Article

A Comprehensive Machine Learning Framework for the Exact Prediction of the Age of Onset in Familial and Sporadic Alzheimer's Disease

Jorge I. Vélez [1,*,†], Luiggi A. Samper [2], Mauricio Arcos-Holzinger [3], Lady G. Espinosa [4], Mario A. Isaza-Ruget [4], Francisco Lopera [5] and Mauricio Arcos-Burgos [3,*,†]

[1] Department of Industrial Engineering, Universidad del Norte, Barranquilla 081007, Colombia
[2] Department of Public Health, Universidad del Norte, Barranquilla 081007, Colombia; luiggis@uninorte.edu.co
[3] Grupo de Investigación en Psiquiatría (GIPSI), Departamento de Psiquiatría, Instituto de Investigaciones Médicas, Facultad de Medicina, Universidad de Antioquia, Medellín 050010, Colombia; oscararcos98@gmail.com
[4] INPAC Research Group, Fundación Universitaria Sanitas, Bogotá 111321, Colombia; lgespinosaar@unisanitas.edu.co (L.G.E.); misaza@unisanitas.edu.co (M.A.I.-R.)
[5] Neuroscience Research Group, University of Antioquia, Medellín 050010, Colombia; floperar@gmail.com
* Correspondence: jvelezv@uninorte.edu.co (J.I.V.); mauricio.arcos@udea.edu.co (M.A.-B.)
† These authors contributed equally to this work.

Citation: Vélez, J.I.; Samper, L.A.; Arcos-Holzinger, M.; Espinosa, L.G.; Isaza-Ruget, M.A.; Lopera, F.; Arcos-Burgos, M. A Comprehensive Machine Learning Framework for the Exact Prediction of the Age of Onset in Familial and Sporadic Alzheimer's Disease. *Diagnostics* **2021**, *11*, 887. https://doi.org/10.3390/diagnostics11050887

Academic Editor: Markos G. Tsipouras

Received: 5 April 2021
Accepted: 29 April 2021
Published: 17 May 2021

Abstract: Machine learning (ML) algorithms are widely used to develop predictive frameworks. Accurate prediction of Alzheimer's disease (AD) age of onset (ADAOO) is crucial to investigate potential treatments, follow-up, and therapeutic interventions. Although genetic and non-genetic factors affecting ADAOO were elucidated by other research groups and ours, the comprehensive and sequential application of ML to provide an exact estimation of the actual ADAOO, instead of a high-confidence-interval ADAOO that may fall, remains to be explored. Here, we assessed the performance of ML algorithms for predicting ADAOO using two AD cohorts with early-onset familial AD and with late-onset sporadic AD, combining genetic and demographic variables. Performance of ML algorithms was assessed using the root mean squared error (RMSE), the R-squared (R^2), and the mean absolute error (MAE) with a 10-fold cross-validation procedure. For predicting ADAOO in familial AD, boosting-based ML algorithms performed the best. In the sporadic cohort, boosting-based ML algorithms performed best in the training data set, while regularization methods best performed for unseen data. ML algorithms represent a feasible alternative to accurately predict ADAOO with little human intervention. Future studies may include predicting the speed of cognitive decline in our cohorts using ML.

Keywords: age of onset; machine learning; Alzheimer's disease; genetic isolates; *PSEN1*; predictive genomics; natural history

1. Introduction

Alzheimer's disease (AD; OMIM 104300) is a neurodegenerative disorder characterized by progressive loss of neurological, mental, and cognitive functions, including memory, changes in judgment, behavior, and emotions [1–4]. AD is the most common cause of dementia and constitutes an increasing challenge due to society's public health and economic costs [5–8]. As of 2016, ~44 million people had AD or related dementia worldwide [9]. Without new medicines to prevent, delay, or stop the disease, this figure is projected to dramatically increase to ~66 million dementia cases by 2030 and ~116 million by 2050 [10]. The financial burden associated with the disease was estimated to be USD 818 billion in 2015 worldwide [11,12].

AD neuropathological damage is characterized by extracellular deposits of the beta-amyloid (Aβ) peptide, the formation of intracellular neurofibrillary tangles of hyperphosphorylated tau protein (p-Tau), and the impairment of neurons and synaptic connections in the cerebral cortex and hippocampus, a key brain region involved in learning and memory processes and emotional control [1,13–15]. Genetically, AD is divided into familial AD (*f*AD), which accounts for <5% of AD cases and is caused by the presence of pathogenic and deleterious mutations harbored in major genes (segregating in a mendelian way) such as *APP*, *PSEN1*, *PSEN2* [9,16–18], *ADAM10*, *AKAP9*, *PICALM*, *PLD3*, *TREM2*, and *UNC5C* [19–27], and sporadic AD (*s*AD), without a clear mendelian pattern of segregation, which accounts for >90% of AD cases. In contrast to *f*AD, mutations in genes associated with *s*AD do not directly cause AD but confer susceptibility [28]. Although *f*AD and *s*AD forms are phenotypically similar, the age of onset (AOO) at which signs and/or symptoms of AD appear for the first time in cases of *f*AD is generally earlier than in cases of *s*AD, with important predictors of ADAOO in *s*AD correspond to several genetic variants of small effect [29,30]. Indeed, it is generally established that *f*AD cases have an AOO before 65 years (ranging from the early 30s to the late 70s), while the AOO in *s*AD generally starts after 65 years [31,32]. Considering that in *s*AD cases are diagnosed later and usually at the later stages of disease compared to familial cases [16,28], developing predictive models of ADAOO will open new possibilities for clinicians, patients, and family members [33–38].

Despite being suggested ~25 years ago as a valuable quantitative phenotype for monitoring AD natural history [39], ADAOO is one of the least studied phenotypes in the epidemiology of AD [39–41]. In fact, recent studies of our group and from other research groups showed that the natural history of AD might lead to the elucidation of new diagnostic, predictive, and therapeutic alternatives while considering interventions to delay the ADAOO [33,42–44].

For more than three decades, our group characterized clinically and genetically the world's most extensive known pedigree with an aggressive form of AD caused by the E280A mutation in the *PSEN1* gene, often referred to as the Paisa mutation [39]. Parallelly, we have characterized other forms of *f*AD and individuals with *s*AD from the same community who share the same genetic background of the E280A pedigree [45,46]. To the best of our knowledge, our group pioneered the exposition of major genetic variants modifying ADAOO in *s*AD [35]. Instead of using a traditional approach where the risk of developing AD is assessed [39,41], we recruited individuals with *s*AD exhibiting ADAOO at the extremes of the AOO distribution in order to identify genetic variants responsible for the wide spread of AOO [33,35,42–44]. Genes harboring these variants play an essential role in cell proliferation, apoptotic and immune dysregulation processes, oligodendrocyte differentiation, protein degradation, neuron apoptosis, cholesterol metabolism, neurogenesis, and inflammatory and memory processes linked to AD [35,36].

Predictive models aim to determine the expected value of an outcome variable Y of interest based on a set of predictors $X = (X_1, X_2, \dots, X_p)^T$. Generally speaking, Y and X can be of any nature (i.e., binary, multinomial, ordinal, or continuous), and the selection of the best predictive model is based on some sort of error-related measure, such as the accuracy, the root mean squared error (RMSE) or the mean absolute error (MAE) [47,48]. Although some predictive models have recently been developed for AD [49–52], the outcome variable is not ADAOO and genetic variants are not included as predictors. We argue that genetic/genomic data will substantially improve AD diagnosis and mitigate the confounding effect of demographic and population structure data while increasing power [53,54].

Machine learning (ML) has attracted the research community's attention for disclosing patterns, detecting objects, and developing predictive frameworks in several diseases [55,56]. AD is one of the most common mental health conditions studied via ML methods [57]. In fact, we and others showed that ML algorithms constitute a promising alternative for assessing AD diagnosis based on prospective clinical, image, and/or biomarker data [18,58–64]. Furthermore, ML algorithms have also proven to be a suitable alternative for the timely diagnosis of late-onset AD based on genetic variation [29,32,65], differentiate AD from other neurological disorders using noninvasive blood markers [65], and predict AD conversion in individuals with Mild Cognitive Impairment (MCI) [66,67]. Interestingly, optimization procedures for tuning the parameters of ML algorithms have been reported to increase the sensitivity, specificity, and accuracy of ML for AD diagnosis [68]. Other ML alternatives include the use of artificial intelligence (AI), namely deep learning (DL), assessing AD diagnosis and progression with brain radiological images [69,70]. Although these results are promising, their main limitation is that the predictive model provided either an estimate of the risk of an individual for developing AD or the range within which the ADAOO may fall with high confidence (i.e., early- or late-onset based on whether the ADAOO was before or after a threshold, respectively), but not an estimate of the actual ADAOO. Moreover, a comprehensive exploration of advanced ML algorithms for ADAOO prediction is yet to be conducted.

In this study, we comprehensively assess ML algorithms' feasibility applied to *f*AD and *s*AD cohorts, with the overarching aims of (1) accurately predicting ADAOO and improving the scope and performance previously reached; and (2) expanding the possibilities of quantifying ADAOO in the clinical setting. Our results suggest that ML constitutes a feasible and easy-to-implement new methodology to predict ADAOO, especially in the clinical setting, while significantly overpowering our previous results and paving the way for new possibilities to define follow-up and counseling strategies for patients and their family members.

2. Materials and Methods

2.1. Subjects

2.1.1. E280A Pedigree

We ascertained 71 patients from the 459 E280A *PSEN1* mutation carriers at the extremes of the ADAOO distribution (44 women [62%] and 27 men [38%]) [33,36]. Detailed clinical assessment and ascertainment procedures of this pedigree have been presented elsewhere [31,71–73].

2.1.2. The Cohort of Sporadic Cases

Fifty-four individuals with *s*AD were included in this study (43 [80%] were women, and 11 [20%] men). Clinical, neurological, and neuropsychological assessment of *s*AD patients has been reported elsewhere [35]. ADAOO was determined during anamnesis with the information provided by patients or their families, with confirmation by several sources. Because some patients started their follow-up during MCI, ADAOO was defined during the follow-up stage based on Petersen's criteria [74]. This strategy was recently proven to be highly accurate [75]. AD affection status was defined based on the DSM-IV criteria [76].

2.2. Variants Associated with ADAOO

We previously studied the association of common exonic functional variants (CEFVs) with ADAOO (Table 1) [35,36] using single- and multi-locus linear mixed-effects models [77] and recursive partitioning ML algorithms [36]. These variants were found to delay ADOO up to 17 years in carriers of the E280A *PSEN1* mutation and accelerate it up to ~14 years in individuals with *s*AD [35,36].

2.3. ADAOO Prediction Using ML

Predictive models of ADAOO were constructed with ML algorithms in individuals carrying the E280A *PSEN1* mutation and individuals with *s*AD. The set of predictor variables consisted of demographic variables (i.e., gender, sex, and years of education) and genomic variants previously identified to be associated as ADAOO modifiers (Table 1). The complete list of ML algorithms is provided in the Supplementary Materials. Construction, parameters tuning, validation, and testing of these predictive models were performed in R version 4.0.2 Patched (2020-06-30 r78761) [80] with the methods implemented in the caret package [47,48] using a 10-fold cross-validation procedure with five repetitions. The training/testing data sets consisted of 70%/30% of individuals per cohort. Given the continuous nature of the outcome variable (i.e., ADAOO), the root mean squared error (RMSE), the R-squared (R^2), and the mean absolute error (MAE) measures were used to evaluate the performance of the ML algorithms. In ML-based predictive models, high values of R^2 and low values of RMSE and MAE indicate good performance. To graphically represent the performance of these ML algorithms and to identify similarities among them, we combined K-means clustering [81] and principal component analysis (PCA) [82,83]; the number of K-means clusters and the number of principal components were determined using the methods implemented in the NbClust [84] and paran [85] packages for R.

Table 1. Common exonic functional variants modifying ADAOO in 125 individuals from the Paisa genetic isolate.

Cohort	Chr	Marker	Position [a]	Gene	Change	$\hat{\beta}$ ($SE_{\hat{\beta}}$) [b]	P_{FDR}
E280A	19	rs7412	45,412,079	*APOE*	p.Arg176Cys	17.45 (0.48)	2.13×10^{-30}
(n = 71)	8	rs36092215	142,367,246	*GPR20*	p.Arg260Cys	12.12 (0.54)	6.58×10^{-22}
	11	rs12364019	5,730,343	*TRIM22*	p.Arg321Lys	−11.64 (0.79)	1.15×10^{-14}
	1	rs16838748	157,508,997	*FCRL5*	p.Asn427Lys	7.14 (0.68)	8.61×10^{-10}
	7	rs12701506	36,566,020	*AOAH*	[c]	−2.75 (0.30)	5.69×10^{-8}
	19	rs2682585	44,081,288	*PINLYP*	p.His6Arg	−1.68 (0.21)	1.67×10^{-6}
	1	rs62621173	159,021,506	*IFI16*	p.Ser512Phe	−2.80 (0.37)	8.63×10^{-6}
	1	rs10798302	173,987,798	*RC3H1*	[d]	1.76 (0.27)	1.86×10^{-4}
	7	rs754554	24,758,818	*DFNA5*	p.Pro142Thr	−1.39 (0.28)	3.62×10^{-2}
Sporadic	2	rs35946826	105,859,249	*GPR45*	p.Leu312fs	−12.67 (0.148)	3.08×10^{-36}
(n = 54)	1	rs61742849	114,226,143	*MAGI3*	p.Gly1318fs	−14.32 (0.199)	4.38×10^{-34}
	6	rs675026	154,414,563	*OPRM1*	p.Ala442fs	5.42 (0.079)	1.15×10^{-33}
	10	rs838759	22,498,468	*EBLN1*	p.Gly149fs	−4.26 (0.092)	3.90×10^{-28}
	17	rs61749930	48,594,691	*MYCBPAP*	p.Arg124fs	−12.08 (0.286)	6.06×10^{-27}
	19	rs7250872	1,811,603	*ATP8B3*	p.Gly45fs	−2.54 (0.088)	9.57×10^{-22}
	16	rs749670	31,088,625	*ZNF646*	p.Lys328fs	−1.52 (0.067)	1.35×10^{-18}
	4	rs7677237	89,306,659	*HERC6*	p.Met123fs	2.14 (0.122)	3.58×10^{-15}
	4	rs6835769	79,284,694	*FRAS1*	p.Ala817fs	−1.11 (0.074)	2.74×10^{-13}
	11	rs4757987	5,906,205	*OR52E4*	p.Arg228fs	1.02 (0.07)	6.86×10^{-13}
	20	rs236150	5,903,141	*CHGB*	p.Lys117fs	−2.14 (0.181)	2.12×10^{-10}
	6	rs3130257	33,256,471	*WDR46*	p.Thr40fs	−2.35 (0.209)	7.92×10^{-10}
	18	rs754093	77,246,406	*NFATC1*	p.Cys751fs	−0.94 (0.094)	1.34×10^{-8}
	3	rs34230332	14,725,878	*C3orf20*	p.Leu84fs	1.59 (0.185)	4.81×10^{-7}
	19	rs867228	52,249,211	*FPR1*	p.Glu346fs	−0.94 (0.115)	1.34×10^{-6}
	4	rs3733251	77,192,838	*FAM47E*	p.Arg166fs	−0.71 (0.127)	2.07×10^{-3}
	16	rs2303772	87,795,580	*KLHDC4*	p.Leu56fs	0.75 (0.135)	2.75×10^{-3}
	16	rs739999	319,511	*RGS11*	p.Met416fs	0.35 (0.075)	3.48×10^{-2}
	16	rs34779002	87,782,396	*KLHDC4*	p.Gly74fs	0.78 (0.172)	4.00×10^{-2}
	15	rs6493068	43,170,793	*TTBK2*	p.Asp9fs	−0.48 (0.107)	4.27×10^{-2}
	16	rs17137138	4,606,743	*C16orf96*	p.Val85fs	1.00 (0.223)	4.40×10^{-2}
	7	rs3823646	99,757,612	*GAL3ST4*	p.Lys468fs	−0.31 (0.069)	4.47×10^{-2}
	13	rs17081389	25,487,001	*CENPJ*	p.Pro55fs	1.00 (0.223)	4.61×10^{-2}
	10	rs78334417	75,071,618	*TTC18*	p.Pro450fs	1.00 (0.223)	4.84×10^{-2}
	7	rs186048202	134,678,273	*AGBL3*	p.Arg52fs	0.61 (0.139)	4.91×10^{-2}

[a] UCSC GRCh37/hg19 coordinates; [b] Markers can accelerate ($\hat{\beta} < 0$) or delay ($\hat{\beta} > 0$) ADAOO according to their effect; [c] Chromatin state segmentation strong enhancer state-5 from ChiP-seq data; [d] CpG islands, DNaseI hypersensitivity uniform peak from ENCODE/analysis. ADAOO = Alzheimer's disease age of onset; Chr: chromosome; $\hat{\beta}$ = Regression coefficient; $\text{SE}_{\hat{\beta}}$ = Standard error of $\hat{\beta}$; P_{FDR} = Corrected P-value using the False Discovery Rate (FDR) [78,79].

To evaluate the stability of each predictor's variable importance, we implemented the following resampling strategy, which is a slight modification of the empirical bootstrap [86,87]. First, we constructed B = 1000 training data sets at random, keeping the 70%/30% proportion for the training/testing data sets initially used to identify the best performing ML model. Secondly, for the b-th training data set (b = 1,2, . . . , B), this model was fitted, and the variable importance measure associated with each predictor was computed. Thus, for any predictor X, we obtained the values $X^{(1)}$, $X^{(2)}$, $X^{(3)}$, . . . , $X^{(B)}$, with $X^{(b)}$ representing the variable importance of X calculated in the b-th randomly generated training data set. Finally, we calculated the bootstrap-based 95% confidence intervals (CIs) based on the 2.5% and 97.5% percentiles of $X^{(1)}$, $X^{(2)}$, $X^{(3)}$, . . . , $X^{(B)}$.

3. Results

3.1. ADAOO Prediction in the fAD E280A Pedigree

Table 2 presents the performance measures for ML algorithms' collection for predicting AOO in the E280A pedigree. The training/testing data sets consisted of 51/20 individuals, respectively. When predicting AOO in the training data set, the xgbLinear ML algorithm outperformed all other algorithms in the RMSE, R^2, and MAE performance measures. When evaluating these ML algorithms' performance for unseen data (i.e., testing data set), the glmboost ML algorithm outperformed all other alternatives.

Table 2. Performance of ML algorithms for predicting ADAOO in the E280A pedigree. RMSE = root mean squared error, lower is better; MAE = mean absolute error, lower is better; R^2 = coefficient of determination, higher is better. Best results are shown in **bold**.

ML Algorithm	**Performance Measure**					
	RMSE		**R^2**		**MAE**	
	Training	**Testing**	**Training**	**Testing**	**Training**	**Testing**
glmboost	3.51	**3.73**	0.62	**0.65**	2.41	**2.86**
bstTree	3.67	6.75	0.59	0.08	3.00	4.52
gbm	4.90	6.68	0.27	0.09	3.86	4.52
glmnet	3.59	3.85	0.62	0.64	2.51	2.89
knn	4.53	6.35	0.39	0.05	3.56	4.13
mlp	6.30	6.62	0.07	0.43	5.64	5.78
qrf	1.35	7.24	0.95	0.03	0.69	4.65
rf	2.14	6.17	0.91	0.12	1.70	3.93
rpart	4.73	6.36	0.31	0.07	3.95	4.51
rpart1SE	4.18	5.89	0.46	0.18	3.35	4.11
rpart2	4.28	6.02	0.43	0.15	3.43	4.11
svmLinear	4.74	6.80	0.43	0.07	2.97	4.21
svmLinear2	4.74	6.80	0.43	0.07	2.97	4.21
svmPoly	3.46	7.30	0.66	0.14	1.86	5.13
svmRadial	5.21	6.50	0.35	0.02	3.43	3.96
treebag	4.26	6.02	0.45	0.16	3.47	4.20
xgbLinear	**0.85**	7.14	**0.98**	0.06	**0.37**	4.28
xgbTree	1.79	7.12	0.90	0.08	1.28	4.65

Following our results, the performance of these ML algorithms can be grouped into three classes. For the training data set, class 1 comprises the rf, xgbTree, xbLinear, and qrf algorithms (Figure 1a; yellow); class 2 is constituted by the mlp, treebag, rpart1SE, rpart2, rpart, knn, gbm, svmRadial, svmLinear, and svmLinear2 algorithms (Figure 1a; red); and class 3 by the bstTree, glmnet, glmboost, and svmPoly algorithms (Figure 1a; blue). In the testing data set, the svmPoly, xgbTree, xgbLinear, gbm, bstTree, rpart, and qrf algorithms belong to class 1 (Figure 1b; yellow); tree bag, rpart1SE, rpart2, svmLinear, svmLinear2, rf, knn, and svmRadial form class 2 (Figure 1a; red); and glmnet and glmboost constitute class 3 (Figure 1b; blue). Overall, the best performing algorithms are grouped into class 1 for the training data set, and into class 3 for the testing data set; the xgbLinear algorithm outperforms all other alternatives in class 1 (Table 2 and Figure 1a), while the glmboost algorithm outperforms those in class 3 (Table 2 and Figure 1b).

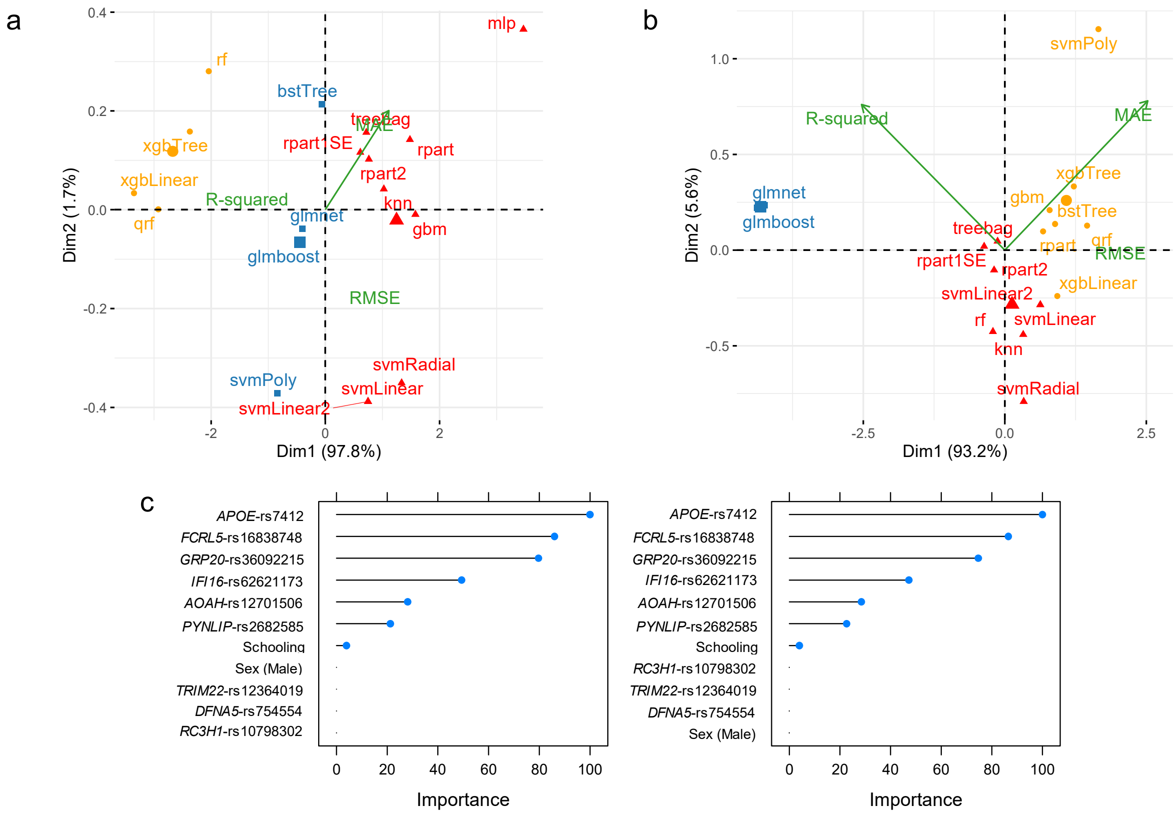

Figure 1. PCA and *K*-means clustering representation of the performance measures for ML algorithms predicting ADAOO in individuals carrying the *PSEN1* E280A mutation when the (**a**) training (n = 51) and (**b**) testing (n = 20) data sets are used. (**c**) Variable importance for the glmnet (left) and glmboost (right) ML algorithms. Here, higher values are better.

Figure 1c depicts variable importance plots for the xgbLinear, glmnet, and glmboost algorithms. Our results suggest that, for the xgbLinear algorithm, which is more suitable for assessing ADAOO in the training data set, years of education (Schooling), genetic variants *GPR20*-rs36092215 and *PYNLIP*-rs2682585, and sex (i.e., being male) are the most important predictors of ADAOO (Figure 1c, left). For the glmnet and glmboost algorithms, which outperform the other alternatives when predicting ADAOO for unseen data, the most important predictors are the genetic variants *APOE*-rs7412, *FCRL5*-rs16838748, *GRP20*-rs36092215, *IFI16*-rs62621173, *AOAH*-rs12701506, and *PYNLIP*-rs2682585, followed by years of education (Figure 1c, center; Figure 1c, right).

3.2. ADAOO Prediction in the Sporadic AD

Table 3 presents the performance measures for collecting ML algorithms used to predict AOO in individuals of the *s*AD cohort. The training and data sets consisted of 40 and 14 individuals, respectively. When predicting AOO in the training data set, the svmLinear and xgbLinear ML algorithms perform reasonably well, with the latter algorithm outperforming all others in terms of the RMSE, R^2, and MAE performance measures. Despite its remarkable performance in the training data set, the predictive power of the xgbLinear algorithm is rather week in unseen data (i.e., possible overlearning). Thus, the svmLinear algorithm seems to be a better alternative than xgbLinear algorithm. On the other hand, when evaluating the performance of these ML algorithms for the testing data set, the lasso outperforms the other alternatives in terms of the RMSE and R^2, while the glmnet algorithm does so in terms of the MAE (Table 3). In contrast, these ML algorithms are strong learners.

Table 3. Performance of ML algorithms for predicting ADAOO in the individuals with sporadic AD from the Paisa genetic isolate. Conventions as in Table 2. Best results are shown in **bold**.

ML Algorithm	Performance Measure					
	RMSE		R^2		MAE	
	Training	Testing	Training	Testing	Training	Testing
bstTree	3.33	5.22	0.83	0.44	2.56	3.75
glmboost	2.32	3.08	0.92	0.84	1.96	2.47
glmnet	0.25	0.52	1.00	0.99	0.17	**0.39**
knn	5.37	6.75	0.48	0.16	3.90	4.98
lasso	0.40	**0.52**	1.00	**1.00**	0.31	0.42
qrf	0.87	5.86	0.99	0.30	0.40	4.57
rf	2.47	5.09	0.94	0.49	1.86	4.15
rpart	5.53	7.69	0.38	0.00	4.46	6.37
rpart1SE	5.53	7.69	0.38	0.00	4.46	6.37
rpart2	5.92	6.98	0.29	0.03	4.63	5.75
svmLinear	0.61	1.11	0.99	0.97	0.57	0.83
svmLinear2	0.61	1.11	0.99	0.97	0.57	0.83
svmPoly	0.75	1.33	0.99	0.96	0.70	1.07
svmRadial	2.57	4.70	0.93	0.51	1.57	3.64
treebag	5.22	7.02	0.48	0.02	4.13	5.54
xgbLinear	**0.03**	4.61	**1.00**	0.67	**0.02**	3.32
xgbTree	1.13	3.98	0.98	0.70	0.93	3.19

Our results indicate that these ML algorithms' performance can be grouped into three classes. For the training data set, class 1 comprises the bstTree, glmboost, rf, and svmRadial algorithms (Figure 2a; yellow); class 2 is constituted by the xgbTree, svmPoly, qrf, svmLinear, svmLinear2, lasso, glmnet, and xbgLinear algorithms (Figure 2a; red); and class 3 by the treebag, knn, rpart1SE, rpart, and rpart2 algorithms (Figure 2a; blue). In the testing data set, the glmboost, xgbTree, rf, svmRadial, and bstTree algorithms belong to class 1 (Figure 2b, yellow); svmPoly, svmLinear, svmLinear2, lasso, and glmnet algorithms belong to class 2 (Figure 2b; red); and treebag, rpart, rpart1SE, rpart2, and qrf constitute class 3 (Figure 2b; blue). Overall, the best performing algorithms are grouped into class 2 for both the training and testing data sets; the xgbLinear algorithm outperforms all other alternatives for the training data set (Table 3 and Figure 2a), while the lasso and glmnet algorithms seem to be the best options for unseen data (Table 3 and Figure 2b).

Figure 2c depicts variable importance plots for the svmLinear, lasso, and glmnet algorithms. We identified that for the svmLinear and lasso algorithms, the most important predictors of ADAOO are variants *HERC6*-rs7677237, years of education, *GPR45*-rs35946826, *NFATC1*-rs754093, *FRAS1*-rs6835769 and *MAGI3*-rs61742849, and *CENPJ*-rs17081389 (Figure 2c, left and Figure 2c, center). Interestingly, under the svmLinear and lasso ML algorithms, sex is a seemingly significant predictor of ADAOO. In terms of variable importance, the glmnet ML algorithm yields similar results to those in the svmLinear and lasso algorithms, but highlights the relevance of variants *GPR45*-rs35946826, *MAGI3*-rs61742849, *C16orf96*-rs17137138, and *C3orf20*-rs34230332, and the small contribution to ADAOO of sex and years of education in unseen individuals with *s*AD (Figure 2c, right).

3.3. Variable Importance: Stability and Relationship with $\hat{\beta}$

Figure 3 shows our implementation results for evaluating variable importance stability for each predictor in the best ML algorithm. When predicting ADAOO in individuals carrying the E280A mutation, the most important predictor is, by far, the *APOE*-rs7412 genetic variant, and the least essential predictors are sex, the genetic variant *RC3H1*-rs10798302, and years of education (Figure 3a).

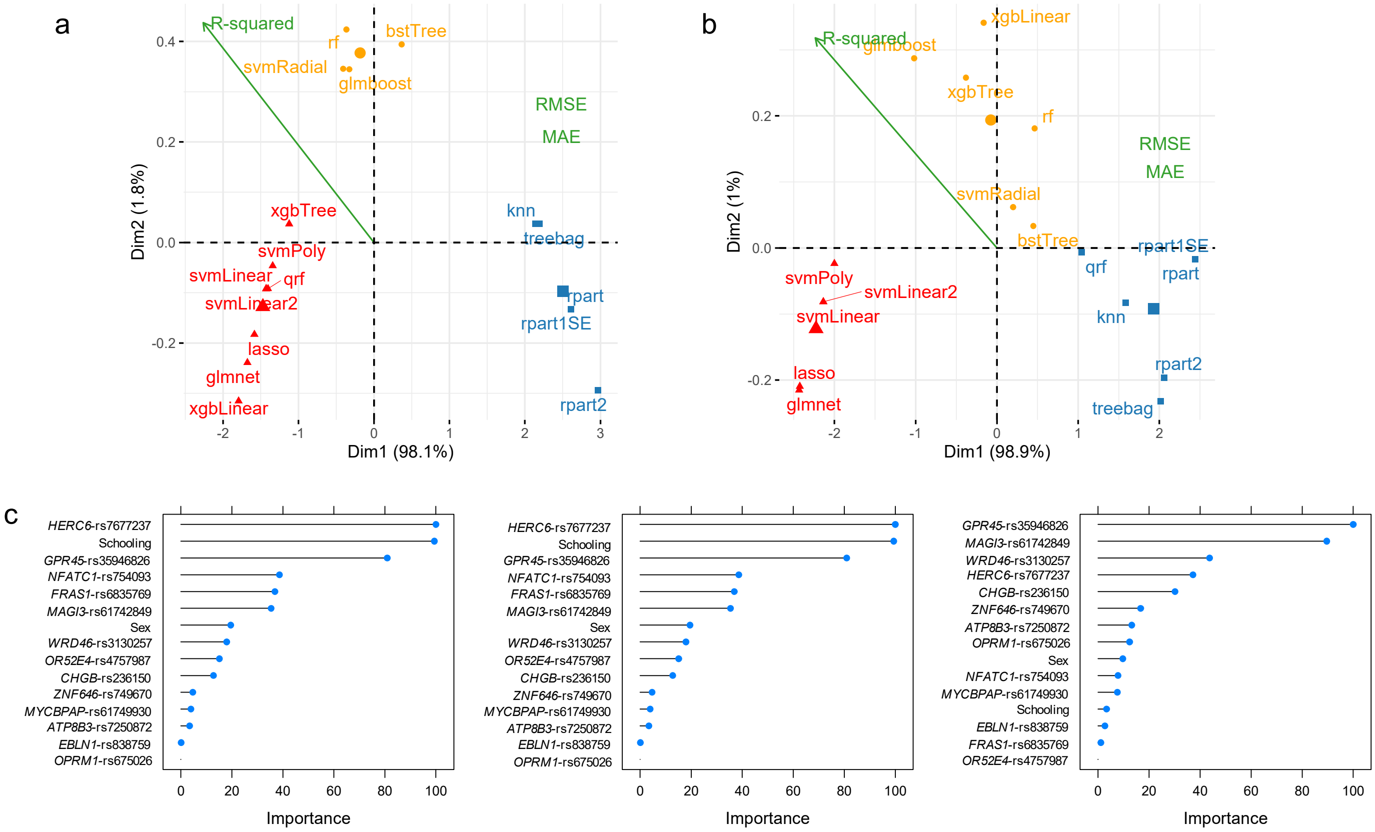

Figure 2. PCA and *K*-means clustering representation of the performance measures for ML algorithms predicting ADAOO in individuals with sporadic AD from the Paisa genetic isolate when the (**a**) training ($n = 40$) and (**b**) testing ($n = 14$) data sets are used. (**c**) Variable importance for the svmLinear (left), lasso (center) and glmnet (right) ML algorithms. Conventions as in Figure 1.

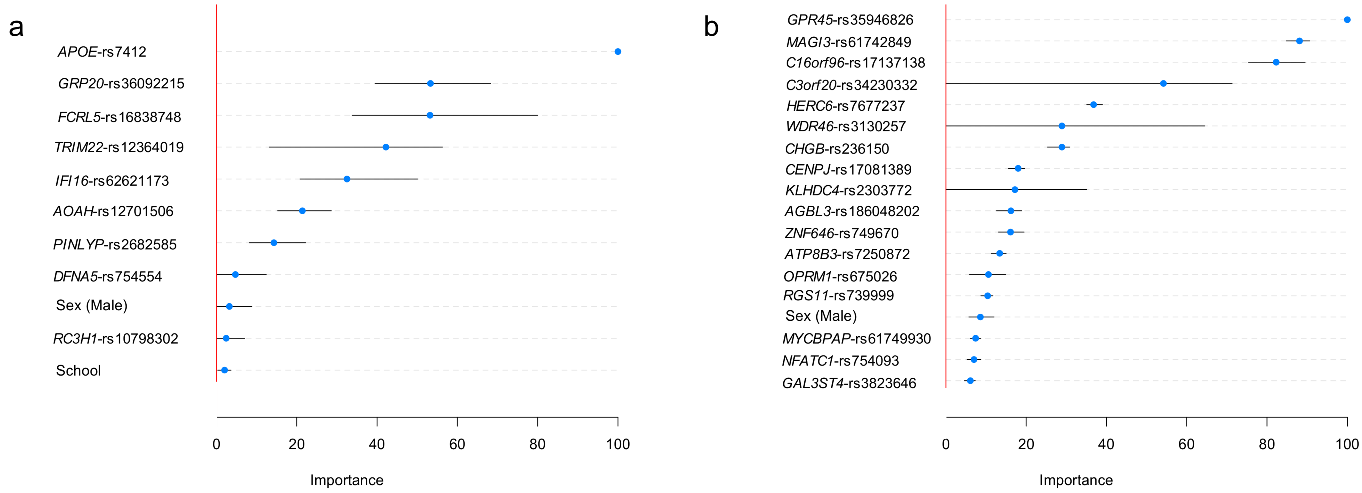

Figure 3. Variable importance for the best ADAOO-predicting ML algorithm in individuals (**a**) carrying the E280A mutation and (**b**) individuals with *s*AD. Blue dots represent the average importance; segments represent 95% bootstrap-based confidence intervals based on $B = 1000$ replicates. Conventions as in Figure 1.

In individuals with *s*AD, the most important ADAOO predictor is the genetic variant *GPR45*-rs359446826, followed by variants *MAGI3*-rs61742849, *C16orf96*-rs17137138, and *C3orf20*-rs34230332. Interestingly, sex and years of education (not shown) are among the least important predictors (Figure 3b). Variable importance bootstrap-based distributions are provided in Figures S1 and S2 (Supplementary Materials). Figure 4 shows scatterplots between $\hat{\beta}$ and their variable importance predicting ADAOO (Tables 2 and 3), confirming that, in contrast to *f*AD, essential predictors of ADAOO in *s*AD correspond to several genetic variants of small effect [29,30].

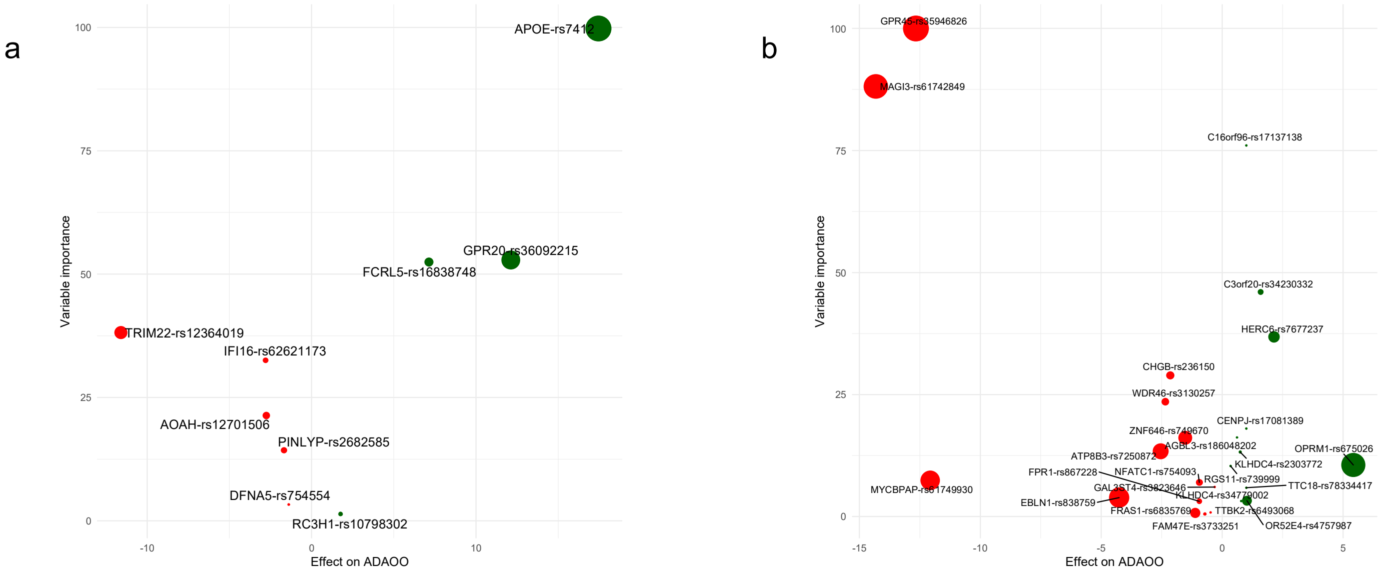

Figure 4. Variable importance vs. effect on ADAOO for genetic variants in individuals with (**a**) E280A *PSEN1* and (**b**) sporadic AD. Protective ($\hat{\beta} > 0$) variants are shown in green, while harmful ($\hat{\beta} < 0$) variants are shown in red. See Table 1 for more details.

4. Discussion

Machine learning (ML) algorithms have recently caught the scientific community's attention because of their flexibility, ease of use, and ability to learn from the data provided [55,56]. Via ML, it has been possible to develop models to identify individuals more susceptible to developing common and rare diseases [58–63,67,88–93] and determine diverse phenotypic response profiles in infectious diseases [94–96]. Considering that ML- and computational-based models have the potential to overcome the limitations of current established clinical models for the diagnosis and follow-up of neurodegenerative diseases, including AD [97], here we studied the feasibility of ML algorithms for predicting Alzheimer's disease age of onset (ADAOO) in individuals from the Paisa genetic isolate. We argue that these ML-based predictive models will improve our understanding of the disease and provide a more accurate and precise definition of the AD natural history landmarks.

We previously identified protective ($\hat{\beta} > 0$; Table 1) and harmful ($\hat{\beta} < 0$; Table 1) ADAOO-modifying variants of significant effect in this community from whole-exome genotyping and whole-exome sequencing data [35,36] using linear-mixed effects models and some ML methods [77]. Thus, the presence of the *APOE*E2* allele alone delays ADAOO up to ~12 years in *PSEN1* E280A mutation carriers. Furthermore, this same allele delays ADAOO up to ~17 years when included in an AD oligogenic model (Table 1) [36]. Subsequent analysis led to the development of a classification tree using advanced recursive partitioning to determine whether individuals carrying this mutation would develop early-onset or late-onset familial AD [36]. Following a similar approach, our group was able to identify ADAOO modifier variants in individuals with sporadic AD (Table 1) [35].

After evaluating several ML-based predictive algorithms for ADAOO in individuals suffering from the most aggressive form of AD (Figure 1 and Table 2) and in individuals with sporadic AD (Figure 2 and Table 3), we identified that the glmboost and glmnet algorithms perform best for predicting ADAOO in unseen data for each cohort, respectively.

These ML-based predictive models showed promising results that can be easily extended to the clinical setting [98]. In particular, the glmboost algorithm in E280A *PSEN1* AD yielded MAE values below 4% and RMSE values of ~4 (Table 2), while the glmnet algorithm yielded MAE values below 1% and RMSE values < 1 in *s*AD (Table 3), suggesting that predicting AOO in these cohorts is feasible. Using these ML-based ADAOO predictive models, AD diagnosis could be made earlier, and potential treatments are provided long before symptoms begin to appear.

Analysis of variable importance shows that the most relevant ADAOO predictors in *f*AD are variants *APOE*-rs7412, *FCRL5*-rs16838748, *GPR20*-rs36092215, *IFI16*-rs62621173, *AOAH*-rs12701506, and *PYNLIP*-rs2682585 (Figures 1b and 3a). Furthermore, protective variants *APOE*-rs7412, *GRP20*-rs36092215, and *FCRL5*-rs16838748 have both the highest effect on ADAOO and are the most important predictors of ADAOO, while variants *TRIM22*-rs12364019, *IFI16*-rs62621173, and *AOAH*-rs12701506 have both the most harmful effect on ADAOO and are among the most important predictors of ADAOO (Figure 4a). Comparing these results with those of previous models predicting AD status (early- vs. late-onset) [36] shows some discrepancies in how the genetic variants are ranked and the relevance of demographic information (i.e., sex and years of education) for predicting AD status. Although predicting AD status may be of interest in some clinical settings, the use of ML-based predictive algorithms for ADAOO is a step forward in both our understanding of the disease and our goal of providing timely clinical care to individuals from this community. While AD cannot be cured and there is no way to stop or slow its progression at the moment, our approach offers the possibility of treating symptoms several years before they begin to appear [4,99,100] under an individually tailored biomarker scheme rather, than using a one-size-fits-all population average strategy [99–101], while taking individual variability into account. Although our results can certainly be used to move AD research in this direction, it is also important to consider the legal implications and the preparation that health providers, neurologists, and centers specializing in AD and neurodegeneration must have in order to interpret these findings and provide proper counseling to patients and their families [102–104]. Another challenge in the years to come is also to significantly reduce the misinformed conclusions produced by ML methods in the absence of clinical domain expertise [105]. In this regard, having a deep understanding of the clinical background in AD, how ML methods operate, and how the results can interpreted and translated to the patient and their relatives is crucial [57].

Variants *GPR45*-rs35946826 and *MAGI3*-rs61742849 have both a more harmful effect on ADAOO and are the most important predictors of ADAOO in individuals with *s*AD (Figure 4b). Interestingly, the harmful effect on ADAOO of variants *MYCBPAP*-rs61749930 and *EBLN1*-rs838759 differs from those of other variants, but their importance for predicting ADAOO is lower, while variants *CHGB*-rs236150 and *WDR46*-rs3130257 accelerate ADAOO and have higher variable importance (Figure 4b). Among protective genetic variants, the highest effect is produced by *OPRM1*-rs675026, followed by *HERC6*-rs7677237 and *C3orf20*-rs34230332, with the former being the less important. Intriguingly, variant *C16orf96*-rs17137138 is the most important ADAOO predictor despite its small effect (Figure 4b).

In summary, here we explore the feasibility of ML algorithms for predicting ADAOO using demographic and genetic data in individuals from the world's most extensive pedigree segregating a severe form of AD caused by a fully penetrant mutation in the *PSEN1* gene and individuals with *s*AD inhabiting the same geographical region. Based on the RMSE, MAE, and R^2 performance measures, our results indicate that ML algorithms are a feasible and promising alternative for assessing ADAOO in these individuals. Interestingly, the most important predictors in these ML-based predictive models were genetic variants, which makes it possible to assess ADAOO at the individual level and opens new personalized medicine and predictive genomic alternatives for AD [98–101].

Future studies should assess the ability of the ML-based predictive models for ADAOO presented herein with out-of-sample data (i.e., determine how close the model is to predicting ADAOO in a patient with known genetic data that was not part of our cohorts) and the development of ML-based models of disease progression [38,50,51,60]. Ultimately, these models could help us to provide an easy-to-use platform, with potential application in the clinical setting, to provide early and accurate estimates of ADAOO and the evolution of AD in individuals with a family history of the disease.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/diagnostics11050887/s1, Table S1: ML algorithms used to predict ADAOO; Figure S1: Variable importance bootstrap-based density distribution for ADAOO predictors in E280A AD; Figure S2: Variable importance bootstrap-based density distribution for ADAOO predictors in *s*AD.

Author Contributions: Conceptualization, J.I.V. and M.A.-B.; methodology, J.I.V. and L.A.S.; software, J.I.V. and L.A.S.; validation, M.A.-H., L.G.E., M.A.I.-R., F.L. and M.A.-B.; formal analysis, J.I.V. and L.A.S.; investigation, J.I.V., F.L. and M.A.-B.; resources, J.I.V. and L.A.S.; data curation, J.I.V., F.L. and M.A.-B.; writing—original draft preparation, J.I.V. and L.A.S.; writing—review and editing, J.I.V., M.A.-H., L.G.E., M.A.I.-R., F.L. and M.A.-B.; visualization, J.I.V. and L.A.S.; supervision, J.I.V., F.L. and M.A.-B.; project administration, J.I.V. and M.A.-B.; funding acquisition, J.I.V., L.A.S. and M.A.-B. All authors have read and agreed to the published version of the manuscript.

Funding: L.A.S. was financed by COLCIENCIAS and the Department of Atlántico, project "Desarrollo, Implementación y Comparación de Modelos Genómicos de Predicción para la Edad de Inicio de la Enfermedad de Alzheimer", grant #809-2018. L.A.S. is a MSc in Epidemiology student at Universidad del Norte. Some of this work is to be presented in partial fulfilment of the requirements for the MSc degree. J.I.V. and M.A.-B. have full access to all the data in the study. J.I.V. and M.A.-B. are responsible for submitting this work for publication.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of University of Antioquia, Medellín, Colombia (protocol #1115- 408-20543).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Acknowledgments: We express our highest appreciation to all patients and their relatives enrolled in this study for more than 30 years.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Caruso, A.; Nicoletti, F.; Mango, D.; Saidi, A.; Orlando, R.; Scaccianoce, S. Stress as risk factor for Alzheimer's disease. *Pharmacol. Res.* **2018**, *132*, 130–134. [CrossRef]
2. Fonteijn, H.M.; Modat, M.; Clarkson, M.J.; Barnes, J.; Lehmann, M.; Hobbs, N.Z.; Scahill, R.I.; Tabrizi, S.J.; Ourselin, S.; Fox, N.C.; et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* **2012**, *60*, 1880–1889. [CrossRef] [PubMed]
3. Ghanemi, A. Alzheimer's disease therapies: Selected advances and future perspectives. *Alex. J. Med.* **2015**, *51*, 1–3. [CrossRef]
4. Sancesario, G.M.; Bernardini, S. Alzheimer's disease in the omics era. *Clin. Biochem.* **2018**, *59*, 9–16. [CrossRef]
5. Barber, I.S.; Braae, A.; Clement, N.; Patel, T.; Guetta-Baranes, T.; Brookes, K.; Medway, C.; Chappell, S.; Guerreiro, R.; Bras, J.; et al. Mutation analysis of sporadic early-onset Alzheimer's disease using the NeuroX array. *Neurobiol. Aging* **2017**, *49*, 215-e1. [CrossRef]
6. Bialopiotrowicz, E.; Kuzniewska, B.; Kachamakova-Trojanowska, N.; Barcikowska, M.; Kuznicki, J.; Wojda, U. Cell cycle regulation distinguishes lymphocytes from sporadic and familial Alzheimer's disease patients. *Neurobiol. Aging* **2011**, *32*, 2319-e13. [CrossRef] [PubMed]
7. Cruchaga, C.; Del-Aguila, J.L.; Saef, B.; Black, K.; Fernandez, M.V.; Budde, J.; Ibanez, L.; Deming, Y.; Kapoor, M.; Tosto, G.; et al. Polygenic risk score of sporadic late-onset Alzheimer's disease reveals a shared architecture with the familial and early-onset forms. *Alzheimer's Dement.* **2018**, *14*, 205–214. [CrossRef]
8. Reiman, E.M.; Langbaum, J.B.; Tariot, P.N.; Lopera, F.; Bateman, R.J.; Morris, J.C.; Sperling, R.A.; Aisen, P.S.; Roses, A.D.; Welsh-Bohmer, K.A.; et al. CAP-advancing the evaluation of preclinical Alzheimer disease treatments. *Nat. Rev. Neurol.* **2016**, *12*, 56. [CrossRef]
9. Feigin, V.L.; Nichols, E.; Alam, T.; Bannick, M.S.; Beghi, E.; Blake, N.; Culpepper, W.J.; Dorsey, E.R.; Elbaz, A.; Ellenbogen, R.G.; et al. Global, regional, and national burden of neurological disorders, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **2019**, *18*, 459–480. [CrossRef]

10. Prince, M.; Bryce, R.; Albanese, E.; Wimo, A.; Ribeiro, W.; Ferri, C.P. The global prevalence of dementia: A systematic review and metaanalysis. *Alzheimer's Dement.* **2013**, *9*, 63–75. [CrossRef]
11. Wimo, A.; Guerchet, M.; Ali, G.C.; Wu, Y.T.; Prina, A.M.; Winblad, B.; Jönsson, L.; Liu, Z.; Prince, M. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's Dement.* **2017**, *13*, 1–7. [CrossRef] [PubMed]
12. Association, A. 2018 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **2018**, *14*, 367–429. [CrossRef]
13. Dubois, B.; Feldman, H.H.; Jacova, C.; Cummings, J.L.; DeKosky, S.T.; Barberger-Gateau, P.; Delacourte, A.; Frisoni, G.; Fox, N.C.; Galasko, D.; et al. Revising the definition of Alzheimer's disease: A new lexicon. *Lancet Neurol.* **2010**, *9*, 1118–1127. [CrossRef]
14. Musardo, S.; Marcello, E. Synaptic dysfunction in Alzheimer's disease: From the role of amyloid β-peptide to the α-secretase ADAM10. *Eur. J. Pharmacol.* **2017**, *817*, 30–37. [CrossRef] [PubMed]
15. Selkoe, D.J.; Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol. Med.* **2016**, *8*, 595–608. [CrossRef] [PubMed]
16. Dorfman, V.B.; Pasquini, L.; Riudavets, M.; López-Costa, J.J.; Villegas, A.; Troncoso, J.C.; Lopera, F.; Castaño, E.M.; Morelli, L. Differential cerebral deposition of IDE and NEP in sporadic and familial Alzheimer's disease. *Neurobiol. Aging* **2010**, *31*, 1743–1757. [CrossRef] [PubMed]
17. Haass, C.; De Strooper, B. The presenilins in Alzheimer's disease—Proteolysis holds the key. *Science* **1999**, *286*, 916–919. [CrossRef] [PubMed]
18. Jiao, B.; Tang, B.; Liu, X.; Xu, J.; Wang, Y.; Zhou, L.; Zhang, F.; Yan, X.; Zhou, Y.; Shen, L. Mutational analysis in early-onset familial Alzheimer's disease in Mainland China. *Neurobiol. Aging* **2014**, *35*, 1957-e1. [CrossRef]
19. Yuan, X.Z.; Sun, S.; Tan, C.C.; Yu, J.T.; Tan, L. The Role of ADAM10 in Alzheimer's Disease. *J. Alzheimer's Dis.* **2017**, *58*, 303–322. [CrossRef] [PubMed]
20. Xu, W.; Tan, L.; Yu, J.T. The Role of PICALM in Alzheimer's Disease. *Mol. Neurobiol.* **2015**, *52*, 399–413. [CrossRef] [PubMed]
21. Baig, S.; Joseph, S.A.; Tayler, H.; Abraham, R.; Owen, M.J.; Williams, J.; Kehoe, P.G.; Love, S. Distribution and expression of picalm in alzheimer disease. *J. Neuropathol. Exp. Neurol.* **2010**, *69*, 1071–1077. [CrossRef]
22. Cruchaga, C.; Karch, C.M.; Jin, S.C.; Benitez, B.A.; Cai, Y.; Guerreiro, R.; Harari, O.; Norton, J.; Budde, J.; Bertelsen, S.; et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* **2014**, *505*, 550–554. [CrossRef] [PubMed]
23. Hooli, B.V.; Lill, C.M.; Mullin, K.; Qiao, D.; Lange, C.; Bertram, L.; Tanzi, R.E. PLD3 gene variants and Alzheimer's disease. *Nature* **2015**, *520*, E7–E8. [CrossRef]
24. Ulland, T.K.; Colonna, M. TREM2—A key player in microglial biology and Alzheimer disease. *Nat. Rev. Neurol.* **2018**, *14*, 667–675. [CrossRef] [PubMed]
25. Carmona, S.; Zahs, K.; Wu, E.; Dakin, K.; Bras, J.; Guerreiro, R. The role of TREM2 in Alzheimer's disease and other neurodegenerative disorders. *Lancet Neurol.* **2018**, *17*, 721–730. [CrossRef]
26. Gratuze, M.; Leyns, C.E.G.; Holtzman, D.M. New insights into the role of TREM2 in Alzheimer's disease. *Mol. Neurodegener.* **2018**, *13*, 1–16. [CrossRef]
27. Wetzel-Smith, M.K.; Hunkapiller, J.; Bhangale, T.R.; Srinivasan, K.; Maloney, J.A.; Atwal, J.K.; Sa, S.M.; Yaylaoglu, M.B.; Foreman, O.; Ortmann, W.; et al. A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death. *Nat. Med.* **2014**, *20*, 1452–1457. [CrossRef] [PubMed]
28. Piaceri, I.; Nacmias, B.; Sorbi, S. Genetics of familial and sporadic Alzheimer's disease. *Front. Biosci.* **2013**, *5*, 167–177. [CrossRef]
29. Vélez, J.I.; Chandrasekharappa, S.C.; Henao, E.; Martinez, A.F.; Harper, U.; Jones, M.; Solomon, B.D.; Lopez, L.; Garcia, G.; Aguirre-Acevedo, D.C.; et al. Pooling/bootstrap-based GWAS (pbGWAS) identifies new loci modifying the age of onset in PSEN1 p.Glu280Ala Alzheimer's disease. *Mol. Psychiatry* **2013**, *18*, 568–575. [CrossRef]
30. Vélez, J.I.; Lopera, F.; Creagh, P.K.; Piñeros, L.B.; Das, D.; Cervantes-Henríquez, M.L.; Acosta-López, J.E.; Isaza-Ruget, M.A.; Espinosa, L.G.; Easteal, S.; et al. Targeting Neuroplasticity, Cardiovascular, and Cognitive-Associated Genomic Variants in Familial Alzheimer's Disease. *Mol. Neurobiol.* **2019**, *56*, 3235–3243. [CrossRef]
31. Vélez, J.I.; Lopera, F.; Patel, H.R.; Johar, A.S.; Cai, Y.; Rivera, D.; Tobón, C.; Villegas, A.; Sepulveda-Falla, D.; Lehmann, S.G.; et al. Mutations modifying sporadic Alzheimer's disease age of onset. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2016**, *171*, 1116–1130. [CrossRef] [PubMed]
32. Vélez, J.I.; Lopera, F.; Sepulveda-Falla, D.; Patel, H.R.; Johar, A.S.; Chuah, A.; Tobón, C.; Rivera, D.; Villegas, A.; Cai, Y.; et al. APOE∗E2 allele delays age of onset in PSEN1 E280A Alzheimer's disease. *Mol. Psychiatry* **2016**, *21*, 916–924. [CrossRef] [PubMed]
33. Vélez, J.I.; Lopera, F.; Silva, C.T.; Villegas, A.; Espinosa, L.G.; Vidal, O.M.; Mastronardi, C.A.; Arcos-Burgos, M. Familial Alzheimer's Disease and Recessive Modifiers. *Mol. Neurobiol.* **2020**, *57*, 1035–1043. [CrossRef]
34. Vélez, J.I.; Rivera, D.; Mastronardi, C.A.; Patel, H.R.; Tobón, C.; Villegas, A.; Cai, Y.; Easteal, S.; Lopera, F.; Arcos-Burgos, M.A. Mutation in DAOA Modifies the Age of Onset in PSEN1 E280A Alzheimer's Disease. *Neural. Plast.* **2016**. [CrossRef]
35. Naj, A.C.; Schellenberg, G.D. Genomic variants, genes, and pathways of Alzheimer's disease: An overview. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.* **2017**, *174*, 5–26. [CrossRef]
36. Reitz, C.; Mayeux, R. Use of genetic variation as biomarkers for alzheimer's disease. *Proc. Ann. N. Y. Acad. Sci.* **2009**, *1180*, 75. [CrossRef]

37. Acosta-Baena, N.; Sepulveda-Falla, D.; Lopera-Gómez, C.M.; Jaramillo-Elorza, M.C.; Moreno, S.; Aguirre-Acevedo, D.C.; Saldarriaga, A.; Lopera, F. Pre-dementia clinical stages in presenilin 1 E280A familial early-onset Alzheimer's disease: A retrospective cohort study. *Lancet Neurol.* **2011**, *10*, 213–220. [CrossRef]
38. Braak, H.; Del Tredici, K. Where, when, and in what form does sporadic Alzheimer's disease begin? *Curr. Opin. Neurol.* **2012**, *25*, 708–714. [CrossRef] [PubMed]
39. Lopera, F.; Ardilla, A.; Martínez, A.; Madrigal, L.; Arango-Viana, J.C.; Lemere, C.A.; Arango-Lasprilla, J.C.; Hincapié, L.; Arcos-Burgos, M.; Ossa, J.E.; et al. Clinical features of early-onset Alzheimer disease in a large kindred with an E280A presenilin-1 mutation. *J. Am. Med. Assoc.* **1997**, *277*, 793–799. [CrossRef]
40. Kessler, R.C.; Angermeyer, M.; Anthony, J.C.; DE Graaf, R.; Demyttenaere, K.; Gasquet, I.; DE Girolamo, G.; Gluzman, S.; Gureje, O.; Haro, J.M.; et al. Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry* **2007**, *6*, 168.
41. Kessler, R.C.; Amminger, G.P.; Aguilar-Gaxiola, S.; Alonso, J.; Lee, S.; Üstün, T.B. Age of onset of mental disorders: A review of recent literature. *Curr. Opin. Psychiatry* **2007**, *20*, 359. [CrossRef]
42. Beyer, K.; Lao, J.I.; Latorre, P.; Ariza, A. Age at onset: An essential variable for the definition of genetic risk factors for sporadic Alzheimer's disease. *Ann. N. Y. Acad. Sci.* **2005**, *1057*, 260–278. [CrossRef] [PubMed]
43. Kamboh, M.I.; Barmada, M.M.; Demirci, F.Y.; Minster, R.L.; Carrasquillo, M.M.; Pankratz, V.S.; Younkin, S.G.; Saykin, A.J.; Sweet, R.A.; Feingold, E.; et al. Genome-wide association analysis of age-at-onset in Alzheimer's disease. *Mol. Psychiatry* **2012**, *17*, 1340–1346. [CrossRef]
44. Naj, A.C.; Jun, G.; Reitz, C.; Kunkle, B.W.; Perry, W.; Park, Y.S.; Beecham, G.W.; Rajbhandary, R.A.; Hamilton-Nelson, K.L.; Wang, L.S.; et al. Effects of multiple genetic loci on age at onset in late-onset Alzheimer disease: A genome-wide association study. *JAMA Neurol.* **2014**, *71*, 1394–1404. [CrossRef]
45. Arcos-Burgos, M.; Muenke, M. Genetics of population isolates. *Clin. Genet.* **2002**, *61*, 233–247. [CrossRef]
46. Bravo, M.L.; Valenzuela, C.Y.; Arcos-Burgos, O.M. Polymorphisms and phyletic relationships of the Paisa community from Antioquia (Colombia). *Gene Geogr. Comput. Bull. Hum. Gene Freq.* **1996**, *10*, 11–17.
47. Kuhn, M. Package 'caret'—Classification and Regression Training. R Package Version 6.0-86. 2020. Available online: https://cran.r-project.org/package=caret (accessed on 21 January 2021).
48. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
49. Hall, A.; Pekkala, T.; Polvikoski, T.; van Gils, M.; Kivipelto, M.; Lötjönen, J.; Mattila, J.; Kero, M.; Myllykangas, L.; Mäkelä, M.; et al. Prediction models for dementia and neuropathology in the oldest old: The Vantaa 85+ cohort study. *Alzheimer's Res. Ther.* **2019**, *11*, 1–12. [CrossRef] [PubMed]
50. Qiu, R.G.; Qiu, J.L.; Badr, Y. Predictive modeling of the severity/progression of Alzheimer's diseases. In Proceedings of the 2017 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2017), Stockholm, Sweden, 8–11 August 2017.
51. 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]
52. Nori, V.S.; Hane, C.A.; Crown, W.H.; Au, R.; Burke, W.J.; Sanghavi, D.M.; Bleicher, P. Machine learning models to predict onset of dementia: A label learning approach. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2019**, *5*, 918–925. [CrossRef] [PubMed]
53. Porto, A.; Peralta, J.M.; Blackburn, N.B.; Blangero, J. Reliability of genomic predictions of complex human phenotypes. *BMC Proc.* **2018**, *12*, 157–161. [CrossRef]
54. Spiliopoulou, A.; Nagy, R.; Bermingham, M.L.; Huffman, J.E.; Hayward, C.; Vitart, V.; Rudan, I.; Campbell, H.; Wright, A.F.; Wilson, J.F.; et al. Genomic prediction of complex human traits: Relatedness, trait architecture and predictive meta-models. *Hum. Mol. Genet.* **2015**, *24*, 4167–4182. [CrossRef]
55. Dey, A. Machine Learning Algorithms: A Review. *Int. J. Comput. Sci. Inf. Technol.* **2016**, *7*, 1174–1179.
56. Dhall, D.; Kaur, R.; Juneja, M. Machine learning: A review of the algorithms and its applications. *Proc. ICRIC* **2020**, *2019*, 47–63.
57. Shatte, A.B.R.; Hutchinson, D.M.; Teague, S.J. Machine learning in mental health: A scoping review of methods and applications. *Psychol. Med.* **2019**, *49*, 1426–1448. [CrossRef] [PubMed]
58. Stamate, D.; Kim, M.; Proitsi, P.; Westwood, S.; Baird, A.; Nevado-Holgado, A.; Hye, A.; Bos, I.; Vos, S.J.B.; Vandenberghe, R.; et al. A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort. *Alzheimer's Dement. Transl. Res. Clin. Interv.* **2019**, *5*, 933–938. [CrossRef]
59. Bryan, R.N. Machine learning applied to Alzheimer disease. *Radiology* **2016**. [CrossRef] [PubMed]
60. Fisher, C.K.; Smith, A.M.; Walsh, J.R.; Simon, A.J.; Edgar, C.; Jack, C.R.; Holtzman, D.; Russell, D.; Hill, D.; Grosset, D.; et al. Machine learning for comprehensive forecasting of Alzheimer's Disease progression. *Sci. Rep.* **2019**, *9*, 1–14.
61. Liu, L.; Zhao, S.; Chen, H.; Wang, A. A new machine learning method for identifying Alzheimer's disease. *Simul. Model. Pract. Theory* **2020**, *99*, 102023. [CrossRef]
62. Fulton, L.V.; Dolezel, D.; Harrop, J.; Yan, Y.; Fulton, C.P. Classification of alzheimer's disease with and without imagery using gradient boosted machines and resnet-50. *Brain Sci.* **2019**, *9*, 212. [CrossRef]
63. Khan, A.; Usman, M. Early diagnosis of Alzheimer's disease using machine learning techniques: A review paper. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Lisbon, Portugal, 12–14 November 2015.

64. Londono, A.C.; Castellanos, F.X.; Arbelaez, A.; Ruiz, A.; Aguirre-Acevedo, D.C.; Richardson, A.M.; Easteal, S.; Lidbury, B.A.; Arcos-Burgos, M.; Lopera, F. An 1H-MRS framework predicts the onset of Alzheimer's disease symptoms in PSEN1 mutation carriers. *Alzheimer's Dement.* **2014**, *10*, 552–561. [CrossRef]
65. De Velasco Oriol, J.; Vallejo, E.E.; Estrada, K.; Taméz Peña, J.G.; The Alzheimer's Disease Neuroimaging Initiative. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinform.* **2019**, *20*, 1–17. [CrossRef]
66. Palmqvist, S.; Janelidze, S.; Quiroz, Y.T.; Zetterberg, H.; Lopera, F.; Stomrud, E.; Su, Y.; Chen, Y.; Serrano, G.E.; Leuzy, A.; et al. Discriminative Accuracy of Plasma Phospho-tau217 for Alzheimer Disease vs Other Neurodegenerative Disorders. *JAMA J. Am. Med. Assoc.* **2020**, *324*, 772–781. [CrossRef] [PubMed]
67. Grassi, M.; Perna, G.; Caldirola, D.; Schruers, K.; Duara, R.; Loewenstein, D.A. A clinically-translatable machine learning algorithm for the prediction of Alzheimer's disease conversion in individuals with mild and premild cognitive impairment. *J. Alzheimer's Dis.* **2018**, *61*, 1555–1573. [CrossRef] [PubMed]
68. Moradi, E.; Pepe, A.; Gaser, C.; Huttunen, H.; Tohka, J. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* **2015**, *104*, 398–412. [CrossRef] [PubMed]
69. Ezzati, A.; Zammit, A.R.; Harvey, D.J.; Habeck, C.; Hall, C.B.; Lipton, R.B. Optimizing Machine Learning Methods to Improve Predictive Models of Alzheimer's Disease. *J. Alzheimer's Dis.* **2019**, *71*, 1027–1036. [CrossRef] [PubMed]
70. Ding, Y.; Sohn, J.H.; Kawczynski, M.G.; Trivedi, H.; Harnish, R.; Jenkins, N.W.; Lituiev, D.; Copeland, T.P.; Aboian, M.S.; Aparici, C.M.; et al. A deep learning model to predict a diagnosis of Alzheimer disease by using 18 F-FDG PET of the brain. *Radiology* **2019**, *290*, 456–464. [CrossRef]
71. Jo, T.; Nho, K.; Saykin, A.J. Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. *Front. Aging Neurosci.* **2019**, *11*, 220. [CrossRef]
72. Fleisher, A.S.; Chen, K.; Quiroz, Y.T.; Jakimovich, L.J.; Gomez, M.G.; Langois, C.M.; Langbaum, J.B.S.; Ayutyanont, N.; Roontiva, A.; Thiyyagura, P.; et al. Florbetapir PET analysis of amyloid-β deposition in the presenilin 1 E280A autosomal dominant Alzheimer's disease kindred: A cross-sectional study. *Lancet Neurol.* **2012**, *11*, 1057–1065. [CrossRef]
73. Reiman, E.M.; Langbaum, J.B.S.; Fleisher, A.S.; Caselli, R.J.; Chen, K.; Ayutyanont, N.; Quiroz, Y.T.; Kosik, K.S.; Lopera, F.; Tariot, P.N. Alzheimers prevention initiative: A plan to accelerate the evaluation of presymptomatic treatments. *J. Alzheimer's Dis.* **2011**, *26* (Suppl. S3), 321–329. [CrossRef]
74. Reiman, E.M.; Quiroz, Y.T.; Fleisher, A.S.; Chen, K.; Velez-Pardo, C.; Jimenez-Del-Rio, M.; Fagan, A.M.; Shah, A.R.; Alvarez, S.; Arbelaez, A.; et al. Brain imaging and fluid biomarker analysis in young adults at genetic risk for autosomal dominant Alzheimer's disease in the presenilin 1 E280A kindred: A case-control study. *Lancet Neurol.* **2012**, *11*, 1048–1056. [CrossRef]
75. Petersen, R.C.; Smith, G.E.; Waring, S.C.; Ivnik, R.J.; Tangalos, E.G.; Kokmen, E. Mild cognitive impairment: Clinical characterization and outcome. *Arch. Neurol.* **1999**, *56*, 303–308. [CrossRef]
76. Aguirre-Acevedo, D.C.; Jaimes-Barragán, F.; Henao, E.; Tirado, V.; Muñoz, C.; Reiman, E.M.; Tariot, P.N.; Langbaum, J.B.; Lopera, F. Diagnostic accuracy of CERAD total score in a Colombian cohort with mild cognitive impairment and Alzheimer's disease affected by E280A mutation on presenilin-1 gene. *Int. Psychogeriatr.* **2016**, *28*, 503. [CrossRef] [PubMed]
77. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, 4th ed.; American Psychiatric Association: Washington, DC, USA, 2000.
78. Segura, V.; Vilhjálmsson, B.J.; Platt, A.; Korte, A.; Seren, Ü.; Long, Q.; Nordborg, M. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* **2012**, *44*, 825. [CrossRef] [PubMed]
79. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
80. Vélez, J.I.; Correa, J.C.; Arcos-Burgos, M. A New Method for Detecting Significant *p*-values with Applications to Genetic Data. *Rev. Colomb. Estadística* **2014**, *37*, 69–78. [CrossRef]
81. R Core Team. R: A Language and Environment for Statistical Computing. 2021. Available online: https://www.R-project.org/ (accessed on 21 January 2021).
82. MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Berkeley, CA, USA, 1967.
83. Ringnér, M. What is principal component analysis? *Nat. Biotechnol.* **2008**, *26*, 303–304. [CrossRef]
84. Lever, J.; Krzywinski, M.; Altman, N. Points of Significance: Principal component analysis. *Nat. Methods* **2017**, *14*, 7. [CrossRef]
85. Charrad, M.; Ghazzali, N.; Boiteau, V.; Niknafs, A. Nbclust: An R package for determining the relevant number of clusters in a data set. *J. Stat. Softw.* **2014**, *61*, 1–36. [CrossRef]
86. Dinno, A. Paran: Horn's Test of Principal Components/Factors. R Package Version 1.5.2. 2018. Available online: https://cran.r-project.org/web/packages/paran/index.html (accessed on 2 March 2021).
87. Efron, B. Bootstrap Methods: Another Look at the Jackknife. In *Breakthroughs in Statistics*; Springer: New York, NY, USA, 1979.
88. Efron, B.; Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat. Sci.* **1986**, *1*, 54–75. [CrossRef]
89. Tenev, A.; Markovska-Simoska, S.; Kocarev, L.; Pop-Jordanov, J.; Müller, A.; Candrian, G. Machine learning approach for classification of ADHD adults. *Int. J. Psychophysiol.* **2014**, *93*, 162–166. [CrossRef] [PubMed]

90. Kautzky, A.; Vanicek, T.; Philippe, C.; Kranz, G.S.; Wadsak, W.; Mitterhauser, M.; Hartmann, A.; Hahn, A.; Hacker, M.; Rujescu, D.; et al. Machine learning classification of ADHD and HC by multimodal serotonergic data. *Transl. Psychiatry* **2020**, *10*, 1–9.
91. Jamal, S.; Khubaib, M.; Gangwar, R.; Grover, S.; Grover, A.; Hasnain, S.E. Artificial Intelligence and Machine learning based prediction of resistant and susceptible mutations in Mycobacterium tuberculosis. *Sci. Rep.* **2020**, *10*, 1–16.
92. Goldenberg, S.L.; Nir, G.; Salcudean, S.E. A new era: Artificial intelligence and machine learning in prostate cancer. *Nat. Rev. Urol.* **2019**, *16*, 391–403. [CrossRef] [PubMed]
93. Zhu, W.; Xie, L.; Han, J.; Guo, X. The application of deep learning in cancer prognosis prediction. *Cancers* **2020**, *12*, 603. [CrossRef] [PubMed]
94. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
95. Agrebi, S.; Larbi, A. Use of artificial intelligence in infectious diseases. In *Artificial Intelligence in Precision Health*; Academic Press: Cambridge, MA, USA, 2020.
96. Bhardwaj, T.; Somvanshi, P. Machine learning toward infectious disease treatment. In *Machine Intelligence and Signal Analysis*; Springer: Singapore, 2019.
97. Vidal, O.M.; Acosta-Reyes, J.; Padilla, J.; Navarro-Lechuga, E.; Bravo, E.; Viasus, D.; Arcos-Burgos, M.; Vélez, J.I. Chikungunya outbreak (2015) in the colombian caribbean: Latent classes and gender differences in virus infection. *PLoS Negl. Trop. Dis.* **2020**, *14*, e0008281. [CrossRef]
98. Golriz Khatami, S.; Mubeen, S.; Hofmann-Apitius, M. Data science in neurodegenerative disease: Its capabilities, limitations, and perspectives. *Curr. Opin. Neurol.* **2020**, *33*, 249. [CrossRef]
99. Mihaescu, R.; Detmar, S.B.; Cornel, M.C.; Van Der Flier, W.M.; Heutink, P.; Hol, E.M.; Rikkert, M.G.M.O.; Van Duijn, C.M.; Janssens, A.C.J.W. Translational research in genomics of Alzheimer's disease: A review of current practice and future perspectives. *J. Alzheimer's Dis.* **2010**, *20*, 967–980. [CrossRef]
100. Freudenberg-Hua, Y.; Li, W.; Davies, P. The role of genetics in advancing precision medicine for Alzheimer's Disease—A narrative review. *Front. Med.* **2018**, *5*, 108. [CrossRef]
101. Hampel, H.; Vergallo, A.; Perry, G.; Lista, S. The Alzheimer Precision Medicine Initiative. *J. Alzheimer's Dis.* **2019**, *68*, 1–24. [CrossRef] [PubMed]
102. Collins, F.S.; Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **2015**, *372*, 793–795. [CrossRef] [PubMed]
103. Sabau, M.; Bungau, S.; Buhas, C.L.; Carp, G.; Daina, L.G.; Judea-Pusta, C.T.; Buhas, B.A.; Jurca, C.M.; Daina, C.M.; Tit, D.M. Legal medicine implications in fibrinolytic therapy of acute ischemic stroke. *BMC Med. Ethics* **2019**, *20*, 1–9. [CrossRef]
104. Dindelegan, C.; Faur, D.; Purza, L.; Bumbu, A.; Sabau, M. Distress in neurocognitive disorders due to Alzheimer's disease and stroke. *Exp. Ther. Med.* **2020**, *20*, 2501–2509. [CrossRef] [PubMed]
105. Bone, D.; Goodwin, M.S.; Black, M.P.; Lee, C.C.; Audhkhasi, K.; Narayanan, S. Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises. *J. Autism Dev. Disord.* **2015**, *45*, 1121–1136. [CrossRef] [PubMed]