Medical neuroimage consolidation from a grid of hospitals for the external validation of predictive models

Rui Nóbrega
Instituto Superior Técnico
Lisbon, Portugal
Ruiqnobrega@tecnico.ulisboa.pt

ABSTRACT

The diagnosis of Alzheimer's disease is only certain with a detailed post-mortem microscopic examination of the brain. Machine learning approaches are increasingly used in the development of predictive models for the early diagnosis of Alzheimer´s disease. The major issues with such models are the lack of interpretability at the clinical end and the lack of generalization of said models due to the heterogeneity of the data sources (instrumentation, monitoring protocol, individual demographics). To tackle these issues, this work proposes a multi-diagnostic, multi-site, clinically interpretable tool using MRI imaging. Furthermore, it presents the steps for the data consolidation where the MRIs are extracted from heterogeneous sources and are anonymized in order to maintain the anonymity of the patients subjected to the study. In addition, the performance of the models is externally validated on data obtained independently according to temporal, geographic, and/or domain differences. The models could not generalize well for the target population as they generalized for the testing partitions of the original data. Out of the three possible class labels, class Control showed the worst results, returning 100% of precision yet significantly low levels of recall. MCI and AD classes returned similar results of precision, 29% and 30% respectively, however, AD had 83% of recall whereas MCI only 43%. The gathered observations confirm the difficulty of performing neuroimaging diagnostics under the different monitoring protocols, medical classifications, and population demographics.

Author Keywords

Medical Resonance Imaging - Alzheimer´s disease - Mild Cognitive Impairment - Predictive modeling - External Validation - Data Consolidation.

INTRODUCTION

Dementia is a class of diseases associated with losses of memory and thinking abilities considerable enough to interfere with the daily life of a person. Dementia associated diseases include Alzheimer's disease, Vascular dementia, Lewy body dementia, Parkinson's disease and others.The work here presented, focus on a specific dementia disease, Alzheimer's disease, representing two thirds of the total cases of dementia [12]. Currently, to diagnose such disease with total certainty is only possible with a detailed post-mortem microscopic examination of the brain [8]. The fact that, for the time being, it is not always easy to diagnose a patient with Alzheimer's disease while still alive or even at an early stage of progression does not mean that we should not discard the presence of more robust diagnostic methods to be discovered. In fact, it is possible to diagnose patients with Alzheimer's with around 95 percent accuracy by using different types of tools for the purpose. The tools that might be used to diagnose a patient are based on studying the history of the patients and their families and with that, it is then possible to assess cognitive function by neuropsychological tests. The biggest problem with such solution is that is highly dependable on medical professionals to determine the diagnose and such diagnose might take several weeks to be accomplished. In addition, the diagnose may only be performed already in a later stage of the disease when it is harder to delay or reverse the development of the disease.

Problem Description

More and more approaches based on machine learning have been used in order to develop models capable of providing an early and accurate diagnosis of Alzheimer's disease or even of a preliminary state of cognitive impairment preceding Alzheimer's at a later time of life. The biggest setback of such models is the need to guarantee their interpretability in face of the complex data available (combining imagiology, cognitive scoring exams, demography, and clinical records) and the need to guarantee their adequate generalization ability on external data i.e data the model has never seen.

At the moment, Magnetic Resonance Imaging based personalized diagnostic tools for dementia are still scarce due to the several difficulties that arise when handling such models. The acquired data to feed the models for classification is massive and heterogeneous in nature. When an MRI is performed, the output of such exam is a compilation of images displaying the brain of the patient in 3 dimensions, with a general resolution of over X thousands voxel [17]. In addiction, each image of the exam combines medical data, with static demographic information concerning the patient and the physician involved in the exam.

Such data must be properly processed for research ends. As previously mentioned, these Magnetic Resonance images contain information about the patient and some of it must be anonymized due to the patient anonymity that must be maintained. The anonymization must be performed in ways that it makes the identification of the patient impossible to the researchers and easy for the hospital or clinic, once it receives an output from the models.

MRIs can be acquired using different technologies and protocols [13] so, it is only expectable that, before such images are handled by the models, they must be pre-processed. There are several protocols of acquisition, although, the two structural protocols of relevance for this paper are *Magnetization prepared rapid gradient echo (MP-RAGE)* and *Spoiled gradient recalled echo (SPGR)*, explained in the Background section.

Research contributions

The main goal is to develop a multi-diagnostic, multi-site, and clinically interpretable tool for early diagnosis of AD using MRI imaging initially collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and later, from several different hospitals or clinics. The solution proposed will also extend learning and assessment to new populations (new cohorts).

In accordance, the major contributions placed by our work are:

1. Validate the predictive power of the developed MRI diagnostic models for dementia in a Portuguese population using retrospective data.
   (a) Collect anonymized retrospective clinical and neuroimaging data from Portuguese hospitals of the consorcim study NEUROBIOAI.
   (b) Build a database capable of storing the images received by the hospitals.
   (c) Test existing models with such data.
   (d) Perform an external validation on the models with information the models have not yet seen.
2. Build an interface capable of visualizing important data from the database and manipulate it at will.

RELATED WORK

The consolidation of data and its external validation are not novel topics in computer science, although, the purpose of the application domain and the unique challenges associated with the available data sources make a given project unique based on its own constraints. Therefore, the purpose of this section is to present a compilation of related work.

Data Consolidation

Volosnikov et al. [18] proposed a tool capable of allowing unified access to heterogeneous and distributed data. According to the paper, the heterogeneity of data sources increases the difficulty to perform comprehensive research. Furthermore, the data presenting the heterogeneous indicators of medical exams range in type, some even might be considered heavy, for instance, MRI or fMRI. Such images demand an intensive preprocessing phase in order to use them in a research analysis. To tackle the problems that arise due to the heterogeneity of data sources and the the required preprocessing of the images, the committee of authors then introduces the developed tool. Such implementation makes use of a service-oriented architecture, commonly known as SOA, preventing a series of problems that, otherwise, would have arisen. Compliance to the law when it comes to handling the personal information of each patient subjected to the study along with scaling difficulties and the use of new resources are examples of problems could have appeared. The tool developed uses python libraries to access and store the heterogeneous data and the interface and work environment of the tool was implemented using MEAN stack or Mongo, Angular, Express, Node. Similar to the work previously mentioned [18], the solution presented in this paper must handle the consolidation of medical imaging and all the problems that may emerge with it. The data used to preform our own analysis is handed by a grid of hospitals hence the heterogeneity of the data sources. Since the files gathered and being used are raw MRIs, it is also necessary for them to undergo a preprocessing phase in order to analyse them. Since each hospital is a unique case, each one of them demands a different way to extract required data from the servers into our workstation. For the time being, web scrapers are being used to access and extract data from the PACS system of each hospital. The interface of the solution here proposed is implemented in python as for all the access and storage of content in the database created in our workstation. Since the anonymization of data is performed by the hospital by a script developed alongside the proposed solution, compliance to the law in terms of handling personal information does not raise a problem since the anonymization process was accepted by each supplier of data and the research work is compliant with HIPAA, GDPR and other data privacy regulations.

Data Warehousing, as the name suggests, it is used to store data from disparate sources. The work of Saliya Nugawela[11] identifies the main obstacles of data integration of healthcare data and the proposal of a data warehousing model capable of integrating fragmented data in a cardiac surgery unit. The work proposes a star schema to organise the data collected along with an enterprise architecture. The main difference between such solution and the one presented in this thesis, is that the solution presented here follows a snowflake schema. The less space it is wasted, the more information can be stored. In a star schema a lot of the information turns out to be redundant whereas in the snowflake there is almost no redundancy.

Biomarkers Discovery

Bocchetta et al. [3] studied the relevance of AD biomarkers such as cerebrospinal fluid (CSF), medial temporal atrophy (MTA), fluorodeoxyglucose positron emission tomography (FDG-PET) and amyloid-PET by AD European consortium centers, obtained by inspecting MRIs in the diagnosis of MCI. According to the article, the most used biomarker is clearly MTA with 75% of the respondents claiming to always or at least frequently use it. The second most used is CSF markers with 22% of respondents using it, followed by FDG-PET with 16% and finally amyloid-PET with 3%. In terms of confidence in the use of such markers in the early diagnosis of MCI, only 45% of the consortium centers that answered the survey considered that MTA had a "moderate" contribution to the diagnosis whereas 79% felt "very/extremely" confident in a diagnosis of early MCI due to AD when levels of amyloid and neural injury biomarkers were abnormal, especially when the measurement of the levels of both were simultaneously abnormal, thus, being an indicative of AD signature.

Other literature corroborates conclusions as the ones, aforementioned, for instance, the work by Jack et al. [5] in which, the authors provide a framework developed with the purpose of testing hypothesis presenting correlations between changes in AD biomarkers throughout time and clinical diseases stages or even between temporal changes in AD biomarkers themselves.It is possible to understand that biomarkers, detected through the use of structural MRI might not be as relevant as one would predict despite the frequent use of said markers due to abnormalities only presenting at a later stage of the disease. On the other hand β-amyloid abnormalities seem to appear at an early stage of the disease, thus, corroborating the highly confidence level in a diagnosis where amyloid levels were abnormal.

External validation

Despite the fact that a predictive model is validated internally, an external validation is required and essential since, in that way, it is possible to test the model on a population acquired in an independent way. By doing so, external validation allows for the assessment on the generalization of a predictive model, allowing for a better understanding on how the model performs on a new population.

Most of the predictive models used in a Alzheimer's disease related issues make use of deep learning techniques. According to the work of Qiu et al. [14] there is a lack of external validation methods being implemented in deep learning techniques based predicted models since such models are developed, i.e. trained and tested, with data from a single group of subjects who share a defining characteristic. The fact that a lack of external validation methods exists, deep learning models applied to AD tend to fall short on the expected outcome considering the fact that such models have a decrease on performance and their comprehensibility is limited since these models work as a "black-box" and provide no elucidate diagnostic review.

Furthermore, external validation is necessary in prediction research. The work of Bleeker et al. [2] elucidates the fact that predictive models tend to perform better when facing data used to train and develop the model rather than when facing data new to the model. The results from predictive models tend to be considered with regard to the internal validation and with almost no regard for the external one. Bleeker et al. [2] present the limitations to internal validation, therefore, expressing the importance of external validation. The predictive model used in the paper aims at classifying the presence of serious bacterial infections in children with fever (total amount of 376). Internal evaluated performance on average of 0.83 for the apparent area under the receiver operating characteristic curve and 0.76 after applying a bootstrapping method to provide bias-corrected estimates of model performance. After validating the model internally, a small set of 179 individuals was validated externally and the authors obtained a performance of 0.57 proving that only validating a small data set internally is not enough and in the future models who do it, tend to fall short on performance. External validating is, therefore, considered essential and vital to be performed on a model before inserting it in clinical practices.

To summarise, let us consider the work of Siontis et al. [15] where the goal of the authors was to evaluate how often newly developed risk prediction models undergo external validation and how well they perform in such validations. The method used to try and find an answer was to evaluate 127 new prediction models. Only in about 25% of the models, an external validation was encountered and that the probability of having such validation method to be performed by different authors was 16% proving that external validation of predictive models in different studies is uncommon and, therefore, their performance might be considerably lower when facing said validation.

To perform a clear external validation on a predictive model, it is necessary to expose such model to different data, that it has not encountered before. The difference in the data has to be, according to Moons et al. [7], in these parameters:

1. Temporal differences so that a temporal external validation might be performed since the individuals presented on the data that the model is facing belong to the same cohort but to different time periods.

2. Geographical difference in the data allow for a geographical external validation considering new individuals from different locations, this is, patients subject to prediction by the model are from a different clinic or hospital.

3. Domain differences express new individuals who are considerably different from the individuals from which the model was developed representing, thus, a domain validation.

The procedure, then, consists in applying the models to the data with the aforementioned differences and recalculating the performance of the model based on discrimination, calibration and classification measures.

Internal and internal-external validation

Depending on the literature and on the predictive model case where such model is inserted, external validation may or may not be essential to correct the model due to low values in perfomance. However, internal validation and, in some cases, internal-external validation are some types of validations that are present in the developmento of the models. The work by Steyerberg et al. [16] expresses the fact that internal validation is essential and the preferred method for validation is the bootstrapping aproach to estimate the performance of the model. Since some type of external validation might be considered in time of development, the authors also recommend an internal-external validation. That way, the model is tested with a different sample, although with the same characteristics, keeping the model from returning overly optimistic performance values, thus, offering a more realistic assessment.

PROPOSED SOLUTION

To better understand the solution at hand, first, we need to acknowledge the functional requirements of said solution. Such requirements aim to represent what the developed solution must be able to do and how it performs a certain task given a specific input by the user. Depending on the goal, different tasks must be performed to obtain the correct output for the user query. For that reason, the objectives of this dissertation are:

1. Guarantee proper anonymization and privacy of neuroimaging data.

2. Receive imaging from the hospitals/clinics.

3. Easily and quickly handle input from users.

4. Store required data in a database.

5. Return analysis based on the individuals from an hospital e.g. diagnosis.

6. Perform external validation on the new population showing the generalization guarantees and vulnerabilities.

For the better understanding of the reader, it is important to note that the solution here presented was developed with the purpose of aiding the Institute of Biophysics and Biomedical Engineering under the project NEUROBIOAI. The need for a solution capable of storing the data acquired from the partner hospitals as well as a critical analysis of how the predictive models behave under a new population, the Portuguese one, resulted in the solution presented in the next sections. The project presented consists of four main subgroups:

1. **Data consolidation** - Consists of acquiring and handling images from the partner hospitals related to the patients at a specific hospital. Such images must be anonymized and, later, stored in order for the predictive models to have access to this new information. Such data is consolidated in the developed database.

2. **Service layer** - This layer is meant for handling all sorts of requests to access and alter the database if necessary.

3. **External validation** - The models, after proper training, must undergo a critical analysis so that it is possible to assess how the models handle new data.

4. **Graphical user interface** - In order for the users to have access to the database, a centralized app (GUI) was developed in order to manipulate the database as the user sees fit.

Data Consolidation

For a better classification of future patients, the predictive models need not only images from ADNI but also require images from hospitals or clinics. The preprocessing and spatial alignment on new data is essential to make images more easily comparable, but not necessarily similar. In that matter, before receiving such images, it is necessary to prepare them and, only after, extract such data from the partner hospitals and clinics.

Data anonymization

The first stage of the extraction of the images is to anonymize the information that might be identifiable of the patient. Inside each DICOM file, besides the image itself, there are several tags with information regarding the patient. The main goal is to de-identify or remove data from the DICOM files from the hospitals/clinics, thus, enabling the sharing of such images to outside of the hospital guard without breaking any security and data privacy protocols.

With this process, the data is anonymized to the entity receiving the images, as it is not possible for the receiving end to obtain the original values from the images or find the original person behind the anonymization as the data is either removed or de-associated from the patient as a consequence of using hash keys to replace ids. The data is de-identified to the hospital/clinic end since the alterations to the image are stored, thus, enabling the hospital/clinic to identify the patient once a diagnose has been made by the receivers end (IBEB).

The process of anonymization consists in deleting or replacing with random values DICOM's header tags that allow for identification of the patient. Tags that are anonymized are permanently de-identified from their source. On the other hand, de-identification of some tags replaces the tags' values with artificial identifiers, random key of 10 characters, that can still be used to re-identify the patient, but only by authorized personnel of the hospital sharing the data.

Authors in different literature [1, 10, 9] provide different advice regarding the removal of some of the tags kept within the scope of the research project. It must be emphasized, though, that several tags, present in the appendix of the dissertation, are either de-identified – and only re-identifiable by the hospital - or completely anonymized as their nature does not allow them to be used in patient re-identification efforts. The decisions regarding the anonymization of the images are considered and thought of under the hospitals/clinics supervision. Before implementing such a script, the project's partners must accept and agree on the process explained above.

Data Storage

After the extraction of the images from the hospitals, it is required to have someplace to store the content of said images. For that reason, a local database must be created with the required tables to store the data that feeds the models for classification.

The solution that fits best the requirements is a relational model due to several reasons such as:

1. All the information can be stored in a single database so, OLAP functionality would not be that much of an asset.

2. Since one of the main reasons is for the scientists at IBEB to consult the data as it is stored in the database without any integration performed to it, a relational model suits the problem better.

3. Each image contains a high amount of data and that amount must be multiplied by hundreds of thousand other images, queries would take substantial time to run.

4. Future users of the database are not experienced in this matter so there is a need for a simple, efficient, and free way of inserting and manipulating the data.

5. Several data that was not relevant to the project was discarded in the data anonymization process so, the data being stored is of the utmost importance and must not be summarized as the user may need to see the raw content of each entry.

Although a multidimensional approach would benefit the project, after careful consideration and since the database would run at a local level with limited access, a relational approach was a more suitable way to store the data as it also leaves space for a multidimensional approach in the future if the project has such necessities, through the use of a ROLAP (Relational On-line Analytical Processing) method creating a new layer on top of the relational one.

Graphical user interface

The primary goal taken into account during the development of the GUI was to allow the user to insert new data into the database without having to write any SQL query in the console. The GUI allows for the easy and fast insertion of new patients into the database with very few interactions or effort.

To sum up, the GUI must perform the following requisites:

1. Add new patients by providing the age of the first visit of the patient at the hospital and the preliminary diagnosis. When adding a patient, it is also possible to select the Dicom images of the said patient from a directory.

2. Access and display the tables with the data from a patient or all patients, among other relevant data.

The GUI was developed in python with several libraries, including Dash, the framework where the interface is built on.

Input files

As previously mentioned, the main goal of the GUI is to allow for the insertion of new patients. The hospitals and clinics send a CSV file with minimal information regarding the patients together with all the DICOM files concerning the patients' MRI.

For that matter, the developed app is prepared for receiving simultaneously a CSV and all the images the user wants to. The service layer, then, checks the database for duplicates and in case it finds, it does not insert the content of said MRIs into the database.

View Data

Not only the insertion of new data was considered in the development of the database. One great asset of the Dash library is that, since it uses the Plotly library, it is capable of displaying several visualization tools that able the user in terms of getting to know the population present in the database.

In order for a better understanding, the figures ahead display the database populated with only a few patients.

Figure 1 displays a parallel coordinates chart capable of displaying vital information about the subjects such as age, gender, the hospital where the images were taken, and the diagnosis attributed to the patient. This visualization tool also represents the lines in such a way that it displays the diagnosis by color.

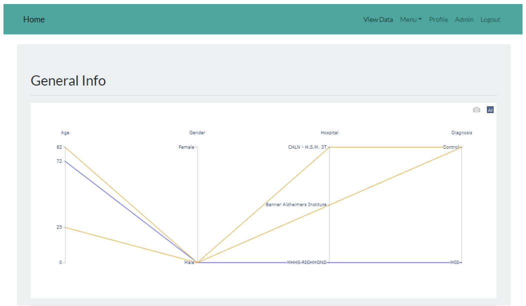


Figure 1: Parallel coordinates displaying a general view of the database

One great concern was to enable the user to quickly get statistical values regarding different aspects of the database population. Figure 2 shows data about the Diagnosis gender, imaging protocols of acquisition, and the image source.

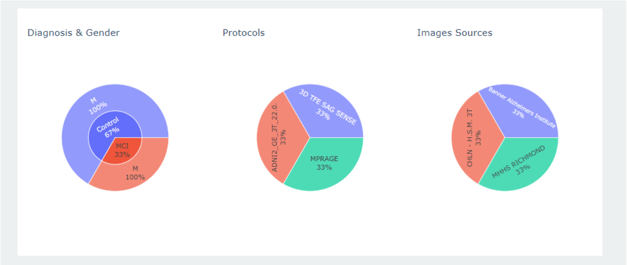


Figure 2: One sunburst and two pie charts displaying relevant information about the images and patients

The user is also able to get an idea of how each class of interest is affecting the subject of several ages. The scatter plot in 3 shows just that since it plots for each age and diagnose how many patients there are with that same criteria.

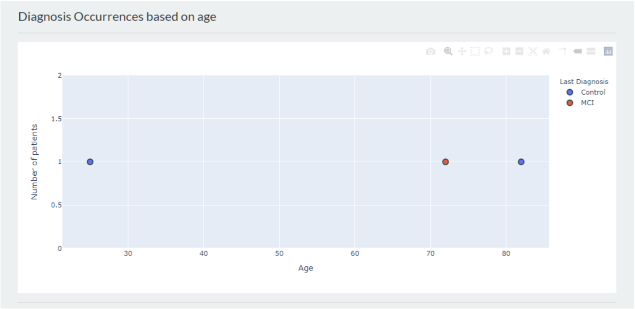
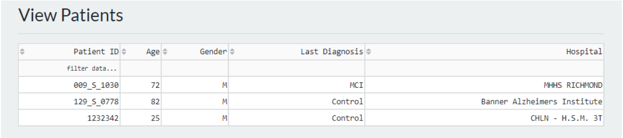


Figure 3: Number of patients / Age

Lastly, the user is able to consult all the tables as he/she sees fit as the example on Figure 4 shows. The only table that is not being displayed is the table regarding the usernames and other details of each profile of a user in the database. Only a user assigned with such privileges can access such data.



View Patients

Patient ID	Age	Gender	Last Diagnosis	Hospital
filter data...				
009_S_1030	72	M	MCI	MHHS RICHMOND
129_S_0778	82	M	Control	Banner Alzheimers Institute
1232342	25	M	Control	CHLN - H.S.M. 3T

Figure 4: Example of the table regarding some information about a patient

External Validation

The goal of this section is to describe the process that took place in order to assess the performance of the models in a real-life situation. For that measure, the models were executed using data from hospitals that are partners of the project. Several hospitals are joining but in the time this thesis was developed only two hospitals, Hospital Vila Franca de Xira and Hospital Fernando Fonseca, were able to supply medical images in the available time span.

Measures in training

In order to assess how well the models are generalizing we considered the use of learning curves and calculation of the bias and variance which allowed understand the following:

1. The variation of performance by varying the number of patients used in the training process;

2. If the models are properly fitted or if they are overfitting or underfitting;

3. If the dataset used in the training and in the validation is representative of the population;

With that said, let us first get into the dataset used to train the models. Such dataset is composed of patients classified as Control (651 patients), MCI (191 patients) or AD (241 patients). There are seven models for each one of the three different scenarios. The dataset that contains the patients represents the original population which is then split so that 70% of the available data is used in the training leaving the other 30% for testing. These models were trained with patients who were classified with either one of the two classes in a scenario, i.e.:

1. **Control vs MCI**: Consists on a dataset of patients classified by the hospitals as Control or MCI

2. **MCI vs AD**: Consists on a dataset of patients classified by the hospitals as MCI or AD

3. **Control vs AD**: Consists on a dataset of patients classified by the hospitals as Control or AD.

Each dataset of each scenario is then divided in order to plot the learning curves, i.e. 80% is used to plot the training error curve and the or 20% is used for the validation curve so, in order to assess the different learning rates of the models by the number of observations (patients), the learning curves were plotted for the group sizes presented in table 1. It is important to note that the maximum value in each group size represents the entire 80% mentioned before:

Model	Group sizes
Control vs MCI	[1, 40, 80, 120, 160, 200, 240, 280, 312]
MCI vs AD	[1, 40, 80, 120, 160, 200, 240, 280, 296]
Control vs AD	[1, 40, 80, 120, 160, 200, 240, 280, 303]

Table 1: Different dataset sizes for the learning curves plotting.

In addition, in order to complement the analysis of the learning curves, the bias-variance trade-off was another metric that was implemented. Such metric allows us to get a better insight on whether or not the model is overfitting or underfitting so, for that measure, the average expected loss, average bias, and the average variance were calculated for each one of the models.

After all the models were trained, they were compared with each other in order to analyze the variance thus, allowing us to understand if there is any difference overall between the models. To evaluate such difference, it was used the One-way ANOVA instead of a t-test since it is a parametric test that tests for statistically significant differences between three or more models whereas a t-test allows for just two. The data analyzed by the One-way ANOVA were the values from the accuracy from each one of the five folds of each model.

It is important to know that before running the One-way ANOVA, there are some assumptions that were verified as the One-way ANOVA depends on such dependencies to works which are:

1. The distribution of the values from all five folds in each model must be normal, a condition verified using the Shapiro-Wilk test.

2. All models must be independent of each other.

3. All models must have equal variances.

Despite analyzing if there is a statistical significant difference between the predictive models, it is also necessary to see which models differ or not from other models. That way it is possible to compare models in pairs by conducting a Post-Hoc testing using the Bonferroni correction.

Testing the models on a Portuguese population

After the analysis performed in the training dataset , the next stage of the work is to run the models on a Portuguese population. As mentioned before, there were hospitals that provided MRIs for their patients. Such hospitals were Hospital Vila Franca de Xira (HVFX) and Hospital Fernando Fonseca (HFF).

Target population

The target population is composed of HVFX and HFF patients. Such population contains very few patients diagnosed with AD (18 patients) but, in contrast, there is a great number of Control patients(43 patients). Such imbalance can be explained by two factors:

1. When providing the images, the hospitals prepared patients with dementia and not AD, exclusively so, the patients with AD are a fraction of the whole that is patients with Dementia.

2. When pre-processing the images, the protocols of acquisition of the images (ex: MP-Rage, SGPR, Sag) did not match any protocol accepted by the models.

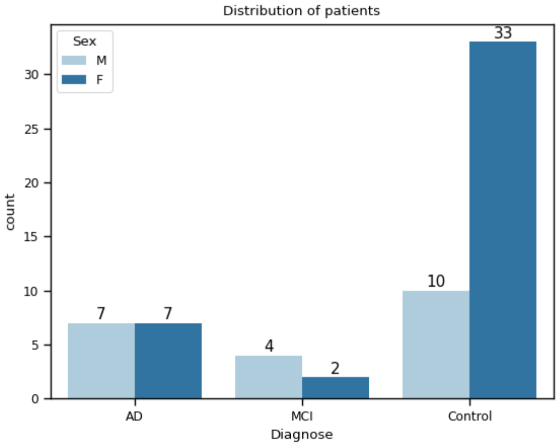


Figure 5: Distribution of the target population by gender and diagnose

Regarding the age group of the population, patients diagnosed as AD or MCI range from 55 years old onwards whereas patients diagnosed as Control range from 25 years old onwards which, once again is expected since AD is a disease that occurs, mostly, at later stages of life. After getting a general view of the population, it is now time to present the measures that are implemented in order to interpret the performance of the models on said population.

The first evaluation step to be applied is the confusion matrix analysis since it allows us to have a generalized and summarized view of the performance of the models for the multiclass problem. Having calculated the confusion matrix is then possible to get the values of the precision, sensitivity, and specificity, as well as the balanced accuracy of the model. Although accuracy is calculated too, relying just on such metric may be incredibly misleading when handling an imbalanced dataset. The confusion matrix and the balanced accuracy come in handy to solve just that as they account for both the positive and negative predicted classes without misleading with imbalanced data.

After having calculated the specificity and the sensitivity, the Area under the ROC (receiver operating characteristic) curve was plotted to assess how well the model is capable of distinguishing between HC, MCI, and AD. With precision and sensitivity calculated as well, the precision-recall curve was plotted too.

RESULS AND DISCUSSION

Data anonymization

The developed script receives, as mentioned before, DICOM files that contain confidential information from the patient. After anonymization the script produces 3 files and the anonymized image itself.

The first produced file, Keys.csv, contains the original identifier from the patient and the new identifier generated by the script. This way, once there is a new classification for the patient, the hospital can re-identify the patient. The second generated file, PhysicianName.csv, stores the real name of the physician that performed the exam as well as its new ID. Lastly, AccessionNumber.csv is the last file that is produced, storing the accession number of the exams, which is basically the ID of a specific exam, as well as the new ID created by the script. All these files are only in the possession of the hospital/clinic so, IBEB has no knowledge of the content of such files.

STATISTICAL VALIDATION

This section has the purpose of presenting and discussing the results obtained on the train as well as the test. Section presents the results obtained for the learning curves, bias-variance trade-off and the ANOVA tests implemented in the models as soon as they were trained. Section presents the results of the validation performed on the models. Such validation was performed on the original data available and on the the target data which represents the the patients from Hospital Vila Franca de Xira and Hospital Fernando Fonseca.

Validation using the original (heterogeneous) population

In order to assess how the models learn, i.e how they change in terms of performance over different sizes of the training dataset as well as seeing if any model is underfitting or overfitting or in the best scenario, they are well fitted, learning curves were plotted for each one of the seven models in each one of the three possible scenarios (Control vs MCI, Control vs AD and MCI and AD). The plotted figures display the mean square error for both the validation set and the training set.

Similar to the learning curves, the bias and variance are also calculated in order to complement the analysis of the models regarding how well-fitted, or not, are the models. Tables 2, 3, 4 show the results obtained for each one of the seven models in each one of the three possible scenarios (Control vs MCI, Control vs AD, MCI and AD) of the bias, variance and the expected loss.

	Control vs MCI						
	SVM-Linear	DT	RF	ET	LR	LDA	LR-SGD
Average expected loss	0.376	0.426	0.390	0.339	0.288	0.333	0.200
Average bias	0.294	0.299	0.257	0.199	0.170	0.151	0.060
Average variance	0.081	0.127	0.134	0.141	0.118	0.182	0.140

Table 2: Bias and variance for each model in the Control vs MCI scenario.

	Control vs AD						
	SVM-Linear	**DT**	**RF**	**ET**	**LR**	**LDA**	**LR-SGD**
Average expected loss	0.386	0.545	0.465	0.201	0.366	0.352	0.231
Average bias	0.253	0.210	0.313	0.057	0.238	0.086	0.057
Average variance	0.133	0.335	0.152	0.144	0.128	0.267	0.173

Table 3: Bias and variance for each model in the Control vs AD scenario

	MCI vs AD						
	SVM-Linear	**DT**	**RF**	**ET**	**LR**	**LDA**	**LR-SGD**
Average expected loss	0.179	0.252	0.180	0.190	0.187	0.193	0.167
Average bias	0.100	0.148	0.091	0.142	0.108	0.086	0.068
Average variance	0.079	0.104	0.089	0.049	0.079	0.106	0.099

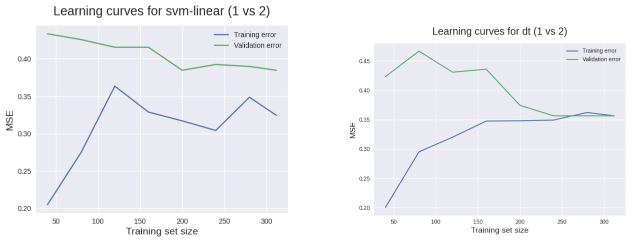
Table 4: Bias and variance for each model in the MCI vs AD scenario

The analysis of the learning curves and tables 2, 3, 4 allows us to see that for scenario:

1. **Control vs MCI**, decision trees will not be benefit from the increase of instances in the training set since the validation and training error curves have already converged. LR-SGD may be overfitting since the validation error is high wheres as the training error is much lower resulting in a high variance;

2. **Control vs AD**, LDA is overfitting as a result of the decreasing training error and the increasing of the validation error. A case of overfitting may be identified in the extra trees due to the low bias and the reletivetly higher variance;

3. **MCI vs AD**, models seem to present lower levels of bias and a higher value of variance, with the exception of SVM-linear.



(a) (b)

Figure 6: Example of the Learning curves in the Control vs MCI scenario.

The overall conclusion, is that the models may not generalise as well as they could as seen by the low bias and higher variance, hence the much higher validation error when compared against the training error. The training set sizes do not allow for an extensive analysis of the models so the best recommendation would be to increase the instances available in the training, i.e. gather a higher number of patients which can be used to re-train the models so that the learning curves could show the validation and training error curves converged which it not happening for the most cases.

Comparing the models with ANOVA

After the aforementioned metrics were calculated and the training process was concluded, it was also necessary to compare the models with each other and see if they are equal in any way. Since we need our data to follow a normal distribution and the variance must be the same for all the data [6], table 5 displays the results from the Saphiro-Wilk test where it compares the balanced accuracy from the folds of each model. The results prove that the data follows as normal distribution as the p-value is above 0.05 in all cases for each one of the model.

Model	***svm-linear***	***dt***	***rf***	***et***	***lda***	***lr***	***lr-sgd***
p-value (Control vs MCI)	0.109	0.637	0.557	0.967	0.669	0.794	0.515
p-value (MCI vs AD)	0.771	0.062	0.763	0.437	0.592	0.147	0.196
p-value (Control vs AD)	0.414	0.399	0.071	0.071	0.918	0.348	0.155

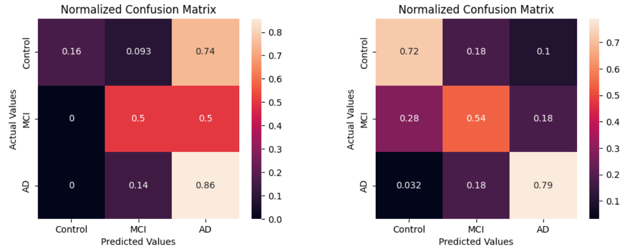
Table 5: Bias and variance for each model in the MCI vs AD scenario

Regarding the homogeneity of variance the p-values obtained, after comparing the models in each one of the three scenarios under the Levene's test, are not significant since the p-values are 0.969, 0.99 and 0.834 for the Control vs MCI, Control vs AD and MCI vs AD, respectively, (p-values > 0.05) thus, concluding there is no statistical difference in the variability of the models within a scenario.

After verifying the assumptions above, a Post-Hoc Test [4] was performed to see which models significantly differ from each others. The Post-Hoc test with the Bonferroni correction returned false for all the pairs of models comapred which means that, no model differs significantly from other models.

Generalization analysis in a Portuguese population

When running the models for the target population (Hospital Vila Franca de Xira + Hospital Fernando Fonseca), the confusion matrix in Figure 7a is obtained. Table 6 complements the confusion matrix since it allows us to know the values of precision, recall/sensitivity, F1-score, and the number of patients that support such calculus.



(a) (b)

Figure 7: Confusion matrix of the target population 7a and the original population 7b

Starting with the class Control, it is clear that the models do not classify any patients as Control when they should not be classified as such hence the precision value of 1 (100%)

which in other words mean that the fraction of instances correctly predicted as Control is 100% out of the total classified instances as Control. On the other hand, the value of the recall is only 0.16 which means that out of the 43 patients, the models might not have wrongly classified AD or MCI patients as Control but, the low value of recall means that 84% (36 patients) of the Control patients were classified as either MCI or AD patients.

In the case of the MCI patients, the precision value decreases substantially from 1 to 0.30 but on the other hand, the recall value increased from 0.16 to 0.43. One might say that these values are preferable when looking at f1-score which is higher. Out of all the patients classified as MCI, the models lacked the ability to accurately find all the MCI patients since 57% of the MCI patients were classified as AD. Such low values may be explained by the low support value of patients (only 7 patients in the entire target population).

Lastly, looking at the AD patients, the precision value is 0.29, which represents that out of all the patients classified as AD only 29% of those were correctly classified as AD. In contrast, 83% of the patients with AD were correctly classified as AD.

	Target Population			
	Precision	**Recall**	**F1-score**	**Support**
Control	1.00	0.16	0.28	43
MCI	0.30	0.43	0.35	7
AD	0.29	0.83	0.43	18

Table 6: Different metric results for all of the classes in the target population

	Original Population			
	Precision	**Recall**	**F1-score**	**Support**
Control	0.88	0.72	0.79	651
MCI	0.39	0.54	0.45	191
AD	0.66	0.79	0.72	247

Table 7: Different metric results for all of the classes in the original population

The ROC curves displayed in figure 8a represent the trade-off between sensitivity and specificity. Such curves are useful since they do not rely on the distribution of classes which comes in handy considering the number of control patients is not balanced with the number of AD patients (43 control to 14 AD patients) and allows for the better interpretation of the MCI class (6 patients). The models can be interpreted by comparing their performance against a baseline which is the FPR = TPR diagonal that represents the expected values a random classifier would return. The models' performance is considered low since the curves are closer to the 45 degrees diagonal when they should be closer to the top-left corner of the graph as it is the case on the original population.

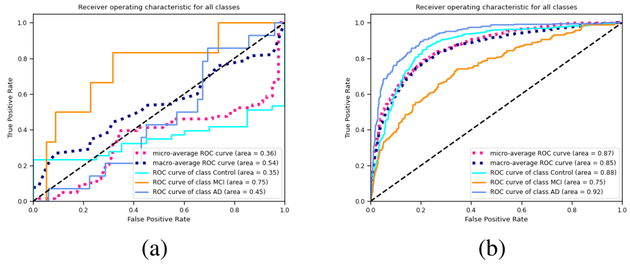


Figure 8: ROC curves obtained in the target population obtained in the target population 8a and the original population 8b

In addition, to obtain a better view of the models output, figure 9 displays the distribution of probabilities for the patients diagnosed by the hospitals as Control, MCI and AD. The main goal of said figure is to show that, for instance, Control patients are classified as AD as proved by the high probability in the AD column, hence the lack of ability to predict control patients. The same event occurs for the MCI patients as the class with higher probabilities is AD instead of MCI. As mentioned before, 83% of the patients classified as AD were in fact AD patients.

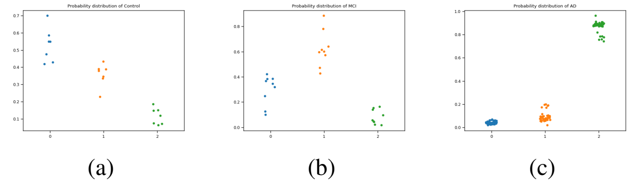


Figure 9: Distribution of calculated probabilities for patients diagnosed by the hospitals as Control 9a, MCI 9b or AD 9c

Overall, it is possible to see that the models can not generalise well for the target population as it generalized for the original data. Such disparity in results may be explained by the relatively small size of training set in which the models were trained or even by the fact that the exams used when validating the models presented new protocols of acquisition of MRIs.

CONCLUSION

This dissertation aimed to validate models under a heterogeneous population to ensure adequate representation of the Portuguese population and guarantee sufficient generalization capability of the models. In face of all the requirements, a relational database was developed in order to store the content of the MRIs received from the hospital. Prior to the reception of said data, a script was developed so that the MRIs could be anonymized and later sent. The existence of a database resulted in the development of Graphic User Interface capable of manipulating the database by allowing the user to insert new data or view data and statistics from the database content. In addiction, a validation of the predictive models was performed, being the latter the main focus of this dissertation. To assess the generalization ability of the models, these are tested on a target population consisting of patients from Hospital Vila Franca de Xira and Hospital Fernando Fonseca. Before running the models on the target population, learning curves are plotted which together with the bias and variance calculus allow to understand whether or not the models are adequately fit on the data. ANOVA and a Post-Hoc test are used in order to compare the models with each other and see if they were equal in any way. To figure how well the models were generalizing for the target population, commonly used measures were calculated in order to extract statistics on the capabilities of the models.

The learning curves showed that the models are not yet at a point of maturity. Both the learning curves and the bias-and-variance calculus allowed to understand that some models could be facing an underfitting or overfitting problem when handling Control vs MCI patients, as is the case of the models with Logistic Regression or with Linear Discriminant Analysis due to high values of bias. In addition, the learning curves showed, in the case of Linear Discriminant Analysis, a decreasing training error and an increasing validation error for an increased data size. The Post-Hoc test showed that all the pairs of models compared presented no significant difference in the variability of the models within each scenario. Regarding the results on the target population, the models showed that they lack the ability to generalize well on the new population as they did on the original one. For class Control, 100% was achieved on precision whereas in the case of Recall or F1-score only 16% and 28%, respectively, was achieved. Class MCI had slightly different results with 30% of precision, 43% of recall and 35% of F1-score while the AD class presented a 29% of precision, 83% of recall and 43% of F1-score. Furthermore, the Area Under the Receiver Operating Characteristics for class Control showed an area of 36%, followed by MCI area of 74% and AD with 0.47%. Overall, the results showed that the models do not have the desired ability to generalize well for a new population. Although the results for the AD class were better and no false negative were returned for the control class, i.e. no patient with Alzheimer's disease was classified as control, the models did not perform well on the population. The main hypothesized reason for this inability to generalize well is the low volume of available patients used in the training of the models. One would recommend adding more patients to the training process so that the models could achieve optimal performance in training. The learning curves showed the training and validation error curve did not converge due to the low volume of instances so, by adding more patients this problem could be solved. Another reason for the low generalization capacity is considered to be the different image acquisition protocols of the MRIs in the target population.

FUTURE WORK

Despite the relevance of the produced results from the targeted models, there are some measures that could be implemented in order to improve the overall performance of the models. First, the extension of the predictive models to another population since new populations result in different data that might correlate better with the models' requirements, so that models could be retrained under a more heterogeneous populations and, consequently, generalize better on a new target population. In addition, it would be interesting to see alternative supervised classification principles to assess whether the performance of the models. Third, the proposed data consolidation and external validation principles can be considered to expand the scope of the work in order to handle new neurological diseases. Finally, regarding the database and the GUI, it would be interesting to deploy the database onto a remote server so that several entities could access it and easily insert new data with the necessary guarantees of security, privacy and usability.

REFERENCES

1. Aryanto, K. Y., Oudkerk, M., and van Ooijen, P. M. Free DICOM de-identification tools in clinical research: functioning and safety of patient privacy. *European Radiology 25*, 12 (2015), 3685–3695.
2. Bleeker, S., Moll, H., Steyerberg, E., Donders, A., Derksen-Lubsen, G., Grobbee, D., and Moons, K. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology 56*, 9 (2003), 826–832.
3. Bocchetta, M., Galluzzi, S., Kehoe, P. G., Aguera, E., Bernabei, R., Bullock, R., Ceccaldi, M., Dartigues, J.-F., De Mendonca, A., Didic, M., et al. The use of biomarkers for the etiologic diagnosis of mci in europe: An eadc survey. *Alzheimer's & Dementia 11*, 2 (2015), 195–206.
4. Hilton, A., and Armstrong, R. A. Statnote 6: post-hoc anova tests. *Microbiologist 2006* (2006), 34–36.
5. Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade. *The Lancet Neurology 9*, 1 (2010), 119–128.
6. Kim, T. K. Understanding one-way anova using conceptual figures. *Korean journal of anesthesiology 70*, 1 (2017), 22.
7. Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., and Woodward, M. Risk prediction models: Ii. external validation, model updating, and impact assessment. *Heart 98*, 9 (2012), 691–698.
8. Mucke, L. Alzheimer's disease. *Nature 461*, 7266 (2009), 895–897.
9. Newhauser, W., Jones, T., Swerdloff, S., Newhauser, W., Cilia, M., Carver, R., Halloran, A., and Zhang, R. Anonymization of dicom electronic medical records for radiation therapy. *Computers in Biology and Medicine 53* (2014), 134 – 140.
10. Noumeir, R., Lemay, A., and Lina, J. M. Pseudonymization of radiology data for research purposes. *Journal of Digital Imaging 20*, 3 (2007), 284–295.
11. Nugawela, S. *Data warehousing model for integrating fragmented electronic health records from disparate and heterogeneous clinical data stores*. PhD thesis, Queensland University of Technology, 2013.
12. Nussbaum, R. L., and Ellis, C. E. Alzheimer's disease and parkinson's disease. *New england journal of medicine 348*, 14 (2003), 1356–1364.
13. Prayer, D., Brugger, P. C., and Prayer, L. Fetal mri: techniques and protocols. *Pediatric radiology 34*, 9 (2004), 685–693.
14. Qiu, S., Joshi, P. S., Miller, M. I., Xue, C., Zhou, X., Karjadi, C., Chang, G. H., Joshi, A. S., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y. J., Swaminathan, A., Kedar, S., Saint-Hilaire, M.-H., Auerbach, S. H., Yuan, J., Sartor, E. A., Au, R., and Kolachalama, V. B. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. *Brain 143*, 6 (05 2020), 1920–1933.
15. Siontis, G. C., Tzoulaki, I., Castaldi, P. J., and Ioannidis, J. P. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology 68*, 1 (2015), 25 – 34.
16. Steyerberg, E. W., and Harrell, F. E. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology 69* (2016), 245–247.
17. Ung, H., Brown, J. E., Johnson, K. A., Younger, J., Hush, J., and Mackey, S. Multivariate classification of structural mri data detects chronic low back pain. *Cerebral cortex 24*, 4 (2014), 1037–1044.
18. Volosnikov, V. I., Korkhov, V. V., Vorontsov, A. O., Gribkov, K. V., Degtyarev, A. B., Bogdanov, A. V., Zalutskaya, N. M., Neznanov, N. G., and Ananyeva, N. I. Data consolidation and analysis system for brain research. *CEUR Workshop Proceedings 2267*, Grid (2018), 388–392.