Evolving complex yet interpretable representations: application to Alzheimer's diagnosis and prognosis

Jean-Philippe Kröll
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
j.kroell@fz-juelich.de

Simon B. Eickhoff
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
s.eickhoff@fz-juelich.de

Felix Hoffstaedter
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
f.hoffstaedter@fz-juelich.de

Kaustubh R. Patil
Inst. of Neurosci. and Medicine, INM-7, Forschungszentrum Jülich, and Inst. of Systems Neurosci., HHU Düsseldorf
Germany
k.patil@fz-juelich.de

***Abstract*—With increasing accuracy and availability of more data, the potential of using machine learning (ML) methods in medical and clinical applications has gained considerable interest. However, the main hurdle in translational use of ML methods is the lack of explainability, especially when non-linear methods are used. Explainable (i.e. human-interpretable) methods can provide insights into disease mechanisms but can equally importantly promote clinician-patient trust, in turn helping wider social acceptance of ML methods. Here, we empirically test a method to engineer complex, yet interpretable, representations of base features via evolution of context-free grammar (CFG). We show that together with a simple ML algorithm evolved features provide higher accuracy on several benchmark datasets and then apply it to a real word problem of diagnosing Alzheimer's disease (AD) based on magnetic resonance imaging (MRI) data. We further demonstrate high performance on a hold-out dataset for the prognosis of AD.**

Keywords — grammar evolution, feature representation, interpretability, Alzheimer's disease, machine learning

I. Introduction

Application of machine learning and artificial intelligence (AI) methods in medical and clinical problems has gained increasing attention in recent years [1][2]. These methods can find patterns in high-dimensional data and thus have the potential to provide gains in diagnostic and prognostic accuracy. However, there are also skepticisms and societal concerns, especially regarding the explainability of the models and their predictions [1][3][4]. According to the latest EU guidelines for trustworthy AI, transparency is one of the main requirements for the application of machine learning algorithms [5]. Importantly, fostering trust between clinicians assisted by ML/AI methods and patients by communicating reasons behind decisions and uncertainties associated with options is crucial for the acceptance of ML methods [6], [7]. In addition, explainable/human-interpretable models are inherently beneficial in a clinical setting as they can help understand the biology underlying disease mechanisms and disease progression. It is, therefore, important to develop methods and frameworks that can simultaneously provide high accuracy and interpretability.

Feature engineering is one of the key concepts to improve model performance: Processing the available features in such a way that they are easily learnable by a classifier is arguably one of the most important parts of machine learning [8]. Single features may seem irrelevant until considered in combination with others. Often, exhaustively exploring the complete range of possible feature combinations is computationally too expensive, due to the high dimensionality of the data. Evolutionary algorithms can improve the search in such combinatorial problems by systematically searching the space guided by the usefulness of the candidate solutions. Previous work utilizing evolutionary algorithms have shown promise in various research areas. Some of these approaches have relied on grammatical evolution (GE) [9] for feature selection and generation. For example, Silva et al. employed GE to select and generate features for the prediction of the daily peak electricity load in planning of power systems [10]. Implementing a combination of GE and neural networks, Gavrilis et al. generated new features and could thereby improve performance on nine out of ten classification datasets [11]. Demonstrating its suitability for medical purposes, Smart et al. similarly utilized GE to select the best subset of features as well as to generate new features for detecting epileptic oscillations in patients with epileptic seizures [12]. Motsinger et al. proposed a combination of GE and neural networks to perform automatic feature selection in genetic epidemiology [13]. In a study by Georgulas et al., GE was utilized to improve the classification of pathological fetal heart rate where artificial features were derived from the 19 original features and used to train a neural network [14]. These studies show that the models generally benefited from the constructed features (CF). If the generated features and the model are restricted to retain a human-interpretable form, such a feature generation framework can be leveraged to promote both accuracy as well as interpretability. Towards this goal, we propose a framework based on GE, which achieves a good trade-off between these two goals.

Building upon its promise in engineering new and useful features, here we use GE to evolve new feature representations as combinations of the original or base features which are then used as a basis for classification. Our motivation for using GE was to test a feature construction method that can produce human-interpretable features that meet the requirements for trustworthy AI. Although there exist other feature extraction/construction methods (e.g. PCA) the resulting features are often not interpretable.GE can limit the search space and efficiently construct new features by incorporating domain-specific knowledge and user expectations through a pre-defined

978-1-7281-6929-3/20/$31.00 ©2020 IEEE

set of rules, the so-called 'grammar'. By restricting the grammar to basic arithmetic operations, we enforce the expectation on the engineered features to be human-interpretable. We then use the naïve Bayes (NB) classifier as a model. We first demonstrate utility of evolved feature representation on eight benchmark datasets. We then apply our framework to the clinical problem of diagnosis of the Alzheimer's disease —i.e. AD versus healthy control (HC) classification—using base features derived from structural MRI (sMRI) data. We expected our approach to generate human-interpretable features which include information about the interactions between brain regions. Additionally, we apply the AD vs. HC model to a hold-out set to probe its prognostic capacity—i.e. to predict whether a person with mild cognitive impairment (MCI) will develop AD or not.

Taken together, the main contributions of our work are: (1) we propose a GE framework to construct arithmetic combinations of base features which improves accuracy; and (2) by applying it to the real-world clinical problems of diagnosis and prognosis of AD, we demonstrate that the proposed framework can uncover complex yet interpretable interactions between brain regions.

This paper is structured as follows: Section II lays out the background of AD and briefly showcases current ML-based diagnostic approaches. Section III gives a brief introduction to GE and the general workflow. Section IV gives a detailed description of the feature construction method. In section V, the results are presented and discussed. Section VI presents the conclusions of our work.

II. Alzheimer's Disease and its Diagnosis and Prognosis

Among the estimated 50 million people suffering from dementia worldwide, AD is the most common form [15]. Disturbances in memory, language and higher executive functions lead to severe obstruction of a patient's life. With high prevalence in the elderly, AD has become a major public health problem, due to the increasing life expectancy of the population. It is, therefore, important to develop accurate and interpretable methods for early diagnosis of AD. One approach which has shown a good diagnostic promise—i.e. AD versus HC classification—is using sMRI derived features in combination with machine learning algorithms. Since the progression of AD is highly associated with loss of brain volume detectable in sMRI images, various algorithms capitalizing on atrophy in AD patients have shown good classification accuracy. Furthermore, as sMRI is routinely acquired in many clinics, a highly accurate and interpretable method using sMRI data has a high translational potential.

Utilizing support vector machine (SVM), Klöppel et al. classified grey matter segments of 20 pathologically proven AD patients and matched healthy controls with 96% accuracy [16]. On a larger dataset of 652 subjects, Liu et al. employed an ensemble method, based on sparse representation-based classifiers with an accuracy of 91% [17]. Lebedev et al. proposed random forest based ensembles and were able to differentiate AD from HC with an accuracy of 90% [18]. All of these approaches rely either on whole-brain analysis or atlas derived features. In most cases, classification is based on grey matter volumes of individual brain regions and benefits from the fact that areas highly affected in AD, like the hippocampus, provide good discrimination. In addition to high accuracy, it is desirable to have interpretable models that can help uncover the brain regions involved in AD and interactions between them. This, in turn, can help understand the disease mechanisms and progression leading to better treatment and care. However, a trade-off exists between these two goals such that the implicitly interpretable methods (e.g. linear SVM or logistic regression) do not implicitly take complex interactions between features into account, while other models do so with reduced implicit interpretability (e.g. RBF kernel SVM and neural networks).

III. Grammatical Evolution Framework

We consider the binary supervised learning problem where given a labeled dataset $D = \{(x_i, y_i)\}_{i=1}^{n}, x \in \mathbb{R}^d, y \in \{0,1\}$, we want to learn a mapping function $f: x \rightarrow y$ such that f generalizes on unseen data. Here, we use GE to evolve feature combinations using CFG to learn $f': x' \rightarrow y$, where $x' = \mathrm{CFG}(x)$, $x' \in \mathbb{R}^p, p \leq d$. Our aim is to identify f' such that it performs better than f and is still interpretable. We propose to use grammatical evolution for this.

Fig. 1 shows the general workflow of the proposed method. The initial population of chromosomes is translated into expressions using the production rules of the CFG. Note that each chromosome can result in a different number of expressions. Subsequently, new features are constructed by combining the base features according to the defined expressions (see section IV or details). The constructed features are used to train and evaluate a classifier in a cross-validated (CV) fashion to estimate generalization performance [19].

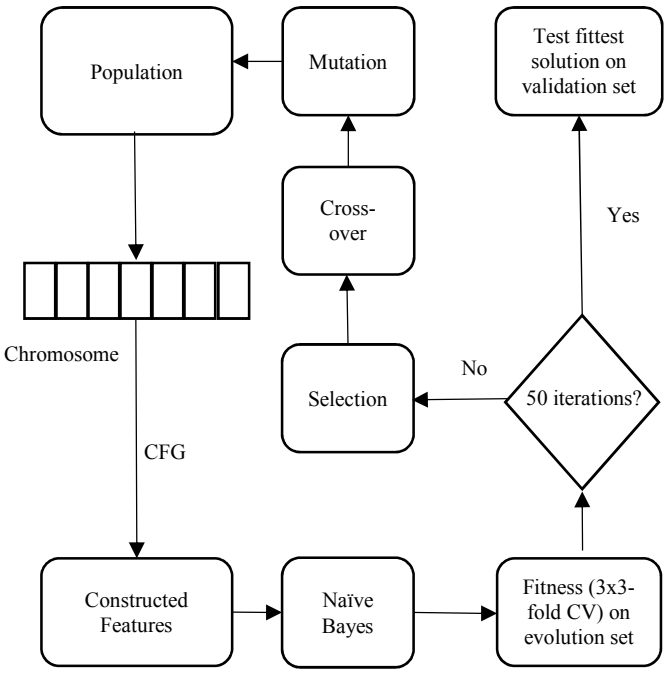

Fig.1 Workflow of the proposed framework

In CV, the data is randomly split into k equally sized subsets. One subset is retained as validation data, while the other k-1

folds are used to train the model. Consequently, each of the subsets becomes the validation data once, resulting in k different estimations. The mean of those results gives an estimate of how well the model will generalize on new data. Chromosomes then undergo selection, cross-over and mutation and are evolved to maximize their fitness. Even with the use of CV, studies have shown that optimization based solutions are prone to over-fit [20], [21]. To assess the true generalization performance of the constructed features, using CV is not enough as it is a part of the optimization process and can thus lead to overly optimistic estimates [22]. It is, therefore, necessary to assess the generalization performance outside the optimization procedure. To achieve this, the data is randomly split into two sets, 80% evolution set and 20% validation set. The evolution set is used to evaluate the features constructed by GE expressions during evolution. This is done using 3-times 3-fold CV. The fittest solution is then evaluated on the hold-out 20% validation set.

We selected NB as it provides two desirable properties, 1) its low variance and high bias which makes it less prone to over-fitting and therefore counteracts the susceptibility to overfitting within the GE optimization iterations, and 2) its probabilistic output which makes the predictions easier to communicate. Furthermore, NB is negatively affected by redundant and irrelevant features [23], so we expect the evolved feature representations to mostly contain relevant and non-redundant features.

We use the Brier score as cost (negative fitness) measure. Brier score was chosen as it is a proper scoring rule and hence can be used to rank solutions.

A. Grammatical Evolution

In this study, the gramEvol R-package (https://github.com/fnoorian/gramEvol) was used [24]. GE combines CFG and genetic algorithms to optimize programs towards a specific task. CFG is used to generate patterns of strings according to a set of recursive rules. The notation technique used here is the Backus-Naur form (BNF). The CFG is described by the tuple (T,N,R,S), where T is the set of terminal symbols, N is the set of non-terminals with $N \cap T = \emptyset$, R the set of production rules and S the start symbol, $S \in N$. Non-terminal symbols can be replaced by other non-terminal or terminal symbols whereas terminal symbols are literals. N and T together build the lexical elements which are used in the production rules R. R is defined as relations in the form of $x \rightarrow \alpha$ with $x \in N$, $\alpha \in (N \cup T)$. The user-defined grammar is utilized to impose a set of grammatical production rules which determine the chromosomes. Each gene denotes a production rule of the CFG. Following the predefined set of rules, genotypic integer strings are translated into functional phenotypic programs, a process which is called genotype-to-phenotype mapping. The mapping function is the mod rule, defined as:

$$R = B \bmod RN$$

Where B is the codon integer value, mod is the modulus operator and RN is the number of rules for the current non-terminal. Mapping begins at S and subsequently replaces each non-terminal element N, according to the production rule determined by the mapping function. Mapping continues until every non-terminal element is replaced by a terminal. If the chromosome runs out of codons before a valid expression could be produced, wrapping is applied. By reusing the codons, the mapping process continues. To prevent infinite recursions, wrapping is limited to a certain number and will result in a poor fitness score if the limit is reached. Details of the settings that we used and feature construction are provided in the next section.

The evolution was performed with a genetic algorithm (GA) [25]. GA is an optimization algorithm inspired by evolution in which generations of chromosomes, representing the genotype—i.e., candidate solutions, are successively optimized and evaluated based on a fitness measure. The chromosomes are then subject to selection, cross-over and mutation, producing the next generation. This process is repeated until terminal criteria like a certain threshold of fitness or the predefined number of generations are reached.

IV. Experimental Setup

The production rules of the CFG were defined as the grammar shown in Table I. Non-terminal symbols are expression, operator and variable and are enclosed by angle brackets. On the other side, terminals are the actual mathematical operators and original features. Thereby, the resulting expressions are arithmetic combinations of the original features. The feature construction process takes the values of each chromosome and applies the CFG rules from Table I to the base features (Tab. IIB).

To reduce computational cost whilst preserving the diversity of the solutions, population size was set to 20 chromosomes. Additionally, the number of generated features was fixed to be equal to or less than the number of original features of the given dataset, with 14 codons per expression. The mutation chance for each codon was set to $1/(\text{genomeLength}+1)$ and single-point cross-over was used. The initial population included the base model (all original features by themselves). Other chromosomes in the initial population were randomly created in the range of [0, d-1]. Evolution was terminated after 50 generations.

The cost (negative fitness) of each chromosome was calculated as stratified Brier score [26] to take the imbalanced nature of some datasets into account. Using the constructed features, an NB model was fit to the two training folds within the evolution set and used to predict the held-out fold. The predicted assignment probabilities were used to calculate the Brier score for each class separately. The two Brier scores were then averaged to get the cost value, with lower values indicating better performance. The settings of the GE are shown in Tab. IIA.

The optimized *GE model* using the 80% evolution set —i.e. the NB model on the constructed features—was evaluated on the 20% validation set. The same evolution set was used to build a *base model* using the original features and evaluated on the validation set. To consider the randomness in the evolution set-validation set split and the GE initialization, we ran the GE framework five times for each dataset. Four evaluation metrics are reported: area under the ROC curve (ROC), balanced accuracy (Acc), F1-score (F1) and stratified Brier score (Brier).

TABLE I. *GRAMMAR USED*

Rule			Rule number
S	::=	<expr>	0
<expr>	::=	<expr> <op> <expr>	0
		<var>	1
<op>	::=	+ | - | * | /	0|1|2|3
<var>	::=	$X_1 \mid X_2 \mid ... \mid X_n$	0|1|...|n-1

TABLE II. FEATURE CONSTRUCTION

A) SETTINGS OF THE GE

Parameters	Value
Number of individuals	20
Number of generations	50
Chromosome length	[0, d-1]
Mutation rate	1/(d+1)

B) EXAMPLE FEATURE CONSTRUCTION

String	Chromosome	Operation
<expr>	8,9,14,3,6,11,7,6,13,4	8 mod 2 = 0
<expr> <op> <expr>	9,14,3,6,11,7,6,13,4	9 mod 2 = 1
<var> <op> <expr>	14,3,6,11,7,6,13,4	14 mod 14 = 0
X_1 <op> <expr>	3,6,11,7,6,13,4	3 mod 4 = 3
X_1 * <expr>	6,11,7,5,13,4	6 mod 2 = 0
X_1 * <expr> <op><expr>	11,7,5,13,4	11 mod 2 = 1
X_1 * <var> <op> <expr>	7,5,13,4	7 mod 14 = 7
X_1 * X_8 <op> <expr>	5,13,4	5 mod 4 = 1
X_1 * X_8 + <expr>	13,4	13 mod 2 = 1
X_1 * X_8 + <var>	4	4 mod 14 = 4
X_1 * X_8 + X_5	**constructed feature**	

The original chromosome is [8,9,14,3,6,11,7,6,13,4]. The process starts with the first integer of the chromosome, in this case, eight. Since the start symbol is <expr>, which has two different rules, the first operation is 8 mod 2 = 0. Consequently, rule number 0 is selected and <expr> is translated into <expr> <op> <expr>. After that, the leftmost non-terminal is selected and the next integer is used to determine the following rule. The process is repeated until every non-terminal element is substituted by a terminal. The final expression is 'X_1 * X_8+ X_5'.

A. Datasets

We used eight real-world benchmark datasets from UCI (http://www.wisostat.uni-koeln.de/de/forschung/software-und-daten/data-for-classification/) and two real-world clinical datasets.

1) Breast Cancer Wisconsin: The sample contains 569 patients with breast cancer. The objective is to differentiate malignant and benign cases using 30 features computed from a fine needle aspirate of a breast mass, describing characteristics of the cell nuclei of the image. The database contains 257 benign and 212 malignant cases.

2) Pima Indians diabetes: This dataset contains 768 females of Pima Indian heritage. The objective is to predict diabetic status using eight diagnostic measurements. Variables include the number of pregnancies, glucose concentration in plasma, blood pressure, skin thickness, insulin concentration, BMI, age and Diabetes Pedigree Function. 268 of the subjects are diagnosed as diabetics.

3) Heart Disease: The sample contains 270 participants with 120 patients with diagnosed heart disease. The objective is to classify the absence or presence of heart disease using 13 features with various diagnostic measurements.

4) Irish: The dataset contains 500 instances of Irish school children. The objective here is to classify into male and female, based on five features dealing with the educational status of the children.

5) Image Segmentation: The dataset contains 660 outdoor images. The images were hand segmented to create a classification for every pixel. In this case, images are classified into "containing window" and "containing cement". 330 examples are available for each class.

6) Tennis: The dataset contains 87 instances of subjects under pain medication. Based on 15 features dealing with experienced drug efficacy, the objective is to classify into male and female.

7) Diabetes: The dataset contains 112 instances of diabetics. The objective is to differentiate the diabetic type based on five metabolic variables.

8) Crabs: The dataset contains 200 instances of Leptograpsus crabs. Based on 5 features describing physical attributes, the objective is to classify into male and female.

9 and 10) Alzheimer's Disease Neuroimaging Initiative (ADNI). We derived two datasets from the ADNI database [27]. (A) The AD diagnosis dataset contains 459 subjects with 3T scans with the objective to classify them as AD or HC. Structural (T1-weighted) MRI images of 153 AD patients and 306 HC are extracted. Utilizing the CAT toolbox (http://dbm.neuro.uni-jena.de/cat), voxel-based morphometry (VBM) is performed to estimate local grey matter volume. Subsequently, a brain atlas is applied which partitions the brain into 173 parcels. The brain atlas contains 100 Schaefer atlas parcels covering the cortex [28], complemented by 36 subcortical regions from Brainnetome [29] and 36 cerebellum parcels from Buckner et al [30]. The average grey matter volume within each of the 173 parcels is calculated as base features for each subject. (B) The MCI to AD prognosis dataset contains similarly derived 173 features for 267 subjects of which 138 later converted to AD. The objective here is to classify converters and non-converters.

V. EXPERIMENTAL RESULTS

In this section, we discuss the results obtained on the benchmark datasets as well as the two clinical datasets. Overall, we observed that both the GE constructed features as well as the base features combined with the naïve Bayes classifier were able

to classify most datasets with high accuracy. Notably, feature construction via GE resulted in superior performance in most cases.

A. Benchmark UCI Datasets

In Table III, the results from the application of both base and GE models on several test datasets are listed. We observed that for six out of eight datasets the NB model using GE constructed features outperformed the NB model using base features in at least four out of five runs on the stratified Brier score which was optimized by GE. Importantly, the constructed features also benefited other performance metrics with an increase in area under ROC, balanced accuracy, and F1-Score. Interestingly, for two datasets, Diabetes and Crabs, all the metrics improved considerably. For the two datasets with no clear improvement, Irish and Image, the performance of NB with constructed features was similar to the base model. These results suggest that our framework was able to evolve feature representations which improved overall classification efficacy. Incorporating information about the interactions between features seems to increase the discriminative validity of the evolved representation, in comparison to base features. Furthermore, the versatile set of classification tasks at hand suggests that our framework is generally applicable in diverse research domains.

TABLE III. *AVERAGE PERFORMANCE ON UCI DATASETS*

Dataset	Model	ROC	Acc	F1	Brier	Runs
Breast	Base	0.992	0.955	0.969	0.034	4/5
	GE	**0.997**	**0.973**	**0.980**	**0.026**	
Pima	Base	0.800	0.693	0.798	0.215	4/5
	GE	**0.805**	**0.709**	**0.804**	**0.193**	
Heart	Base	0.884	0.787	0.798	0.162	5/5
	GE	0.873	**0.822**	**0.847**	**0.153**	
Irish	Base	0.640	0.600	0.542	0.241	2/5
	GE	0.596	0.570	0.504	0.247	
Image	Base	0.965	0.917	0.915	0.065	2/5
	GE	**0.977**	0.915	0.914	0.066	
Tennis	Base	0.497	0.451	0.526	0.425	4/5
	GE	0.467	**0.489**	**0.546**	**0.369**	
Diabet.	Base	0.960	0.925	0.881	0.073	4/5
	GE	**1**	**1**	**1**	**0.007**	
Crabs	Base	0.646	0.590	0.591	0.295	5/5
	GE	**0.982**	**0.900**	**0.897**	**0.066**	

B. Alzheimer's diagnosis

Results on the clinical problem of AD diagnosis confirmed our observations on the benchmark datasets. In Table IV the results from the application of our framework to the classification of AD vs HC are presented. Remarkably, the representation evolved by our framework exceeded the baseline performance on all four metrics in all five runs. It is evident that the constructed features greatly benefited classification. Therefore, it is plausible to suggest, that the consideration of feature interaction provides additional discriminative information.

TABLE IV. DIAGNOSIS RESULTS ON AD DATASET

Dataset	Model	ROC	Acc	F1	Brier	Runs
AD	Base	0.850	0.782	0.818	0.214	5/5
	GE	**0.913**	**0.815**	**0.859**	**0.178**	

In addition to our first goal of improving classification accuracy, our second goal was to maintain the explainability/human-interpretability of the evolved representations. To establish the interpretability of the GE constructed features, we investigated their biological relevance based on known results from the literature. Exemplary, two of the CF are discussed here. The first feature was constructed during the fourth run of our approach and represents a combination of three base features $CF_1 = x_{29} - x_{116} * x_{136}$. Although this CF integrates information about complex interactions into the model, the relation between the constituent base features is still understandable. The underlying brain regions were the left temporal pole (TmP, x_{29}), the ventromedial putamen (vmPu, x_{116}), and the right lateral prefrontal thalamus (lpThal, x_{136}). The location of these regions is depicted in Fig. 2. CF_1 suggests an interaction between TmP, lpThal and the vmPu and all are known to be affected in AD [31], [32], [33]. A second constructed feature, CF_2, represented a combination of additional three regions, such that $CF_2 = X_{135} + X_{104} * X_{128}$ (Fig.3). In this case, the underlying regions were the left lateral prefrontal thalamus (lpThal, X_{135}), the right lateral amygdala (lAmyg, X_{104}) and the right rostral temporal thalamus (rTThal, X_{128}). Apart from lpThal and rTThal for which we have already shown involvement in AD, the lAmyg is another region that is highly affected during the disease [34], [35]. The association of regional atrophy or co-atrophy in different brain regions may hint at the underlying biological mechanisms playing a role in development and course of AD. On average, the five runs on the ADNI data produced 130 features. A set of selected expressions from the runs are shown in Table V. Put in a clinical context, our approach is well suited to identify disease-relevant patterns. It is evident, that our proposed method is not only interpretable, but the basis on which classification is performed can be easily explained by analyzing the CFs.

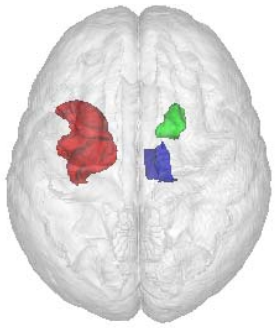

Fig. 2 Superior view of the brain. Depicted are TmP (red), vmPu (green) and lpThal (blue).

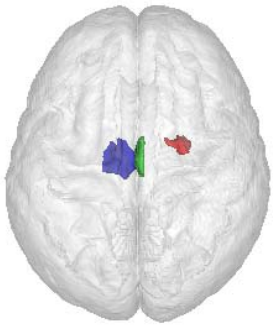

Fig. 3 Superior view of the brain. Depicted are lpThal (blue), lAmyg (red) and rTThal (green).

TABLE V. SELECTED EXPRESSIONS FROM 5 GE RUNS

Expression
$CF_1 = X_{29} - X_{116} * X_{136}$
$CF_2 = X_{135} + X_{104} * X_{128}$
$CF_3 = X_{38} - X_{59} / X_{122}$
$CF_4 = X_{76} - X_{132} * X_{83}$
$CF_5 = X_{136} - X_{53} * X_{74}$
$CF_6 = X_{101} * X_{119} / X_{66}$
$CF_7 = X_{48} / X_{23} * X_{56} / X_{101}$
$CF_8 = X_{110} + X_{23} - X_{158}$
$CF_9 = X_{51} - X_{46} * X_{65}$
$CF_{10} = X_{63} / X_{45} * X_{56}$

C. *Alzheimer's prognosis*

Since AD is marked by a continuous loss of neurons, early detection will play a vital role in future therapeutic methods. To this end, we tested if the diagnostic models using the features constructed for AD vs HC classification in the previous section would also be suitable for prognosis—i.e. to detect if MCI patients will later on convert to AD. If confirmed, it will indicate the generalizability of the constructed features and speak for their biological meaningfulness.

MCI is a neurological disorder that involves cognitive decline beyond what is expected for a person's age. It is generally seen as a prodromal stage of dementia, especially of AD [36]. Since not all MCI patients transition to dementia, it is a constant endeavor to differentiate between subjects on the verge of transitioning to AD (so-called converters), from stable MCI patients (non-converters). As the constructed features of our approach were able to pick up disease-relevant patterns in AD, we hypothesized that the same patterns could be useful to differentiate MCI-converters (MCIc) from stable MCI (MCIs) patients. Therefore, we extracted the same 173 features from 138 MCIc and 138 MCIs subjects' sMRI images from ADNI. The classification was performed first with the base model (trained on AD vs HC diagnostic data) and then using each of the five grammar models separately (again, trained on AD vs HC). Before applying the GE derived NB models, base features were transformed to match the constructed features of the respective model. In Table VI, the results of base and GE models are shown. The results of both models are comparable to those found in recent literature, although on the lower end of performance [37][38][39]. Nevertheless, our GE models could improve classification performance in comparison to the base model on all four metrics. Since GE is not limited to naïve Bayes classifiers, but well compatible with more sophisticated learning algorithms, future applications might yield even better results.

TABLE VI. PROGNOSIS RESULTS ON MCI DATASET

Dataset	Model	ROC	Acc	F1	Brier	Runs
ADNI	Base	0.717	0.680	0.699	0.316	5/5
	Grammar	**0.744**	**0.688**	**0.707**	**0.305**	

VI. CONCLUSION

We presented a simple GE based framework to evolve complex yet interpretable feature representations and showed its effectiveness on several benchmark datasets. We then tested the framework on two clinically relevant problems, diagnosis and prognosis of AD. In both cases, GE constructed features provided improved classification over base features. Moreover, the constructed features were interpretable. Our framework could prove useful in translational applications like the ones showcased here by providing both accuracy and interpretability.

Our framework is not without limitations. Firstly, we only considered the NB classifier primarily for its desirable property of low variance. However, other algorithms could provide higher accuracy if their variance can be properly controlled and should be tested. Secondly, we did not take special precautions to avoid overfitting in the optimization process itself [21], though our framework validates optimized models on hold-out data to avoid optimistic estimates. Heuristics such as early stopping could be investigated possibly further improving performance. Additionally, there can be multiple evolved solutions that perform equally well, which was the case for our results. In such cases, it is important to choose the most interpretable representation, which can be challenging.

Taken together, our simple framework can be useful for generating complex yet interpretable feature representations that can help improve both accuracy and interpretability.

ACKNOWLEDGMENT

This study was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 785907 (HBP SGA2) and Grant Agreement No. 7202070 (HBP SGA1). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.;

Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

[1] V. H. Buch, I. Ahmed, and M. Maruthappu, "Artificial intelligence in medicine: current trends and future possibilities," *Br J Gen Pract*, vol. 68, no. 668, pp. 143–144, Mar. 2018, doi: 10.3399/bjgp18X695213.

[2] F. Jiang *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke Vasc Neurol*, vol. 2, no. 4, pp. 230–243, Dec. 2017, doi: 10.1136/svn-2017-000101.

[3] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv:1702.08608 [cs, stat]*, Mar. 2017, Accessed: Dec. 10, 2019. [Online]. Available: http://arxiv.org/abs/1702.08608.

[4] F. K. Dosilovic, M. Brcic, and N. Hlupic, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, May 2018, pp. 0210–0215, doi: 10.23919/MIPRO.2018.8400040.

[5] "EU guidelines on ethics in artificial intelligence: Context and implementation," Sep. 2019, [Online]. Available: http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2019)640163.

[6] B. Heinrichs and S. B. Eickhoff, "Your evidence? Machine learning algorithms for medical diagnosis and prediction," *Hum Brain Mapp*, p. hbm.24886, Dec. 2019, doi: 10.1002/hbm.24886.

[7] Anirban Mukhopadhyay, David Kügler, Andreas Bucher, Dieter Fellner, Thomas Vogl, "Putting Trust First in the Translation of AI for Healthcare," Jan. 22, 2019.

[8] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012, doi: 10.1145/2347736.2347755.

[9] M. O'Neill and C. Ryan, "Grammatical evolution," *IEEE Trans. Evol. Computat.*, vol. 5, no. 4, pp. 349–358, Aug. 2001, doi: 10.1109/4235.942529.

[10] A. M. D. Silva, F. Noorian, R. I. A. Davis, and P. H. W. Leong, "A Hybrid Feature Selection and Generation Algorithm for Electricity Load Prediction Using Grammatical Evolution," in *2013 12th International Conference on Machine Learning and Applications*, Miami, FL, USA, Dec. 2013, pp. 211–217, doi: 10.1109/ICMLA.2013.125.

[11] D. Gavrilis, I. G. Tsoulos, and E. Dermatas, "Selecting and constructing features using grammatical evolution," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1358–1365, Jul. 2008, doi: 10.1016/j.patrec.2008.02.007.

[12] O. Smart, I. G. Tsoulos, D. Gavrilis, and G. Georgoulas, "Grammatical Evolution for Features of Epileptic Oscillations in Clinical Intracranial Electroencephalograms," *Expert Syst Appl*, vol. 38, no. 8, pp. 9991–9999, Aug. 2011, doi: 10.1016/j.eswa.2011.02.009.

[13] A. A. Motsinger, D. M. Reif, S. M. Dudek, and M. D. Ritchie, "Understanding the Evolutionary Process of Grammatical Evolution Neural Networks for Feature Selection in Genetic Epidemiology," in *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, Toronto, Ont., Sep. 2006, pp. 1–8, doi: 10.1109/CIBCB.2006.330945.

[14] G. Georgoulas, D. Gavrilis, I. G. Tsoulos, C. Stylios, J. Bernardes, and P. P. Groumpos, "Novel approach for fetal heart rate classification introducing grammatical evolution," *Biomedical Signal Processing and Control*, vol. 2, no. 2, pp. 69–79, Apr. 2007, doi: 10.1016/j.bspc.2007.05.003.

[15] C. Patterson, "World Alzheimer Report 2018 - The state of the art of dementia research."

[16] S. Kloppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681–689, Feb. 2008, doi: 10.1093/brain/awm319.

[17] M. Liu, D. Zhang, D. Shen, and Alzheimer's Disease Neuroimaging Initiative, "Ensemble sparse classification of Alzheimer's disease," *Neuroimage*, vol. 60, no. 2, pp. 1106–1116, Apr. 2012, doi: 10.1016/j.neuroimage.2012.01.055.

[18] A. V. Lebedev *et al.*, "Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness," *Neuroimage Clin*, vol. 6, pp. 115–125, 2014, doi: 10.1016/j.nicl.2014.08.023.

[19] Trevor Hastie Robert Tibshirani Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, Second Edition. 2009.

[20] E. M. Dos Santos, R. Sabourin, and P. Maupin, "Overfitting cautious selection of classifier ensembles with genetic algorithms," *Information Fusion*, vol. 10, no. 2, pp. 150–162, Apr. 2009, doi: 10.1016/j.inffus.2008.11.003.

[21] J. Loughrey and P. Cunningham, "Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets," in *Research and Development in Intelligent Systems XXI*, M. Bramer, F. Coenen, and T. Allen, Eds. London: Springer London, 2005, pp. 33–43.

[22] Gavin C. Cawley, Nicola L.C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *JMLR*.

[23] Joaquín Abellán and Javier Castellano, "Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy," *Entropy*, vol. 19, no. 6, p. 247, May 2017, doi: 10.3390/e19060247.

[24] F. Noorian, A. M. de Silva, and P. H. W. Leong, "**gramEvol** : Grammatical Evolution in *R*," *J. Stat. Soft.*, vol. 71, no. 1, 2016, doi: 10.18637/jss.v071.i01.

[25] John H. Holland, "Genetic Algorithms," *Scientific American*, vol. 267, pp. 66–73, Jul. 1992.

[26] Glenn W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Wea. Rev.*, vol. 78, pp. 1–3, 1950.

[27] R. C. Petersen *et al.*, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010, doi: 10.1212/WNL.0b013e3181cb3e25.

[28] A. Schaefer *et al.*, "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI," *Cerebral Cortex*, vol. 28, no. 9, pp. 3095–3114, Sep. 2018, doi: 10.1093/cercor/bhx179.

[29] L. Fan *et al.*, "The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture," *Cereb. Cortex*, vol. 26, no. 8, pp. 3508–3526, Aug. 2016, doi: 10.1093/cercor/bhw157.

[30] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. T. Yeo, "The organization of the human cerebellum estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 5, pp. 2322–2345, Nov. 2011, doi: 10.1152/jn.00339.2011.

[31] L. W. de Jong *et al.*, "Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study," *Brain*, vol. 131, no. 12, pp. 3277–3285, Dec. 2008, doi: 10.1093/brain/awn278.

[32] M. Zarei *et al.*, "Combining shape and connectivity analysis: An MRI study of thalamic degeneration in Alzheimer's disease," *NeuroImage*, vol. 49, no. 1, pp. 1–8, Jan. 2010, doi: 10.1016/j.neuroimage.2009.09.001.

[33] J. Hänggi, J. Streffer, L. Jäncke, and C. Hock, "Volumes of Lateral Temporal and Parietal Structures Distinguish Between Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease," *JAD*, vol. 26, no. 4, pp. 719–734, Oct. 2011, doi: 10.3233/JAD-2011-101260.

[34] C.-A. Cuénod, "Amygdala Atrophy in Alzheimer's Disease: An In Vivo Magnetic Resonance Imaging Study," *Arch Neurol*, vol. 50, no. 9, p. 941, Sep. 1993, doi: 10.1001/archneur.1993.00540090046009.

[35] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, and B. C. Dickerson, "Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity," *Psychiatry Research: Neuroimaging*, vol. 194, no. 1, pp. 7–13, Oct. 2011, doi: 10.1016/j.pscychresns.2011.06.014.

[36] S. Gauthier *et al.*, "Mild cognitive impairment," *The Lancet*, vol. 367, no. 9518, pp. 1262–1270, Apr. 2006, doi: 10.1016/S0140-6736(06)68542-5.

[37] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, and J. Tohka, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," *NeuroImage*, vol. 104, pp. 398–412, Jan. 2015, doi: 10.1016/j.neuroimage.2014.10.002.

[38] H.-I. Suk and D. Shen, "Deep Learning-Based Feature Representation for AD/MCI Classification," in *Advanced Information Systems Engineering*, vol. 7908, C. Salinesi, M. C. Norrie, and Ó. Pastor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 583–590.

[39] S. Basaia *et al.*, "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage: Clinical*, vol. 21, p. 101645, 2019, doi: 10.1016/j.nicl.2018.101645.