applied sciences

MDPI

Article

Multi-Task Deep Evidential Sequence Learning for Trustworthy Alzheimer's Disease Progression Prediction

Zeyuan Zhao [1], Ping Li [2,*], Yongjie Dai [2], Zhaoe Min [2] and Lei Chen [2]

[1] Bell Honor School, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; b20070223@njupt.edu.cn
[2] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; b20030332@njupt.edu.cn (Y.D.); minzhaoe@njupt.edu.cn (Z.M.); chenlei@njupt.edu.cn (L.C.)
* Correspondence: ping.li@njupt.edu.cn

Abstract: Alzheimer's disease (AD) is an irreversible neurodegenerative disease. Providing trustworthy AD progression predictions for at-risk individuals contributes to early identification of AD patients and holds significant value in discovering effective treatments and empowering the patient in taking proactive care. Recently, although numerous disease progression models based on machine learning have emerged, they often focus solely on enhancing predictive accuracy and ignore the measurement of result reliability. Consequently, this oversight adversely affects the recognition and acceptance of these models in clinical applications. To address these problems, we propose a multi-task evidential sequence learning model for the trustworthy prediction of disease progression. Specifically, we incorporate evidential deep learning into the multi-task learning framework based on recurrent neural networks. We simultaneously perform AD clinical diagnosis and cognitive score predictions while quantifying the uncertainty of each prediction without incurring additional computational costs by leveraging the Dirichlet and Normal-Inverse-Gamma distributions. Moreover, an adaptive weighting scheme is introduced to automatically balance between tasks for more effective training. Finally, experimental results on the TADPOLE dataset validate that our model not only has a comparable predictive performance to similar models but also offers reliable quantification of prediction uncertainties, providing a crucial supplementary factor for risk-sensitive AD progression prediction applications.

Keywords: Alzheimer's disease; disease progression prediction; deep sequence learning; trustworthy evidential learning; adaptive weighting scheme

check for updates

Citation: Zhao, Z.; Li, P.; Dai, Y.; Min, Z.; Chen, L. Multi-Task Deep Evidential Sequence Learning for Trustworthy Alzheimer's Disease Progression Prediction. *Appl. Sci.* **2023**, *13*, 8953. https://doi.org/10.3390/app13158953

Academic Editor: Jan Egger

Received: 1 July 2023
Revised: 27 July 2023
Accepted: 2 August 2023
Published: 3 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Alzheimer's disease (AD) is a progressive degenerative disease of the nervous system that has a profound impact on the health of millions of people worldwide, while also imposing a significant burden on society and the economy. Given the absence of a definitive cure for AD, timely intervention in the early stages of the disease is considered an effective therapeutic strategy [1]. Additionally, the ADAS-Cog13 scores, the most widely utilized cognitive measurement method, contribute significantly to the treatment process of patients. The ADAS-Cog13 consists of thirteen tasks including word recall, naming objects and fingers, commands, constructional praxis, ideational praxis, orientation, word recognition, language, comprehension of spoken language, word finding difficulty, remembering test instructions, number cancellation, and delayed free recall. Usually, studies administer all thirteen tasks and score them on a single scale from 0 to 85 [2]. Monitoring changes in ADAS-Cog13 scores among patients not only helps doctors to gain better insights into disease progression but also holds significant promise in augmenting both the effectiveness and efficiency of the assessment protocol employed in clinical treatments, which is typically time-consuming and expensive. Consequently, research focusing on early diagnosis and

cognitive score prediction holds a particular significance in clinical practice, as it enables physicians to make more accurate assessments of disease progression and provide patients with better treatment plans.

Recently, the rapid progress of deep learning has attracted a growing number of researchers to explore its applications in the diagnosis of AD and the prediction of cognitive scores. Wang et al. [3] utilized recurrent neural networks (RNNs) with LSTM cells to predict future cognitive scores. Hong et al. and Cui et al. [4,5] employed variants of RNNs to predict AD diagnosis. Although both achieved promising results in their respective tasks, they focused solely on completing individual prediction tasks, disregarding the existing correlations among different tasks. Studies investigated by Liang et al. [6] and Jung et al. [7] have demonstrated that the simultaneous learning of multiple tasks during the AD disease modeling process benefits the model's ability to capture underlying disease characteristics and enhances the generalization performance of each task. Therefore, establishing a multi-task learning model that simultaneously handles disease diagnosis and cognitive score prediction proves to be an effective strategy for improving model performance.

However, determining the appropriate loss weights for each task becomes a new unresolved issue when introducing multi-task learning. The allocation of loss weights directly impacts the model's performance and generalization ability across different tasks. Currently, most AD progression prediction models overlook the significance of loss weights among tasks and either assign equal weights or rely on laborious manual adjustments. Hence, there is a need to explore more effective methods for adjusting loss weights to construct superior multi-task learning models.

Meanwhile, despite the high accuracy of existing AD clinical diagnosis and cognitive score prediction models, their recognition and acceptance in clinical applications remain limited. Trust in the outcomes of deep learning models is hindered in the healthcare domain due to their complex and black-box characteristics, presenting a challenge for decision makers. Therefore, there is a pressing need for a method that can measure the reliability of prediction results, aiding decision makers in better understanding the model's predictions and making more accurate and reliable decisions.

To effectively solve these challenges, we propose a multi-task evidential sequence learning model for reliable AD clinical diagnosis and cognitive score prediction. Specifically, we introduce recurrent neural networks to the multi-task learning framework, utilizing longitudinal time series data from individuals at risk of AD to simultaneously predict their future clinical diagnosis and ADAS-Cog13 cognitive scores. We then introduce evidential learning to provide uncertainty estimates for the model's predictions, thereby quantifying the confidence level of the model's outcomes. Furthermore, we integrate the Dynamic Weight Average (DWA) scheme to automatically allocate learning resources for different tasks in the training process, aiding in the improvement of model performance. Experiments conducted on the TADPOLE dataset [8] substantiate that our proposed model not only achieves satisfactory performance in disease progression prediction but also provides effective uncertainty estimation. Furthermore, the core idea inherent in our model is not only applicable to AD progression prediction, but it also offers the potential for extension to various deep learning applications in the medical domain. By effectively modeling the uncertainty of model results, the credibility of most existing deep learning models in the medical field can be enhanced, thereby contributing to the advancement of trustworthy and reliable methodologies in healthcare settings. In conclusion, this paper makes the following contributions:

- We propose a multi-task evidential sequence learning model for the trustworthy prediction of disease progression, which performs AD clinical diagnosis and cognitive score predictions while providing reliable uncertainty quantification of each prediction.
- To achieve balanced learning of various tasks, we introduce the Dynamic Weight Average (DWA) weighting scheme into our model to automatically adjust the weights of each task, resulting in improved prediction accuracy.

- We conduct extensive experiments on the TADPOLE dataset which validate the superior effectiveness and accuracy of our model thanks to the promising uncertainty estimation and DWA.

The structure of this paper is as follows: the related works are discussed in Section 2, while Section 3 expounds upon the formulation of the problem. Section 4 describes the implementation of our proposed model. Section 5 is dedicated to the verification and evaluation of the model's performance, and Section 6 provides a more comprehensive analysis of the model. Section 7 summarizes this paper.

2. Related Works

Deep learning is a rapidly advancing field that has gained wide application in recent years. Research teams are actively exploring its use in addressing issues related to Alzheimer's disease progression modeling (DPM). DPM involves predicting the time of clinical state transitions within the AD spectrum, indicating when patients progress to the next stage, as well as changes in AD-related features. Predicting clinical state transition time is crucial for DPM as it facilitates early identification of AD patients and provides information about disease progression and severity. Furthermore, studying AD-related feature changes helps decision makers better understand the trends in disease development. Over time, AD features exhibit varying degrees of change, such as accelerated brain volume reduction and decreased cognitive scores. By capturing these changes in longitudinal data, one can enhance the understanding of the underlying temporal characteristics of the disease. Significant progress has been made in this area. For instance, Ghazi et al. [9] utilized six volumetric magnetic resonance imaging (MRI) biomarkers to establish a model based on LSTM for predicting the progression of AD. Xu et al. [10] proposed a multi-modal sequence learning model which provides a new solution for handling incomplete data while modeling AD disease progression.

As crucial components of DPM and clinical applications, cognitive scores and disease clinical diagnosis have attracted significant attention from researchers. Utsumil et al. [11] proposed a personalized model based on the Gaussian process to forecast the progression of ADAS-Cog13 scores. Ning et al. [12] examined the correlation between ADAS-Cog13 and brain image characteristics by employing a 3D convolutional neural network. Jung et al. [7] developed a variant of the recurrent network which extracts features from multiple data sources, such as clinical assessments and brain imaging, to predict patients' cognitive abilities and biomarker changes. Liang et al. [6] developed a novel multi-task learning framework that provides more accurate results for disease diagnosis, cognitive scores, and biomarker prediction. Nguyen et al. [13] utilized MinimalRNN for modeling and introduced the method of model filling to handle missing data.

Liu et al. [14], inspired by GradNorm [15], addressed the issue of determining the loss weights for each task in multi-task learning by proposing the Dynamic Weight Average (DWA) method. This approach dynamically adjusts the loss weights of each task during each training iteration, ensuring equal importance across different tasks. The DWA method has been successfully applied in various domains, including fault diagnosis [16] and weather recognition [17].

Furthermore, despite the remarkable achievements of existing models, decision-makers find it challenging to wholeheartedly trust the outcomes due to the opaque nature of deep learning algorithms. Thus, in domains with high safety requirements such as modern healthcare, there is an urgent need to establish an interpretable model that can quantify the reliability of prediction results. Uncertainty theory serves as a method for assessing the credibility of model outputs, and common techniques for uncertainty modeling include Bayesian neural networks [18], dropout [19], and deep ensemble [20]. However, these approaches possess inherent limitations; Bayesian neural networks incur significant computational costs during operations, while dropout and deep ensemble require substantial additional time for sampling and model training.

In recent years, the evidential deep learning (EDL) theory introduced by Amini et al. and Sensoy et al. [21,22] has exhibited great potential in the field of uncertainty. It directly models the uncertainty of outcomes by placing evidence distributions on the original likelihood function. This methodology eliminates the need for integration or sampling. Specifically, EDL employs Dirichlet distribution on the classification distribution for classification tasks and normal-inverse-gamma distribution on the Gaussian distribution for regression tasks. This approach has yielded unprecedented success in detecting out-of-distribution (OOD) samples and resisting adversarial perturbations. Furthermore, it has been applied in action recognition [23], medical image diagnosis [24], and predicting molecular structural properties [25].

3. Problem Formation

For the prediction of AD progression, this study focuses on predicting the clinical diagnoses which are classified into cognitively normal (CN), mild cognitive impairment (MCI), and AD, as well as the ADAS-Cog13 cognitive score, for any future time interval based on variable length and incomplete historical data of individuals at risk of AD. Assuming the total sample size is N, we represent the population sample as $\chi = \{x_i\}_{i=1,2,\cdots,N}$, where x_i denotes the i-th sample. In this paper, each month is considered as a time point, and the longitudinal features of the i-th sample at different time points are denoted as $\{\boldsymbol{x}_{it}\}_{t=1,2,\cdots,T^{(i)}}$, where $T^{(i)}$ represents the temporal length of the i-th sample. Suppose a sample's data collection begins in June 2006, and in addition to that, data points are also collected in November 2006, June 2007, and November 2007. In this case, the temporal length T of that sample would be 18. Each sample's features are composed of two parts: the first part is a clinical diagnosis vector $\boldsymbol{s}$ represented by one-hot encoding, and the second part is a vector $\boldsymbol{g}$ consisting of continuous biomarkers. The final representation of the i-th sample's features at time point t is denoted as $\boldsymbol{x}_{it} = [\boldsymbol{s}_{it}; \boldsymbol{g}_{it}]$. If any feature is missing, the corresponding element in the feature vector will be represented as a null value.

4. The Proposed Model

4.1. Model Overview

In this work, a multi-task evidential sequence learning model is proposed for AD progression prediction as shown in Figure 1. At each time point, the missing values of the input are first imputed by the prediction from the previous time point. For missing values at the initial time point, the imputation method employed in this study involves using the mean value derived from all available time points of the training subjects. After that, the deep evidential sequence module, which is composed of sequence learning modules and evidential deep learning modules as shown in Figure 2, receives the entirety of the observations, wherein missing values are substituted with imputed counterparts. Through this process, the proposed deep evidential sequence module effectively captures the latent temporal characteristics while simultaneously modeling the evidential distributions inherent in the provided longitudinal data. Subsequently, the output of the evidential deep learning module undergoes further transformations to forecast both the clinical status and ADAS-cog13 cognitive score while providing uncertainty estimation.

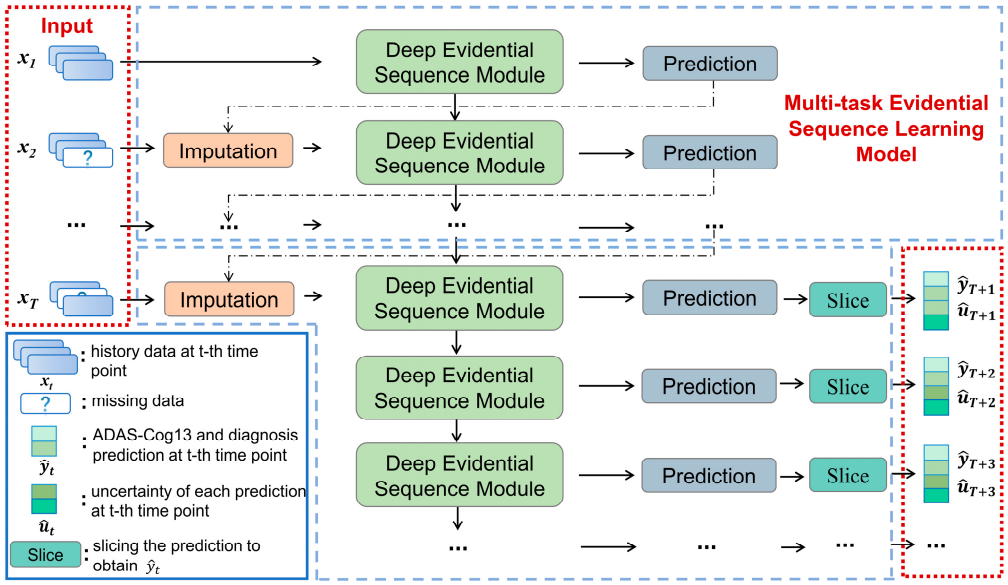


Figure 1. Architecture of our proposed multi-task evidential sequence learning model.

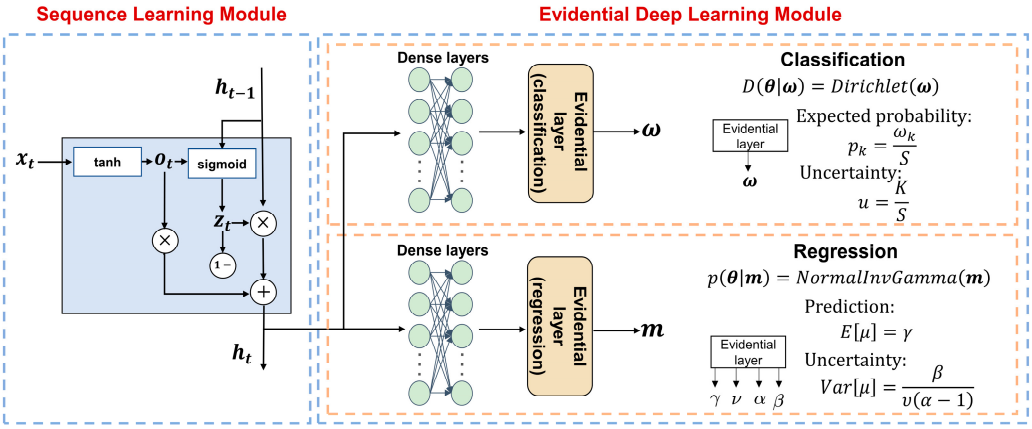


Figure 2. Illustration of the deep evidential sequence module.

4.2. Sequence Learning Module

The data of AD patients exhibit temporal characteristics and are prone to varying degrees of missing values. To address this, recurrent neural networks (RNNs) are an effective method as they can capture correlations between different time points and handle missing data efficiently. Therefore, in this study, we employ RNNs for AD progression prediction. Among the diverse range of RNN models available, we choose MinimalRNN [26] as the backbone of our model. MinimalRNN has a simpler structure compared to other networks and utilizes stronger constraints on the update rules. Additionally, MinimalRNN utilizes fewer parameters, which helps prevent overfitting issues caused by excessive parameterization. It is crucial to emphasize that the calculation of loss in the model does not take into account any missing data. Figure 2 illustrates the structure of MinimalRNN, where $\boldsymbol{x}_t$ represents the model input, $\boldsymbol{h}_t$ represents the hidden state, $\boldsymbol{O}_t$ represents the transformed input, and $\boldsymbol{Z}_t$ represents the forget gate. The update equation for the hidden state is as follows:

$$\boldsymbol{O}_t = tanh(\boldsymbol{W}_x \boldsymbol{x}_t) \tag{1}$$

$$\boldsymbol{Z}_t = sigmoid(\boldsymbol{U}_h \boldsymbol{h}_{t-1} + \boldsymbol{W}_O \boldsymbol{O}_t) \tag{2}$$

$$\boldsymbol{h}_t = \boldsymbol{Z}_t \odot \boldsymbol{h}_{t-1} + (1 - \boldsymbol{Z}_t) \odot \boldsymbol{O}_t \tag{3}$$

where W_x, U_h, and W_O represent weight matrices, and $\odot$ denotes element-wise product.

4.3. Evidential Deep Learning Module

4.3.1. Evidential Deep Learning for Classification

In deep learning related classification tasks, it is common to utilize the softmax function to transform the models' output values into class probabilities. However, the softmax function can lead to overconfidence in the model even when it produces incorrect results [27]. Recently, evidential deep learning (EDL) based on second-order uncertainty modeling to parameterize the Dirichlet distribution of class probabilities provides the model with the ability to "know unknown". Specifically, given a K-class classification problem, EDL assumes that the class probabilities, denoted as $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, follow a prior Dirichlet distribution $D(\boldsymbol{p} \mid \boldsymbol{\omega})$, where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_k)$. Building upon the Dempster–Shafer Theory (DST) [28] and Subjective Logic (SL) [29], the $\boldsymbol{\omega}$ is connected to the acquired evidence $\boldsymbol{e}$ through the equation $\boldsymbol{\omega} = \boldsymbol{e} + 1$. It is worth noting that to ensure a non-negative output that serves as the evidence vector for the predicted Dirichlet distribution, the previous softmax layer is substituted with a ReLU layer. Then, the predicted uncertainty u and the class probability of category j are obtained as:

$$u = \frac{K}{S} \tag{4}$$

$$p_j = \frac{\omega_j}{S} \tag{5}$$

where $S = \sum_i^K (e_i + 1)$ is also referred to as the Dirichlet strength.

For each i-th sample, we adopt the following form of the cross-entropy loss function during the training process:

$$\mathcal{L}_i^{CE} = \int \left[\sum_{j=1}^{3} -s_{ij} log(p_{ij}) \right] \frac{1}{B(\boldsymbol{\omega_i})} \prod_{j=1}^{3} p_{ij}^{\omega_{ij}-1} d\boldsymbol{p}_i \tag{6}$$

where p_{ij} denotes the predicted probability of the i-th sample being diagnosed as the j-th stage, s_{ij} is the j-th component of the disease diagnosis vector represented by one-hot encoding, and $B(\boldsymbol{\omega_i})$ represents the K-dimensional multinomial beta function. In this task, the value of K is 3, indicating the three stages of diagnosis.

In addition, to address the issue of misleading evidence generated for incorrect labels, we propose the introduction of a KL divergence term. This term aims to enforce the cumulative evidence to converge towards 0 when a sample is misclassified.

$$\mathcal{L}_i^{KL} = KL[D(\boldsymbol{p}_i \mid \tilde{\boldsymbol{\omega}}_i) \parallel D(\boldsymbol{p}_i \mid 1)] \tag{7}$$

where $\tilde{\boldsymbol{\omega}}_i = \boldsymbol{s_i} + (1 - \boldsymbol{s_i}) \odot \boldsymbol{\omega_i}$ is the adjusted Dirichlet distribution parameter.

In conclusion, for the classification problem, the overall loss function for the i-th sample can be defined as follows:

$$\mathcal{L}_i^c = \mathcal{L}_i^{CE} + \lambda_{epoch} \mathcal{L}_i^{KL} \tag{8}$$

where $\lambda_{epoch} = \min(1.0, epoch/10) \in [0, 1]$ is the annealing coefficient. $epoch$ is the index of the current training epoch. By gradually increasing λ_{epoch}, the network can assign relatively lower importance to the KL divergence term at the beginning of training, which has a positive effect on the exploration of the parameter space and avoids outputting a flat uniform distribution.

4.3.2. Evidential Deep Learning for Regression

In the case of regression tasks, EDL adopts a slightly different approach. Given a dataset $D = \{x_i, y_i\}_{i=1}^N$, where the labels y_i are assumed to follow a Gaussian distribution

with unknown parameters $\{\mu, \sigma^2\}$, to estimate model uncertainty, the evidential model places a prior distribution on the parameters $\{\mu, \sigma^2\}$. Specifically, the mean μ follows a Gaussian distribution, while its variance σ^2 follows an inverse gamma distribution. These resulting higher-order distributions, also known as evidence distributions, can be represented using the Gaussian conjugate prior to normal-inverse-gamma (NIG) distribution $p(\mu, \sigma^2|\boldsymbol{m})$, where the parameters $\boldsymbol{m} = \{\gamma, \upsilon, \alpha, \beta\}$.

$$\begin{array}{c}(y_1, y_2, \ldots, y_N) \sim \mathcal{N}(\mu, \sigma^2) \\ \mu \sim \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1}) \sigma^2 \sim \Gamma^{-1}(\alpha, \beta)\end{array} \tag{9}$$

where $\Gamma(\cdot)$ represents the gamma function, $\gamma \in R, \upsilon > 0, \alpha > 1, \beta > 0$.

In the regression task, the model outputs four values per target through the evidential layer, corresponding to $\{\gamma, \upsilon, \alpha, \beta\}$. From these values, we can obtain the predicted value γ and uncertainty $\frac{\beta}{\upsilon(\alpha-1)}$.

According to Bayesian probability theory, we can obtain the likelihood of the observed predicted targets by marginalizing the parameters of the Gaussian distribution. Additionally, leveraging the properties of the NIG distribution, we derive the following marginal likelihood:

$$p(y_i \mid \boldsymbol{m}) = \int_{\sigma^2=0}^{\sigma^2=\infty} \int_{\mu=-\infty}^{\mu=\infty} p\left(y_i \mid \mu, \sigma^2\right) p\left(\mu, \sigma^2 \mid \boldsymbol{m}\right) d\mu d\sigma^2 \tag{10}$$

where y_i represents the ADAS-cog13 cognitive score of the i-th sample.

To learn the model parameters, the training process involves minimizing the negative log marginal likelihood (NLL) loss. The NLL loss function is defined as follows:

$$\mathcal{L}_i^{NLL} = -log(p(y_i \mid \boldsymbol{m})) \tag{11}$$

In addition, we also introduce a regularization term $\mathcal{L}_i^R$ to minimize evidence on errors. The regularization term $\mathcal{L}_i^R$ is defined as follows:

$$\mathcal{L}_i^R = |y_i - \mathbb{E}[\mu_i]| \cdot \Phi = |y_i - \gamma| \cdot (2\upsilon + \alpha) \tag{12}$$

where $\Phi = 2\upsilon + \alpha$ represents the total evidence. This regularization term serves to penalize errors in model predictions, and the magnitude of this penalty is directly proportional to the total evidence.

In conclusion, for the regression problem, the overall loss function for the i-th sample can be defined as follows:

$$\mathcal{L}_i^r = \mathcal{L}_i^{NLL} + \lambda_r \mathcal{L}_i^R \tag{13}$$

where λ_r is a hyperparameter to adjust regularization.

4.4. Dynamic Weight Average Weighting Scheme

In multi-task learning, model performance is significantly influenced by the weights assigned to each task. If the weight of a task's loss is too high, the model will focus more on that task while neglecting others. Conversely, if the weight is too low, the model cannot effectively utilize the information from that task. Both situations can result in a decline in the model performance. However, manually adjusting the optimal weights for each task is often expensive and challenging [30]. Therefore, this paper introduces the Dynamic Weight Average (DWA) weighting scheme [14] on top of the existing multi-task learning model. During the training process, this scheme dynamically adapts the weights associated with each task, allowing the model to better balance different tasks and achieve optimal performance.

The relative descending rate of the k-th task can be calculated as:

$$\theta_k(c-1) = \frac{\mathcal{L}_k(c-1)}{\mathcal{L}_k(c-2)} \tag{14}$$

where c is an iteration index. $\mathcal{L}_k(c)$ represents each epoch's average loss calculated over several iterations. The smaller $\theta_k(c-1)$ indicates that the value of the task's loss decreases more compared to the previous training round, which also means that the task gets better learning, so the attention to this task can be reduced appropriately in the c-th iteration. The weight assigned to the k-th task is calculated as:

$$\lambda_k(c) = \frac{Mexp(\theta_k(c-1)/T)}{\sum_i exp(\theta_i(c-1)/T)} \tag{15}$$

where T denotes a temperature parameter that controls the smoothness of the task weighting. Here the value of T is set to 2 following Liu et al. [14]. To ensure consistency of loss value, the value of M is set to 3.

For $c = 1, 2$, we initialize the value of $\theta_k(c)$ to 1. Finally, the DWA technique is employed to ascertain each task's weight in the multi-loss function.

4.5. Loss Function

The loss functions for disease clinical diagnosis and ADAS-cog13 cognitive scores have been presented in Sections 4.3.1 and 4.3.2. For continuous variables other than ADAS-Cog13 cognitive scores, this paper uses mean absolute error (MAE) to calculate the loss, and MAE of the i-th sample is calculated as follows:

$$\mathcal{L}_i^{mae} = \frac{1}{h}\sum_{j=1}^{h}\left|g_{ij} - \hat{g}_{ij}\right| \tag{16}$$

where h is the number of remaining continuous variables, g_{ij} represents the true value of the j-th continuous variable, and $\hat{g}_{ij}$ represents the corresponding predicted value.

Extending the loss function to all time points of every sample, the total loss function is obtained as:

$$\mathcal{L}_{overall} = \sum_{t>1}\sum_{i=1}^{N}(\lambda_1\mathcal{L}_{it}^c + \lambda_2\mathcal{L}_{it}^r + \lambda_3\mathcal{L}_{it}^{mae}) \tag{17}$$

where t represents the time point index, N denotes the total number of samples, and λ_1, λ_2, and λ_3 correspond to the weights assigned to each task determined by DWA.

5. Experiments

5.1. Datasets

The data utilized in this paper are provided by the TADPOLE challenge [8], which is designed to identify the data, features and approaches that are most predictive of the future progression of subjects at risk of AD. The TADPOLE dataset comprises 1737 patient samples sourced from the ADNI database [31], encompassing a total of 12,741 recorded time points. The time point of the first data collection for each sample is called the baseline, and the subsequent time points are named according to the interval with baseline. Each sample contains a different number of time points, and there is missing time-series data in the dataset.

Although the TADPOLE challenge provides researchers with a wide range of biomarkers to choose from, this paper, similar to previous studies [6,13], only selects 22 variables among them:

- Cognitive tests: Rey's Auditory Verbal Learning Test (RAVLT)_forgetting, ADAS-Cog13, RAVLT_learning, ADAS-Cog11, RAVLT_perc_forgetting, Functional Activities Ques-

tionnaire (FAQ), Mini-Mental State Examination (MMSE), Clinical dementia rating sum of boxes (CDRSB), Montreal cognitive assessment (MOCA), RAVLT_immediate.
- PET measures: Fluorodeoxyglucose (FDG), AV45.
- MRI measures: Hippocampus, WholeBrain, Fusiform, Intracranial Volume (ICV), Ventricles, Entorhinal, MidTemp.
- CSF measures: Beta-amyloid (CSF), Total tau, Phosphorylated tau.

We also take the disease clinical diagnosis into consideration. In the dataset, disease diagnosis is divided into AD, CN, early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI), and significant memory concern (SMC). Consistent with prior studies [6,7,13], the CN and SMC groups are consolidated into a unified CN group, whereas the EMCI and LMCI groups are amalgamated to form the MCI group. As a result, the dataset is categorized into three distinct groups: CN, MCI, and AD.

5.2. Experimental Setting

In this paper, the dataset is randomly divided into 10 subsets, each time using 8 subsets as the training set, 1 subset as the validation set, and 1 subset as the test set. To ensure result stability, the random division is repeated 10 times. All variables, except for the diagnostic category, undergo z-normalization. In both the validation and test sets, each sample's first half of time points are utilized to forecast the second half of time points for the same subject. If the number of time points is odd, the last time point is excluded from the prediction. Performance evaluation metrics are computed for each division, and the average of these metrics over the 10 repetitions is considered as the final result.

In line with the TADPOLE challenge, we employ two metrics, namely multiclass area under the operating curve (mAUC), and balanced class accuracy (BCA), to assess the accuracy of diagnosis classification. A higher value of BCA and mAUC indicates superior performance in terms of classification accuracy.

mAUC can be calculated as:

$$mAUC = \frac{2}{K(K-1)} \sum_{i=2}^{K} \sum_{j=1}^{i} \hat{A}(c_i, c_j) \tag{18}$$

where K denotes the number of classes, $\hat{A}(c_i, c_j)$ is the mean AUC of any two classes, and $\hat{A}(c_i, c_j)$ can be calculated as:

$$\hat{A}(c_i, c_j) = \frac{\hat{A}(c_i|c_j) + \hat{A}(c_j|c_i)}{2} \tag{19}$$

$\hat{A}(c_i|c_j)$ is the AUC of a class c_i against another class c_j, its calculation formula is as follows:

$$\hat{A}(c_i|c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j} \tag{20}$$

where n_i and n_j represent the quantities of data points associated with classes i and j, respectively. S_i denotes the cumulative sum of the ranks assigned to the test points from class i after arranging all the data points from classes i and j in ascending order based on the probability of belonging to class i.

BCA can be obtained as:

$$BCA = \frac{1}{K} \sum_{k=1}^{K} BCA_k \tag{21}$$

where BCA_k is the balanced accuracy for class k:

$$BCA_k = \frac{1}{2} \left[\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right] \tag{22}$$

where TP, FP, TN, and FN represent the quantities of true positives, false positives, true negatives, and false negatives, respectively, for classifying as class k.

The MAE can be calculated as:

$$MAE = \frac{1}{N_p}\sum_{i=1}^{N_p}\left|\tilde{M}_i - M_i\right| \tag{23}$$

where N_p is the number of data points acquired by the time the forecasts are evaluated, M_i is i-th individual's true value, and $\tilde{M}_i$ is the corresponding predicted score.

It is important to mention that all experiments were conducted on a personal computer equipped with an AMD Ryzen 7 4800 CPU and an NVIDIA RTX 2060 GPU.

5.3. Uncertainty Estimation

To evaluate the estimated uncertainty, we visualized the distribution of uncertainty under different noises. Specifically, we introduced Gaussian noise with varying standard deviations to the test samples and obtained different uncertainty distributions. Then, we conducted a comparative analysis between these distributions. The results are shown in Figure 3 (to improve clarity, any uncertainties in the cognitive scores that exceed 1 are set to 1 in the visualization).

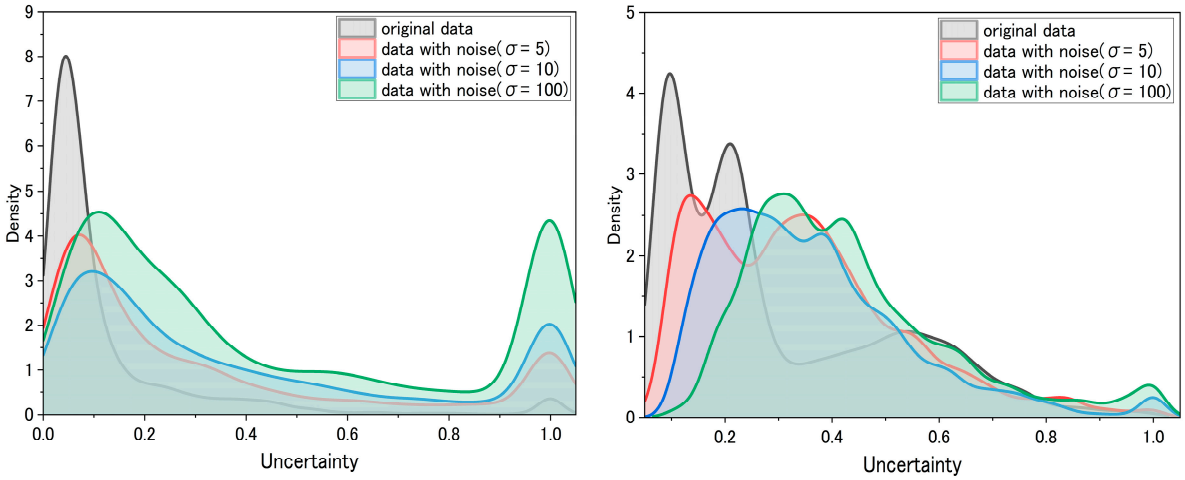


Figure 3. The uncertainty distribution of ADAS13 and diagnosis after adding noise. (**a**) ADAS-Cog13. (**b**) Diagnosis.

From Figure 3, we can analyze that the diagnostic uncertainty and ADAS-Cog13 uncertainty of the samples generally increase with the addition of Gaussian noise with different standard deviations. This is statistically manifested through a reduction in the count of low-uncertainty results and an increase in the count of high-uncertainty results, which is visually represented by a rightward shift in the overall distribution. These observations indicate the effectiveness of our model in estimating uncertainty, as it can effectively capture anomalous data and reflect low confidence in the result by assigning high uncertainty. This experiment also demonstrates the practical significance of our proposed model, which can immediately identify and submit patients with abnormal clinical data to decision makers for further judgment in real-world applications.

Figure 4 illustrates the change in the accuracy of the test sample after changing the uncertainty threshold. It shows that the lower the threshold value is set, the more accurate the results obtained by the model. In particular, when the threshold is set to 0.1, there is a huge improvement in the performance of all three metrics. This implies the trusted decisions are supported on the basis of the output (prediction and its corresponding

uncertainty) of our model. In practice, decision makers can obtain higher accuracy and reduce the risk of misclassification by moderately adjusting the uncertainty threshold.

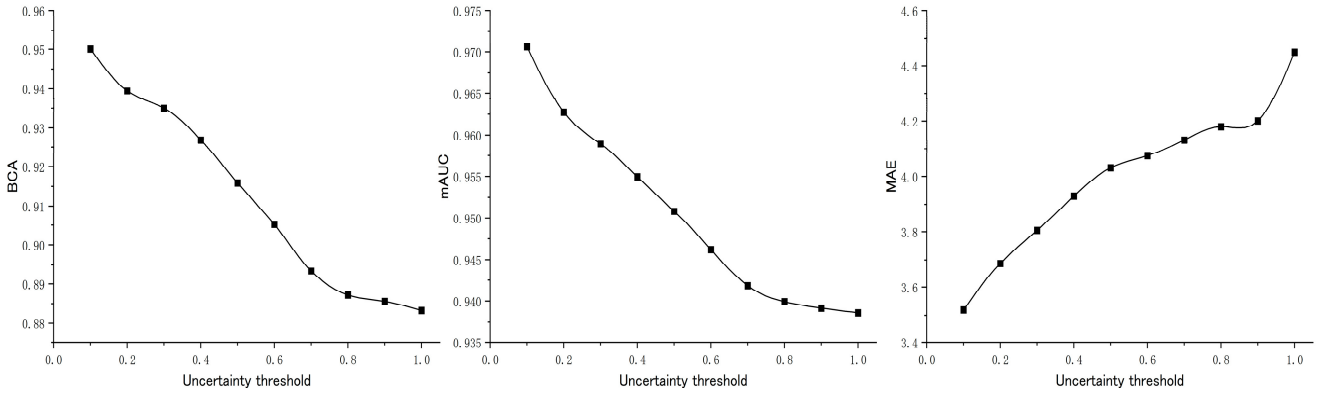


Figure 4. The variation in accuracy concerning the uncertainty threshold for EDL.

5.4. Uncertainty Benchmarking

We performed a comparative analysis between our method and the following two approaches which are designed for modeling uncertainty: (a) dropout [19] and (b) deep ensemble [20]. As shown in Table 1, EDL does not show superiority in accuracy. While EDL performs best in terms of the mAUC metric, it slightly lags behind dropout and deep ensemble in terms of BCA and adasMAE. However, when it comes to inference speed, EDL outperforms both dropout and deep ensemble, being, respectively, 70% and 74% faster. This outcome is expected when compared with dropout, which requires multiple samples and deep ensemble, which combines multiple models, EDL does not have excessive computational burden and does not require extensive modifications to the backbone network. Therefore, EDL clearly surpasses the other two methods in terms of operational efficiency and convenience, rendering it applicable to the existing deep learning models. Consequently, EDL emerges as an effective solution to the constrained acceptance of deep learning models within the medical domain.

Table 1. Performance comparison of different uncertainty methods.

Methods	mAUC (↑)	BCA (↑)	adasMAE (↓)	Inference Speed (ms) (↓)
EDL	**0.942 ± 0.0091**	0.879 ± 0.0098	4.280 ± 0.40	**0.78**
Dropout	0.939 ± 0.0135	**0.885 ± 0.0224**	4.518 ± 0.36	2.57
Deep Ensemble	0.933 ± 0.0202	0.863 ± 0.0162	**4.263 ± 0.38**	2.99

5.5. Prediction Performance

To ascertain the prediction performance of the proposed model, we conducted a comparative analysis against the following four methods:

- LSTM-F [32]: LSTM-F is an AD progression model based on a standard LSTM network. It utilizes the most recently observed values to impute missing data.
- LSTM-M [5]: LSTM-M is a variant of LSTM-F that differs in its approach to imputing missing data as it uses the mean values from the training set for imputation.
- MinRNN [13]: MinRNN is a model based on MinimalRNN which employs a technique called 'Model Filling' to impute missing data for predicting AD progression.
- GRU-MF [33]: GRU-MF is the variant of MinimalRNN in which the backbone is replaced with GRU.

Table 2 presents a comparison of various methods in clinical diagnosis and cognitive score prediction tasks. A higher value of BCA and mAUC indicates better performance of the model in classification task, while adasMAE represents the mean absolute error of

cognitive score prediction, with lower values indicating better performance in regression task. Among all the methods compared, our approach outperforms most existing methods and is only slightly inferior to MinRNN. This demonstrates that our method can accurately predict clinical diagnosis and cognitive scores, with a predictive performance comparable to similar models. Additionally, considering that our method provides a reliable measure of uncertainty in predicting results, which is lacking in similar models, it offers valuable auxiliary decision making support for AD progression prediction. Therefore, we believe that the slight shortfall in performance is acceptable, given the additional benefits our method provides.

Table 2. Model performance comparison with other methods.

Methods	mAUC (↑)	BCA (↑)	adasMAE (↓)
LSTM-F	0.922 ± 0.0124	0.869 ± 0.0117	4.299 ± 0.25
LSTM-M	0.916 ± 0.0134	0.864 ± 0.0136	4.284 ± 0.37
MinRNN	**0.944 ± 0.0080**	**0.882 ± 0.0143**	**4.160 ± 0.39**
GRU-MF	0.927 ± 0.0140	0.874 ± 0.0145	4.536 ± 0.72
Ours	0.942 ± 0.0091	0.879 ± 0.0098	4.280 ± 0.40

5.6. Effect of RNN Backbones

To examine the influence of various backbone networks on the performance of our proposed model, we replaced MinimalRNN with GRU and LSTM for our experiments. It should be noted that the loss weights in these experiments were selected without using DWA but using grid search. As shown in Table 3, among the three models, MinRNN-EDL demonstrates the best performance in terms of mAUC and BCA, while slightly trailing behind LSTM-EDL in adasMAE. This highlights the superiority of MinimalRNN as the backbone network. Moreover, the better performance of GRU and MinimalRNN in comparison to LSTM suggests that backbone networks with fewer parameters are more suitable, as they are less prone to overfitting.

Table 3. Model performance of different backbones.

Methods	mAUC (↑)	BCA (↑)	adasMAE (↓)
LSTM-EDL	0.914 ± 0.0129	0.842 ± 0.0110	**4.432 ± 0.47**
GRU-EDL	**0.937 ± 0.0068**	0.846 ± 0.0662	4.720 ± 0.83
MinRNN-EDL	**0.937 ± 0.0140**	**0.879 ± 0.0128**	4.449 ± 0.42

5.7. Ablation Study

To assess the impact of the DWA scheme, we conducted a comparative analysis against the following methods which determined the weight of the muti-loss function: (a) Equal weighting (EW), and (b) Grid search (GS). It is worth noting that we selected the best relative performance among all parameter combinations as the final result of the grid search. As shown in Table 4, compared to EW, GS's performance has an improvement in mAUC and BCA and a decrease in adasMAE. However, DWA improved in all metrics and performed best among all methods. DWA improved over EW by 0.6%, 0.8%, and 3.3% in each metric, respectively. This indicates that despite utilizing grid search, achieving a balance between different tasks with fixed weights still remains challenging. Consequently, for multi-task learning, utilizing a dynamic weight scheme may be a more favorable approach which can also be employed in other health care scenarios.

Table 4. Model performance of different weighting methods.

Methods	mAUC (↑)	BCA (↑)	adasMAE (↓)
EW	0.936 ± 0.0139	0.872 ± 0.0114	4.421 ± 0.32
GS	0.937 ± 0.0140	0.879 ± 0.0128	4.449 ± 0.42
DWA	**0.942 ± 0.0091**	**0.879 ± 0.0098**	**4.280 ± 0.40**

6. Discussion

More recently, we have witnessed the potential of deep learning methods for AD progression prediction. Wang et al. [3] used RNNs with LSTM cells for predicting future cognitive scores, while Hong et al. and Cui et al. [4,5] employed RNN variants for AD diagnosis. However, these studies focused on individual prediction tasks and did not consider the correlations among different tasks. In contrast, Jung et al. and Nguyen et al. [7,13] demonstrated the advantages of the simultaneous learning of multiple tasks, which resulted in improved prediction accuracy.

Borrowing the idea of multi-task learning, our proposed model performs AD clinical diagnosis and cognitive score predictions simultaneously. To address the unresolved issue of selecting the appropriate loss weight for each individual task in the methods of Jung et al. and Nguyen et al. [7,13], we introduce the DWA scheme. This scheme automatically adjusts each task's weight, enhancing the overall model performance. Furthermore, unlike previous methods for AD progression prediction that focused solely on enhancing the accuracy and ignored the measurement of result reliability, our model introduces the theory of evidential deep learning to provide uncertainty estimates for the model's predictions. By utilizing this uncertainty awareness mechanism, our model can help foster trust in deep learning methods among clinicians.

Despite achieving promising results, our model also has some limitations that require addressing in the future. Firstly, while the predictive performance of our model is notable, there is still room for improvement. To enhance prediction accuracy further, incorporating the attention mechanism into the model could prove beneficial. Secondly, utilizing only the clinical diagnosis and 22 continuous variables as features is insufficient to fully characterize the course of AD. There likely exist supplementary features that could aid in detecting and discerning subtle changes in a patient's disease progression. Consequently, future research efforts should pay attention to better extracting AD features from the dataset, thereby facilitating a deeper comprehension of the disease's evolution and enhancing the model's predictive capacities.

7. Conclusions

In this paper, we present a novel approach for the trustworthy prediction of AD progression by proposing a multi-task evidential sequence learning model. Specifically, given the longitudinal historical information of 22 AD multimodal biomarkers and clinical diagnosis, we select MinimalRNN as the backbone for multi-task sequence learning to simultaneously predict future clinical diagnoses and ADAS-Cog13 cognitive scores at multiple time points. Meanwhile, we introduce evidential deep learning theory to provide an uncertainty estimation for each diagnosis and a cognitive score prediction which can measure the reliability of model results. Additionally, we introduce the Dynamic Weight Average (DWA) weighting scheme to automatically allocate learning resources to different tasks in the process of training. Finally, the experimental results of the TADPOLE challenge dataset validate the effectiveness of our proposed model. Notably, the core idea of our proposed approach exhibits a generalizability that extends to various deep learning applications, thereby contributing to the advancement of trustworthy methodologies in the healthcare domain of deep learning models.

Author Contributions: Conceptualization, L.C. and Z.Z.; methodology, L.C. and Z.Z.; software, Z.Z.; validation, Z.Z., P.L., Y.D. and Z.M.; formal analysis, Z.Z. and P.L.; investigation, Z.Z., P.L. and Y.D.; resources, Z.Z., P.L. and L.C.; data curation, Z.Z. and Y.D.; writing—original draft preparation, Z.Z.; writing—review and editing, P.L. and L.C.; visualization, Z.Z.; supervision, P.L., Z.M. and L.C.; project administration, L.C.; funding acquisition, P.L. and L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded partially by the National Natural Science Foundation of China under grant 62006126, Key Research and Development Program of Jiangsu Province under Grant BE2021093, Natural Science Foundation of Jiangsu Province under grant BK20200740, Natural Science Foundation of the Jiangsu Higher Education Institutions of China under grant 20KJB520004, Natural Science Research Start-up Foundation of Recruiting Talents of Nanjing University of Posts and Telecommunications under grant NY219150.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Scheltens, P.; Blennow, K.; Breteler, M.M.; De Strooper, B.; Frisoni, G.B.; Salloway, S.; Van der Flier, W.M. Alzheimer's disease. *Lancet* **2016**, *388*, 505–517. [CrossRef] [PubMed]
2. Mohs, R.C.; Knopman, D.; Petersen, R.C.; Ferris, S.H.; Ernesto, C.; Grundman, M.; Sano, M.; Bieliauskas, L.; Geldmacher, D.; Clark, C. Development of cognitive instruments for use in clinical trials of antidementia drugs: Additions to the Alzheimer's Disease Assessment Scale that broaden its scope. *Alzheimer Dis. Assoc. Disord.* **1997**, *11*, 13–21. [CrossRef]
3. Wang, T.; Qiu, R.G.; Yu, M. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. *Sci. Rep.* **2018**, *8*, 9161. [CrossRef] [PubMed]
4. Hong, X.; Lin, R.; Yang, C.; Zeng, N.; Cai, C.; Gou, J.; Yang, J. Predicting Alzheimer's disease using LSTM. *IEEE Access* **2019**, *7*, 80893–80901. [CrossRef]
5. Cui, R.; Liu, M.; Li, G. Longitudinal analysis for Alzheimer's disease diagnosis using RNN. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1398–1401.
6. Liang, W.; Zhang, K.; Cao, P.; Liu, X.; Yang, J.; Zaiane, O. Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network. *Comput. Biol. Med.* **2021**, *138*, 104935. [CrossRef]
7. Jung, W.; Jun, E.; Suk, H.-I.; Alzheimer's Disease Neuroimaging Initiative. Deep recurrent model for individualized prediction of Alzheimer's disease progression. *NeuroImage* **2021**, *237*, 118143. [CrossRef]
8. Marinescu, R.V.; Oxtoby, N.P.; Young, A.L.; Bron, E.E.; Toga, A.W.; Weiner, M.W.; Barkhof, F.; Fox, N.C.; Klein, S.; Alexander, D.C. TADPOLE challenge: Prediction of longitudinal evolution in Alzheimer's disease. *arXiv* **2018**, arXiv:1805.03909.
9. Ghazi, M.M.; Nielsen, M.; Pai, A.; Cardoso, M.J.; Modat, M.; Ourselin, S.; Sørensen, L.; Alzheimer's Disease Neuroimaging Initiative. Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. *Med. Image Anal.* **2019**, *53*, 39–46. [CrossRef]
10. Xu, L.; Wu, H.; He, C.; Wang, J.; Zhang, C.; Nie, F.; Chen, L. Multi-modal sequence learning for Alzheimer's disease progression prediction with incomplete variable-length longitudinal data. *Med. Image Anal.* **2022**, *82*, 102643. [CrossRef]
11. Utsumil, Y.; Rudovicl, O.O.; Petersonl, K.; Guerrero, R.; Picardl, R.W. Personalized gaussian processes for forecasting of alzheimer's disease assessment scale-cognition sub-scale (adas-cog13). In Proceedings of the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 18–21 July 2018; pp. 4007–4011.
12. Ning, K.; Cannon, P.B.; Yu, J.; Shenoi, S.; Wang, L.; Alzheimer's Disease Neuroimaging Initiative; Sarkar, J. Characterizing brain imaging features as-sociated with ADAS-Cog13 sub-scores with 3D convolutional neural networks. *bioRxiv* **2022**. [CrossRef]
13. Nguyen, M.; He, T.; An, L.; Alexander, D.C.; Feng, J.; Yeo, B.T.; Alzheimer's Disease Neuroimaging Initiative. Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage* **2020**, *222*, 117203. [CrossRef] [PubMed]
14. Liu, S.; Johns, E.; Davison, A.J. End-to-end multi-task learning with attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1871–1880.
15. Chen, Z.; Badrinarayanan, V.; Lee, C.-Y.; Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 794–803.
16. Xie, Z.; Chen, J.; Feng, Y.; Zhang, K.; Zhou, Z. End to end multi-task learning with attention for multi-objective fault diagnosis under small sample. *J. Manuf. Syst.* **2022**, *62*, 301–316. [CrossRef]
17. Xie, K.; Huang, L.; Zhang, W.; Qin, Q.; Lyu, L. A CNN-based multi-task framework for weather recognition with multi-scale weather cues. *Expert Syst. Appl.* **2022**, *198*, 116689. [CrossRef]
18. Jospin, L.V.; Laga, H.; Boussaid, F.; Buntine, W.; Bennamoun, M. Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Comput. Intell. Mag.* **2022**, *17*, 29–48. [CrossRef]
19. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
20. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–12.
21. Amini, A.; Schwarting, W.; Soleimany, A.; Rus, D. Deep evidential regression. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 14927–14937.

22. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 01768.
23. Bao, W.; Yu, Q.; Kong, Y. Evidential deep learning for open set action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13349–13358.
24. Ghesu, F.C.; Georgescu, B.; Gibson, E.; Guendel, S.; Kalra, M.K.; Singh, R.; Digumarthy, S.R.; Grbic, S.; Comaniciu, D. Quan-tifying and leveraging classification uncertainty for chest radiograph assessment. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Proceedings, Part VI 22, Shenzhen, China, 13–17 October 2019; pp. 676–684.
25. Soleimany, A.P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S.N.; Coley, C.W. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **2021**, *7*, 1356–1367. [CrossRef]
26. Chen, M. Minimalrnn: Toward more interpretable and trainable recurrent neural networks. *arXiv* **2017**, arXiv:1711.06788.
27. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
28. Dempster, A.P. A generalization of Bayesian inference. *J. R. Stat. Soc. Ser. B (Methodol.)* **1968**, *30*, 205–232. [CrossRef]
29. Jsang, A. *Subjective Logic: A Formalism for Reasoning under Uncertainty*; Springer: Berlin/Heidelberg, Germany, 2018.
30. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7482–7491.
31. Jack, C.R., Jr.; Bernstein, M.A.; Fox, N.C.; Thompson, P.; Alexander, G.; Harvey, D.; Borowski, B.; Britson, P.J.; Whitwell, J.L.; Ward, C.; et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging Off. J. Int. Soc. Magn. Reson. Med.* **2008**, *27*, 685–691. [CrossRef] [PubMed]
32. Lipton, Z.C.; Kale, D.C.; Wetzel, R. Modeling missing data in clinical time series with RNNs. *Mach. Learn. Healthc.* **2016**, *56*, 253–270.
33. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. mboxempharXiv **2014**, arXiv:1412.3555.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.