ideal value, or the value that decreases loss at the highest rate [27].

2.6. Equipment Used

2.6.1. OASIS Brains Dataset

Primarily, the Open-access Series of Imaging Studies (OASIS) 3 data set was accessed via the XNAT central database, which provides a platform for the distribution and storage of scientific imagery. The OASIS 3 dataset is a publicly accessible repository of neuroimaging data that was provided by the principal investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly. Using the database service, a comma separated value (.csv) file was created containing the patient IDs associated with the MRI files desired for study, which was then prepared for use through the addition of UNIX formatting. As a result of the data's immense size, an XNAT API-based open source software[1] was used to automate the downloading process. The program receives the Patient ID .CSV file, an XNAT central username, as well as the desired output directory and scan types, and then pulls files directly from the repository's servers and stores them within the local hard drive. Specifically, the script was used to download T1-weighted neurological scans, which were stored in the three-dimensional NIFTI file format. Additionally, the OASIS-2 demographic dataset was downloaded as an Excel spreadsheet from the OASIS Brains official website.

2.6.2. Jupyter Notebooks

The majority of data preprocessing and augmentation was carried out using the Jupyter Notebooks programming environment. This integrated development environment (IDE) allows for the simple subdivision of processes and scripts between separate "cells" of code that can be run independently, thus saving time and computational expense. Additionally, Jupyter Notebooks allows for the integration of statistics and data science into python scripting by offering native support of data science packages, such as anaconda, which is an essential feature for machine learning engineering [28].

2.6.3. Microsoft Visual Studio Code (VScode)

The majority of model development, implementation, and analysis was conducted within the VScode environment. This programming utility was used for the training and testing of machine learning models due to the fact that it is better able to handle RAM-intensive processes than Jupyter Notebooks. Since the latter application is run on a web browser, built-in memory thresholds limit the computations that can be performed by the

[1] *The software repository can be accessed via http://github.com/NrgXnat/oasis-scripts. The download_oasis_scans.sh script was utilized*

algorithms. Since VScode is a local IDE, it was able to properly carry out standard AI operations.

2.7. Data Analysis

The metrics of accuracy, recall, and precision score were used to assess the ability of both CNN complexes to perform precise classifications upon out of sample T1-weighted MRI scans, as seen in Equations 1-3.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Number\ of\ Predictions} \tag{1}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2}$$

$$Precision = \frac{True\ Positives}{All\ Predicted\ Positives} \tag{3}$$

Additionally, the loss function outputs and accuracies associated with each model were monitored and graphed over time during training using the python package MatPlotLib. Independent samples T-tests ($p<0.05$) were used to test for statistically significant differences in the mean accuracy, precision, and recall metrics of both models. Lastly, one-sample T-tests ($p<0.05$) were used to test for significance between model accuracies and the hypothesized value of 85%, as well as test for significance between the CNN model's accuracy and numerous estimates of clinical diagnostic accuracy.

2.8. Controlled Variables

To allow for the fair comparison of model accuracies, numerous methods and variables were controlled. For instance, neither of the model's were trained on any data that was used for training and an equal number of cross-validation folds were used to validate each model. This means that the data used to assess the models was completely out of sample, as to prevent the over-estimation of accuracy, and that each model was tested on an amount of data in equal proportions to their respective dataset sizes. This is important due to the fact that a model tested on a smaller proportion of its total dataset would return more variable accuracy metrics, due to the fact that the standard deviation of the sampling distribution of accuracy increases as sample size decreases [29].

3. Results

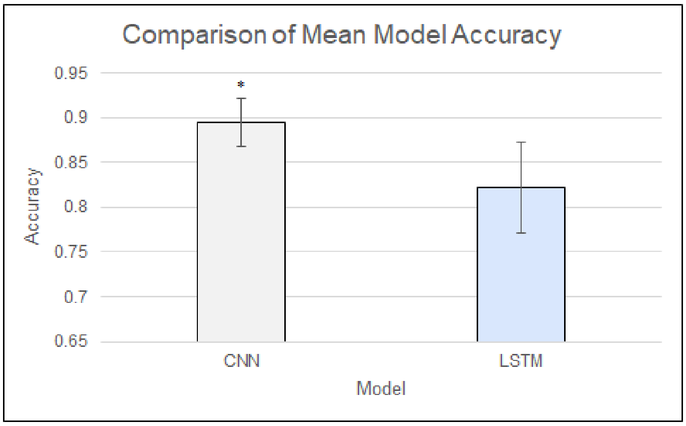

Figure 2. **Comparison of LSTM and CNN model accuracies over 6 cross validation iterations.**

Within Fig. 2, a direct comparison of the mean model accuracies of both the CNN and LSTM algorithm are displayed. The x-axis is defined by the independent variable, model type, while the y-axis represents mean accuracy, defined as the ratio between correct predictions and total predictions. Each group consisted of 1 neural network model, and 6 independent accuracy scores were collected by exposing the model to 6 unique sets of out of sample data. The CNN model developed within this study was able to achieve a mean accuracy of approximately 0.895 ± 0.027, or $89.5\% \pm 2.7\%$, over 6 independent trials. Similarly, the LSTM neural network returned a mean accuracy metric of 0.823 ± 0.051, or $82.3\% \pm 5.1\%$. Using an independent samples t-Test ($p<0.05$), a p-value of 0.016 was returned denoting a statistically significant difference between the sample means. This statistical analysis method uses the standard t-distribution to model the sampling distributions of two given populations by factoring in sample sizes and sample means. Then, a p-value is determined which denotes the area of the overlap of the two distributions. The p-value of 0.016 suggests that the convolutional network was better able to autonomously classify demented and non-demented individuals with clear significance. Additionally, the standard deviation of the LSTM accuracy metrics was much greater than that of the CNN. This suggests that the LSTM obtained a lesser-quality fit to the input data, as its ability to form correct predictions was greatly impacted by which set of out-of-sample data it received. This may mean, for example, that the model was able to recognize a pattern abundant in one cross-validation fold, resulting in an accuracy score much greater than the mean, while it may have been unable to comprehend a trend abundant in another fold, resulting in abnormally poor performance.

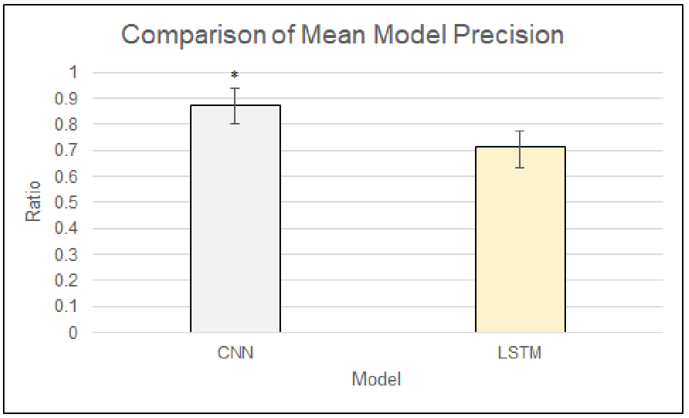

Figure 3. **Comparison of LSTM and CNN model precision scores over 6 cross validation iterations.**

Within Fig. 3, a comparison of mean algorithm precision over 6 trials of cross validation is displayed. The x-axis denotes model type, while the y-axis represents the unitless precision ratio. Each sample within each trial for the two models consisted of one accuracy score. The CNN model was able to achieve a mean precision score of 0.873 ± 0.067, or $87.3\% \pm 6.7\%$, while the LSTM network returned a precision of $0.717 \pm .059$, or $71.7\% \pm 5.9\%$. An independent samples t-Test ($p<0.05$) was conducted to examine statistical significance between the two means, and a p-value of 0.002 was returned. This suggests that the CNN model is significantly more precise than its recurrent counterpart. The distributions of the metric samples for both models possessed similar standard deviations.

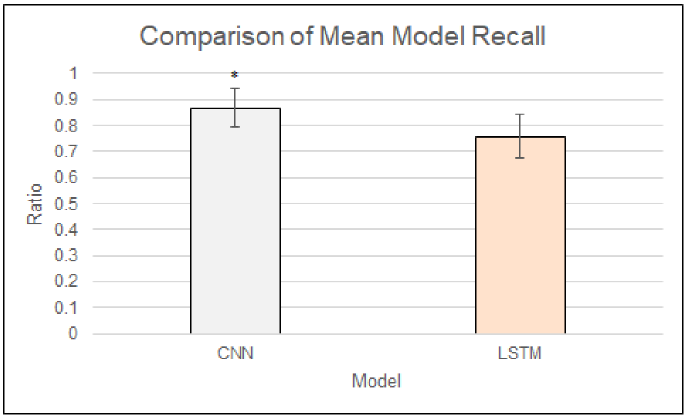

Figure 4. **Comparison of LSTM and CNN model recall scores over 6 cross validation iterations.**

Fig. 4 denotes the mean recall metrics of the LSTM and CNN models over 6 independent trials of testing. The x-axis represents the two different model types, while the y-axis represents the recall ratio. The CNN model achieved a mean score of 0.868 ± 0.072, or $86.8\% \pm 7.2\%$, while the LSTM returned a value of 0.758 ± 0.084, or $75.8\% \pm 8.4\%$. An independent samples t-Test ($p<0.05$) was utilized to assess for statistical significance between the population means, and a p-value of 0.035 was returned. This suggests that the CNN model was significantly more capable at identifying relevant data within the testing set. Once again, the standard deviations of the two samples were quite similar.

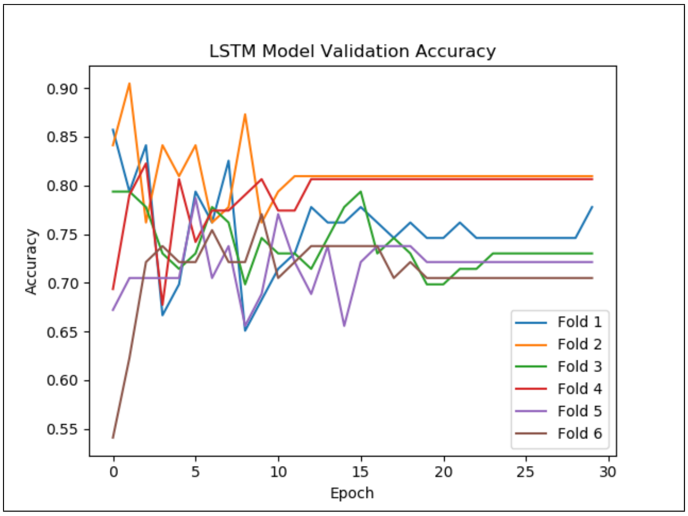

Figure 5. **LSTM Model validation accuracy over 30 epochs of testing.**

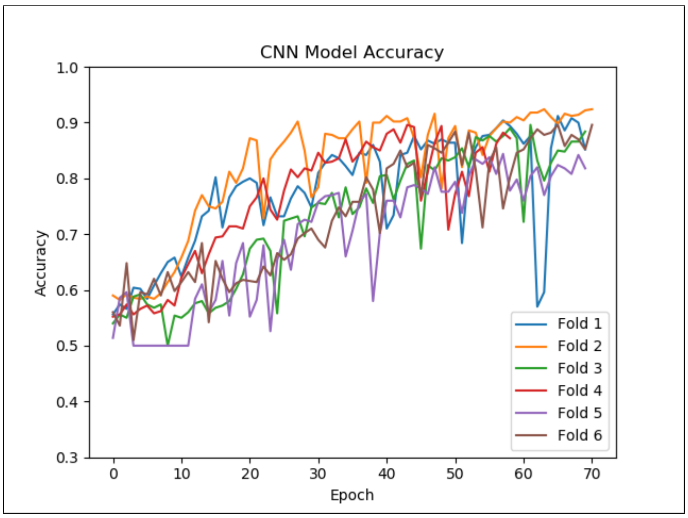

Figure 6. **CNN Model validation accuracy over 70 epochs of testing.**

Within figures 5 and 6, the x-axis represents epochs, or iterations, of testing, while the y-axis depicts model accuracy. Each of the separate colored lines represent independent cross-validation folds, or trials. In Fig. 5, LSTM accuracy seems to rapidly incline for the first 2-3 epochs of testing, at which point each fold varies sporadically for the succeeding 10-12 iterations. However, at roughly the 15th epoch, the accuracies cease their approach upon a fixed value and plateau. Within Fig. 6, the accuracy of the CNN for each validation fold seems to steadily incline throughout the duration of testing. In spite of increasing with acute variations, the graph exemplifies a strong, upward trend in accuracy for each of the trials. At no point do the accuracy values plateau, but rather each metric continually increases with each epoch, suggesting that an increase in training iterations may have improved model performance.

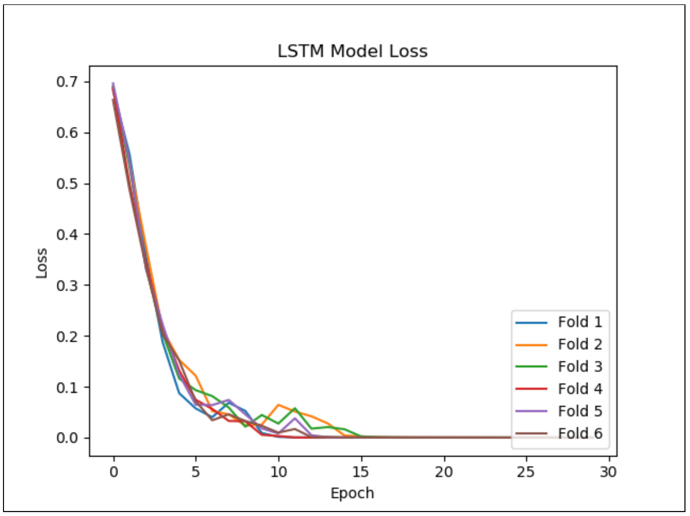

Figure 7. **LSTM model training loss over 30 training epochs.**

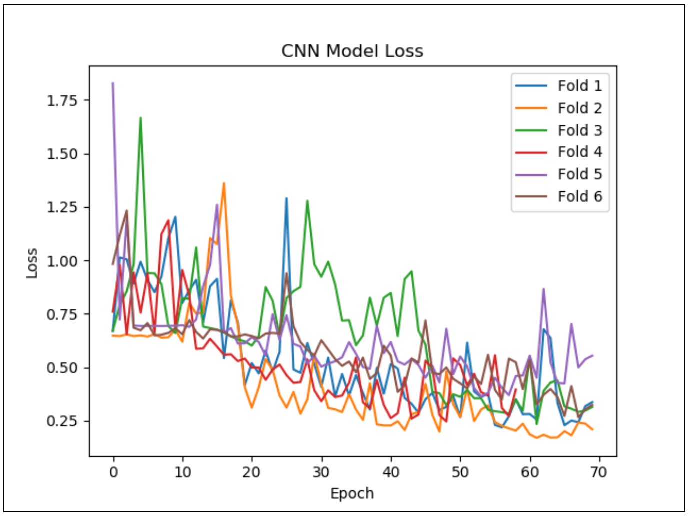

Figure 8. **CNN model training loss over 70 training epochs.**

Within Figures 7 and 8, the x-axis represents the model evaluation epoch, while the y-axis represents the unitless loss output value. Each line denotes an independent cross-validation fold. Fig. 7 depicts the loss of the LSTM model during training. For each fold, loss seems to decline at a rapid and consistent rate for the first 4 epochs of training, at which point the trials begin to diverge. Then, for the proceeding 10 epochs, training varies, until it is able to converge to approximately zero by the 15th iteration. In Fig. 8, the CNN model's loss over training is depicted. For the entirety of the training process, loss seems to generally trend downward, but in a highly variable fashion. While the loss of the model for each trial approaches 0 throughout training, the process is gradual, occurring with many seemingly random spikes and declines throughout.

4. Discussion

To restate, the purpose of this study was to determine the ability of LSTM and Convolutional Neural Network models to classify demented and non-demented individuals, based on both visual neuroimaging and scalar demographical data. As demonstrated by Fig. 6, the convolutional neural network was significantly more accurate than the recurrent model. Although accuracy is not an all-encompassing metric, it is highly relevant when determining the model's clinical practicality, which means that the CNN model would most likely be the most-applicable to real world utilization. More specifically, the CNN achieved an accuracy score of 89.5% ± 2.7%, which is greater than the 75.31% and 86.6% mean accuracies achieved by the CAD systems developed by Dua et al (2020) [30] and Alakwaa et al (2017) [31], respectively. For this reason, the CNN developed within this study may be a viable alternative for manual radiology. However, the recall metrics obtained by both the CNN and LSTM models may be of even greater importance when the purpose of the model is contextualized. When creating a system to be used for the diagnosis of a chronic illness, it is generally agreed upon that a false negative is much more hazardous than a wrongful positive diagnosis [32]. For this reason, recall can be used to assess a model's ability to identify and flag dementia-affected individuals, regardless of false positive rate. The CNN returned the high value of 86.8% ± 7.2%, while the LSTM underperformed at a mere 75.8% ± 8.4%. This value of .868 outperforms numerous estimates of clinical recall, such as 0.64 [33] and 0.49 [34]. This metric suggests that only 14% of dementia-affected subjects encountered by the algorithm are unnoticed, which, as previously stated, outperforms the clinical gold standard. In addition, the precision metric refers to the probability that a patient truly has dementia given a positive diagnosis. This is of great importance as well, bearing in mind that low precision results in improper medication and misguided treatment. The CNN achieved a value of 87.3% ± 6.7%, while the LSTM returned a metric of 71.7% ± 5.9%. Due to the fact that clinical precision is nearly impossible to estimate, the score achieved by the higher-performing model developed within this study can be compared to that of other CAD systems to assess its viability. The most-successful 3 models implemented by Roy et al (2019) [15] achieved testing precisions of 0.7143, 0.7334, and 0.8334, all much lower than the value returned by the proposed CNN. This suggests that the model possesses a profound ability to pinpoint dementia within MR imagery.

In order to test for statistically significant differences between mean model metrics, Independent Samples t-Tests ($p<0.05$) were run using IBM SPSS v26. This statistical analysis tool models sample distributions using the standard t distribution, and determines the area under area of the union of the two density curves, which denotes the probability that the two means are equal. When conducted for the CNN and LSTM mean accuracy, precision, and recall metrics, p-values of 0.016, 0.002, and 0.035 were returned, respectively. Since a threshold probability of 0.05 was set, and each of the p-values lie below 0.05, the differences between the mean metric values for each model were all statistically significant. Moreover, a 1-tailed Single Sample t-Test ($p<0.05$) was utilized to determine whether the CNN's achieved accuracy score of 89.5% $\pm$ 2.7% was significantly greater than the benchmark value of 85% delineated in the alternate hypothesis. A p-value of 0.010 was returned, which denotes that the CNN achieved an accuracy score significantly greater than 85%. For the LSTM model, a 2-tailed Single Sample T-test ($p<0.05$) was run to determine whether the achieved accuracy value was statistically equitable to 85%. Since a p-value of 0.241 was returned, it can be concluded that the accuracy score of 82.3% $\pm$ 5.1% was not significantly different from 85%, meaning that the alternative hypothesis was supported since both models achieved greater than or equal to 85% accuracy.

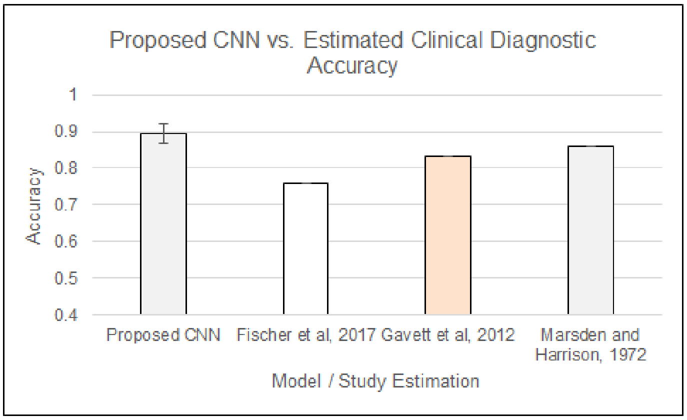

Figure 9. **CNN model accuracy compared to modern estimates of clinical diagnostic accuracy.**

In terms of relevant implications, the accuracy of both the CNN and LSTM models developed within this study exceeded numerous estimated clinical diagnosis accuracies with statistical significance, as can be seen in Fig. 9. 2-tailed Single Sample t-Tests ($p<0.05$) were utilized to determine statistical significance between the proposed CNN and the researched estimations. These values were obtained via the analysis of thorough dementia-prevalence databases, such as National Alzheimer's Coordinating Center Database, and can, thus, be assumed to be the most accurate assessments of accuracy available. This demonstrates that the proposed CNN model not only achieves greater accuracy, precision, and recall metrics than contemporary CAD systems, but also significantly outperforms current gold-standard dementia diagnostic practices. For this reason, it can be concluded that the incorporation of the developed CNN model into the process of AD screening would undoubtedly decrease misdiagnosis and the utilization of inappropriate treatment methods. Roughly 40,000 to 80,000 deaths occur as a direct result of medical misdiagnosis each year, with an additional 80,000-160,000 being seriously harmed [35]. For this reason, it is imperative that modern methods of dementia diagnosis are accepted in order to save lives and limit unnecessary and preventable injury.

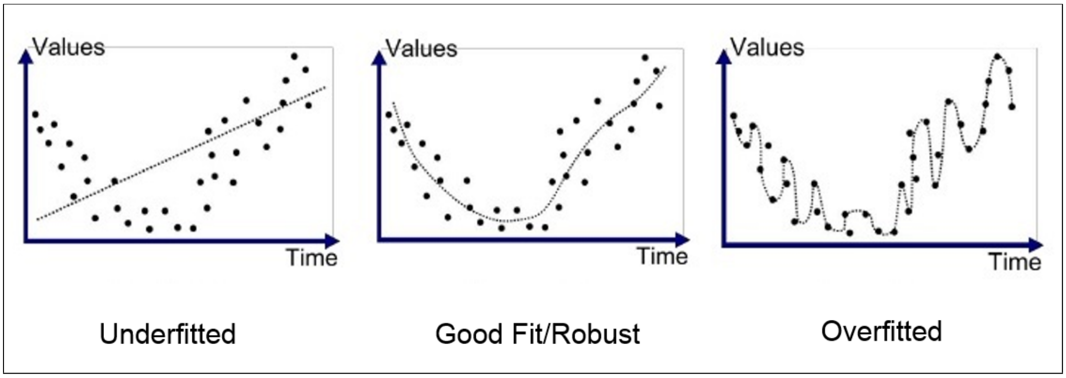

Figure 10. **Visual depiction of model fitting levels [36].**

Within this study, the superior performance of the CNN was due to the superior depth of the convolutional model [37]. This neural network possessed a total of 977,986 tunable parameters that could all be adjusted in order to contribute to the development of a successful computer aided diagnosis model. This large potential for tuning compensated for the vast scope of the input data, consisting of 8,000 MRI scans of size 19,000 square pixels. Contrastingly, the LSTM algorithm consisted of a lesser 667,854 tunable parameters and only 373 input sequences, resulting in overfitting in spite of protections such as dropouts. Overfitting refers to when a machine learning algorithm fits too specifically to a set of training data, making it unable to abstractualize and make correct predictions on out of sample data [38]. The model's overfit is supported by figures 5 and 7, as there is a weak correlation between model training loss and validation accuracy. This suggests that the model fit far too specifically to the training dataset, which hindered its ability to perform out of sample classification. due to the fact that and this occurrence resulted from the difficulty of deciphering general delineations between dementia-affected and healthy brains from such a small sample size. As a result, the CNN was able to be significantly more accurate.

Additionally, as seen in figures 7 and 8, the optimal quantity of training iterations for the LSTM and CNN models appeared to be 16 and 70 epochs, respectively. Once again, the low value for the LSTM model can be attributed to overfitting. In spite of utilizing two dropout units with large probability parameters, training was unable to be prolonged due to a low quantity of available data. The CNN model, however, experienced gradual increasing accuracy throughout the entirety of training. This demonstrates that the algorithm achieved a good fit, and could have possibly achieved greater metrics if left to train for a longer period of time. Moreover, necessary training length is proportional to model complexity, as more iterations are required to tune and adjust a greater quantity of parameters. For this reason, it is clear that the CNN necessitated a greater training period due to its possession of more parameters.

Lastly, this study was limited by several factors — the first of which being a lack of available computer memory. This directly hindered progress by not allowing for the use of the grid-searching technique in order to properly tune model hyperparameters, such as learning rate and batch size. The operation simply required too much random access memory to be carried out, which meant that basic trial and error and comparisons to literature had to be used to find optimal values. Secondly, the process of model training was held-back greatly by a lack of available Alzheimer's disease data. The two major datasets available for dementia computer vision studies are OASIS and ADNI. Since the ADNI set is only available for collegiate or doctorate level research, the only possibility remaining was OASIS. Although this imaging initiative provides high quality neuroimaging, the number of participants is quite small, resulting in the necessity of data augmentation to recompense. Lastly, the small size of the dataset is confounded by its incompleteness and possession of null values and unfilled data indexes in the demographic set. These gaps need to be replaced by mean values, or voided entirely, which reduces the purity of the data and may contribute to a lack of model performance. However, in spite of the limitations encountered, two viable machine learning CAD systems were able to be developed that, if utilized in clinical practice, could hopefully save thousands of lives from the troubles of misdiagnosis.

5. Conclusion

This study consisted of the implementation and assessment of two machine learning algorithms for AD diagnosis: the CNN and the LSTM. These two models used the differing deep learning approaches of convolution and recurrence in order to extract information from MRI brain volumes and demographic sequences. Both datasets underwent preprocessing in order to eliminate unusable elements, while a data-augmentation pipeline was only applied to the neuroimagery to allow for lengthened training duration. This data augmentation pipeline consisted of random rotation, scaling, gamma correction, and gaussian blur, yielding a 10,800-fold increase in dataset size. The raw MRI brain volumes were originally selected through the use of a data wrangling script that cross-referenced OASIS spreadsheet data in order to identify viable MRI sessions for study. Then, the BET was used to eliminate skeletal features, and numpy was used to slice the volumes along the sagittal axis. For the LSTM model, the demographic data was processed through the substitution of null values with category means and the elimination of unnecessary criteria. For the training of the final recurrent network, only the characteristics of gender, age, education level, SES, Mini-Mental State Exam (MMSE) score, CDR score, estimated total intracranial volume (eTIV), normalized whole brain volume (nWBV), and axis scaling factor (ASF) were utilized. Ultimately, the accuracy, precision, and recall metrics returned by the CNN and LSTM were 0.895 ± 0.027, 0.873 ± 0.067, and 0.868 ± 0.072, and 0.823 ± 0.051, 0.717 ± 0.059, and 0.758 ± 0.084, respectively. The CNN significantly outperformed the LSTM in each category, as well as numerous estimated clinical AD diagnostic accuracies. The superior performance of the CNN can most-likely be attributed to its superior depth and complexity, as well as the simpler class divisibility of the imaging dataset. Evidently, the use of the proposed CNN algorithm, in tandem with the professional insight of radiologists and physicians, could decrease the rampant amount of dementia misdiagnosis plaguing America. In addition, the alternate hypothesis was partially supported due to the fact that, as demonstrated by statistical analysis, both models achieved greater than or equal to 85% accuracy. However, the alternate hypothesis cannot be entirely supported due to the fact that CNN significantly outperformed the LSTM algorithm, which was not predicted prior to model implementation.

6. Future Studies

To further the developments of this study, additional machine learning algorithms could be developed for AD diagnosis using both demographic and imaging data, as well as auditory. More specifically, a ResNET-RNN hybrid or LSTM-CNN hybrid could be trained on spontaneous speech recordings, longitudinal demographic sequences, and MRI volumes in order to provide a more holistic assessment of subjects. The ResNET-RNN hybrid would offer a computationally efficient solution to the issue of AD diagnosis, focusing on numerical patterns to provide classifications, while the LSTM-CNN model would mainly utilize longitudinal imagery to sequentially analyze visual data. These networks would make use of the OASIS and ADNI datasets, as well as the ComParE and eGeMaps sets for acoustic features. The viability of dementia-classification via speech recordings has been demonstrated by Haider et al (2020) [39], where an Active Data Representation (ADR) model was able to achieve 77.44% accuracy when classifying out of sample data. This method of expanding input data scope would dilute inconsistencies in visual and demographic data, allowing for a stark decline in model error.

Acknowledgments

I would like to sincerely thank Mrs. Alison Huenger and the Manhasset Science Research community for giving me the aid and support necessary to conduct this study.

Conflict of Interest

The author of this article declares that he has no conflict of interest.

Human Studies/Informed Consent

No human studies were carried out by the author for this article.

Animal Studies

No animal studies were carried out by the author for this article.

References

[1] Greb, Erik (2013) Misdiagnosis of Alzheimer's Disease Is Linked to Less Severe Dementia Profile, *Neurology Reviews*: 21(10), 14.

[2] Alzheimer's Association (AA) (2018) How is Alzheimer's disease Diagnosed?: Why Get Checked?, https://www.alz.org/alzheimers-dementia/diagnosis/why-get-checked. Retrieved July 7, 2020. WEB

[3] National Institute on Aging (NIA) (2017) Health Information: What is Dementia? Symptoms, Types, and Diagnosis, https://www.nia.nih.gov/health/what-dementia-symptoms-types-and-diagnosis. Retrieved July 6, 2020. WEB

[4] Butterfield, Allan, Aaron Swomley, and Rukhsana Sultana (2013) Amyloid ß-Peptide (1–42)-Induced Oxidative Stress in Alzheimer Disease: Importance in Disease Pathogenesis and Progression, *Antioxidants and Redox Signaling*: 19(8), 823-835.

[5] Gella, Alejandro (2009) Oxidative Stress in Alzheimer's Disease, *Cell Adhesion and Migration*: 3(1), 88-93.

[6] Radiological Society of North America (RSNA) (2020) Alzheimer's Disease: How is Alzheimer's disease diagnosed and evaluated?, https://www.radiologyinfo.org/en/info.cfm?pg=alzheimers#:~:text=In%20the%20earl\y%20stages%20of,the%20temporal%20and%20parietal%20lobes). Retrieved July 9, 2020. WEB

[7] Brown, Sara (2021) Machine learning, explained, https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained. Retrieved April 25, 2021. WEB

[8] Korolev, Sergey, Amir Safiullin, Mikhail Belyaev, and Yulia Dodonova (2017) Residual and plain convolutional neural networks for 3D brain MRI classification, *IEEE 14th International Symposium on Biomedical Imaging*: 835-838.

[9] Snavely, Noah (2013) CS114 Section 6: Convolution, https://www.cs.cornell.edu/courses/cs1114/2013sp\/sections/S06_convolution.pdf. Retrieved July 31, 2021. WEB

[10] Layton, Oliver (2019) CS343: Neural Networks - Convolution and Max Pooling, https://cs.colby.edu/courses/F19/cs343/lectures/lecture11/Lecture11Slides.pdf. Retrieved April 25, 2021. WEB

[11] Olah, Christopher (2015) Understanding LSTM Networks, https://web.stanford.edu/class/cs379c/archive/20\18/class_messages_listing/content/Artificial_Neural_Network_Technology_Tutorials/OlahLSTM-NEURAL-NETWORK-TUTORIAL-15.pdf. Accessed April 5, 2021. WEB

[12] Di Caro, Gianni (2016) Deep Learning and Vision: Convolutional neural networks, http://www.cs.cmu.edu/~arielpro/15381f16/c_slides/781f16-18.pdf. Retrieved April 25, 2021. WEB

[13] Makin, J.G. (2006) Backpropagation, http://www.cs.cornell.edu/courses/cs5740/2016sp/resources/backprop.\pdf. Retrieved April 25, 2021. WEB

[14] Ieracitano, Cosimo, Nadia Mammone, Alessia Bramanti, Amir Hussain, Francesco C. Morabito (2018) A Convolutional Neural Network approach for Classification of Dementia Stages based on 2D-Spectral Representation of EEG recordings, *Neurocomputing*: 323, 96-107

[15] Roy, Sanjiban Sekhar, Raghav Sikaria, and Aarti Susan (2019) A deep learning based CNN approach on MRI

for Alzheimer's disease detection, *Intelligent Decision Technologies*: 13(4), 495-505

[16] Zhang, Yudong, Shui-hua Wang, Yuxiu Sui, Ming Yang, Bin Liu, Hong Cheng, Junding Sun, Wenjuan Jia, Preetha Phillips, and Juan Manuel Gorriz (2018) Multivariate Approach for Alzheimer's Disease Detection Using Stationary Wavelet Entropy and Predator-Prey Particle Swarm Optimization, *Journal of Alzheimer's Disease*: 65(3), 855-869

[17] Smith, Steve (2013) FLIRT: Contents: Introduction, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT. Retrieved August 16, 2020. WEB

[18] Hughes, CP, L Berg, WL Danziger, LA Coben, and RL Martin (1982) A new clinical scale for the staging of dementia, *The British Journal of Psychiatry: The Journal of Mental Science*: 140, 566-572

[19] Smith, Stephen (2002) Fast robust automated brain extraction, *Human Brain Mapping*: (17)3, 143-155

[20] Wang, Shui-hua, Preetha Phillips, Yuxiu Sui, Bin Liu, Ming Yang, Hong Cheng (2018) Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling, *Journal of Medical Systems*: (42)5, 85

[21] Preston, David (2006) Magnetic Resonance Imaging (MRI) of The Brain and Spine: Basics, https://casemed.case.edu/clerkships/neurology/Web%20Neurorad/MRI%20Basics.htm. Retrieved August 16, 2020. WEB

[22] Dua, Mohit, Drishti Makhija, P.Y.L. Manasa, and Prashant Mishra (2020) A CNN-RNN-LSTM Based Amalgamation for Alzheimer's Disease Detection, *Journal of Medical and Biological Engineering*: 40, 688-706

[23] Srivistava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov (2014) Dropout: A Simple Way To Prevent Neural Networks From Overfitting, *Journal of Machine Learning Research*: (15)1, 1929-1958

[24] Cui, Ruoxuan and Manhua Liu (2019) RNN-based longitudinal analysis for diagnosis of Alzheimer's disease, *Computerized Medical Imaging and Graphics*: 73, 1-10

[25] Schneider, Jeff (1997) Cross Validation, https://www.cs.cmu.edu/~schneide/tut5/node42.html. Retrieved April 25, 2021. WEB

[26] Dozat, Timothy (2015) Incorporating Nesterov Momentum into Adam, http://cs229.stanford.edu/proj2015/054_report.pdf. Retrieved April 25, 2021. WEB

[27] Winovich, Nick (2021) Deep Learning, https://www.math.purdue.edu/~nwinovic/deep_learning_optimization.html. Retrieved April 25, 2021. WEB

[28] Project Jupyter (2021) Project Jupyter — About Us, https://jupyter.org/about. Retrieved April 25, 2021. WEB

[29] Altman, Douglas, and Martin Bland (2005) Standard deviations and standard errors, *The British Medical Journal*: (331)7521, 903

[30] Dua, Mohit, Drishti Makhija, P.Y.L. Manasa, and Prashant Mishra (2020) A CNN-RNN-LSTM Based Amalgamation for Alzheimer's Disease Detection, *Journal of Medical and Biological Engineering*: 40, 688-706

[31] Alakwaa, Wafaa, Mohammad Nassef, and Amr Badr (2017) Lung Cancer Detection and Classification with 3D Convolutional Neural Network (3D-CNN), *International Journal of Advanced Computer Science and Applications*: (8)8, 409-417

[32] Health News Review (2020) Understanding medical tests: sensitivity, specificity, and positive predictive value, https://www.healthnewsreview.org/toolkit/tips-for-understanding-studies/understanding-medical-tests-sensitivity-specificity-and-positive-predictive-value/. Retrieved April 21, 2021. WEB

[33] Hoffman, J. M., K.A. Welsh-Bohmer, M Hanson, B Crain, C Hulette, N Earl, and R.E. Coleman (2000) FDG PET imaging in patients with pathologically verified dementia, *Journal of Nuclear Medicine*: (41)11, 1920-1928

[34] Jobst, Kim, Lin Barnetson, Basil Shepstone (2005) Accurate Prediction of Histologically Confirmed Alzheimer's Disease and the Differential Diagnosis of Dementia: The Use of NINCDS_ADRDA and DSM-III-R Criteria, SPECT, X-Ray CT, and ApoE4 in Medial Temporal Lobe Dementias, *International Psychogeriatrics*: (10)3, 271-302

[35] Newman-Toker, David, Adam Schaffer, Winnie Yu-Mode, Najlla Nassery, Ali Saber Tehrani, Gwendolyn Clemens, Zheyu Wang, Yuxin Zhu, Mehdi Fanai, and Dana Siegal (2019) Serious misdiagnosis-related harms in malpractice claims: The "Big Three" - vascular events, infections, and cancers, *Diagnosis*: (6)3, 227-240

[36] Bhande, Anup (2018) What is underfitting and overfitting in machine learning and how to deal with it, https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76. Retrieved July 31, 2021. WEB

[37] LeCun, Yann, Yoshua Bengio, and Geoffery Hinton (2015) Deep Learning, *Nature*: 521, 436-444

[38] Dietterich, Tom (1995) Overfitting and undercomputing in machine learning, *ACM Computer Surveys*: (27)3, 326-327

[39] Haider, Fasih, Saturnino Luz, Sofia de la Fuente, Davida Fromm, Brian MacWhinney (2020) Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge, *arXiv*: 3