Received: January 19, 2024. Revised: February 24, 2024. 

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

# Random Forest-Based Survival Analysis for Predicting the Future Progression of Brain Disorder from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD)

**Nour Zawawi**[1] **Nermin Negied**[2]*

[1]*Faculty of Computer Science, October University for Modern Science and Arts, Giza, Egypt*
[2]*School of Information Technology and Computer Science, Nile University, Giza, Egypt*
* Corresponding author's Email: nnegied@nu.edu.eg

**Abstract:** The race to halt Alzheimer's disease (AD) in its tracks demands an early warning system. By predicting which mild cognitive impairment (MCI) patients are likely to decline into AD, clinicians can intervene while the window of opportunity remains open. But how to separate the MCI patients bound for AD from those with more benign forms of impairment? The key lies in examining the factors that influence disease progression. While prior studies have scratched the surface, a comprehensive analysis has proven elusive. Enter the Alzheimer's Disease Neuroimaging Initiative database, which tracks AD progression through a wealth of patient characteristics. Leveraging these rich data, our hybrid approach combines survival analysis with machine learning to generate dynamic predictions of time to AD onset. Rather than merely detecting AD early or diagnosing its current state, our model gazes into the future, forecasting progression from MCI to AD before the disease fully erupts. Among similar efforts, the proposed approach stands apart in scale and accuracy, validated on more patients and with higher predictive power than earlier attempts. Even cognitive tests or brain scans alone can foretell decline, with the proposed work achieving a remarkable C-index of 0.85 when evaluated using the whole ADNI dataset not only a sample from it. By revealing who is likely to convert to AD and when, this work enables clinicians to intervene at the critical junction where MCI transitions to inevitable decline. The future of AD treatment may hinge on such early warnings.

**Keywords:** Alzheimer's disease, Prediction of future AD, Survival analysis, Disease progression, Random forest.

## 1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative disorder with mental, cognitive, and structural deteriorations that accounts for 60% to 80% of dementia cases [1]. Alzheimer's disease causes 5% of deaths in the U.S. It's the seventh leading cause of mortality for 65-and-overs [2]. There is currently no treatment that can reverse the effects of Alzheimer's disease (AD), so there is a significant amount of focus on research aimed at developing a deeper understanding of the condition as well as techniques for determining who is at risk for developing Alzheimer's disease before the onset of symptoms. Because of the mentioned facts, it is extremely important to recognize patients who are at a high risk of acquiring Alzheimer's disease [1]. In the same vein, early detection is necessary for the development of a treatment strategy that would slow down the advancement of the disease. It is when one symptom leads to another symptom in the progression of the illness.

There are numerous causes for the onset and progression of Alzheimer's disease; however, it is unknown how much each one contributes to the disease. Consequently, it is essential to study exhaustively the implications of data emanating from diverse sources with varying statistical characteristics and missing data. Recent research has proven that many sources of clinical data might provide complementary information, such that combining multiple sources of data improves the

prediction of cognitive decline over utilizing a single source [3]. It presents analytical challenges, such as the risk of overfitting the approach to the data, the approach's inability to generalize to fresh data, and the large variance of approaches fitted to this data. Therefore, picking the appropriate characteristics can assist the inductive learner in enhancing their learning velocity, generalization ability, and induced approach simplicity [4].

In the recent years, considerable effort has been invested in the development of machine learning algorithms that can measure multiple predictors for a more precise risk assessment. Most of the machine learning has been accomplished through classification algorithms [5-7], which divide subjects into two groups, such as stable or progressive Mild Cognitive Impairment (MCI). On the other hand, classification methods cannot deal well with missing data, and there are various ways to define who belongs in a class. Survival analysis (SA) has recently been utilized as a superior method for calculating the risk of AD in MCI [8].

Random Forest is a powerful and versatile machine learning algorithm that has found significant applications in survival analysis. Survival analysis deals with predicting the time until an event of interest occurs, such as death or failure, and understanding the factors that influence it. Random Forest's importance in survival analysis stems from its ability to handle high dimensional data, handle missing values, and capture complex nonlinear relationships between predictors and survival outcomes. Moreover, Random Forest can effectively handle interactions and non-proportional hazards, which are common challenges in survival analysis. By aggregating multiple decision trees and using ensemble learning, Random Forest can provide more accurate and robust predictions, reducing the risk of overfitting. Additionally, Random Forest offers feature importance measures, enabling researchers to identify the most influential predictors for survival outcomes. This information is crucial for understanding the underlying mechanisms and making informed decisions in various fields, including healthcare, finance, and engineering. Overall, Random Forest plays a pivotal role in survival analysis by providing reliable predictions, interpretability, and valuable insights into the factors influencing survival outcomes.

The objective of using Random Forest in survival analysis for predicting the conversion from MCI to Alzheimer's Disease (AD) is to develop a predictive model that can effectively identify individuals who are at a higher risk of progressing to AD. The conversion from MCI to AD is a critical stage in the disease progression, and early detection plays a crucial role in providing appropriate care and intervention. By utilizing Random Forest in survival analysis, we aim to leverage its strengths in handling complex and high-dimensional data to capture the multifaceted nature of MCI-to-AD conversion. Random Forest's ability to handle interactions, non-linear relationships, and missing values allows for the inclusion of various predictors such as demographic information, cognitive assessments, genetic markers, and imaging data. By aggregating multiple decision trees, Random Forest can provide reliable predictions while mitigating the risk of overfitting. Additionally, by assessing the feature importance within the Random Forest model, we can identify the key predictors that contribute significantly to the prediction of conversion. Ultimately, the objective is to develop a robust and accurate predictive model that can aid clinicians and researchers in identifying individuals at higher risk of MCI-to-AD conversion, enabling targeted interventions and personalized care to potentially delay or prevent the onset of Alzheimer's disease. The rest of the work is organized as follows: Section 2 discusses the related work presented in the literature for identifying AD. Section 3 describes the scientific approaches, methods, and data. Section 4 explores the experimental work to illustrate the results and discussing them. Finally, Section 5 concludes the findings and results of the paper.

## 2. Related work

Medical imaging and its applications have been widely attracted the attention of researchers for decades. Different types of scanning techniques and the analysis of them can help physicians in accurately diagnose the disease, not only that but also aid in early detection of dangerous diseases like tumours and AD [9-12]. The emergence and development of machine learning and deep learning techniques have also shared in the development of this research area. The following subsections focus on the work done in literature in the field of AD detection from medical scans.

### 2.1 Prediction based on MRI only

Most Alzheimer's disease (AD) research has relied only on medical imaging [13]. Using structural MRI data, Liu et al [14] proposed a multi-model deep learning framework based on convolutional neural networks (CNN) for automated hippocampal segmentation and AD classification. The authors achieved an accuracy rate of 88% in

AD classification, and 77% in MCI classification, but the authors limited their study to the hippocampus area only. Liu et al. [15] proposed a novel method for extracting ROI features and interregional features based on multiple measures from MRI images. The authors used six different anatomical measures to obtain six node feature sets and six edge feature sets, then they applied MKBoost algorithm to obtain the best classification accuracy from each feature set to select the best feature set, however, the work is very time and resource consuming with no remarkable enhancements in accuracy rate.

Basheera and Ram [16] proposed a framework to predict differentiate between MCI and CN in order to detect AD at an early stage and they confirmed that they obtained high accuracy, but the noticeable drawback is the small size of the used dataset as they used only 4463 images for both training and testing. Basheer et al. [17] suggested a method based on deep learning to automatically measure the hippocampus volume without prior segmentation of the volumetric MRI scans. The authors developed a 2D convolutional neural network (CNN) model that uses 3-channel 2-D patches to predict the number of voxels that belong to the hippocampus to detect AD accordingly, but the results didn't exceed 84% for the right hippocampus and 83% for the left hippocampus. Castro et al. [18] main goal was to make a system that automatically finds the disease in sagittal magnetic resonance images (MRI). The authors confirmed that Deep Learning models and transfer learning could be effective in this field, but nothing clear was mentioned about the results. Zao et al. [19] tried to find out if the radiomic features of the hippocampus are useful in reliably classifying MRI markers for (AD) using multivariate support vector machine (SVM). The authors succeeded to reach an accuracy rate of 88.2%, but the number of patients used to evaluate this study was small, as they used scans of 261 patients only.

In summary, while MRI is a valuable tool in AD diagnosis, relying only on MRI for prediction has limitations in terms of specificity, sensitivity in early stages, subtype differentiation, cost and accessibility, lack of functional information, and the need for longitudinal data. Integrating MRI with other biomarkers and clinical assessments can enhance the accuracy of AD prediction and improve detection outcomes.

### 2.2 Prediction based on multiple features

Cai et al. [20] also used SVM and least squares loss function to select the optimal subset of embedded features, but the authors only reached an accuracy rate of 71.17%. Shikalgar and Sonavane [21] employed a multimodal data classifier using a hybrid deep neural network (NNs) that takes EEG inputs to classify MRI images. The authors' aimed at improving the learning process by incorporating the weight components that uncover relationships between brain areas and genes to the NNs. The authors confirmed that they achieved an accuracy rate that exceeded 98% but the dataset used was too small to validate the results, as they only used 512 MRI scans, which suggests an overfitting.

Bi et al. [22] proposed "brain region-gene pairs" as the sample's multimodal characteristics for detecting relationships between brain regions and genes. In addition, cluster evolutionary random forest (CERF), a novel data analysis technique suitable for "brain region-gene pairs," was introduced, but the authors only used a small sample from ADNI dataset. Qiu et al. [23] explained how the data obtained from MRI scans is useful in improving the accuracy of diagnoses, especially for the Mini-Mental State Examination (MMSE) and logical memory (LM) tests, but the authors confirmed that their work was just proof of principle that multimodal fusion of models developed using MRI scans, and still needs to be validated in large number of scans.

Alexiou et al. [24] discussed how the abnormal testing in one or more biomarkers can cause the development or presence of Alzheimer's disease, but the authors confirmed that a validation on a large dataset is still needed to measure how effective is their approach. Venugopalan et al. [25] used deep learning (DL) to integrate MRI, genetic and clinical test data to classify patients into AD and MCI. The authors used 3D-convolutional neural networks (CNNs) for imaging data and extract features from clinical and genetic data, the authors just confirmed that their multimodal approach has outperformed other single models in literature. Zawawi et al. [26] developed a new hybrid model for extracting features from medical data by combining three different types of features: MRI, seven different types of neurologic tests, and baseline diagnosis. The authors confirmed that the three used features outperformed most of the well-known features, the authors succeeded in achieving an accuracy that exceeded 90%, but the main drawback was in the long time needed for diagnoses. Later than that, the same authors improved their results and reduced

time needed for features' extraction, but they confirmed that they used only a small sample of instances [27].

In summary, while combining MRI with other biomarkers or assessments can enhance AD prediction, there are limitations in terms of increased complexity and cost, lack of standardized protocols, unclear added value, limited availability or invasiveness of complementary techniques, interpretation challenges, and limited prediction accuracy. These drawbacks highlight the need for ongoing research, standardization efforts, and careful consideration of the practicality and clinical utility of combining multiple techniques in AD prediction.

### 2.3 Prediction based on time series

Hong et al. [28] used a deep learning approach to differentiate between MCI and AD. The authors relied on LSTM, and they succeeded in differentiating between MCI and AD with an accuracy rate of only 78%. Yang et al. [29] went for using the linear regression, and they succeeded to discover that the corpus callosum (CC) atrophy increases AD development, but they confirmed that they have a drawback which is the poor prediction time. The authors also confirmed that their suggested approach could appropriately restore the failure time in right censoring. Zawawi et al. [30] proposed a Prediction Model to capture the conditions between characteristics and the next stage. The authors confirmed that their model successfully recognized the affected brain regions across both MRI and Neurological data sets, but the authors mentioned that their system needs to be validated using actual data.

Mirabnahrazam et al [31] used a deep learning approach to predict the time to AD conversion. The authors used survival analysis model that extends the traditional Cox regression model. They also discovered that the genetic factors had the least impact on survival analysis, while cognitive tests, demographics, and CSF features had the greatest impact. The authors used large sample of ADNI dataset to validate their work, and they succeeded to measure disease progression with a C-index of 80%.

Compared with methods that focus on the goal of binary classification accuracy at a fixed threshold, our interest lies in modelling an approach that dynamically predicts the progression from MCI to AD over next eight years. Therefore, this approach can be transferred into clinically prediction approach.

The last research conducted by Sarica et al [32] proposed an intriguing Random Survival Forests model for survival prediction. They found that optimizing hyperparameters yielded strong c-index scores of 0.798 with a Cox model and an even better 0.85 with their proposed Random Survival Forest. The drawbacks of Sarica's work include the small size of the used dataset, and most importantly they calculated the C-index individually for every parameter then they calculated the average which might be not an accurate metric for such a problem.

Time series analysis can be a valuable approach for AD prediction, it has limitations related to data availability and quality, variability, and noise in longitudinal data, limited temporal resolution, the non-linear and dynamic nature of AD progression, generalizability and external validity, and interpretability and clinical utility. Addressing these challenges requires careful consideration of study design, data collection protocols, modelling techniques, and validation strategies to enhance the accuracy and applicability of AD prediction using time series analysis.

Using survival analysis is important in AD prediction as it handles censoring, accommodates time-dependent covariates, estimates cumulative probabilities, incorporates competing risks, enables individualized risk prediction, and facilitates longitudinal data analysis. By utilizing survival analysis techniques, researchers and clinicians can gain valuable insights into the timing, risk factors, and progression of AD, leading to improved prediction models and personalized approaches for early detection and intervention.

This paper addresses the limitations of all previous approaches found in literature, beside that it presents a novel way in the disease progression detection and analysis by studying cases that have been monitored for seven years not only four like other methodologies in literature work (See sections 3 & 4). At the end from all the work done in literature related to AD, Khajehpiri et al [8], Mirabnahrazam et al [31], and Sarica et al [32] are the only state-of-the-art works used to compare survival analysis for predicting the future progression of MCI to AD, which is the main target of this work also. Because of that, a comparison between the proposed approach's results and their results can be found in section 5 (see table 4).

## 3. Materials & methods

This section explains the proposed approach, how and why is has been selected, the dataset used, and how it was prepared and processed. The following subsection demonstrates the proposed work in this paper step by step.

### 3.1 Dataset description

Data from the Alzheimer Disease Neuroimaging Initiative (ADNI) was used in this work as this is the largest publicly available dataset for this kind of problems [33]. The ADNI clinical dataset comprises clinical information about each subject including recruitment, demographics, physical examinations, and cognitive assessment data. The dataset also contains variety of scans and data representations like MRI, PET scans. Genetic data, cognitive tests data, demographics, etc. (see fig.1). Amongst all the previously mentioned scans and data representations, only MRIs and Neurological data were useful in our problem (see fig. 2 and 3). This work, unlike state-of-the-art work done in literature used the whole dataset not only a sample of it. The whole dataset contains a time series of measurements from 1,737 patients at each visit was used to evaluate the proposed approach. The used dataset contains at least 9 different scans for every patient, which means 1,737 x 9 = 15633 MRI images with their corresponding neurological data. The ADNI dataset has been instrumental in advancing our understanding of AD, identifying biomarkers, developing diagnostic criteria, and facilitating the development and evaluation of new therapeutic interventions. Its rich and diverse data continue to support research efforts aimed at improving early detection, monitoring disease progression, and developing effective treatments for AD.

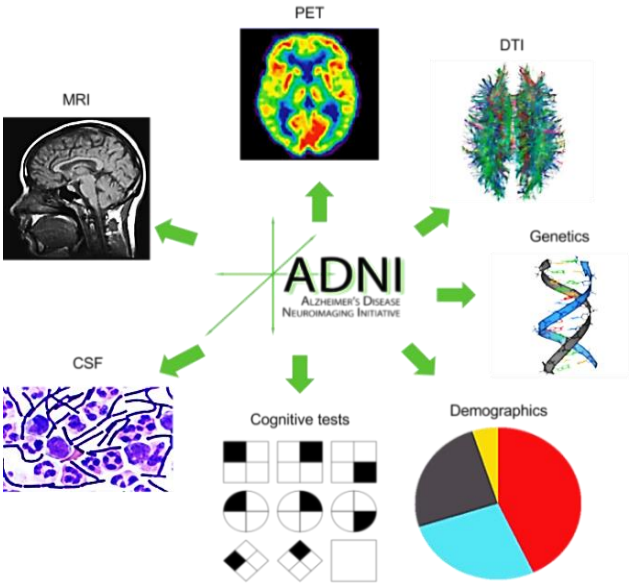


Figure.1 Different data representations in ADNI dataset [33]

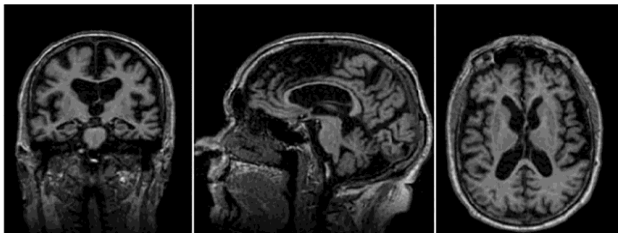

Figure. 2 Sample MRI scans from ADNI dataset showing coronal, sagittal, and axial plans

| | ADNI dataset | | | | |
|---|---|---|---|---|---|
| Scan type | Ax | Cor | Sag | 3D | Total |
| T1w | 0 | 0 | 276 | 2380 | 2656 |
| T1wC | 0 | 0 | 0 | 0 | 0 |
| T2w | 1725 | 488 | 5 | 0 | 2218 |
| PDw | 1069 | 0 | 0 | 0 | 1069 |
| T2w-FLAIR | 1 | 0 | 3 | 488 | 492 |
| DWI | 558 | 0 | 2 | 0 | 560 |
| PWI-DSC | 0 | 0 | 0 | 0 | 0 |
| Derived | 183 | 0 | 2 | 47 | 232 |
| Total | 3536 | 488 | 288 | 2915 | 7227 |

Figure. 3 Sample Neurological data from ADNI dataset

### 3.2 Subjects

The suggested prediction approach was trained and tested on ADNI data using 18-month longitudinal trajectories of 900 cases per class (norm); Contains a total of 1800 items. Including 24 neurological tests associated with MRI. Each patient profile consisted of 24 tests and 7 image files. Patient trajectories described temporal changes in falls variables over intervals that are three months each. The detailed data processing steps are described in the following sections. The inclusion criteria used in this study was as follows: 1) 55 to 90 years; 2) Education level from primary education to higher education institutions; 3) People of all colours and strokes. The different types of data used in this study is as follows: 1) Neurological examination (neuropsychologist); 2) Criteria: initial testing and patient diagnosis; 3) Brain imaging techniques (MRI only).

### 3.3 Data preprocessing

Data pre-processing step is done to get the data ready and in suitable format to be inputted to the machine learning classifiers. We assumed that some data are missing at random, and that filtering is not informative because missing values are frequently seen in medical datasets. After that, imputation is used to replace the missing values with the correct ones. Missing data is often imputed using multiple

imputations, which maintain the relationships between the data and the uncertainty in those relationships.

The dataset used here included variety of cases such as patients with cognitive typical Cognitively Normal (NL), patients with mid-cognitive impairment (MCI), and patients with Alzheimer's Disease (AD) who underwent six-month follow-up exams after being recruited from more than fifty different US and Canadian centers. When it comes to brain state analysis there are typically nine classes, however, only two classes can effectively exhibit the concept of AD survival time prediction. Those two classes are namely: the AD and the MCI.

The assessment data was primarily made up of neurological tests, medical scan dates, and MRI results. Imputation was carried out during the cross-validation cycle, but only on the training set, using the prediction matrix that was created on that set.

The features obtained from the ADNI dataset after preprocessing are 12 features. Those 12 features can be categorized as follows: one categorical, three ordinals, and the rest are numeric representations of the images. The last step in the data preprocessing was the normalization step, where any feature with more than 60% of its values missing would be excluded from the dataset.

### 3.4 Method

There are several reasons for selecting Random Forest for survival analysis in predicting the conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD). Following is a list of those reasons:

1) Handling Complex and Nonlinear Relationships: Random Forest can capture complex and nonlinear relationships between predictors and survival outcomes. This is particularly important in analysing medical data where the relationship between variables can be intricate and non-linear.
2) Robustness to Overfitting: Random Forest has built-in mechanisms to reduce the risk of overfitting, which occurs when a model performs well on the training data but fails to generalize to new data. By aggregating multiple decision trees and using ensemble learning, Random Forest provides more robust predictions and reduces the risk of overfitting.
3) Dealing with High-Dimensional Data: Random Forest can effectively handle high-dimensional data, allowing for the inclusion of a wide range of predictors such as demographic information, clinical assessments, genetic markers, and imaging data. This enables a comprehensive analysis that considers multiple factors influencing the conversion from MCI to AD.
4) Handling Missing Values: Random Forest is capable of handling missing values in the dataset, reducing the need for extensive data imputation or exclusion of samples with missing values. This is particularly advantageous in real-world scenarios where missing data is common.
5) Variable Importance Measures: Random Forest provides variable importance measures, allowing researchers to identify the most influential predictors for the conversion from MCI to AD. This information can provide valuable insights into the underlying mechanisms and help prioritize specific factors for further investigation.
6) Flexibility and Adaptability: Random Forest can be applied to various types of survival data, including right-censored data commonly encountered in survival analysis. It can handle time-dependent covariates, time-varying effects, and non-proportional hazards, making it suitable for analysing dynamic and complex survival data.

Overall, it can be said that Random Forest is a powerful and versatile algorithm for survival analysis, offering robust predictions, handling complex relationships, accommodating high-dimensional data, and providing valuable insights into variable importance. Those factors make it a suitable choice for predicting the conversion from MCI to AD, where accuracy and interpretability of results are crucial for early identification and diagnosis.

Detecting the true survival time for a patient is of ultimate importance in the field of healthcare. Accurately determining how long a patient is expected to survive can have significant implications for treatment decisions, care planning, and patient outcomes. By identifying the true survival time, healthcare professionals can tailor interventions and therapies to meet the individual needs of the patient, ensuring the most effective and appropriate care. Additionally, accurate survival time prediction enables healthcare providers to offer patients and their families' realistic expectations, enabling them to make informed decisions regarding end-of-life care, advanced directives, and putting future plans. Detecting the true survival time empowers medical professionals to optimize care, provide compassionate support, and enhance the overall quality of life for the patient during their coming lifetime.

After data preprocessing and normalization, the procedure starts with the analysis of the raw data and guides the reader through the process of assessing the outcomes using the cross-validation penalty method. The following subsections explain the general steps of our approach.

### 3.4.1. Inputting pre-processed data

The suggested approach uses three types of data: The MRI image, the results of a neurological examination, and the diagnosis at the first visit. It includes 68 features (26 neurological tests, 7 MRI and 35 key indices). Each represents a patient record. The first step of the proposed approach used patient data as input. Missing values exclusion is handled at this stage.

### 3.4.2. Feature selection

Such massive data recording requires a significant investment in computing power and money. This fact indicates that the suggested method chooses the best attributes to enhance performance. The factor that most effectively explains patient survival is the result of this stage.

### 3.4.3. Running fandom forest algorithm

The final step is the estimation of the prognosis of Alzheimer's disease patients using random forests for Alzheimer's disease analysis. The methodological approach presented here offers a new scientific approach to the study and interpretation of other methods.

Most machine learning algorithms contain one or more hyperparameters that should be selected to optimize the performance of the approach. In this work, a 10-fold nested cross-validation loop was used to automatically tune these hyperparameters. In the inner loop, the hyperparameter values were selected by a random search over 25 iterations. Performance was evaluated on an outdoor cycle. Therefore, all sampling processes were performed for each combination of training and testing data.

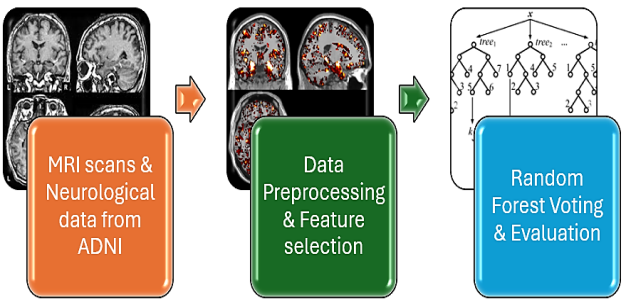


Figure. 4 The architecture of the proposed approach

Figure 4 shows an overall architecture of the proposed approach with its three main layers.

### 3.4.4. Survival analysis

The main purpose of survival analysis is to estimate when an event is likely to occur or to predict time until an event, such as when mild cognitive impairment (MCI) will progress to Alzheimer's disease (AD). Survival analysis can handle right-censored data which occurs when the event of interest is not observed because the study ends first. This happens with stable mild cognitive impairment patients. The waiting time until an event happens is defined as a positive random variable T. Given T's probability density function f(t), the cumulative distribution function is:

$$F(t) = \Pr[T < t] = \int_{-\infty}^{t} f(u)du \qquad (1)$$

The survival probability S(t) that the event of interest has not occurred by some time t is:

$$S(t) = 1 - F(t) = \Pr[T < t] \qquad (2)$$

The hazard function h(t) represents the approximate probability that an event will occur in the small interval [t, t + dt]. Meanwhile, the cumulative hazard function H(t) equals the integral of h(t) over the interval [0, t]. For a discrete time interval subdivided into J parts, the risk score for a sample x is calculated as:

$$r(x) = \sum_{j=1}^{J} H(t_j, x) \qquad (3)$$

The Cox proportional hazard model (CPH) [34] is a semi-parametric approach that makes parametric assumptions about how predictors affect the hazard function but does not assume any particular form for the baseline hazard function itself. The Cox model hazard function h(t) can be estimated as:

$$h(t, \vec{x}i) = h0(t)\eta(\vec{x}i) \qquad (4)$$

The variable h0(t) denotes the unknown baseline hazard function, which represents the hazard when all predictor values are zero. The risk function,

typically defined as a linear representation, is written as $\eta(\vec{x}i)$:

$$\eta(\vec{x}i) = e^{\sum_{j=1}^{p} x_j^i w_j} \quad (5)$$

where $w_j$ are the coefficients to determinate and $\vec{x}i$ is the observed feature vector.

The Cox proportional hazards (CPH) model estimates parameters by maximizing the partial likelihood function, allowing predictors to have a multiplicative effect on the hazard function. A major advantage of CPH is the ability to interpret results similarly to regression models. However, CPH can yield inaccurate standard deviations for estimators when faced with high-dimensional data and few observations.

### 3.5 Approach evaluation

The evaluation of the proposed approach involves assessing its performance and reliability in accurately identifying MCI patients who are at risk of MCI to AD conversion. Several evaluation measures can be used to evaluate the model's predictive capabilities. The Concordance Index (C-index) [34,35] is a common metric used in survival analysis to measure the model's ability to rank individuals based on their predicted risk of conversion.

$$C - index = \frac{N.ConcordantPair}{N.ComparablePair} \quad (6)$$

A higher C-index indicates better discriminatory power, with values closer to 1 indicating stronger predictions. Additionally, time-dependent measures such as the time-dependent Area Under the Curve (AUC) can be employed to assess the model's performance at different time points during the follow-up period [36,37]. This provides insights into the model's predictive accuracy over time and its ability to capture changes in risk over the disease progression. Calibration plots can also be used to assess the calibration of the model by comparing the predicted probabilities of conversion with the observed conversion rates across different risk groups. Furthermore, internal, and external validation techniques can be employed to evaluate the model's generalizability across different datasets and populations. Internal validation, such as cross-validation, can assess the model's performance on the same dataset used for training, while external validation involves testing the model on independent datasets to verify its performance in real-world scenarios. The evaluation of a Random Forest survival analysis model for predicting conversion from MCI to AD involves a comprehensive assessment of its discriminative power, calibration, and generalizability using appropriate evaluation metrics and validation techniques.

Good classification results should produce two non-overlapping classes or sets of probabilities, namely positive class, and adverse events class. Calibration is the degree to which the expected probability matches the actual events quantitatively [38]. When the observed and expected values for every conceivable grouping of the data, arranged by increasing predicted values, concur, the technique is considered well-calibrated. Statistics that split a data set into classes and contrast the average expected probability with the result's prevalence in each class are known as calibration measures.

The null hypothesis uses a paired t test to test whether the means of two sets of values are equal. Since many training and test sets may overlap, the assumption of the t-test is that the two sets of values in question are randomly selected to test the performance of both approaches (e.g., k-fold repeated cross-validation).

The experiment was carried out under the following conditions. The R package Machine Learning in R (MLR) was used as a framework to perform the comparison, and all code for the experiments was written in R. AD patient survival analysis additionally makes use of survminer and survival rates from the R package. Using five iterations of 10-fold stratified cross-validation, all resampling was done. 70% of the data are training data, while the remaining 30% are test data.

## 4. Results and discussion

In this paper, a new survival analysis approach to detect the progression probability of an MCI patient to be converted to an AD patient was proposed and discussed. In this work the shape of the survival function is not as important as the probability of progression.

The proposed work computes the predicted survivor function for AD patient using RF survival approach. Table 1 discusses the danger of the progression the patients are prone to over 48 months. It illustrates the following criteria:

- n.risk: Number of patients at risk of being diagnosed with AD each time.
- n.event: There are 238 events that occurred initially. This number then begins to decrease until it finally reaches 23.

Table 1. Patients' Proportional Hazards Approach

| | Initial examination | After 48 months |
|---|---|---|
| **n.risk** | 1741 | 56 |
| **n.event** | 238 | 23 |
| **Survival** | 86% | 14% |
| **Standard error** | 0.8% | 1.8% |
| **Means value** | -0.852 and 0.883 | -0.126 and 0.200 |
| **Concordance index** | 85% | 15% |

- Survival: Estimates the probability of survival. In the total sample of patients, 86% of disease progressed. That number starts to drop to 14% after 48 months.
- std.err: existence standard error (Patient had a low life expectancy).
- Lower and upper 95% CI: lower bound of the confidence interval. About 85% of the generated intervals contain the true population. This means that the sampling process is repeated several times.

As a result, the sample size increases and the range of interval values decreases, making the mean more accurate than with smaller samples. The probability that the population mean is -0.852 and 0.883 standard deviations (z-score) from the sample mean is 85% for time interval 0. The 48-month interval is -0.126 to 0.200. As a result, there is a risk that 15% of the population mean is outside the upper and lower bounds of the confidence interval.

In survival analysis, the time it takes for events to occur—also known as survival time—is researched and modelled. The most popular technique for analysing the correlation between predictor factors and survival time is the Cox proportional hazards regression methodology. Table 2 shows the effect of each feature on the survival analysis. It consists of the following:

1. Coefficient: Measures the effect of a covariate (log hazard ratio).
2. exp(coef): Displays the hazard ratio, the effect size of a covariate.
3. se(coef): standard error.
4. z: Gives the Wald statistic corresponding to the standard error ratio of each regression coefficient (z = coef/se(coef)).
5. Pr(>|z|): The probability of the Wald statistical value. Review what each feature means.
6. Hazard ratio confidence interval: The hazard ratio (exp(coef)) has an upper and lower 95% confidence interval on the sum.

Three tests that can be used to determine the overall importance of an approach are the likelihood ratio test, the Wald test, and the log-rank statistic. Asymptotically, these three tests are equivalent. If N is large enough, we get similar results. For small N, it may be slightly different. Likelihood ratio tests work best with the small sample sizes for which they are commonly used. The results of the suggested work were as follows:

- Agreement = 0.78 (se = 0.009)
- Likelihood ratio test = 815.3/3 df
- Wald test = 524.3 p = 2e-16
- Test score (log rank) = 549.7 in 3 DF, p = 2e-16

In general, the most used evaluation measure for survival approaches is the goodness-of-fit index (c-index, c-statistic). The goodness-of-fit index (c-index) is a measure for evaluating the predictions of an approach. It can be expressed as the ratio of matching pairs to the total number of assessment pairs that can be made [39]. It ranges from 0.5 to 1 and is equivalent to the area under the receiver operating characteristic (ROC) curve. Sub-0.5 values signify inadequate approximation performance. When the method's prediction is less accurate than chance, it has a value of 0.5. Ultimately, performance above 0.7 is considered good. Details of the results are presented in Table 3. Six months apart, patients were reassessed. After two years, the patient's chance of developing the illness dropped as a result. Although 1741 instances were used to begin this procedure, 56 individuals eventually did not convert to AD. Table 4 represents a comparison between the results of the proposed approach and the best state-of-the-art approaches in literature.

From the above table it can be seen clearly that the proposed approach is evaluated using a huge number of patients compared to the number of patients used in literature to make accurate analysis avoiding any possibility of overfitting. It can also be seen that the proposed work outperformed other state-of-the-art approaches addressing the same problem with the largest C-index. Only Sarica et al

Table 2. Random Forest Survival Approach Summary

| | coef | exp(coef) | se(coef) | z | Pr(>\|z\|) | Lower0.95 |
|---|---|---|---|---|---|---|
| CDRSB | -4.666e-01 | 6.271e-01 | 2.898e-02 | -16.100 | <2e-16 | 0.6514 |
| MMSE | 4.216e-02 | 1.043e+00 | 1.429e-02 | 2.951 | 0.00317 | 1.0143 |
| Ventricles | 1.991e-06 | 1.000e+00 | 1.448e-06 | 1.375 | 0.16922 | 1.0000 |

Table 3. Interpretation of the Proposed Approach

| Time | n.risk | n.event | Survival | std.err | lower 95% CI | upper95% CI |
|---|---|---|---|---|---|---|
| 0 | 1741 | 238 | 0.863 | 0.00823 | 0.847 | 0.880 |
| 6 | 1260 | 210 | 0.719 | 0.01137 | 0.697 | 0.742 |
| 12 | 863 | 166 | 0.581 | 0.01332 | 0.556 | 0.608 |
| 18 | 536 | 120 | 0.451 | 0.01471 | 0.423 | 0.481 |
| 24 | 376 | 77 | 0.359 | 0.01500 | 0.330 | 0.389 |
| 36 | 156 | 51 | 0.241 | 0.01683 | 0.211 | 0.277 |
| 48 | 56 | 23 | 0.142 | 0.01871 | 0.110 | 0.184 |

Table 4. Comparison between the results of this work and the state-of-the-art results

| Authors | Year | Dataset | No. of Patients | C-index |
|---|---|---|---|---|
| Khajehpiri et al [8] | 2022 | ADNI | 882 | 73.3% |
| Mirabnahrazam et al [31] | 2022 | ADNI | 401 | 80% |
| Sarica et al [32] | 2023 | ADNI | 387 | 85% |
| This work | 2023 | ADNI | 1741 | 85% |

[32], the last work done in this area obtained comparable results, but their C-index is a calculated average of individual C-indexes for different parameters which cannot be considered an accurate metric for such a problem as mentioned before. The small sample selected from ADNI also suggests overfitting in their results.

## 5. Conclusion

This paper introduces a new survival analysis approach for detecting the progression from MCI to AD using Random Forest. Because the condition is essentially progressive, the approach considers the timing data gathered from the cases. This work is predicting the future progression of the MCI to AD, unlike most of the work done in literature which only focus on the classifying the state of the current diagnosis. Amongst all the work done in literature in this field only two attempts were done to address the problem of the prediction of the possible future progression of the disease, and the proposed work outperforms both of them in terms of number of patients studied, and C-index, where the proposed approach succeeded to achieve a C-index of 85% when evaluated on the whole dataset which contains 1741 patients compared to 401, 397, and 882 patients (see table 4). Our approach is not limited to the prediction of Alzheimer's Disease, but it also identifies the relative feature that affects progression. This is a step forward in this field to be applied in real scenarios. The results of the proposed experiments are as follows: Concordance = 0.78, likelihood ratio test = 815.3 on 3 df, Wald test = 524.3, Test score (log-rank) = 549.7 on 3 DF, and p = 2e-16.

Improving the proposed approach's performance in future will require further research in other regression and deep learning methods. Personal data could improve the performance and efficiency of AD prediction at an earlier stage.

## Conflicts of Interest

The authors declare no conflict of interest.

## Author Contributions

Conceptualization, Nour Zawawi and Nermin Negied; methodology, Nour Zawawi; software, Nour Zawawi; validation, Nour Zawawi and Nermin Negied; formal analysis, Nermin Negied; investigation, Nermin Negied; data curation, Nour Zawawi; writing—original draft preparation, Nour Zawawi; writing—review and editing, Nermin Negied; visualization, Nermin Negied; supervision, Nermin Negied; project administration, Nermin Negied.

## References

[1] A. Association, "2020 Alzheimer's disease facts and figures", *Alzheimer's & Dementia*, Vol. 16, No. 3, pp. 391–460, 2020.

[2] M. Heron, "Deaths: Leading causes for 2019", *National Vital Statistics Reports*, Vol. 70, No. 9, pp. 1–114, 2021.

[3] A. Rouhi and H. Nezamabadi-Pour, "Feature Selection in High-Dimensional Data", *Cham: Springer International Publishing*, Ch. 5, pp. 85–128, 2020.

[4] V. Bol ´on-Canedo, N. S´anchez, and A. Alonso-Betanzos, "Foundations of Feature Selection", *Cham: Springer International Publishing*, Ch. 2, pp. 13–28, 2015.

[5] M. Rambhajani, W. Deepanker, and N. Pathak, "A survey on implementation of machine learning techniques for dermatology diseases classification", *International Journal of Advances in Engineering & Technology*, Vol. 8, No. 2, pp. 194–202, 2015.

[6] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic", *Journal of Intelligent Learning Systems and Applications*, Vol. 9, pp. 1–16, 2017.

[7] K. D. Tzimourta, V. Christou, A. T. Tzallas, N. Giannakeas, L. G. Astrakas, P. Angelidis, D. Tsalikakis, and M. G. Tsipouras, "Machine Learning Algorithms and Statistical Approaches for Alzheimer's Disease Analysis Based on Resting-State EEG Recordings: A Systematic Review", *International Journal of Neural Systems*, Vol. 31, No. 5, 2021, doi: 10.1142/S0129065721300023.

[8] B. Khajehpiri, H. Moghaddam, M. Forouzanfar, R. Lashgari, J. Ramos-Cejudo, R. S. Osorio, and B. Ardekani, "Survival analysis in cognitively normal subjects and in patients with mild cognitive impairment using a proportional hazards model with extreme gradient boosting regression", *Journal of Alzheimer's Disease*, Vol. 85, pp. 837–850, 2022.

[9] N. Negied, "Infrared Thermography Based Breast Cancer Detection –Comprehensive Investigation", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 33, No.06, 2018.

[10] R. Ali, A. Abbas, and H. Daway, "Medical Images Enhanced by Using Fuzzy Logic Depending on Contrast Stretch Membership Function", *International Journal of Intelligent Engineering and Systems*, Vol.14, No.1, 2021, doi: 10.22266/ijies2021.0228.34.

[11] R. Kalyani, P. Sathya, and V. Sakthivel, "Multilevel Thresholding for Medical Image Segmentation Using Teaching Learning Based Optimization Algorithm", *International Journal of Intelligent Engineering and Systems*, Vol.14, No.2, 2021, doi: 10.22266/ijies2021.0430.02.

[12] N. Negied and A. Serag-eldin, "Automatic Detection of Alzheimer Disease from 3D MRI Images using Deep CNNs", *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 12, 2022, doi:10.14569/IJACSA.2022.0131258.

[13] G. Juan, G. Guell, and G. Piellaa, "Martí-Juan G, Sanroma-Guell G, Piella G. A survey on machine and statistical learning for longitudinal analysis of neuroimaging data in Alzheimer's disease", *Computational Methods Programs Biomed*, Vol. 189, 2020, doi: 10.1016/j.cmpb.2020.105348.

[14] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, A. D. N. Initiative, L. Shen, and M. Xu, "A multi-model deep convolutional neural network for automatic Hippocampus segmentation and classification in Alzheimer's disease", *NeuroImage*, Vol. 208, 2020.

[15] J. Liu, J. Wang, Z. Tang, B. Hu, F. Wu, and Y. Pan, "Improving alzheimer's disease classification by combining multiple measures", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 15, No. 5, pp. 1649–1659, 2018.

[16] S. Basheera and M. Ram, "A novel CNN based Alzheimer's disease classification using hybrid enhanced Ica segmented Gray Matter of MRI", *Computerized Medical Imaging and Graphics*, Vol. 81, 2020.

[17] A. Basher, B. Kim, K. Lee, and H. Jung, "Automatic localization and discrete volume measurements of hippocampi from MRI data using a Convolutional neural network", *IEEE Access*, Vol. 8, pp. 725–739, 2020.

[18] A. Castro, E. Blanco, A. Pazos, and C. Munteanu, "Automatic assessment of Alzheimer's disease diagnosis based on deep learning techniques", *Computers in Biology and Medicine*, Vol. 120, No. 103764, 2020.

[19] K. Zhao, Y. Ding, Y. Han, Y. Fan, A. F. Alexander-Bloch, T. Han, D. Jin, B. Liu, J. Lu, C. Song, P. Wang, D. Wang, Q. Wang, K. Xu, H. Yang, H. Yao, Zheng, C. Yu, B. Zhou, X. Zhang, Y. Zhou, T. Jiang, X. Zhang, and Y. Liu, "Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis", *Science Bulletin*, Vol. 65, No. 13, pp. 1103 – 1113, 2020.

[20] J. Cai, L. Hu, Z. Liu, K. Zhou, and H. Zhang, "An embedded feature selection and multi-class classification method for detection of the progression from mild cognitive impairment to Alzheimer's disease", *Journal of Medical Imaging and Health Informatics*, Vol. 10, No. 2, pp. 370–379, 2020.

[21] A. Shikalgar and S. Sonavane, "Hybrid deep learning approach for classifying Alzheimer disease based on multimodal data", In: *Proc. Of Computing in Engineering and Technology*, Eds, Singapore, Springer Singapore, pp. 511–520, 2020.

[22] X. Bi, X. Hu, H. Wu, and Y. Wang, "Multimodal data analysis of Alzheimer's disease based on clustering evolutionary random forest", *IEEE Journal of Biomedical and Health Informatics*, Vol. 24, No. 10, 2020, DOI: 10.1109/JBHI.2020.2973324.

[23] S. Qiu, G. H. Chang, M. Panagia, D. M. Gopal, R. Au, and V. B. Kolachalama, "Fusion of deep learning models of mri scans, mini-mental state Examination, and logical memory test enhances diagnosis of mild cognitive impairment",

*Alzheimers Dement (Amst)*, Vol. 28, No. 10, pp. 737–749, 2018.

[24] A. Alexiou, V. Mantzavinos, N. Greig, and M. Kamal, "A Bayesian model for the prediction and early diagnosis of Alzheimer's disease", *Frontiers on Aging Neuroscience*, Vol. 9, No. 77, 2017.

[25] J. Venugopalan, L. Tong, H. Hassanzadeh, and M. Wang, "Multimodal deep learning models for early detection of Alzheimer's disease stage", *Scientific Reports*, Vol. 11, No. 3254, 2021.

[26] N. Zawawi, H. Saber, M. Hashem, and T. Gharib, "Predicting Alzheimer's disease progression by combining multiple measures", In: *Proc. of CS & IT conference*, Vol. 11, No. 19, 2021.

[27] N. Zawawi, H. Saber, M. Hashem, and T. Gharib, "An efficient hybrid approach for diagnosis high dimensional data for alzheimer's diseases using machine learning algorithms", *International Journal of Intelligent Computing and Information Sciences*, Vol. 22, No. 2, pp. 97–111, 2022.

[28] X. Hong, R. Lin, C. Yang, N. Zeng, C. Cai, J. Gou, and J. Yang, "Predicting Alzheimer's disease using LSTM", *IEEE Access*, Vol. 7, No. 80, pp. 893 901, 2019.

[29] S. Yang, H. Shin, S. Lee, and H. Lee, "Functional linear regression model with randomly censored data: Predicting conversion time to Alzheimer 's disease", *Computational Statistics & Data Analysis*, Vol. 150, 2020.

[30] N. Zawawi, H. Saber, M. Hashem, and T. Gharib., "A new neural network model for prediction next stage of Alzheimer's disease", In: *Proc. of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, pp. 689–696, 2022.

[31] G. Mirabnahrazam, D. Ma, C. Beaulac, S. Lee, K. Popuri, and H. Lee, "Predicting time-to-conversion for dementia of Alzheimer's type using multi-modal deep survival analysis", *Neurobiology of Aging*,pp. 139-156, 2023, doi: 10.1016/j.neurobiolaging.2022.10.005.

[32] A. Sarica, F. Aracri, M. Bianco, F. Arcuri, and A. Quattrone, "Explainability of random survival forests in predicting conversion risk from mild cognitive impairment to Alzheimer's disease", *Brain Informatics*, Vol. 18, No.1, 2023, doi: 10.1186/s40708-023-00211-w.

[33] https://adni.loni.usc.edu/about/

[34] S. Devlin and G. Heller, "Concordance probability as a meaningful contrast across disparate survival times", *Statistical Methods in Medical Research*, Vol. 30, No. 3, pp. 816–825, 2021.

[35] B. Jing, T. Zhang, Z. Wang, Y. Jin, K. Liu, W. Qiu, L. Ke, Y. Sun, C. He, D. Hou, L. Tang, X. Lv, and C. Li, "A deep survival analysis method based on ranking", *Artificial Intelligence in Medicine*, Vol. 98, pp. 1–9, 2019.

[36] Y. Li, J. Wang, J. Ye, and C. K. Reddy, "A multi-task learning formulation for survival analysis", In: *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York*, USA, pp. 1715–1724, 2016, doi: org/10.1145/2939672.2939857.

[37] N. Obuchowski, "A roc-type measure of diagnostic accuracy when the gold standard is continuous-scale", *Statistics in medicine*, Vol. 25 No. 3, pp. 481–493, 2006.

[38] R. Agostino and B. Nam, "Evaluation of the performance of survival analysis models: Discrimination and calibration measures", In: *Proc. of Advances in Survival Analysis, ser. Handbook of Statistics, Elsevier*, Vol. 23, pp. 1–25, 2023, https://www.sciencedirect.com/science/article/pii/S0169716103230017

[39] A. Brentnall and J. Cuzick, "Use of the concordance index for predictors of censored survival data", *Statistical Methods in Medical Research*, Vol. 27, No. 8, pp. 2359–2373, 2018, doi: 10.1177/0962280216680245.