To Recall or Not to Recall: That Is the Question for CNN-Based Alzheimer's Disease Detection Using RecallNET and MRI Imaging

Rahil Radia[1], Amy Scholl[2]

[1,2]*Department of Information Sciences and Technology, Pennsylvania State University*
State College, PA, USA
[1]rnr5129@psu.edu
[2]ats5585@psu.edu

***Abstract*— Alzheimer's Disease (AD) is one of the most prevalent neurodegenerative disorders, impacting approximately 1 in 9 individuals aged 65 and older. This condition is characterized by progressive memory loss, impaired spatial and visual perception, and challenges in daily task performance. Despite ranking as the sixth leading cause of death in the United States and being the most common form of dementia, clinical trials aiming to develop an effective treatment for Alzheimer's Disease have remained unsuccessful. Experts hypothesize that early detection could be pivotal in addressing this issue. Studying the disease in its early stages may offer insights that are challenging to obtain once significant neurodegenerative damage has occurred. Early diagnosis is also crucial, granting patients more time and flexibility to develop a proper care plan with their physicians. However, current methods of detecting AD are expensive, time-consuming, and invasive. Diagnoses occur after neurological exams and brain imaging, such as MRIs, are reviewed by radiologists. This presents a significant challenge in early detection, as options for assessing the disease's state are limited before clinical symptoms manifest. In this paper, we introduce an AI deep-learning approach to this problem, using computer vision techniques proposed by Fareed et al. to detect early signs of dementia in the brain from MRI scans. The dataset employed in this project is the same that Fareed et al. used in their project and comes from Kaggle. The dataset includes MRIs from various sources that are categorized into four classes of dementia: NonDemented, VeryMildDemented, MildDemented, and ModerateDemented. Additionally, we delve into a detailed discussion of the Fareed et al. paper and two others, outlining how and why researchers developed their early detection models. Our implementation of the RecallNET model establishes a state-of-the-art method for early AD detection, addressing performance issues associated with unbalanced datasets present in previous works. Our model achieves impressive metrics, including accuracy, AUC, F1-score, precision, and recall values of 99.29%, 99.99%, 99.27%, 99.29%, and 99.18%, respectively.**

***Keywords*— Alzheimer's Disease, mild cognitive impairment, computer vision, early detection, image classification, deep learning, imbalanced dataset, convolutional neural network, instance normalization,**

I. Introduction

Alzheimer's Disease (AD) is a profound medical condition demanding meticulous attention for diagnosis, treatment, and passionate care. As the most common form of dementia, AD involves brain atrophy, leading to memory loss, challenges in daily tasks, and behavioral changes. While rare, AD can affect young people, though these cases are treated as anomalies. Genetics play a role in AD development, yet adopting a healthy lifestyle can help mitigate the risk [1].

The urgency of early detection becomes apparent when considering the persistent struggles in developing a cure for AD. Thus far, clinical trials focusing on AD reversal have faced setbacks and failure, potentially due to late-stage detection. The CDC notes that changes in the brain can begin years before symptoms manifest [1], emphasizing the need for early AD detection. With an aging US population, more individuals receive AD diagnoses, increasing the demand for caregivers, who are more unattainable as the population grows.

Our investigation begins with three pivotal research papers. The first research paper [2] that we will review is considered our "parent paper" for the remainder of this document. We will be replicating the processes, models, and results of this methodology later in this paper. Our parent paper is titled, "ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans" and serves as the foundation. These researchers created a hand-crafted convolutional neural network (CNN) utilizing MRI images to classify brain scans into four different categories of AD. These categories are NonDemented, VeryMildDemented, MildDemented, and ModerateDemented. This research team was able to create a state-of-the-art approach to classifying a balanced MRI dataset into four different stages of AD using a CNN dedicated to detecting early stages of dementia.

The second research paper [3] is titled, "On the design of convolutional neural networks for automatic detection of Alzheimer's disease." This paper employed a 3D CNN model to detect and classify AD stages using MRI scans. They utilized instance normalization instead of batch normalization and techniques to avoid downsampling, innovations that we

plan to incorporate into our own methods. This paper acted as research for our own implementation techniques and a comparison to our parent paper.

The third research paper [4] that we will discuss is titled, "Enriching Neural Models with Targeted Features for Dementia Detection." This study focuses on speech patterns of patients with and without dementia through a manually transcribed interview. Their hybrid CNN-LSTM model detects AD utilizing targeted and implicitly-learned features and represents the new state-of-the-art technique for the dataset they used.

Our interest in harnessing deep learning and computer vision techniques for detecting early stages of AD stems from our shared commitment to future healthcare careers. We harbor a personal dedication to contribute to AD research and are completing this paper for a data science course, where we have done extensive research and reviewed a number of peers' research topics that include but are not limited to finance, sports, and predictive analytics. Overall, our primary interest lies in advancing the automatic detection of AD as current means of diagnosing AD and other dementias through MRIs is a costly and time consuming process that may not be feasible in areas with either high elderly populations or understaffed healthcare networks.

The crux of our research involves applying computer vision techniques to detect and classify AD, emphasizing early stages of the disease. Our goal is to create an automatic classification system for patients undergoing MRI scans, streamlining caregiving and contributing to AD reversal efforts. We aim to address challenges of unbalanced real-world data while minimizing false-negative classifications.

Our research into AD is fueled by the urgency of the global challenge and the potential for innovative solutions. Synthesizing insights from seminal papers, coupled with our dedication to healthcare, fuels our pursuit of creating meaningful contributions to automatic AD detection. We hope our research will have a positive impact on those affected by AD and the broader neurodegenerative disease research landscape.

Ultimately, we want to find an automatic way to classify AD using MRI scans so that diagnoses can be simpler and care can be streamlined. Additionally, we want this information to be more useful in clinical trials and attempt to aid in the reversal of AD. Throughout our research, we find ways to address the issues of unbalanced real-world data while accounting for the importance of false negative classifications.

Our proposed CNN model RecallNET achieves significant improvements on the baseline ADD-Net model in terms of accuracy, AUC, precision, recall, and F1-score on both balanced and unbalanced test sets. For the balanced test set, these values were 99.29%, 99.99%, 99.29%, 99.18%, and 99.27% of the aforementioned evaluation metrics. For the unbalanced datasets, these values were 97.66%, 99.93%, 97.66%, 07.66%, and 98.46%, of the above evaluation metrics respectively. Each of these metrics show that RecallNET outperforms the baseline ADD-Net model.

II. Literature Review

In this related works section, we will delve into each of the three research papers introduced in the "Introduction" section of the paper, discussing their significance in classifying and early detection of Alzheimer's Disease. We will provide a detailed examination of each of these papers, including their relevance to classifying and early detection of Alzheimer's Disease. Additionally, we will discuss the datasets used, methodologies, and results of each paper. Furthermore, we will explore how these methods will help in real-world clinical settings and potential limitations to their approach.

[2] *ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans*

Compiling MRI images for dementia datasets poses difficulty due to patient confidentiality terms associated with medical data. Commonly used datasets in the field include the Alzheimer's Disease Neuroimaging Initiative (ADNI) and Open Access Series of Imaging Studies (OASIS) datasets. However, the use of these datasets are restricted behind an application. There are also challenges associated with labeling the data due to ambiguity on where to draw distinctions between different classes. For this paper, we use the *Alzheimer's MRI Preprocessed Data* retrieved from Kaggle [5], which was similarly used by Fareed et al. in our parent paper [2]. This dataset was compiled using a variety of sources including websites, hospitals, and public repositories. It contains four classes of MRI scans that correspond to the severity and stage of dementia, such as NonDemented, VeryMildDemented, MildDemented, and ModerateDemented.

The implementation of ADD-Net [2] first addresses the class imbalance issues present in the data, as classes such as "NonDemented" contain 3200 cases whereas others like "ModerateDemented" contain only 84. This imbalance would result in a heavily biased model, with especially high false negative classification rates which is highly undesirable when it comes to a medical diagnostic model. To prevent this, the authors used the SMOTE-TOMEK algorithm, which combines under and over sampling techniques to rebalance each of the four classes to have 3200 images.

Following this, the MRI scans were passed into the ADD-Net convolutional neural network architecture. The architecture consists of 4 primary convolutional, ReLU activation, and average pooling layers, followed by 2 dense layers and a final softmax output layer. The kernel used for the convolutional layers is GlorotUniform which samples from a uniform distribution with bounds dependent on the dimensions of the input. The distribution was imported from the GlorotUniformV2 package from TensorFlow [6]. The

mechanism described and used in ADD-Net was specifically designed for the purposes of extracting signs of early onset dementia [2] and as such will serve as the foundation for our own project.

The model also performed exceptionally well on the test set, with a reported 98.63% accuracy and F1-score of 98.61%. However, the authors make note that the performance falls significantly when given an unbalanced dataset due to training of the model on SMOTE-TOMEK rebalanced data. This is important to us because in a real world setting, it's likely that the data passed into the model is prone to large amounts of class imbalances. In order to use the model reliably in a clinical setting it's crucial that a high order of precision is maintained even when presented with unbalanced data. In our own implementation, we plan to make changes that can mitigate the performance disparity between balanced and imbalanced datasets, while preserving the overall performance of the base ADD-Net model.

[3] *On the design of convolutional neural networks for automatic detection of Alzheimer's disease*

To do so we looked at various techniques proposed and used by other researchers in the field. One such paper by Sheng Liu et al. [3] proposes a similar model to ADD-Net [2] however, contains adjustments in the architecture which optimizes the model for classification on unbalanced datasets. Their model consists of a 3D CNN trained on T1-weighted structural MRI scans which are scans used for observing the fatty tissues and organs of the body such as the brain. These were obtained from the ADNI dataset. The 3D CNN contains convolutional, normalization, activation, and max-pooling layers which follows a similar pattern to the model in our parents paper.

However, the key differences involve the usage of instance normalization, and small kernel size in the first convolutional layer. These changes would help increase model accuracy and prevent early spatial downsampling which in turn would retain information that could be crucial in the early detection of AD [3]. The findings of their study came to the conclusion that the inclusion of instance normalization and small kernel size did have the intended effects of increasing accuracy on both the test set and an independent dataset as the model achieved an accuracy of 66.9% which was significantly higher than the base ResNet-18 model at 50.1%.

As such, we intend on implementing these changes in RecallNET since they address potential misclassification in smaller classes outside of rebalancing the training data. This paper also introduced the idea of incorporating data outside of MRI images with the addition of an age encoder. Since various demographic information such as age plays a large role in the likelihood of someone having AD, it makes sense to incorporate that information into the classification. However, their findings showed that the addition of the age encoder yielded minimal improvements in the model compared to the version without age encoding, at 66.9% and 68.2% accuracy respectively. With demographic information being difficult to find in publicly available datasets, we decided on forgoing including it in the RecallNET architecture.

[4] *Enriching Neural Models with Targeted Features for Dementia Detection*

The third paper we looked at, written by Palo and Parde [6] took a different approach to the problem than the first two papers did. Instead of using a CNN to analyze MRI scans, they instead used a hybrid CNN-LSTM (Long Short Term Memory) model on text-based data. While the task itself differs from our own, we chose to reference this project to broaden our scope of understanding on the problem of AD detection and the various techniques that could be implemented in order to achieve our goal.

This paper uses the DementiaBank dataset, consisting of text based interview data of dementia patients, including the Cookie Theft scenario [7]. The Cookie Theft scenario is an assessment tool that is commonly used in the evaluation of AD and other types of dementia. In it, patients are presented with an image of a child stealing cookies out of a jar while a woman is working in the kitchen, and asked a set of questions which prompt them to describe the image. Their responses are then reviewed for linguistic patterns that are commonly found in dementia patients.

The model architecture of this paper, as mentioned before, uses a hybrid CNN-LSTM model to classify cases into AD and CT (control). This involves both POS (parts of speech) and embedding data passed through a series of one dimensional convolutional neural networks, of which its outputs are passed into an LSTM and attention layer. This process is repeated a total of four times for the entirety of the model before being passed through a dense output layer. In addition to the deep learning techniques employed, they also hand selected a set of psycho-linguistic features that past studies observed to have had a high correlation to dementia such as age of acquisition, familiarity, sentiment, and encoded those results into the final output.

The models with class weights and attention layer achieved 88.2%, 93.05% accuracy and F1-score respectively. The authors of this paper also ran ablation studies to determine the effects of adding the class weights and an attention mechanism. They found that both improved the performance of the model, but the addition of the class weights increased false negatives in classification. Given the healthcare application, false negatives are particularly undesirable and the false negative rate should be minimized. So, while the accuracy is higher, the overall benefits of the class weights is unclear.

While the design decisions made for this project cannot be directly applied to our own, there are several key components that we will attempt to implement into our computer vision

based approach. The first of these being the addition of an attention layer, which would help mitigate some of the issues surrounding class imbalances in the original dataset. The second being the usage of ablation studies in order to determine which components of the model truly contribute to an improvement in performance in a desirable way. Looking at the model in a more modular way can also allow us to adapt to changes that we notice on both the balanced and unbalanced datasets we will be testing.

III. Methodology

The medical and healthcare field were revolutionized by the advent of Magnetic Resonance Imaging (MRI) technology in the 1970s. By using a combination of strong magnetic fields and radio frequencies to excite and align the rotational axes of protons, MRI sensors were able to detect and map soft tissue and organs in the body [8] with precision not feasible with earlier imaging technologies. Medical symptoms that once may have required surgical intervention to assess could now be surveyed via the non-invasive scans produced by these sensors.

MRIs have become a fundamental tool in medical diagnostics and are widely used by physicians to detect, diagnose, and monitor a variety of conditions. This has in particular been useful for studying the brain and its various neuro-degenerative ailments such as Alzheimer's Disease (AD), which we have elected to focus on in our research. While symptoms of memory loss and other declines in cognitive function have long been observed, pathologic causes outside of aging were not well understood by scientists until the early 20th century when AD was first reported [9].

Since then, the development of MRIs have made it far easier to non-intrusively detect biomarkers of mild cognitive impairment (MCI), i.e. atrophy in the temporal and parietal lobes that occur in the early stages of the disease [10]. While substantial progress in the subfields of medical imaging have been made, the need for a human expert to manually review each scan has remained a resource intensive task that grows ever more challenging in today's day in age. As the proportion of elderly population grows, the number of AD cases are also expected to rise drastically [11]. At the same time, hospital systems across the nation are experiencing strains of being understaffed [12]. Combined, these issues pose a problem for the future of the healthcare system and the quality of care received by AD patients.

The incorporation of modern computer vision techniques into the diagnostic pipeline offers a solution to this problem; it would significantly cut down on the time it takes to assess an image for early-signs of cognitive impairment (CI). Combined with the potential for earlier detection of brain atrophy, such a model would help allocate resources to the patients who are at the highest risk of developing AD. Many researchers have looked to CNN-based computer vision to achieve this goal, but a majority of proposed models have struggled to reach a suitable level of performance for the high rigor standards required in a clinical setting.

In addition, very few papers have addressed the data imbalance issue that plague model performance. Our goal in developing RecallNET is to take the baselines of the ADD-Net architecture, which was specifically designed for the early detection of AD [2] and incorporate additional machine learning techniques to mitigate performance issues when presented with unbalanced datasets.

A. Datasets

The data used for training and testing the RecallNET model, also used by Fareed et al. in the development of ADD-Net, were obtained from the "Alzheimer MRI Preprocessed Dataset" from Kaggle [2] [3]. This data contains longitudinal structural MRI scans of the brain separated into four classes which correspond to the severity and stage of AD, NonDemented, VeryMildDemented, MildDemented, and ModerateDemented. Attempts were made to use the ADNI dataset for independent testing but due to reasons that will be discussed in the Limitations section, we were unable to do so.

B. Baseline ADD-Net Model

To understand RecallNET, the Alzheimer's Disease Detection Network "ADD-Net" model [2] from which it was derived must first be outlined. To address the dataset imbalance issue that resulted in poor performing models in past works, the authors of the paper chose to resample the number of cases in each of the four classes to be equivalent using SMOTE-TOMEK, a random over and undersampling algorithm. The rebalanced dataset was then split into 60-20-20 train, validation, and test sets.

As for the ADD-Net model architecture, it consists of a series of four 2D-convolutional blocks which contain ReLU activation and average pooling layers. In addition to this, the number of filters in each subsequent block increases and the weights are initialized using a kernel following the GlorotUniform distribution. This design specifically allows for the extraction of features in the MRI that may indicate signs of early MCI. The result of the final convolution block is flattened and passed through two dense and dropout layers before a final SoftMax output layer which is used for multi-class classification. An additional "Callback" function is defined to reduce the model's learning rate when loss begins to plateau as a measure to prevent overfitting.

After training and testing the proposed ADD-Net model, the performance was evaluated against the baseline models of DenseNet169, InceptionResNetV2, and VGC19 on both SMOTE-TOMEK balanced and unbalanced datasets. When evaluating the models using rebalanced data, ADD-Net performed marginally better than baselines with a 98.63% accuracy and 97.99% AUC values. In a comparison of F1-scores, the proposed model also performs slightly better

than baselines with 98.6% accuracy. However, significant performance drop offs occurred when ADD-Net was tested using the unbalanced data, where accuracy, AUC, and F1-scores dropped to 66.1%, 69.19%, and 46.04% respectively. The performances of the baseline models were also affected but not to the same extent as that of ADD-Net. Among the evaluation metrics measured, we looked specifically at optimizing recall. This is because given the diagnostic nature of the model we believe it is better to overshoot classes and minimize the number of FN classifications to reduce potential harms. In ADD-Net, recall is 97% and 89% between balanced and unbalanced datasets respectively. In our implementation of RecallNET, we aim to bridge the gap between balanced and unbalanced dataset performances and by extension make the model's performance less dependent on the composition of the dataset it is applied to.

C. *Contributions*

1) *Original Plan for Implementation:*

To address the performance disparities present in the original ADD-Net model, we planned on introducing a number of potential changes to the overall architecture.

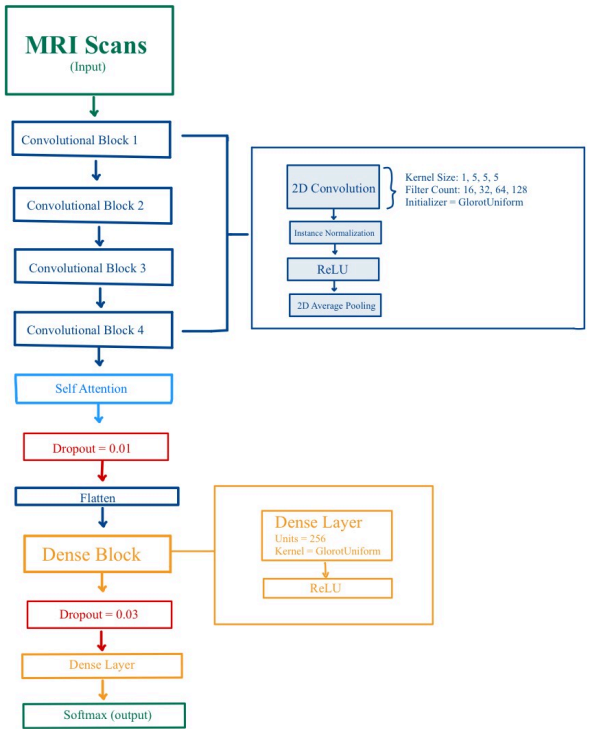


Fig. 1 Diagram of the Proposed RecallNET model for the early detection of Alzheimer's Disease.

RecallNET Convolutional Block

Convolutional Neural Networks (CNNs) are favored in computer vision applications as they allow for better detection of low-level local features present in the image. This is particularly useful when it comes to early detection of AD, as it allows the model to detect patterns not easily discernible in the original image. The design of the convolution block in RecallNET includes 2D convolution, instance normalization, ReLU activation, and 2D average pooling layers structure. The kernel size of the first block's convolution layer is 1 while the remaining three convolutional layers in the block are of size 5. This method aims to reduce early spatial downsampling to improve the accuracy of the overall model.

Self Attention Layer

Differing from ADD-Net [2], RecallNET integrates an attention mechanism to rebalance output weights from the final convolutional block similar to that of the hybrid CNN-LSTM model in our third research paper [4]. This mechanism is used in CNN based models to increase the receptive field of the model [13]. In our application, we will test to see if it leads to better model generalization and improved model accuracy on unbalanced test data.

Dropout Layer

The dropout layer is used to randomly turn nodes on and off to reduce network complexity. This acts as a measure against overfitting the model but still allows the model to learn relevant features as the nodes dropped change from epoch to epoch.

Flatten Layer

The flatten layer is used to reshape the outputs of the convolution layers from a tensor into a one dimensional array, allowing it to be passed into the following dense layers.

Dense Block:

The dense layer is a fully connected layer in which each node of the layer is connected to those of the previous. The dense block, in addition to the dense layer of size 256, also contains a ReLU activation layer.

Softmax output:

The final softmax layer contains four nodes which correspond to the four classes of CI that the MRIs will be classified and output into.

The model schematic in Fig. 1 includes plans to reduce the first convolution layer's kernel size as to prevent early spatial data loss. This method was observed to be successful in the referenced paper "On the design of convolutional neural networks for automatic detection of Alzheimer's disease" [4]. Other modifications to the ADD-Net model include the

addition of an instance normalization layer which was noted to help with overall performance of the model compared to a baseline which used batch normalization. We also took inspiration from "Enriching Neural Models with Targeted Features for Dementia Detection" [5] to add a self-attention mechanism following the last convolutional layer in order to better focus the model on important features from the output. An ablation study will be conducted on the RecallNET model with and without the attention layer (RecallNET and RecallNET+Att) to determine if the inclusion helps the model better extract key features that lead to more accurate classification in both balanced and unbalanced datasets. The same attention mechanism will be applied to the baseline ADD-Net model to draw the same conclusion (ADD-Net and ADD-Net+Att).

To evaluate the model, we plan on using the same metrics as those in "ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans" [2]. The baseline model as previously mentioned will be ADD-Net, with an ablation study to determine the benefits of adding an attention mechanism to the model. This means that the four models tested will be ADD-Net, ADD-Net+Att, RecallNET, and RecallNET+Att. All models will be trained using rebalanced SMOTE-TOMEK data and tested on the rebalanced test data and an independent unbalanced sample from the ADNI dataset [14]. This will allow us to evaluate the models under conditions where the number of cases within classes are equal and when they are not; as this would be likely if the model were used in a clinical setting.

2) Roadblocks:

While we had originally planned on using data from the ADNI dataset [14] for independent testing of all proposed and baseline models, we ran into certain roadblocks that prevented us from doing so. First, obtaining access to the dataset required an application process which had a turnaround time of about a week. Once we did, we soon realized that the scans were of raw 3D MRIs. The process to convert them into 2D scans in the same format as the "Alzheimer MRI preprocessed dataset" [3] would have been a difficult and time consuming task that we were unqualified to complete as individuals with no background in neuroscience or medical imaging.

Another challenge we ran into was trouble implementing an attention layer using the parent paper's pre-existing syntax structure. When using the specific syntax structure, we ran into an error where the attention layer was not being called on a list of inputs, namely query, value, and key.

3) New Implementation Plan:

Due to being unable to utilize the ADNI dataset for independent and unbalanced data testing, we opted to split 10% of the data off from the original dataset prior to SMOTE-TOMEK resampling to use for unbalanced testing. All models will now be tested using both the unbalanced dataset and balanced dataset by the same metrics outlined in the original implementation plan. As for the attention layer, we had to modify the syntax of the original implementation code in order for the dimensions of the input tensor to match what was expected in the attention layer argument. After doing so we were able to successfully implement the attention layer. Table I shows the model summary for our final RecallNET model after our modifications.

TABLE I
MODEL SUMMARY FOR THE PROPOSED RECALLNET MODEL

Model Summary		
Layer Type	**Output Shape**	**Parameters**
Input Layer	(None, 176, 208, 3)	0
RecallNET Convolutional Block 1	(None, 88, 104, 16)	96
RecallNET Convolutional Block 2	(None, 42, 50, 32)	12896
RecallNET Convolutional Block 3	(None, 19, 23, 64)	51392
RecallNET Convolutional Block 4	(None, 7, 9, 128)	205184
Dropout Layer 1	(None, 7, 9, 128)	0
Flatten	(None, 8064)	0
Dense Layer 1	(None, 256)	2064640
Dropout Layer 2	(None, 256)	0
Dense Layer 2	(None, 4)	1028
SoftMax Classification Output	(None, 4)	0
Total Parameters		2335236
Trainable Parameters		2335236
Non-trainable Parameters		0

IV. Results

Table II
Model Performances on SMOTE-TOMEK Rebalanced Test Dataset

Model	Accuracy	AUC	Precision	F1-Score	Recall
ADD-Net	98.42%	99.85%	98.42%	98.38%	98.42%
ADD-Net+ Att	93.13%	99.28%	93.33%	92.92%	93.08%
RecallNET	**99.29%**	**99.99%**	**99.29%**	**99.27%**	**99.18%**
RecallNET +Att	98.96%	99.91%	99.02%	98.94%	98.91%

Table III
Model Performances on Unbalanced Independent Dataset

Model	Accuracy	AUC	Precision	F1-Score	Recall
ADD-Net	94.38%	99.57%	94.52%	96.13%	94.38%
ADD-Net+ Att	84.38%	97.06%	84.60%	89.22%	83.28%
RecallNET	**97.66%**	**99.93%**	**97.66%**	**98.46%**	**97.66%**
RecallNET +Att	96.41%	99.70%	99.40%	97.38%	96.25%

A. Evaluation Metrics

For the evaluation of our proposed models, RecallNET and RecallNET+Att, we took into account several metrics such as accuracy, AUC, precision, recall, and F1-score during and after training. These benchmarks as well as a confusion matrix and ROC curve for the results of the proposed models on our balanced and unbalanced test sets were created. Similarly, for evaluation of our baseline models ADD-Net and ADD-Net+Att, we considered the classification report and confusion matrix for both balanced and unbalanced testing data.

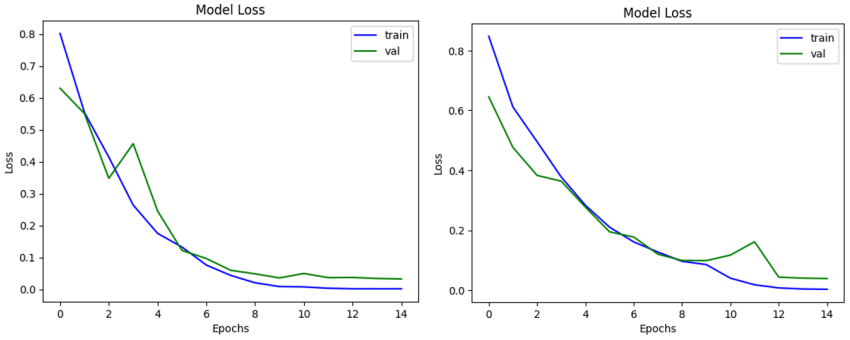
1) Model Training Graphs:

Similarly to our parent paper, we monitored the training of both the RecallNET and RecallNET+Att models through metrics such as accuracy, loss, AUC, precision, recall, and F1-score to prevent overfitting. The performances during training of both models can be observed in Fig. 2, Fig. 3, and Fig. 4 below. These graphs outline the performance of the model at each epoch of training on both the train and validation datasets.

Model loss represents the error in predictions made during training and was evaluated using categorical cross entropy for our case since we were dealing with a classification problem. The loss function was minimized using stochastic gradient descent allowing for greater efficiency of the training. As can be seen in Fig. 2, the loss converged faster at around 10 epochs for the base RecallNET model, whereas RecallNET+Att did not converge till around 12 epochs. While this does not definitively signify anything, it does indicate that the RecallNET+Att model could have required more epochs to properly fit to the data.

Accuracy measures the overall correctness of the model as it is the ratio of true classifications to all classifications. Training accuracy of the RecallNET and RecallNET+Att models can be observed in Fig. 3. Both models performed similarly in terms of accuracy during training.

AUC stands for area under the curve, and measures the area under the Receiver Operating Characteristics (ROC) curve. This is used to represent the model's ability to distinguish between the various classes. As can be seen in Fig. 3, both models' AUC scores plateau around epoch 5.



a) RecallNET b) RecallNET+Att

Fig. 2 Training graphs for loss for RecallNET and RecallNET+Att

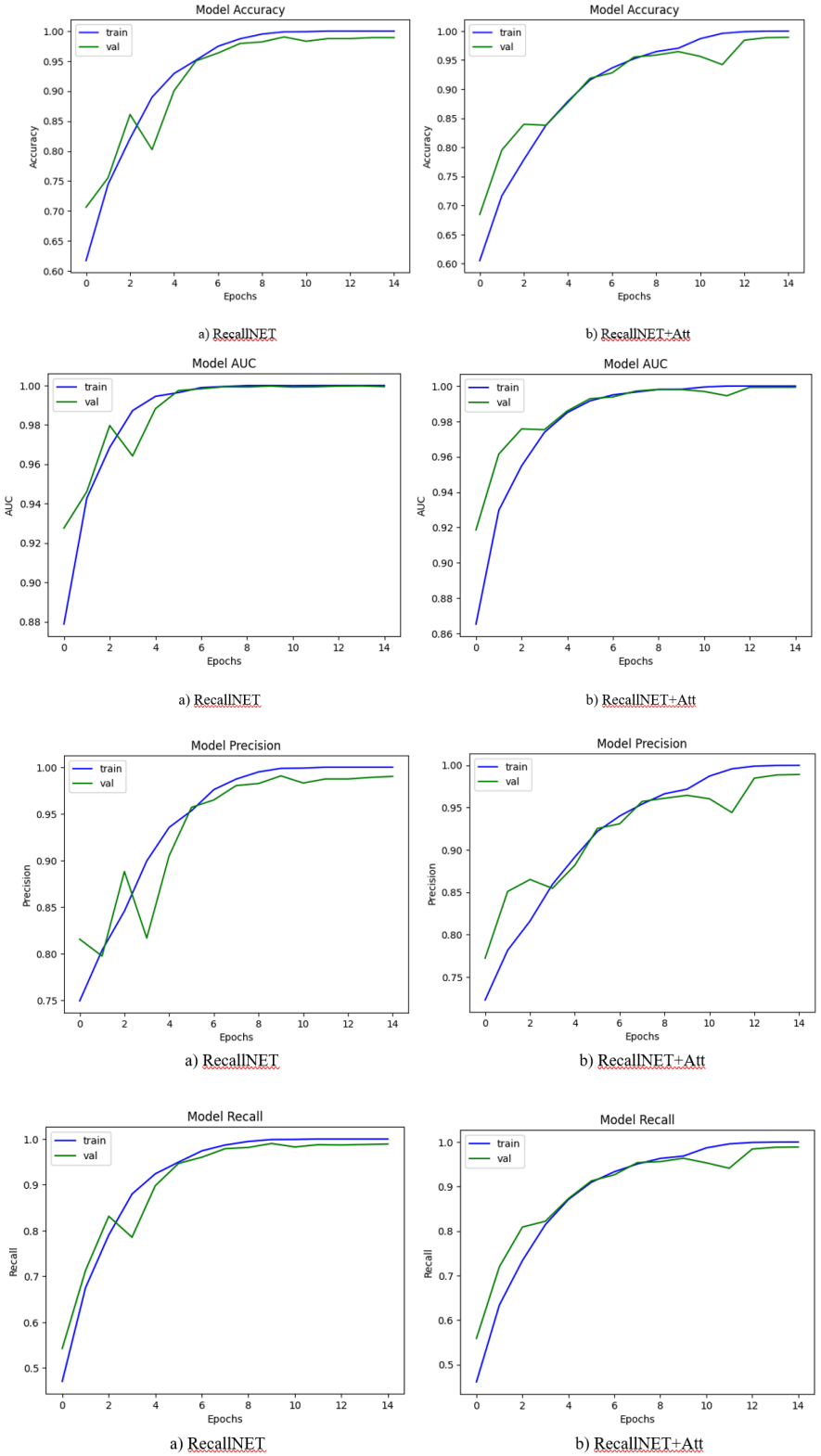
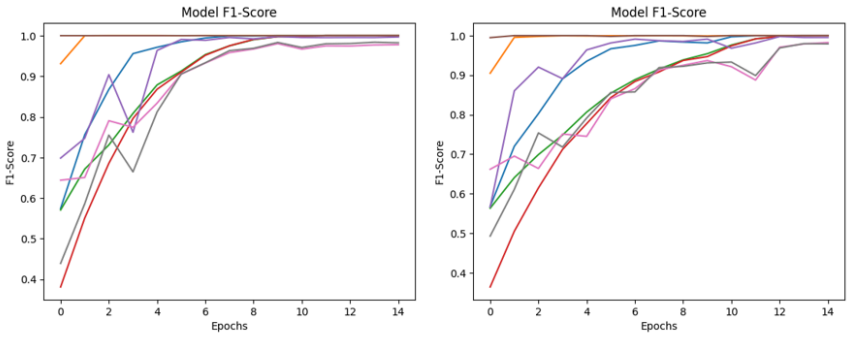


Fig. 3 Training graphs of accuracy, AUC, precision, and recall for RecallNET and RecallNET+Att

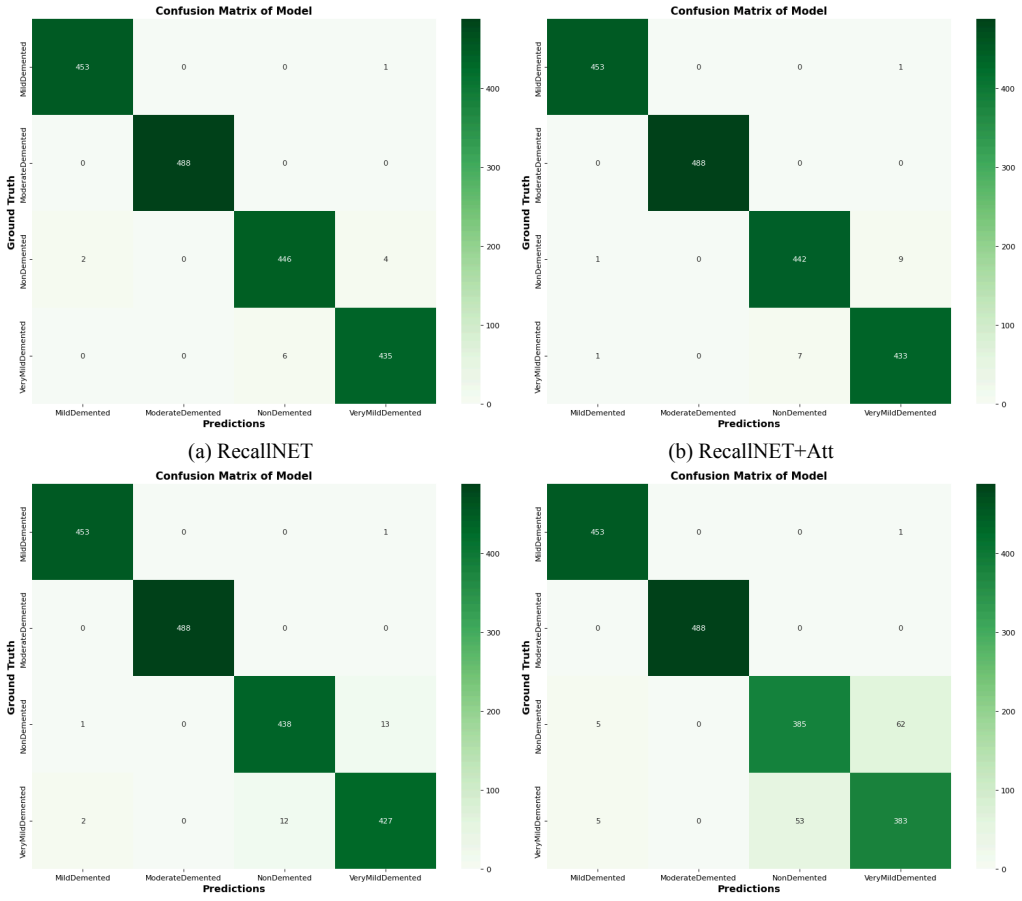
Precision measures the accuracy of positive predictions made by the model by taking the ratio of true positive predictions and total positive predictions. Fig. 3 details the training precision of the models where it can be seen that RecallNET experienced more volatility than RecallNET+Att comparatively.

Recall represents the model's ability to classify all positive instances in the data, which in our case would be the various



a) RecallNET b) RecallNET+Att

Fig. 4 Training graphs of F1-score for RecallNET and RecallNET+Att



(a) RecallNET (b) RecallNET+Att

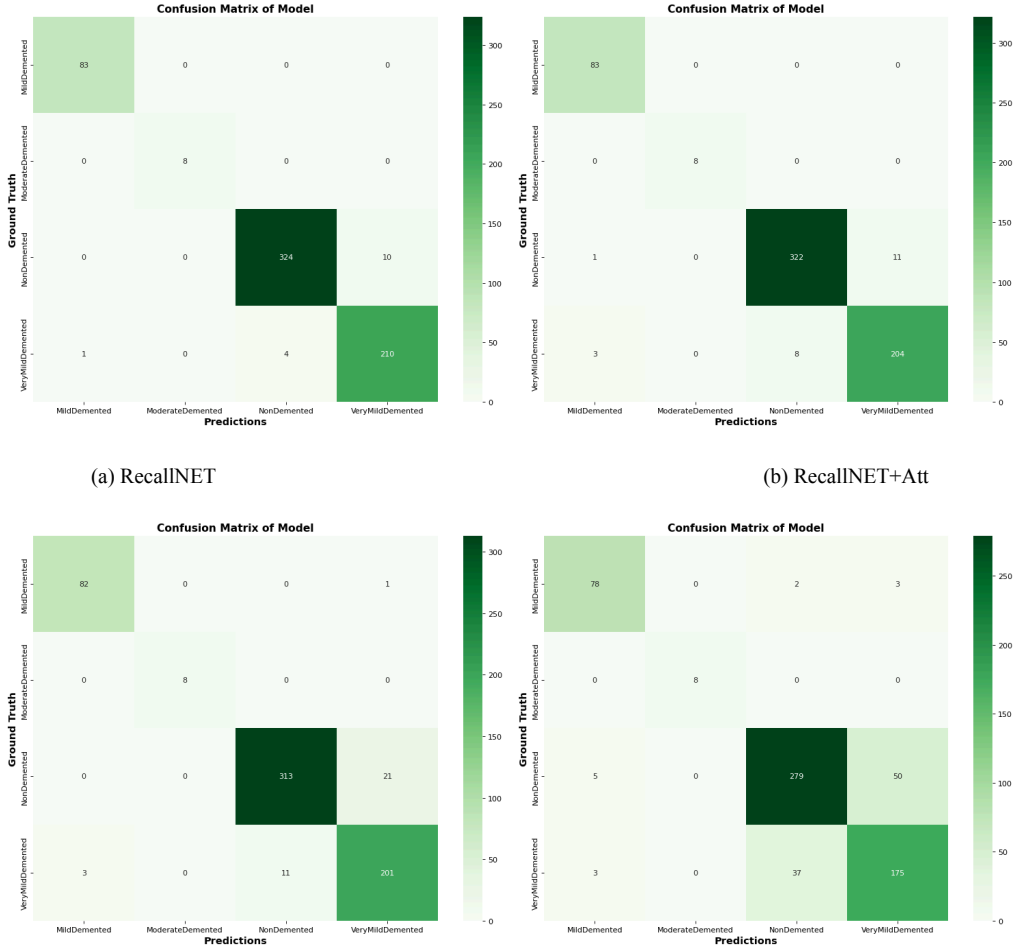
(c) ADD-Net (d) ADD-Net+Att

Fig. 5 Confusion matrices for all models on balanced test set

stages of cognitive impairment. This is important as we aimed to minimize the false negative rate for the RecallNET models as the cost of making false negative classifications in a medical setting is high. Fig. 3 shows the training recall of the models, both of which are relatively similar.

F1-score is a combination of precision and recall that is particularly useful in our case of evaluating unbalanced data. It focuses on class-wise performance rather than overall performance like accuracy. Fig. 4 shows the training F1-score of our RecallNET models.

Overall, all of the model training graphs look very similar between our RecallNET and RecallNET+Att models. However, overall, the RecallNET model seems to have deviations between epochs 2 and 3, while RecallNET+Att looks to have deviations later around epochs 9 to 11.



(a) RecallNET (b) RecallNET+Att

(c) ADD-Net (d) ADD-Net+Att

Fig. 6 Confusion matrices for all models on unbalanced dataset

2) Model Evaluation:

Model evaluation values are shown in Table II and Table III. Overall, RecallNET had the best model performance across the board in every evaluation area. For the rebalanced dataset, RecallNET had accuracy, AUC, precision, F1-score, and recall values of 99.29%, 99.99%, 99.29%, 99.27%, and 99.18%, respectively. These metrics see almost a 1% increase from our baseline model ADD-Net, which is significant given the field of application. The RecallNET+Att model had slightly worse performance than the baseline RecallNET model, but this still performed better than ADD-Net.

For the unbalanced dataset, RecallNET had accuracy, AUC, precision, F1-score, and recall values of 97.66%, 99.93%, 97.66%, 98.46%, and 97.66%, respectively. This shows a significant increase from the baseline ADD-Net model, around a 3% improvement. Overall, our RecallNET model performed the best and represents the new state-of-the-art model for early detection of AD for the Kaggle MRI dataset.

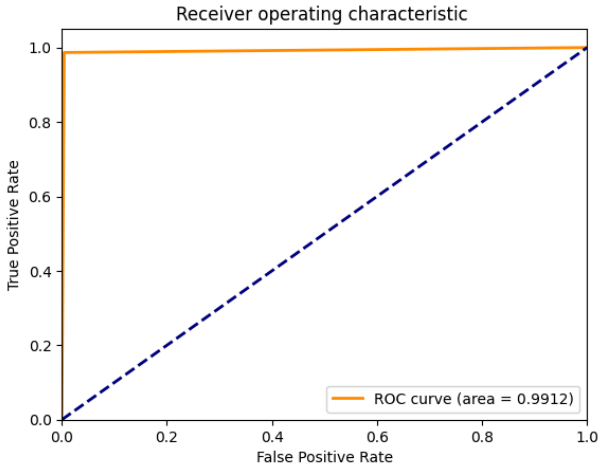
B. Comparison to ADD-Net

For comparison of our proposed models RecallNET and RecallNET+Att to our baseline models ADD-Net and ADD-Net+Att, we consider the confusion matrices, ROC curves, and classification reports in both balanced and unbalanced test sets.

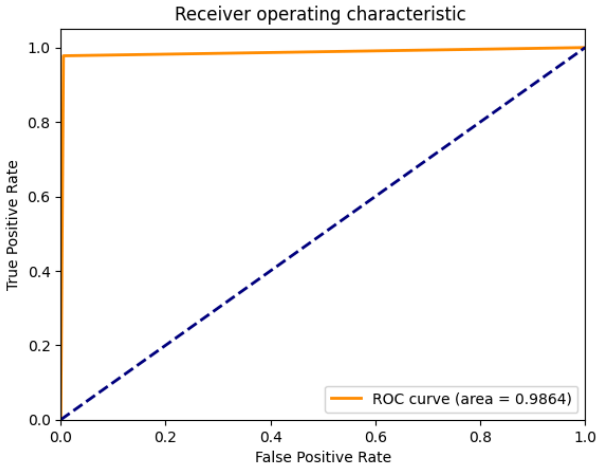
1) Confusion Matrices:

For both the balanced and unbalanced test sets, we visualized the confusion matrices to determine where the errors in our model were made. For this, we compared all four models to observe any possible patterns and reasoning for the misclassifications in Fig. 5 and Fig. 6.

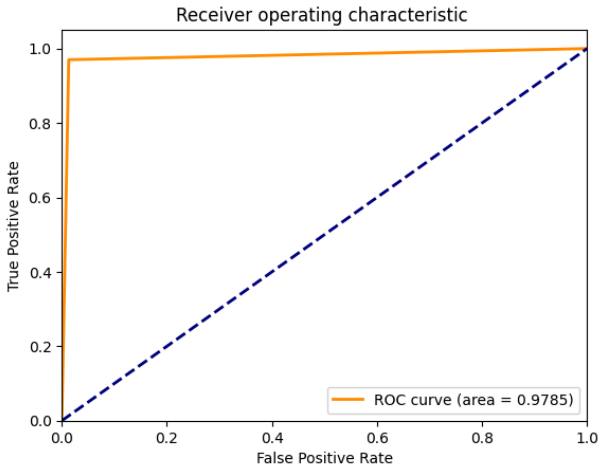
For the balanced test set, there are very few misclassifications in RecallNET, almost all of which were between the NonDemented, and VeryMildDemented classifications. This is likely due to the model having difficulty distinguishing the nuances between these classes since there are minimal variations in the brain at these stages. A similar effect is shown in the RecallNET+Att, ADD-Net,



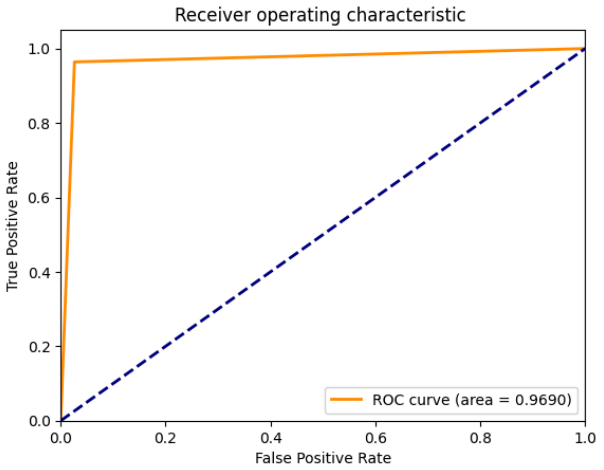
(a) RecallNET balanced



(b) RecallNET+Att balanced



(a) RecallNET unbalanced



(b) RecallNET+Att unbalanced

Fig. 7 ROC curves for RecallNET and RecallNET+Att on balanced and unbalanced test sets

and ADD-Net+Att models. However, within the ADD-Net and ADD-Net+Att models, there are a larger number of misclassifications, indicating there might be a bigger issue than just mislabeling due to subtle differences between the MRI scans in these classes.

For the unbalanced test set, similar patterns were identified where all of the misclassifications were between the NonDemented, VeryMildDemented, and ModerateDemented classes. Again, a larger number of these misclassifications were observed in the ADD-Net and ADD-Net+Att models.

2) Receiver Operating Characteristic Curves:

The ROC curve of the models found in Fig. 7 represent the performance trade-offs between true positive and false positive rates at various decision boundaries. As can be observed in the graphs, all both proposed models perform well on the balanced test set; however, the RecallNET+Att model slightly struggles on the unbalanced set with an AUC score of 0.9690 compared to 0.9785 for the base RecallNET model.

3) Classification Reports:

TABLE IV
RECALLNET UNBALANCED TEST DATA CLASSIFICATION REPORT

	Precision	**Recall**	**F1-Score**
NonDemented	0.99	0.97	0.98
VeryMildDemented	0.95	0.98	0.97
MildDemented	0.99	1.00	0.99
ModerateDemented	1.00	1.00	1.00

TABLE V
RECALLNET+ATT UNBALANCED TEST DATA CLASSIFICATION REPORT

	Precision	Recall	F1-Score
NonDemented	0.98	0.96	0.97
VeryMildDemented	0.95	0.95	0.95
MildDemented	0.95	1.00	0.98
ModerateDemented	1.00	1.00	1.00

TABLE VI
ADD-NET UNBALANCED TEST DATA CLASSIFICATION REPORT

	Precision	Recall	F1-Score
NonDemented	0.97	0.94	0.95
VeryMildDemented	0.90	0.93	0.92
MildDemented	0.96	0.96	0.98
ModerateDemented	1.00	1.00	1.00

TABLE VII
ADD-NET+ATT UNBALANCED TEST DATA CLASSIFICATION REPORT

	Precision	Recall	F1-Score
NonDemented	0.88	0.84	0.86
VeryMildDemented	0.77	0.81	0.79
MildDemented	0.91	0.94	0.92
ModerateDemented	1.00	1.00	1.00

Classification reports for the baseline ADD-Net and proposed RecallNET models were generated, both with and without the added attention layer tested on the unbalanced test set. The results of the classification report can be found in Tables IV - VII. This was done in order to observe each model's performance at classifying each of the four stages of AD: NonDemented, VeryMildDemented, MildDemented, ModerateDemented.

For our research on the early detection of AD, it was important that the model not only performed well overall but achieved exceptional performance at classifying the VeryMildDemented cases. As can be seen in Table IV, this goal was attained as the base RecallNET achieved the highest scores for recall and F1 for classifying VeryMildDemented cases, and it tied with the RecallNET+Att model for highest precision. This shows that in a practical application such as clinical settings, the proposed RecallNET architecture succeeds at detection of early signs of cognitive impairment.

C. *Ablation Studies*

As discussed in our methodology, we conducted an ablation study on both the baseline ADD-Net and proposed RecallNET models to determine if the addition of an attention layer would help improve overall performance. This idea stemmed from the belief that due to potential overfitting resulting in the performance disparities between balanced and unbalanced test datasets, adding some form of regularization may help the models generalize better on data from independent sources. The results of the ablation study can be observed in Fig. 2 and 3 where the "Att" denotes the models with the added attention layer.

Between tests on both balanced and unbalanced data, we noted that models with the added attention layer performed worse than their baselines in all evaluation metrics. The most notable performance drops occurred in the ADD-Net+Att model on unbalanced data where accuracy, precision, recall, and F1 scores all fell significantly in comparison to the baseline ADD-Net. RecallNET's performance also dropped with the added attention layer, however, the differences were far less drastic than observed in ADD-Net. We believe this is due to the inclusion of instance normalization in RecallNET, which resulted in more information retention in the convolutional layers than would comparatively be achieved with batch normalization. This made it so inputs could be better generalized when passed through the attention layer.

Still, a performance drop may be indicative that the inclusion of an attention layer is ultimately unnecessary to the model. This is difficult to say for certain though, as our original plan included testing all models on an external dataset sampled from ADNI [14] to see if the attention layer improved model generalization and therefore performance on independent data. Due to the limitations we faced, we were unable to test this hypothesis on the ADNI dataset. It is possible that the attention layer may have improved performances on external data due to the aforementioned better generalization.

V. DISCUSSION

A. *Limitations*

As mentioned earlier, we were unable to attain access to the ADNI dataset in a timely manner for the scope of this project. With an application and a long turnaround time, it proved inefficient for us to use this dataset and therefore we do not know how RecallNET performs on independent datasets. Additionally, there are a lack of datasets containing 2D longitudinal MRIs that are labeled specifically for AD. Due to this, it is difficult to understand how RecallNET would perform on such datasets. We are looking into other independent datasets to test on RecallNET in the future.

B. Future Work

Through our research of this project and Alzheimer's Disease in general, we encountered various implementations of AD detection tasks. With the time constraint of this project, we were not able to explore every potential avenue of early detection of Alzheimer's that we believe could ultimately help in developing a treatment for its cure. Additionally, as we developed our methodology and executed it, we encountered roadblocks and considered several other implementations that could be considered in our future work on this project.

1) CNN-LSTM

Our first proposal for future work stems from our third research paper and would consist of us implementing a CNN-LSTM model using text and speech based data to detect Alzheimer's Disease. While looking at the current problems associated with diagnosing Alzheimer's, we came to learn that typical diagnosis is time consuming, costly, and invasive. This can in part be attributed to the high healthcare costs in the US where MRI scans are not easily accessible to much of the public. In addition to this, MRIs require radiologists to administer and process them prior to use. Using text-based data eliminates these issues by providing a more accessible way to detect AD, namely through the Cookie Theft scenario interview in the DementiaBank dataset [7]. In our implementation we would have the CNN-LSTM text based model precede RecallNET by flagging patients who are classified as having signs of MCI and recommending them to get an MRI, which will then be processed via RecallNET. We believe that the performance of our original implementation, focused on early detection of AD, might decrease as this model is less nuanced at classifying very mild cases of CI. However, it may be beneficial in broadening access to AI based AD detection in the healthcare field, due to ease of implementation and interpretation.

2) GNN

Our second idea for future work involves using a graph neural network (GNN) instead of our current CNN based approach. GNNs are another class of deep learning models that are designed to operate on graph-structured data. The MRIs in our dataset can be converted into a graph structure and fed into the GNN to perform our computer vision task. GNNs are better at extracting relational information from data, so we suspect it might be better at early detection by understanding the nuances between healthy control and early cognitive impairment cases. Additionally, GNNs could be used to map 3D structural MRIs and use them for classification which was not possible with our CNN based approach. However, the GNN model would likely perform worse at generalizing for an unbalanced dataset since GNN models are prone to overfitting.

3) ViT

For our third area of future work, we would like to explore encoder transformers and their application in MRI scan based AD detection. ViT (Vision Transformers) are entirely transformer and self-attention based deep learning models that have long been used for NLP tasks but have gained recent momentum in the field of computer vision. Because of their efficiency in processing time and lower amounts of data loss some experts believe that they can replace the CNN most commonly used in computer vision today. We believe that using a ViT would lead to shorter training times for the models since no convolutional blocks are required, leading to much faster classification. While this implementation may be more practical, there's also a chance that the model would be underfit since ViT based models require a larger dataset. This poses a problem for our application since there is a lack of available open source MRI data for AD classification. However, if we were able to use the ADNI dataset it is likely that a ViT based model would perform on par with RecallNET with far better practical application because of the improved efficiency.

VI. Conclusion

In this paper, we presented a convolutional neural network model focused on the early detection of Alzheimer's Disease based on the ADD-Net model outlined by Fareed et al [2]. Our proposed model, RecallNET, places emphasis on reducing false negative classifications of AD in a clinical setting and addresses the unbalanced dataset issues that ADD-Net faced. Similar to ADD-Net, RecallNET is also built to classify early stages of AD and does so with unparalleled accuracy. The small kernel size in the first convolutional layer combined with instance normalization in each convolutional block addresses the accuracy loss observed in unbalanced datasets. The RecallNET model achieves evaluation metrics on the balanced dataset with 99.29% accuracy, 99.29% precision, and 99.18% recall. RecallNET has similarly impressive evaluation metrics on an independent unbalanced dataset with 97.66% accuracy, 97.66% precision, and 97.66% recall. In the future, we hope to test other types of models including graph neural networks and vision transformers to address this problem.

References

[1] "Alzheimer's Disease and Related Dementias" Centers for Disease Control and Prevention, https://www.cdc.gov/aging/aginginfo/alzheimers.htm#:~:text=Alzheimer's%20disease%20is%20the%20most,thought%2C%20memory%2C%20and%20language (accessed Nov. 2, 2023).

[2] M. M. S. Fareed et al., "ADD-Net: An Effective Deep Learning Model for Early Detection of Alzheimer Disease in MRI Scans," in *IEEE Access*, vol. 10, pp. 96930-96951, 2022, doi: 10.1109/ACCESS.2022.3204395.

[3] S. Liu, C. Yadav, C. Fernandez-Granda, and N. Razavian , "On the design of convolutional neural networks for automatic detection of Alzheimer's

disease," *Proceedings of Machine Learning Research*, vol. 116, 2020. doi:116:184-201

[4] Flavio Di Palo and N. Parde, "Enriching Neural Models with Targeted Features for Dementia Detection," *INDIGO (University of Illinois at Chicago)*, Jan. 2019, doi: https://doi.org/10.18653/v1/p19-2042.

[5] S. Kumar, "Alzheimer MRI preprocessed dataset," Kaggle, https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset/discussion (accessed Nov. 2, 2023).

[6] "Tf.keras.initializers.GlorotUniform : tensorflow V2.14.0," TensorFlow, https://www.tensorflow.org/api_docs/python/tf/keras/initializers/GlorotUniform (accessed Nov. 3, 2023).

[7] J.T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585-59

[8] National Institute of Biomedical imaging and bioengineering, "Magnetic Resonance Imaging (MRI)," *National Institute of Biomedical Imaging and Bioengineering*, Jul. 17, 2018. https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri

[9] H. D. Yang, D. H. Kim, S. B. Lee, and L. D. Young, "History of Alzheimer's Disease," *Dementia and Neurocognitive Disorders*, vol. 15, no. 4, p. 115, Dec. 2016, doi: https://doi.org/10.12779/dnd.2016.15.4.115.

[10] R. S. of N. A. (RSNA) and A. C. of Radiology (ACR), "Alzheimer's Disease," *Radiologyinfo.org*. https://www.radiologyinfo.org/en/info/alzheimers#:~:text=In%20the%20early%20stages%20of

[11] Alzheimer's Association, "Facts and Figures," *Alzheimer's Disease and Dementia*, 2023. https://www.alz.org/alzheimers-dementia/facts-figures

[12] "The Impact of Hospital Staff Shortages on Patients," *publichealth.tulane.edu*, Nov. 11, 2022. https://publichealth.tulane.edu/blog/hospital-staff-shortages/#:~:text=Hospital%20Staff%20Shortages%20During%20the%20Pandemic&text=Even%20after%20many%20jobs%20came

[13] Ramin, "Self Attention in Convolutional Neural Networks," *MLearning.ai*, Feb. 12, 2022. https://medium.com/mlearning-ai/self-attention-in-convolutional-neural-networks-172d947afc00#:~:text=Self%2Dattention%20is%20described%20in (accessed Nov. 13, 2023).

[14] "ADNI | Alzheimer's Disease Neuroimaging Initiative," *ADNI Alzheimer's Disease Neuroimaging Initiative*. https://adni.loni.usc.edu/ (accessed Nov. 12, 2023).