Expert Systems With Applications 191 (2022) 116221

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Decision tree and artificial immune systems for stroke prediction in imbalanced data

Laércio Ives Santos [a,b], Murilo Osorio Camargos [c], Marcos Flávio Silveira Vasconcelos D'Angelo [d,*], João Batista Mendes [d], Egydio Emiliano Camargos de Medeiros [e], André Luiz Sena Guimarães [f], Reinaldo Martínez Palhares [g]

[a] Graduate Program in Health Sciences, UNIMONTES, Montes Claros, Brazil
[b] Federal Institute of Northern Minas Gerais, Montes Claros, Brazil
[c] Graduate Program in Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil
[d] Department of Computer Science, UNIMONTES, Av. Rui Braga, sn, Vila Mauricéia, Montes Claros, Brazil
[e] City Hall of Pará de Minas, Pará de Minas, Brazil
[f] Department of Dentistry, UNIMONTES, Montes Claros, Brazil
[g] Department of Electronics Engineering, Federal University of Minas Gerais, Belo Horizonte, Brazil

ARTICLE INFO

Keywords:
Stroke prediction
Artificial immune systems
Decision tree
Genetic programming

ABSTRACT

Although cerebral stroke is a important public worldwide health problem with more than 43 million global cases reported recently, more than 90% of metabolic risk factors are controllable. Therefore, early treatment can take advantage of a fast and low-cost diagnosis to minimize the disease's sequels. The use Machine Learning (ML) techniques can provide an early and low-cost diagnosis. However, the performance of these techniques is reduced in problems of prediction of rare events and with class imbalance. We proposed Machine learning approach to cerebral stroke prediction based on Artificial Immune Systems (AIS) and Decision Trees (DT) induced via Genetic Programming (GP). In general, the approaches for stroke prediction presented in the literature do not allow the development of models considered interpretable; our approach, on the other hand, uses a simplification operator that reduces the complexity of the induced trees to increase their interpretability. We evaluated our approach on a highly imbalanced data set with only 1.89% stroke cases and used AIS combined with One Sided Selection (OSS) to create a new balanced data set. This new data set is used by the GP to evolve a population of DTs, and, at the end of this process, the best tree is used to classify new instances. Two experiments are used to test the proposed approach. In the first experiment, our approach achieved, in terms of sensitivity and specificity, are 70% and 78%, respectively, indicating its competitiveness with the state-of-the-art technique. The second experiment evaluates the proposed simplification mechanism in creating rules that can be interpreted by humans. The proposed approach can effectively increase sensitivity and specificity while maintaining accurate prediction using interpretable models, indicating its potential to be clinically used in stroke diagnosis.

1. Introduction

Despite the scientific advances related to the care of stroke patients in recent years, stroke remains a worldwide public health problem and is among the leading causes of adult death and disabilities (Benjamin et al., 2018; Thrift et al., 2014). There are more than 43 million global cases reported in 2015 (Benjamin et al., 2018) and this amount tends to increase with the growth of the elderly population (Simpkins et al., 2020). In addition, the prevalence of stroke has also increased in the younger population (GBD, 2018). Usually, stroke patients undergo an initial period in the hospital for treatment. In the next stage, they remain an extended period at home for recovering their physical, speech, and cognitive functions (Chen et al., 2019), due to sequels of stroke such as depression and imbalance or loss of physical features (Alghwiri, 2016).

The introduction of early treatment is a way for minimizing sequels of stroke once more than 90% of metabolic risk factors are

* Corresponding author.
E-mail addresses: laercio.ives@gmail.com (L.I. Santos), murilo.camargosf@gmail.com (M.O. Camargos), marcos.dangelo@unimontes.br (M.F.S.V. D'Angelo), joao.mendes@unimontes.br (J.B. Mendes), emiliano.camargos@gmail.com (E.E.C.d. Medeiros), andreluizguimaraes@gmail.com (A.L.S. Guimarães), rpalhares@ufmg.br (R.M. Palhares).

https://doi.org/10.1016/j.eswa.2021.116221
Received 23 December 2020; Received in revised form 24 June 2021; Accepted 10 November 2021
Available online 27 November 2021

controllable (O'Donnell et al., 2016). Clinical exams indicate the stroke diagnostics that can be confirmed by a computed tomography scan, where the gold standard to distinguish the disease's subtypes is the non-contrast computed tomography scan (Wardlaw et al., 2004). However, these image exams can be expensive and inaccessible in regions with difficult access such as rural areas (Leira, Hess, Torner, & Adams, 2008); in such cases, it is possible to use weighted clinical score systems to improve the rapid diagnosis of stroke subtypes (Jin et al., 2016). Other alternatives for diagnosis of stroke include increasing state investments or using Machine Learning (ML) techniques to provide an early and low-cost diagnosis (García-Temza, Risco-Martín, Ayala, Roselló, & Camarasaltas, 2019). ML techniques are interesting because they emulate the human way of thinking and making decisions (El Naqa & Murphy, 2015), analyzes large data sets containing many characteristics in a reasonable time, and can handle complex relationships between data sets, making them more accurate than human specialists in some specific situations (Deo, 2015).

The use of ML techniques for health-related diagnostics tasks meets some challenges; one of them resides in the fact that, compared with healthy subjects, patients with a given disease are generally a small part of the total population. This disproportion in the representation of health and non-healthy subjects is known as the problem of imbalanced data sets, where the class with the highest prevalence is called the majority class, while the rarest class is called the minority class (Haixiang et al., 2017). The challenge in applying ML techniques in handling imbalanced data sets is that they tend to rank all instances in the majority class and none in the minority class, which is generally characterized as the event of most significant interest (Li et al., 2017).

Several papers in the literature used ML techniques for predicting stroke. However, most of them ignore the imbalance of the classes while, in clinical practice, the stroke data set is naturally imbalanced (Liu, Fan, & Wu, 2019). In Colak, Karaman, and Turtay (2015), for example, the authors used Artificial Neural Networks (RNA) or Support Vector Machine (SVM) and a knowledge discovery process to predict stroke. A data set with 167 healthy patients and 130 stroke patients, described by eight clinical variables, was used for training and evaluation of the models. SVMs and Margin-based Censored Regression (MCR) are used as learning algorithms for an automatic feature selection procedure proposed in Khosla et al. (2010) to predict stroke. A comparison of several ML methods that have been applied to predict ischemic stroke is made in Arslan, Colak, and Sarihan (2016). The experiments were performed using a data set with 112 healthy patients and 80 sick patients with SVM presenting best accuracy values.

In Liu et al. (2019), a hybrid approach is described for stroke prediction based on physiological data from a highly imbalanced data set (1.18% of cases of stroke). The hybrid approach is executed in three distinct steps: (i) a data imputation process based on Random Forests (Breiman, 2001) is executed; (ii) the data set is balanced using a methodology that combines Principal components Analysis (PCA) and k-Means clustering methods; (iii) the classification operation is performed by a deep Neural Network with hyperparameters automatically adjusted.

The approach detailed in Liu et al. (2019) presented satisfactory sensitivity and poor specificity. Thus, strategies for improving mainly specificity value without reducing sensitivity value should be investigated. Also, the RNA for prediction is not interpretable, i.e., its results present incomprehensible human terms. In health-related applications, it is interesting to adopt interpretable ML techniques, as they facilitate the problem investigation, generate new insights for solving it, and improve specialists' understanding (Caruana et al., 2015).

The adoption of ML tools in clinical practice requires a careful confirmation of their performance before its use. When the results of a diagnosis test are binary, the discrimination performance is usually measured through sensitivity and specificity (Park & Han, 2018). Sensitivity is defined as the proportion of sick individuals correctly identified with the disease. The specificity, on the other hand, refers to the proportion of non-sick people that are correctly identified without the disease (Park, Choi, & Byeon, 2021).

Therefore, in this work, we propose an alternative approach for stroke prediction on highly imbalanced data sets. The approach, illustrated by Fig. 1, combines both Immune/Neural (D'Angelo et al., 2016) and One-Sided Selection (OSS) (Kubat, Matwin, et al., 1997) techniques to balance the training data and uses Decision Trees (DT) induced by Genetic Programming (GP) (Koza, 1992) for the classification operation. In Fig. 1, $\mathcal{X}_{\text{train}}$ identifies the imbalanced training data, which is summarized in $\mathcal{X}^{+}_{\text{train}}$ by the proposed balancing procedure. The GP algorithm uses $\mathcal{X}^{+}_{\text{train}}$ for evolving a population of DTs. The best decision tree (Decision Tree*), returned by the GP algorithm is used to classify unknown instances.

In this work, we use GP in the induction process instead of traditional strategies such as CART (Breiman, Friedman, Stone, & Olshen, 1984) and C4.5 (Quinlan, 2014) due to their ability for global optimization. These traditional strategies use greedy search in the tree generation process which can lead to sub-optimal solutions. Furthermore, the recursive partitioning in the data set can result in data sets too small for attribute selection in deeper nodes of a tree, overfitting the data (Barros, Basgalupp, De Carvalho, & Freitas, 2011).

In summary, this paper focuses on two main challenges. First, in previous studies using ML for stroke prediction, the data sets used do not suffer from class imbalance. In this situation, the performance of the methods in terms of sensitivity and specificity is heavily compromised. In response to this, we propose a new method for balancing the data set through One Sided Selection and Artificial Immune Systems. This new balancing mechanism is associated with Decision Trees to improve the results of stroke prediction in a highly unbalanced data set when compared to the state-of-the-art in terms of specificity and sensitivity. Second, the algorithms generally applied to stroke prediction problem do not allow the development of models considered interpretable; this type of model is important in health problems because it allows the emergence of new hypotheses related to the problem and their validation by specialists knowledge. Thus, we also present a new simplification operator that reduces the complexity of trees induced by GP increasing interpretability in the resulting models. The remainder of this paper is organized as follows. Section 2 describes the new proposed approach. Section 3 presents the experiments and the results as well as the used data set. Finally, the conclusions are presented in Section 4.

2. Proposed approach

2.1. Immune/neural approach

The Artificial Immune Systems (AIS) are adaptive systems whose development is inspired by theoretical immunology and the known immune functions (Timmis, Hone, Stibor, & Clark, 2008). The AIS constitutes an area in the bio-inspired computation in which abstract components of the immune system are proposed to solve engineering problems (Castro, 2002). Among the immune functions implemented by these components, the basic principles of clonal selection can be used for pattern recognition and optimization problems. The ClonALG proposed in De Castro and Von Zuben (2002) considers different immunological aspects of the clonal selection theory, such as maintenance of a specific memory through selecting and cloning the most stimulated antibodies while pruning the non-stimulated; affinity maturation through hypermutation mechanisms; and selecting clones according to their antigenic affinity. The whole principle is simplified in a way only antibodies compose the immune system, while the antigens constitute the individuals to be recognized.

The balancing mechanism proposed in this paper uses an AIS based on the clonal selection theory to obtain more representative instances within the data sets. In this work, we use a modified version of ClonALG in which the affinity maturation process is aided by a Kohonen neural network (Kohonen, 1990); the result is the immune/neural approach

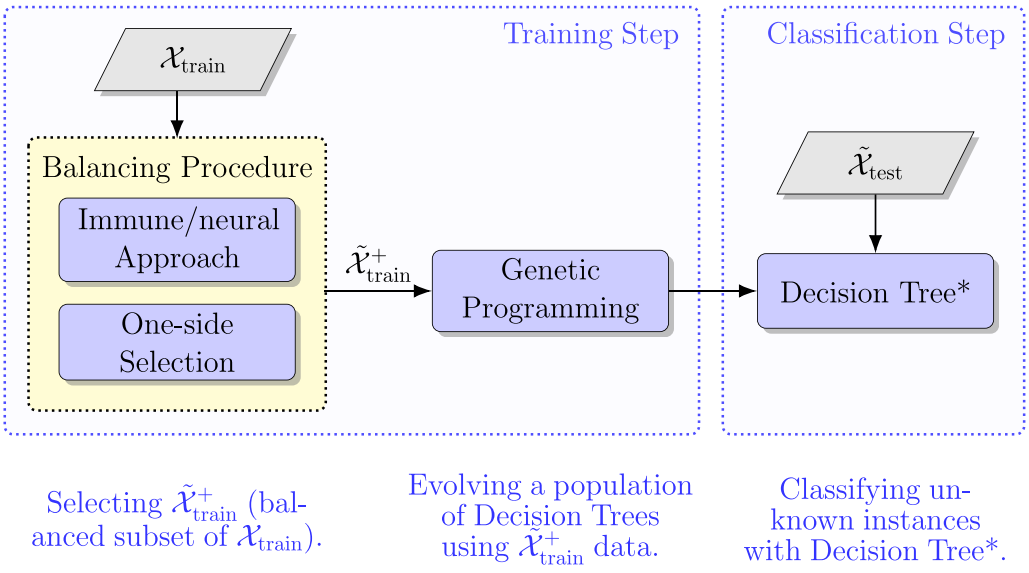


Fig. 1. Overall procedure chart.

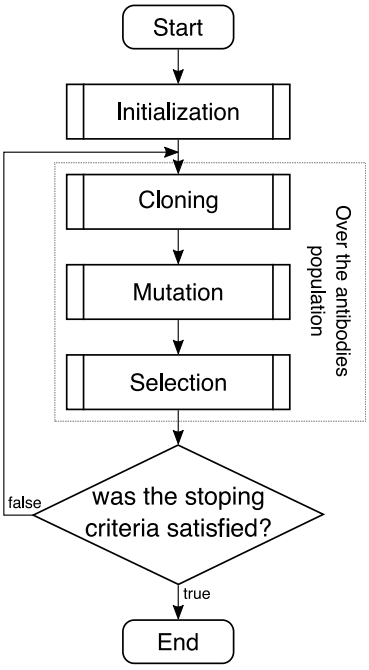


Fig. 2. General flowchart of the used Immune/Neural approach.

described in D'Angelo et al. (2016) and used in Monção et al. (2020) for characterization of salivary gland neoplasms. This structure's training process is made in three steps that are applied to all antibodies in the immune system: cloning, mutation and selection, as shown in Fig. 2. In the cloning step, all antibodies receive two clones that suffer random mutations to ensure diversity in the population. In the hypermutation step, each antibody mutates according to its antigenic affinity; in this step, the weight update procedure of a Kohonen neural network is adopted to increase the antibody affinity to the antigens. In the selection mechanism, the antibodies that recognize the same type of antigen and are close to each other are merged, while antibodies that do not recognize any antigen are pruned. The antibody proximity threshold is computed in each iteration as 25% of the average distance between all antibodies.

The Kohonen network is used in the maturation process of the cloned antibodies. The antibodies are represented by the output units of a Kohonen network in a way their spatial positions are represented by the weight of each unit. The position of the winning unit is adjusted at each iteration with respect to the positions of each antigen (x_i) using the Kohonen network method of weight adjustment. The smallest Euclidean distance between antibodies and antigens defines the winning unit, or winning antibody with respect to that specific antigen. In these terms, the hypermutation mechanism is proportional to the antibodies' affinities since it is directly associated with the distances between antigens and antibodies.

Algorithm 1 Immune/Neural algorithm (IN: $\tilde{\mathcal{X}}^{+}_{\text{train}}$, μ, ψ; OUT: $\mathcal{B}$)

$\mathcal{B} \leftarrow$ Initialize with one centralized antibody
while the stopping criteria is not satisfied **do**
  `// Start cloning section`
  for $\text{ab} \in \mathcal{B}$ **do**
    $\text{ab}^{1}_{\text{tmp}} \leftarrow$ ab clone with random mutation $\mathcal{U}\left(-\frac{\mu}{2}, \frac{\mu}{2}\right)$
    $\text{ab}^{2}_{\text{tmp}} \leftarrow$ ab clone with random mutation $\mathcal{U}\left(-\frac{\mu}{2}, \frac{\mu}{2}\right)$
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{\text{ab}^{1}_{\text{tmp}}, \text{ab}^{2}_{\text{tmp}}\}$
  end for
  `// Start mutation section`
  for $\text{ag} \in \tilde{\mathcal{X}}^{+}_{\text{train}}$ **do**
    $\text{ab}_{\text{win}} \leftarrow$ find the wining antibody in $\mathcal{B}$ with respect to $\tilde{\mathcal{X}}^{+}_{\text{train}}$
    Update ab_{win} position with Kohonen Network updating rule
  end for
  `// Start selection section`
  $\text{affinity} \leftarrow \psi \cdot \text{mean}\left\{\text{norm}(\text{ab}_1, \text{ab}_2) \,|\, (\text{ab}_1, \text{ab}_2) \in \binom{\mathcal{B}}{2}\right\}$
  for $(\text{ab}_1, \text{ab}_2) \in \binom{\mathcal{B}}{2}$ **do**
    if $\text{norm}(\text{ab}_1, \text{ab}_2) < \text{affinity}$ and $\text{ab}_1^{\text{recog AG}} == \text{ab}_2^{\text{recog AG}}$ **then**
      $\text{ab}_{\text{tmp}} \leftarrow$ merges ab_1 with ab_2
      $\mathcal{B} \leftarrow \left(\mathcal{B} \cup \text{ab}_{\text{tmp}}\right) \setminus \{\text{ab}_1, \text{ab}_2\}$
    end if
  end for
end while

Two key parameters are added in the immune/neural approach to allow more liberty in its use for different applications. The random mutation mechanism will be affected by a mutation coefficient $\mu \in [0, 1] \subset \mathbb{R}$ and the proximity threshold is modified by a proximity coefficient

$\psi \in [0,1] \subset \mathbb{R}$. Both parameters will compose the proposed approach's hyperparameters vector and will be optimized in a cross-validation procedure. A short version of the algorithm proposed in D'Angelo et al. (2016) is presented in Algorithm 1.

2.2. One-Sided Selection - OSS

The OSS algorithm is a subsampling method widely used in imbalanced classification problems, i.e., the classification categories are not approximately equally represented (Chawla, 2009). The majority class's most representative instances are selected from a given reference, while all instances of the minority class are preserved. The instance selection from the majority class is made in three steps: (1) selection of a random sample from the majority class; (2) building a data set with all the minority class instances and the instance selected in the first step; (3) the majority class's remaining samples are classified using their nearest neighbor label that belongs to the set constructed in step 2. The instances correctly classified are removed from the data set. The balanced data set is composed of the minority class, the instance selected in step 1, and the instances incorrectly classified in step 3.

2.3. Proposed balancing procedure

Considering an imbalanced data set problem represented by two classes, namely minority and majority classes, the classes set can be defined as $C = \{\min, \text{maj}\}$. The whole data set $\mathcal{X} = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\}$ is composed by n pairs $(\mathbf{x}_i, c_i) \in \mathbb{R}^d \times C$ and can be divided into train and test subsets, such that $\mathcal{X} = \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{test}}$. The set of antibodies $\mathcal{B} = \{(\mathbf{b}_1, c_1), \ldots, (\mathbf{b}_m, c_m)\}$ obtained through the immune/neural approach is composed by m pairs $(\mathbf{b}_j, c_j) \in \mathbb{R}^d \times C$ and can be divided into minority class antibodies and majority class antibodies, such that $\mathcal{B} = \mathcal{B}_{\min} \cup \mathcal{B}_{\text{maj}}$. The recognition function $\rho : \mathbb{R}^d \times \underline{\mathcal{B}} \to C$ that allows the classification of an antigen $(\mathbf{x}_i, c_i) \in \underline{\mathcal{X}} \subseteq \mathcal{X}$ related to a subset of antibodies $\underline{\mathcal{B}} \subseteq \mathcal{B}$ is given as

$$\rho(\mathbf{x}_i, \underline{\mathcal{B}}) = \arg\max_{c_j} \{\sigma(\mathbf{x}_i, \mathbf{b}_j) \mid (\mathbf{b}_j, c_j) \in \underline{\mathcal{B}}\}, \tag{1}$$

where $\sigma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the similarity function, given as

$$\sigma(\mathbf{u}, \mathbf{v}) = \left[\sqrt{(\mathbf{u}-\mathbf{v})^\top (\mathbf{u}-\mathbf{v})}\right]^{-1}. \tag{2}$$

Likewise, the most similar antigen $(\mathbf{x}_i, c_i) \in \underline{\mathcal{X}}$ with respect to a given antibody $\mathbf{b}_j$ can be computed by the function $\beta : \mathbb{R}^d \times \underline{\mathcal{X}} \to \mathbb{R}^d \times C$, in the following way:

$$\beta(\mathbf{b}_j, \underline{\mathcal{X}}) = \arg\max_{(\mathbf{x}_i, c_i)} \{\sigma(\mathbf{x}_i, \mathbf{b}_j) \mid (\mathbf{x}_i, c_i) \in \underline{\mathcal{X}}\}. \tag{3}$$

The proposed balancing strategy is based on a hybrid OSS algorithm aided by the immune/neural approach. A novel scheme for choosing the representative instances in both minority and majority classes is detailed in the following steps:

1. For each antibody in the minority class, obtain the most similar antigen available, such that:

$$\mathcal{S}_{\min} = \{\beta(\mathbf{b}_j, \mathcal{X}_{\text{train}}) \mid (\mathbf{b}_j, k_j) \in \mathcal{B}_{\min}\}. \tag{4}$$

2. For each antibody in the majority class, obtain the most similar antigen available, such that:

$$\mathcal{S}_{\text{maj}} = \{\beta(\mathbf{b}_j, \mathcal{X}_{\text{train}}) \mid (\mathbf{b}_j, k_j) \in \mathcal{B}_{\text{maj}}\}. \tag{5}$$

3. Build a set $\mathcal{S}$ given by the sets found in steps 1 and 2:

$$\mathcal{S} = \mathcal{S}_{\min} \cup \mathcal{S}_{\text{maj}}. \tag{6}$$

4. Classify the other instances in $(\mathbf{x_i}, c_i) \in \mathcal{X}_{\text{train}}$ with the label of its nearest neighbor in $\mathcal{S}$ using the following classification function $\kappa : \mathbb{R}^d \to C$

$$\kappa(\mathbf{x}_i) = \arg\max_{c_s} \{\sigma(\mathbf{x}_i, \mathbf{x}_s) \mid (\mathbf{x}_s, c_s) \in \mathcal{S}\}. \tag{7}$$

5. Build the balanced set composed by the instances incorrectly classified in step 4 and the instances in $\mathcal{S}$, such that:

$$\mathcal{X}^+_{\text{train}} = \{(\mathbf{x}_i, c_i) \mid \kappa(\mathbf{x}_i) \neq c_i, (\mathbf{x}_i, c_i) \in \mathcal{X}_{\text{train}} \backslash \mathcal{S}\} \cup \mathcal{S}. \tag{8}$$

The whole process using a synthetic data set is depicted in Fig. 3. The original data is depicted in Fig. 3(a); the antibodies found by the immune/neural approach are depicted in Fig. 3(b); the most similar antigens from the original data set with respect to the antibodies found are depicted in Fig. 3(c); finally, Fig. 3(d) depicts the balanced data set from (8).

2.4. Decision trees generated by genetic programming

Genetic Programming (GP) is a technique of evolutionary computation that simulates Darwin's principle of natural selection through genetic operators such as reproduction, recombination, and mutation (Banzhaf, 1998). GP systems can represent the candidate solution to a problem in several ways, with trees being quite frequent. In this representation, each individual in the population has ordered ramifications in which the internal nodes are functions while the tree's leaves are the problem's terminals. Each tree is a candidate solution to the problem, and, as in other algorithms in evolutionary computation, they are evaluated with a goodness measure (or fitness) that reflects how good is a solution concerning the others in the same population (Zhao, 2007).

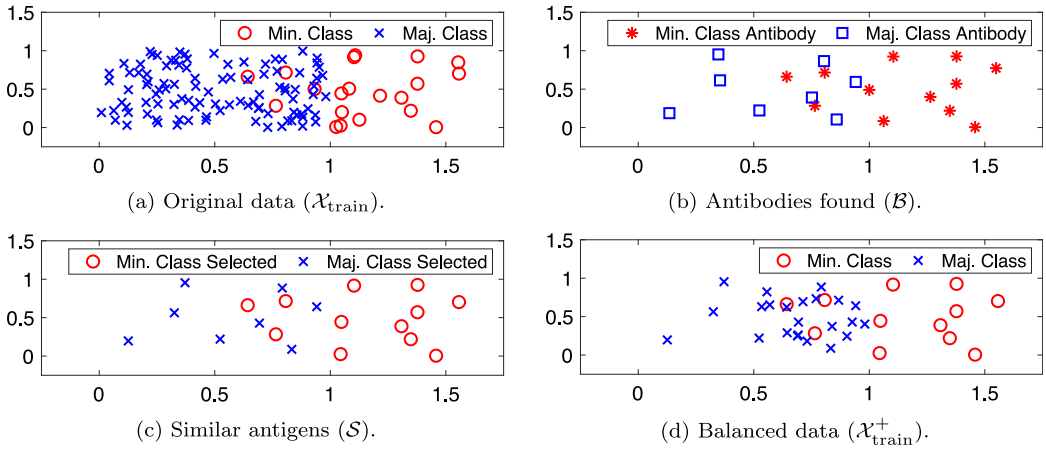
The instance $(\mathbf{x}_i, c_i) \in \mathcal{X}^+_{\text{train}}$, such that $\mathbf{x}_i \in \mathbb{R}^d$, has d attributes that will be preprocessed to assume a finite set of values. The continuous variables will be discretized into N categories using their percentile points while the categorical variables can be re-categorized in the same way when the possible values it can assume are too many, decreasing the final solutions' complexity and overfitting (Saremi & Yaghmaee, 2014). In this paper, we adopt the strategy of limiting the number of percentiles or categories to 4, i.e., $N \in \{2, 3, 4\}$, randomly chosen following a uniform distribution. Therefore, the new training data set $\tilde{\mathcal{X}}^+_{\text{train}}$ is composed by pairs $(\tilde{\mathbf{x}}_i, c_i) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_d \times C$, where $\mathcal{A}_k = \{a^k_1, \ldots, a^k_{n_k}\}$ is the finite set of n_k values that the kth attribute can assume.

In GP's use to induce decision trees, the internal nodes represent the new training data set's attributes while the leaves represent the classes. The attribute test function $\phi_k : \mathcal{A}_k \to \mathcal{Y} \subseteq \tilde{\mathcal{X}}^+_{\text{train}}$ partitions the new training data set such that each partition contains all instances of $\tilde{\mathcal{X}}^+_{\text{train}}$ where the kth attribute is equal to a given value a^k_j

$$\phi_k(a^k_j) = \{(\tilde{\mathbf{x}}_i, c_i) \mid \tilde{x}_{i,k} = a^k_j, (\tilde{\mathbf{x}}_i, c_i) \in \tilde{\mathcal{X}}^+_{\text{train}}\},$$

where $\tilde{x}_{i,k}$ is the kth attribute of vector $\tilde{\mathbf{x}}_i$. When an instance needs to be evaluated, the function in the tree's root tests the corresponding attribute, and if the argument is a terminal, the decision (classification) for this instance will be returned; otherwise, a new attribute will be evaluated.

The Algorithm 2 shows the pseudo-code for the proposed decision tree induced by GP, where the input parameters are: the new training data set $\tilde{\mathcal{X}}^+_{\text{train}}$, the maximum number of generations $\zeta \in \mathbb{N}^*$, the simplification threshold $\varepsilon \in [0,1] \subset \mathbb{R}$, and the simplification frequency $\tau \in \mathbb{N}^*$, where $\mathbb{N}^*$ is the set of natural numbers without zero; and the output parameter is the best Decision Tree (Decision Tree*). The method Initialization (Line 1) produces the initial population based on the training data set, as described in Section 2.4.1; Fitness (Line 2 and Line 11) measures the fitness of each individual in the population with respect to the training data; the method Recombine (Line 6) constructs a child population using the parent select from the Select method (Line 5); Mutate method (Line 7) inserts variability in the child population, as described in Section 2.4.3; the instruction Remainder (Line 8) guarantees the periodicity of the simplification tests at every τ generations; the Simplify method (Line 9) will prune trees with non-expressive internal nodes, controlling their growth, as described in Section 2.4.5; finally, ε represents the predefined threshold for eliminating sub-trees.



(a) Original data ($\mathcal{X}_{\text{train}}$). (b) Antibodies found ($\mathcal{B}$). (c) Similar antigens ($\mathcal{S}$). (d) Balanced data ($\mathcal{X}^{+}_{\text{train}}$).

Fig. 3. Balancing procedure example.

Algorithm 2 Genetic Programming(IN:$\tilde{\mathcal{X}}^{+}_{\text{train}}$, ζ, ε, τ; OUT:Decision Tree*)

1: $P \leftarrow$ INITIALIZE($\tilde{\mathcal{X}}^{+}_{\text{train}}$)
2: $F_P \leftarrow$ FITNESS($\tilde{\mathcal{X}}^{+}_{\text{train}}$, P)
3: gen $\leftarrow 1$
4: **while** gen $< \zeta$ **do**
5: $\quad P_S \leftarrow$ SELECT(P, F_P)
6: $\quad P_C \leftarrow$ RECOMBINE(P_S)
7: $\quad P \leftarrow$ MUTATE(P_C)
8: $\quad$ **if** REMAINDER(gen, τ) $= 0$ **then**
9: $\quad\quad P \leftarrow$ SIMPLIFY(P, ε)
10: $\quad$ **end if**
11: $\quad F_P \leftarrow$ FITNESS($\tilde{\mathcal{X}}^{+}_{\text{train}}$, P)
12: $\quad$ gen $\leftarrow$ gen $+ 1$
13: **end while**

2.4.1. Initial population generation

Considering the tuple of all attributes in the data set $\mathbf{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_d)$, let $\mathbf{A}_p$ be the set of all $d!$ permutations of $\mathbf{A}$ where $\mathbf{A}_p(k)$ is the kth permutation. The initial tree population is constructed by choosing a random permutation $\mathbf{A}^* \in \mathbf{A}_p$, such that

$$\mathbf{A}^* = \mathbf{A}_p(u), \quad \text{with } u \sim \mathcal{U}\{1, d!\}, \tag{9}$$

where $\mathcal{U}\{a, b\}$, with $b > a$, is a discrete uniform distribution with support $s \in \{a, a+1, \ldots, b-1, b\}$. The randomly chosen tuple in (9) can be used to construct a tree $\mathcal{T}_{\mathbf{A}^*} = (t_1, \ldots, t_d)$ whose nodes t_n are elements of $\mathbf{A}$. The first element t_1 is the tree's root while the left and right children of element t_n are given by

$$t_{n,\text{left}} = \begin{cases} t_{2n}, & \text{if } 2n \le d \\ \emptyset, & \text{otherwise,} \end{cases} \tag{10a}$$

$$t_{n,\text{right}} = \begin{cases} t_{2n+1}, & \text{if } 2n+1 \le d \\ \emptyset, & \text{otherwise.} \end{cases} \tag{10b}$$

The following trees in the initial population P (Line 1 of Algorithm 2) will be created by moving the elements of $\mathcal{T}_{\mathbf{A}^*}$ to the first position (root node) one by one. After all rotations to create d different trees from $\mathcal{T}_{\mathbf{A}^*}$, a new instance of $\mathbf{A}^*$ will be drawn if the maximum number of trees $\delta \in \mathbb{N}^*$ in the initial population is not reached. This maximum number is empirically defined for each data set and the process to create the initial population is depicted in Fig. 4. The initial population is the set $P = \{p(1), \ldots, p(\delta)\}$ where $p(k)$ is the tree representation of a tuple $\mathbf{A}^* \in \mathbf{A}_p$ whose fitness $f_p(k)$ composes the fitness vector $F_P = \{f_p(1), \ldots, f_p(\delta)\}$.

2.4.2. Selection operator

In order to select which parents will be responsible to create the next generation, a binary tournament is done. The main idea is to run multiple simulations to select parents with greater fitness values and control the selective pressure. Let $\mathcal{I} = \{1, 2, \ldots, \delta\}$ and $\mathcal{I}_p$ is the set of all $\delta!$ permutations of $\mathcal{I}$ where $\mathcal{I}_p(k)$ is the kth permutation. A random permutation $\mathcal{I}^* = \mathcal{I}_p(v)$ is chosen, with $v \sim \mathcal{U}\{1, \delta!\}$, such that

$$\mathcal{I}^* = (i^*_1, \ldots, i^*_\delta), \tag{11}$$

where $i^*_k \in \mathcal{I}$. The selected population is given as

$$P_S = \{p_s(1), \ldots, p_s(\delta_s)\}, \tag{12}$$

where the individuals are chosen by a binary tournament, such that

$$p_s(k) = \begin{cases} p(i^*_{2k-1}), & \text{if } f_p(i^*_{2k-1}) \ge f_p(i^*_{2k}) \\ p(i^*_{2k}), & \text{otherwise.} \end{cases} \tag{13}$$

The number of individuals in P_S is given by

$$\delta_s = \begin{cases} \frac{\delta}{2}, & \text{if } \delta \text{ is even} \\ \frac{\delta-1}{2}, & \text{otherwise.} \end{cases} \tag{14}$$

2.4.3. Variability operators

The recombination procedure takes the parents in the selected population in pairs. Each pair generates two children that will compose P_C, given as

$$P_C = \{p_s(1), \ldots, p_s(\delta_s), p^1_c(1), p^2_c(1), \ldots, p^1_c(\delta_c), p^2_c(\delta_c)\}, \tag{15}$$

where $p_s(k)$ are individuals from the previously selected population and $(p^1_c(k), p^2_c(k))$ are the children generated by crossing $p_s(k)$ and $p_s(r_\delta - k + 1)$ for $1 \le k < \frac{r_\delta + 1}{2}$. The crossover operator chooses a random node on both parents $p_s(k)$ and $p_s(r_\delta - k + 1)$, except their roots; then, the sub-trees whose roots are the random cutting points are exchanged, creating two children: $p^1_c(k)$ and $p^2_c(k)$. An example of the procedure is depicted in Fig. 5; in Fig. 5(a), the first parent $p_s(1) \in P_S$ is chosen and the attribute node $\mathcal{A}_1$ is selected to be a cut point; in Fig. 5(b), the first parent $p_s(\delta_s) \in P_S$ is chosen and the attribute node $\mathcal{A}_9$ is selected to be a cut point; Figs. 5(c) and 5(d) shows the sub-tree exchange between the parents creating two children.

The mutation operator is applied only to previously continuous attributes at a rate of $\eta \in [0, 1] \subset \mathbb{R}$, i.e., the attribute will suffer mutations with probability η. After deciding that a mutation should happen, a previously continuous attribute will be randomly selected to have its discretization limits changed; according to Saremi and Yaghmaee (2018), this mutation mechanism is competitive with others in terms of simplicity and efficiency. For example, suppose a node that represents

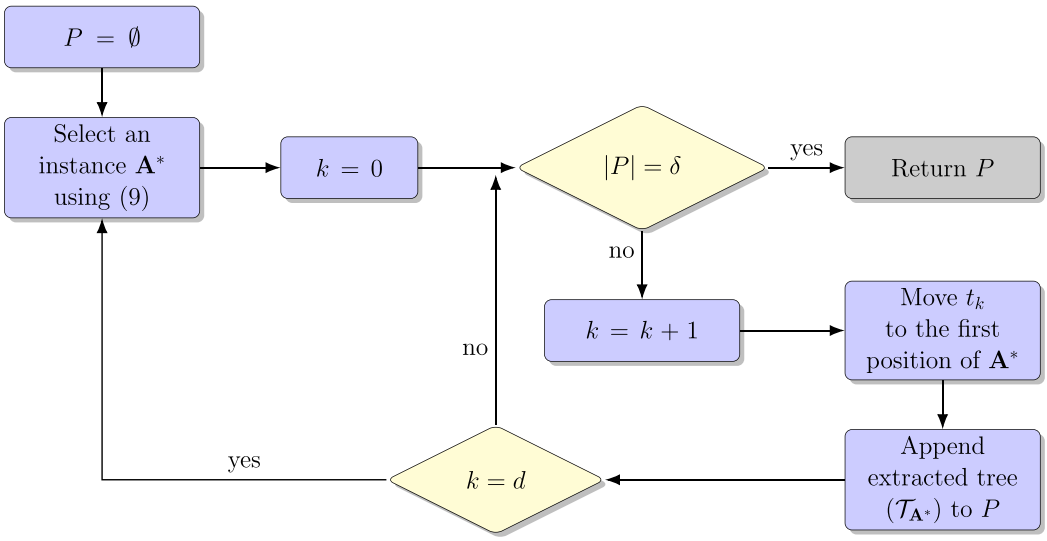
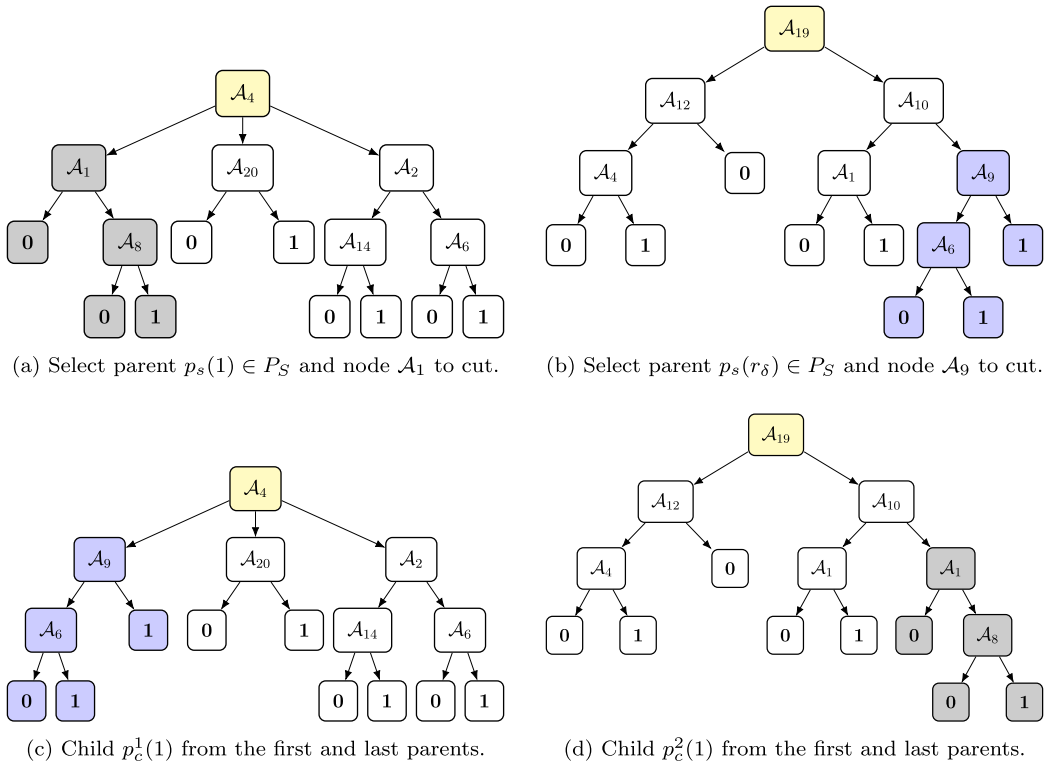


Fig. 4. Initial population generation flowchart.



(a) Select parent $p_s(1) \in P_S$ and node $\mathcal{A}_1$ to cut.

(b) Select parent $p_s(r_\delta) \in P_S$ and node $\mathcal{A}_9$ to cut.

(c) Child $p_c^1(1)$ from the first and last parents.

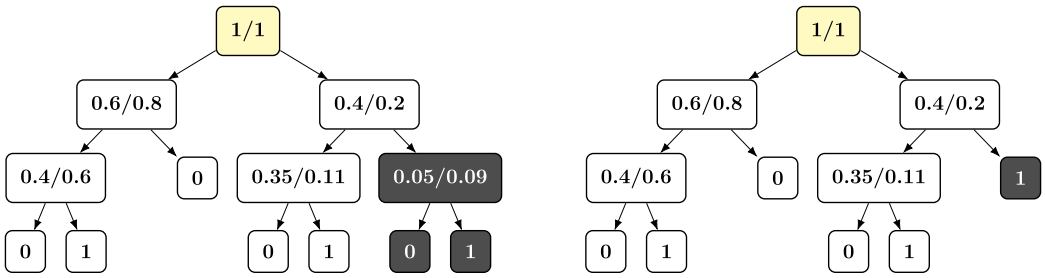
(d) Child $p_c^2(1)$ from the first and last parents.

Fig. 5. Example of the crossover operator behavior applied on a pair of parents $p_s(1), p_s(\delta_s) \in P_S$ generating two children $p_c^1(1)$ and $p_c^2(1)$.

the age attribute in years and the population has individuals whose age lies in the interval $[0, 98]$; a possible discretization of this attribute is to create three categorical ranges, such as $\{[0, 25], [26, 57], [58, 98]\}$. Then a mutation on this discretized attribute could shift the middle interval category in 4 units as, for instance, $\{[0, 25], [26, 61], [62, 98]\}$, creating a new discretization. The new population is assigned to P as the new generation population.

2.4.4. Fitness function

The fitness function evaluates the quality of a given individual with respect to a predefined objective. It is used in the training process to push the population in a given direction aiding in the training process (Devarriya, Gulati, Mansharamani, Sakalle, & Bhardwaj, 2020). In the context of classification tasks, a solution can be evaluated using different metrics such as accuracy, F1-score, G_{mean}, Cohen's kappa, etc (Zhou, Li, & Mitri, 2015). Although accuracy and Choen's kappa are very common evaluating metrics, in imbalanced data sets they can be biased with respect to the majority class, leading to wrongful results (Devarriya et al., 2020). For imbalanced data sets, both F1-score and G-mean metrics minimize the accuracy bias, however, F1-score is preferable when the minority class is more important while G_{mean} is used to maximize the sensitivity of both classes (Al-Badarneh, Habib, Aljarah, & Faris, 2020). Thus, in this paper, the individuals are represented by decision trees and the fitness function is the geometric mean



(a) Selects a sub-tree whose root node represents less than ε% of the training data.

(b) Replaces the sub-tree by a leaf node represented by the most common class in it.

Fig. 6. Example of the simplification operator with $\varepsilon = 0.1$.

between specificity (SPC) and sensibility (SEN), defined as:

$$\text{SPC} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{16a}$$

$$\text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{16b}$$

$$G_{\text{mean}} = \sqrt{\text{SPC} \cdot \text{SEN}}, \tag{16c}$$

where TP, FP, TN and FN denote, respectively, true positives, false positives, true negatives and false negatives. This metric is used due to its property of independence to the distribution of examples from different classes, forcing the learning system to produce correct classifications in a significant fraction of the positive examples (Kubat, Holte, & Matwin, 1998).

2.4.5. Simplification operator

The initialization method creates trees with a fixed amount of internal nodes; however, the recombination procedure can make some trees grow boundless over new generations. As the tree size increases, the results are prone to overfitting and can experience a decrease in its interpretability capabilities (DeLisle & Dixon, 2004). Doerr, Lissovoi, and Oliveto (2019) shows that GP systems for evolving simple Boolean functions formed by the conjunction of some variables will require a logarithmic limit on the tree size. In Lissovoi and Oliveto (2019), it is recommended the use of strategies to reduce the tree's growth by using pre-established limiting thresholds. Thus, a threshold (ε) is defined to remove inexpressive sub-trees, defined as the sub-trees whose root node contains less than ε% of training instances from both classes; each node stores the amount of data of each class it represents concerning the entire data set. This procedure is executed periodically with a period τ empirically defined. An example of this procedure considering $\varepsilon = 0.1$ is depicted in Fig. 6, where the gray sub-tree in Fig. 6(a) represents 5% of class 0 and 9% of class 1. Fig. 6(b) shows the replacement of this sub-tree by a leaf node with the most representative class 1.

3. Experiments and results

3.1. Data set

In order to evaluate the proposed approach, the present work uses the same data set evaluated in Liu et al. (2019). The full data set is provided in Liu (2019). The data set is composed of 43,400 instances with ten features, as described in Table 1. In this work, all cases with missing values for at least one feature were removed. The remaining data set is a typical imbalanced data set containing 29,063 instances, with 1.89% of stroke occurrences.

Table 1
Data set description.

Feature	Values	Feature	Values
Gender (gen)	Male/Female	Hypertension (hyp)	Yes/No
Residence type	Urban/Rural	Age	0.08–82
Avg. glucose (glu)	55–291	Heart disease (htd)	Yes/No
Work type (work)	Private/Employed	BMI	10.1–97.6
Married (mar)	Yes/No	Smoking status	*S/F/N

*S/F/N represents Smoked/Formerly/Never.

3.2. Experimental setup

Two experiments are proposed to evaluate the proposed method. In the first experiment, the decision tree induced by the GP algorithm is compared to a state-of-the-art technique in terms of five metrics: specificity (16a), sensitivity (16b), G_{mean} (16c), Area Under the Curve (AUC) derived from a Receiver Operating Characteristic (ROC) analysis and accuracy (ACC), which is defined as:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}. \tag{17}$$

where TP, FP, TN and FN denote, respectively, true positives, false positives, true negatives and false negatives. The first experiment consists of a 5-fold cross-validation procedure that uses three folds for training, one to adjust the hyperparameters and the last to evaluate the best decision tree obtained. Then, the proposed simplification operator is evaluated in terms of the previous metrics and the average number of internal nodes, referred to as complexity. For these experiments with cross-validation, results are considered statistically significant when the p-value < 0.05 in the comparison, according to the Mann Whitney test (Mann & Whitney, 1947).

In the second experiment, a decision tree obtained is evaluated on a qualitative basis, connecting the current medical knowledge on the subject with the tree generated by data, where 60% of the data set were used for training, 20% to adjust the hyperparameters, and 20% for testing. The hyperparameters vector adjusted on each experiment is given as

$$\boldsymbol{\Theta} = \begin{bmatrix} \mu & \psi & \zeta & \delta & \eta & \tau & \varepsilon \end{bmatrix}^{\top}. \tag{18}$$

3.3. Numerical results and analysis

The proposed balancing procedure performs a sub-sampling in both classes, intending to eliminate less representative instances. This strategy allows the maximization, simultaneously, both sensitivity and specificity. In highly imbalanced data sets, such as the one used in this paper, the proposed mechanism tends to remove more instances in the majority class at the expense of the minority one, as shown in Fig. 7. The average reduction of instances in the majority class is 86%, while in the minority class is 15.5%, considering the five configurations in

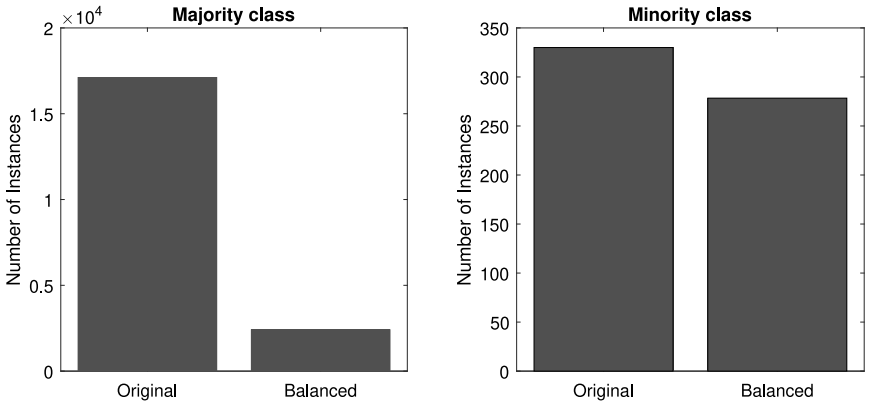


Fig. 7. Instance reduction after balancing.

Table 2
Comparison between the proposed approach and the state-of-the-art technique in terms of four metrics with best values in bold.

Approach	ACC	SPC	SEN	G_{mean}	AUC
Proposed	0.70	**0.70**	**0.78**	**0.74**	**0.74**
AutoHPO Liu et al. (2019)	**0.72**	0.33	0.67	0.47	0.50
p-value	0.713	< 0.002	< 0.002	< 0.002	< 0.002

Table 3
Confusion matrix for proposed approach and state-of-the-art approach.

Approach	Desired output	Predict output	
		Stroke	No stroke
Proposed	Stroke	429	121
	No stroke	8515	19998
AutoHPO	Stroke	368	182
	No stroke	19105	9409

Table 4
Evaluation of the simplification operator in terms of five metrics with best values in bold.

Approach	ACC	SPC	SEN	G_{mean}	AUC	Complexity
With operator	0.70	0.70	**0.78**	**0.74**	0.74	**4.2**
Without operator	**0.76**	**0.76**	0.71	0.73	0.73	17.6
p-value	0.117	0.117	0.14	0.347	0.342	0.009

the cross-validation. Thus, after balancing is done, the minority class represents, on average, 11% of the data set.

The results for the first experiment are compared with the state-of-the-art technique (Liu et al., 2019) in terms of accuracy, sensitivity, specificity, G_{mean} and AUC, as shown in Table 2, where the accuracy was the only metric with no statistical difference to the state-of-the-art (p-value < 0.05). Maximizing sensitivity is of great importance due to the fact that low sensitivity methods can make preventive measures unfeasible, causing irreversible damage to the patients. On the other hand, low specificity can impose a heavy burden to the patients being unfavorable to their health (Liu et al., 2019). According to the results in Table 2, the proposed approach managed to significantly improve sensitivity and specificity in comparison with a state-of-the-art technique; it is essential to emphasize that the proposed method's sensitivity is near 80%, placing it above the second quartile considering the sensitivity from the methods described in a recent systemic review of clinical tools for acute stroke assessment (Antipova, Eadie, Macaden, & Wilson, 2019).

Changing the classification threshold of a model will output multiple pairs of sensitivity and specificity values. The relation between sensitivity and specificity creates a receiver-operating characteristic (ROC) curve, which is an effective method for determining the discrimination performance of the model. The most commonly used summary measure of ROC curves is the area under the ROC curve (AUC) that can take any value between 0 and 1. The closer the AUC is to 1, the better the discrimination performance of the diagnostic test (Park & Han, 2018). The AUC is also widely used to evaluate classifiers in imbalanced data sets. From an statistical point-of-view, the bigger its value, the greater the probability that a randomly chosen diseased person is correctly classified instead of a randomly chosen non-diseased person (Hanley & McNeil, 1982). In Table 2, the AUC of the proposed approach is significantly higher then the AUC of the state-of-the-art approach, implying a greater discriminating power between patients with and without stroke. The low AUC value (0.50) achieved by the state-of-the-art approach is related to its low specificity (0.33).

Another way to show a classifier performance is to use the confusion matrix, in which the principal diagonal indicates the correct classifications (TP and TN). Table 3 shows the confusion matrices of the proposed approach and the state-of-the-art approach. As shown in Table 3, the proposed approach can detect 429 patients with stroke out of 550 patients with the disease; in other words, it detects 61 more patients that can start premature treatment. In short, the proposed approach allows a larger number of interventions in people with some risk of stroke and reduces the false alarm rate, compared to the state-of-the-art approach.

The proposed simplification operator is evaluated in terms of complexity, i.e., the average number of internal nodes in the solution, in Table 4. The simplification operator managed to reduce in almost 80% the number of internal nodes ($p < 0.05$) while maintaining the classification quality concerning sensitivity, specificity, AUC, and G_{mean} ($p > 0.05$). The trees generated when the simplification procedure is active are less complex and more interpretable than trees with boundless growth. In the context of medical practice, the interpretability of a model obtained from ML algorithms has been a topic of interest to establish the trust in these tools, allowing their validation in real environments. Therefore, clinical practitioners and other decision-makers in the health area tend to see the interpretability as a priority in implementing and using these tools (Ahmad, Eckert, & Teredesai, 2018).

In the second experiment, the proposed approach achieved 0.74, 0.74, 0.78 and 0.76 for the metrics accuracy, specificity, sensitivity, and G_{mean}, respectively. The induced tree is composed of three internal nodes, as shown in Fig. 8, where the main reported feature is the Age (the tree's root), followed by Heart Disease (htd) and Avg. Glucose (glu). The age is widely accepted as a strong risk factor for stroke (Gan et al., 2020, 2017; Liu et al., 2019). According to Gan et al. (2020), a person with more the 70 years, in comparison with individuals with 40 to 49 yeas, has approximately 20% more chances of having a stroke. In the induced tree, this knowledge appears in the rule where (Age ≥

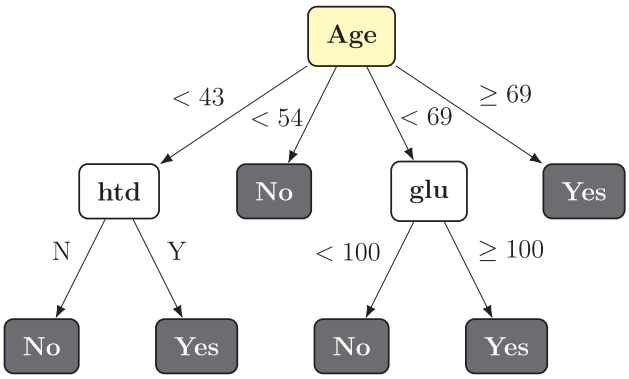


Fig. 8. Tree Induced via GP approach.

69 → Class = yes). Another risk factor associated with stroke, is the presence of Diabetes Mellitus; the presence of this disease can increase in 62% the chances of having a stroke (Gan et al., 2020). This relation is shown in the rule (54 ≤ Age < 69 ∧ glu ≥ 100 → Class = yes), where a high level of glucose is a risk factor for developing diabetes mellitus (Mellitus, 2005). The relationship between blood glucose levels and stoke prediction is also found in the Los Angeles Prehospital Stroke Screen (LAPSS), where blood glucose between 60 and 400 does no exclude true strokes (Kidwell, Starkman, Eckstein, Weems, & Saver, 2000). The occurrence of Heart Disease (htd) as a risk factor in stroke cases was justified in a recent study (Ranganai & Matizirofa, 2020) with private and public hospitals from South Africa where 75.7% of confirmed stroke cases in these hospitals had heart problems and 86.9% had diabetes; this relationship is modeled through the rule (Age < 43 ∧ htd = Y → Class = yes). The presence of glucose and heart diseases as risk factors are important to stroke prediction models because they are controllable factors.

In summary, the proposed balancing strategy eliminates instances considered unrepresentative from both classes; this allows the maximization of sensitivity and specificity, which does not occur in the state-of-the-art technique where the specificity is low. It classifies new instances using a decision tree that can be interpreted by human experts; this interpretability is aided by the complexity reduction achieved through the proposed simplification operator. Using a GP algorithm to induce the creation of decision trees brings two benefits to greedy methods: (i) since GP is a global optimization method, it tends to mitigate the possibility of reach local maxima; (ii) the evolutionary algorithms are more cable of dealing with the complex interactions between attributes and discovering these relations through the concepts of evolution (Barros et al., 2011). It is important to note that the data set used was obtained in a non-invasive way allowing the stroke prediction at a low cost.

4. Conclusion

In this paper, we have presented a novel approach for stroke prediction based on decision trees generated through GP aided by an immune/neural AIS. The proposed approach was evaluated in a highly imbalanced data set composed of sick and non-sick patients' physiological data. The main objective was to present a technique capable of dealing with the imbalance present in the data set while providing a solution that can be interpreted by human specialists. The results have illustrated the achieved improvement in sensitivity and specificity compared to a state-of-the-art model. It also provided an interpretable solution whose complexity management has been done through the proposed simplification operator.

Nevertheless this paper does not address the problem of incomplete data. Unlike the state-of-the-art approach used for comparison, we remove incomplete examples from the data set without any policy to replace missing data with imputed values. Thus, future research directions include studying the proposed approach's robustness to incomplete data sets. Another limitation of the proposed approach is regarding its computational cost; processing times in the order of seconds, minutes, or even days are reported in Espejo, Ventura, and Herrera (2010) for classification tasks using GP. Although high computational costs are common in GP models, due to the individual fitness computation that must be repeatedly evaluated in the evolving process, a parallel version of the proposed algorithm is also an object of future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil (Grant Number s: 307933/2018-0 and 309909/2019-8), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil and the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG), Brazil (Grant Number: PPM-00053-17).

All Authors contributed equally to this work.

References

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics* (pp. 559–560).

Al-Badarneh, I., Habib, M., Aljarah, I., & Faris, H. (2020). Neuro-evolutionary models for imbalanced classification problems. *Journal of King Saud University - Computer and Information Sciences*, http://dx.doi.org/10.1016/j.jksuci.2020.11.005, URL: https://www.sciencedirect.com/science/article/pii/S1319157820305309.

Alghwiri, A. A. (2016). The correlation between depression, balance, and physical functioning post stroke. *Journal of Stroke and Cerebrovascular Diseases*, *25*, 475–479. http://dx.doi.org/10.1016/j.jstrokecerebrovasdis.2015.10.022.

Antipova, D., Eadie, L., Macaden, A., & Wilson, P. (2019). Diagnostic accuracy of clinical tools for assessment of acute stroke: A systematic review. *BMC Emergency Medicine*, *19*, http://dx.doi.org/10.1186/s12873-019-0262-1.

Arslan, A. K., Colak, C., & Sarihan, M. E. (2016). Different medical data mining approaches based prediction of ischemic stroke. *Computer Methods and Programs in Biomedicine*, *130*, 87–92. http://dx.doi.org/10.1016/j.cmpb.2016.03.022.

Banzhaf, W. (1998). *Genetic programming: an introduction on the automatic evolution of computer programs and its applications*. San Francisco, Calif. Heidelburg: Morgan Kaufmann Publishers Dpunkt-verlag.

Barros, R. C., Basgalupp, M. P., De Carvalho, A. C., & Freitas, A. A. (2011). A survey of evolutionary algorithms for decision-tree induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*, 291–312. http://dx.doi.org/10.1109/TSMCC.2011.2157494.

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., et al. (2018). Heart disease and stroke statistics—2018 update: A report from the American heart association. *Circulation*, http://dx.doi.org/10.1161/CIR.0000000000000558.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30 day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721–1730). http://dx.doi.org/10.1145/2783258.2788613.

Castro, L. (2002). *Artificial immune systems: A new computational intelligence approach*. London New York: Springer.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook* (pp. 875–886). Springer US, http://dx.doi.org/10.1007/978-0-387-09823-445.

Chen, Y., Abel, K. T., Janecek, J. T., Chen, Y., Zheng, K., & Cramer, S. C. (2019). Home-based technologies for stroke rehabilitation: A systematic review. *International Journal of Medical Informatics*, *123*, 11–22. http://dx.doi.org/10.1016/j.ijmedinf.2018.12.001.

Colak, C., Karaman, E., & Turtay, M. G. (2015). Application of knowledge discovery process on the prediction of stroke. *Computer Methods and Programs in Biomedicine*, *119*, 181–185. http://dx.doi.org/10.1016/j.cmpb.2015.03.002.

D'Angelo, M. F., Palhares, R. M., Camargos Filho, M. C., Maia, R. D., Mendes, J. B., & Ekel, P. Y. (2016). A new fault classification approach applied to tennessee eastman benchmark process. *Applied Soft Computing*, *49*, 676–686. http://dx.doi.org/10.1016/j.asoc.2016.08.040.

De Castro, L. N., & Von Zuben, F. J. (2002). Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, *6*, 239–251. http://dx.doi.org/10.1109/TEVC.2002.1011539.

DeLisle, R. K., & Dixon, S. L. (2004). Induction of decision trees via evolutionary programming. *Journal of Chemical Information and Computer Sciences*, *44*, 862–870. http://dx.doi.org/10.1021/ci034188s.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, *132*, 1920–1930. http://dx.doi.org/10.1161/CIRCULATIONAHA.115.001593.

Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., & Bhardwaj, A. (2020). Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Systems with Applications*, *140*, Article 112866. http://dx.doi.org/10.1016/j.eswa.2019.112866.

Doerr, B., Lissovoi, A., & Oliveto, P. S. (2019). Evolving boolean functions with conjunctions and disjunctions via genetic programming. In *Proceedings of the genetic and evolutionary computation conference* (pp. 1003–1011). http://dx.doi.org/10.1145/3321707.3321851.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In *Machine learning in radiation oncology* (pp. 3–11). Springer, http://dx.doi.org/10.1007/978-3-319-18305-31.

Espejo, P. G., Ventura, S., & Herrera, F. (2010). A survey on the application of genetic programming to classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*, 121–144. http://dx.doi.org/10.1109/TSMCC.2009.2033566.

Gan, Y., Jiang, H., Room, R., Zhan, Y., Li, L., Lu, K., et al. (2020). Prevalence and risk factors associated with stroke in China: A nationwide survey of 726,451 adults. *European Journal of Preventive Cardiology*, http://dx.doi.org/10.1177/2047487320902324.

Gan, Y., Wu, J., Zhang, S., Li, L., Yin, X., Gong, Y., et al. (2017). Prevalence and risk factors associated with stroke in middle-aged and older Chinese: A community-based cross-sectional study. *Scientific Reports*, *7*, 1–7. http://dx.doi.org/10.1038/s41598-017-09849-z.

García-Temza, L., Risco-Martín, J. L., Ayala, J. L., Roselló, G. R., & Camarasaltas, J. M. (2019). Comparison of different machine learning approaches to model stroke subtype classification and risk prediction. In *2019 spring simulation conference* (pp. 1–10). IEEE, http://dx.doi.org/10.23919/SpringSim.2019.8732846.

GBD (2018). Global, regional, and country-specific lifetime risks of stroke, 1990 and 2016. *New England Journal of Medicine*, *379*, 2429–2437. http://dx.doi.org/10.1056/NEJMoa1804492.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. http://dx.doi.org/10.1016/j.eswa.2016.12.035.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve.. *Radiology*, *143*, 29–36. http://dx.doi.org/10.1148/radiology.143.1.7063747.

Jin, H.-Q., Wang, J.-C., Sun, Y.-A., Lyu, P., Cui, W., Liu, Y.-Y., et al. (2016). Prehospital identification of stroke subtypes in Chinese rural areas. *Chinese Medical Journal*, *129*, 1041–1046. http://dx.doi.org/10.4103/0366-6999.180521.

Khosla, A., Cao, Y., Lin, C. C. -Y., Chiu, H. -K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 183–192). http://dx.doi.org/10.1145/1835804.1835830.

Kidwell, C. S., Starkman, S., Eckstein, M., Weems, K., & Saver, J. L. (2000). Identifying stroke in the field. *Stroke*, *31*, 71–76. http://dx.doi.org/10.1161/01.str.31.1.71.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480. http://dx.doi.org/10.1109/5.58325.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection, volume 1*. MIT Press.

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*, 195–215. http://dx.doi.org/10.1023/a:1007452223027.

Kubat, M., Matwin, S., et al. (1997). Addressing the curse of imbalanced training sets: One-sided selection. *97*, In *International conference on machine learning, volume 97* (pp. 179–186). Morgan Kaufmann.

Leira, E. C., Hess, D. C., Torner, J. C., & Adams, H. P. (2008). Rural-urban differences in acute stroke management practices: A modifiable disparity. *Archives of Neurology*, *65*, 887–891. http://dx.doi.org/10.1001/archneur.65.7.887.

Li, J., Liu, L. -s., Fong, S., Wong, R. K., Mohammed, S., Fiaidhi, J., et al. (2017). Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data. *PLoS One*, *12*, http://dx.doi.org/10.1371/journal.pone.0180830.

Lissovoi, A., & Oliveto, P. S. (2019). On the time and space complexity of genetic programming for evolving Boolean conjunctions. *Journal of Artificial Intelligence Research*, *66*, 655–689. http://dx.doi.org/10.1613/jair.1.11821.

Liu, T. (2019). Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets. Mendeley, http://dx.doi.org/10.17632/X8YGRW87JW.1, URL: https://data.mendeley.com/datasets/x8ygrw87jw/1.

Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial Intelligence in Medicine*, *101*, Article 101723. http://dx.doi.org/10.1016/j.artmed.2019.101723.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60. http://dx.doi.org/10.1214/aoms/1177730491.

Mellitus, D. (2005). Diagnosis and classification of diabetes mellitus. *Diabetes Care*, *28*, S5–S10. http://dx.doi.org/10.2337/diacare.27.2007.s5.

Monção, C. R. L., Santos, E. M., Prates, T. S., de Paula, A. M. B., Cardoso, C. M., Farias, L. C., et al. (2020). Immune/neural approach to characterize salivary gland neoplasms (SGN). *Applied Soft Computing*, *88*, Article 105877. http://dx.doi.org/10.1016/j.asoc.2019.105877.

O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., et al. (2016). Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): A case-control study. *The Lancet*, *388*, 761–775. http://dx.doi.org/10.1016/S0140-6736(16)30506-2.

Park, S. H., Choi, J., & Byeon, J.-S. (2021). Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean Journal of Radiology*, *22*, 442. http://dx.doi.org/10.3348/kjr.2021.0048.

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*, 800–809. http://dx.doi.org/10.1148/radiol.2017171920.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

Ranganai, E., & Matizirofa, L. (2020). An analysis of recent stroke cases in South Africa: Trend, seasonality and predictors. *South African Medical Journal*, *110*, 92. http://dx.doi.org/10.7196/samj.2020.v110i2.013891.

Saremi, M., & Yaghmaee, F. (2014). Evolutionary decision tree induction with multi-interval discretization. In *2014 Iranian conference on intelligent systems* (pp. 1–6). IEEE, http://dx.doi.org/10.1109/IranianCIS.2014.6802543.

Saremi, M., & Yaghmaee, F. (2018). Improving evolutionary decision tree induction with multi-interval discretization. *Computational Intelligence*, *34*, 495–514. http://dx.doi.org/10.1111/coin.12153.

Simpkins, A. N., Janowski, M., Oz, H. S., Roberts, J., Bix, G., Doré, S., et al. (2020). Biomarker application for precision medicine in stroke. *Translational Stroke Research*, *11*, 615–627. http://dx.doi.org/10.1007/s12975-019-00762-3.

Thrift, A. G., Cadilhac, D. A., Thayabaranathan, T., Howard, G., Howard, V. J., Rothwell, P. M., et al. (2014). Global stroke statistics. *International Journal of Stroke*, *9*, 6–18. http://dx.doi.org/10.1111/ijs.12245.

Timmis, J., Hone, A., Stibor, T., & Clark, E. (2008). Theoretical advances in artificial immune systems. *Theoretical Computer Science*, *403*, 11–32. http://dx.doi.org/10.1016/j.tcs.2008.02.011.

Wardlaw, J. M., Seymour, J., Cairns, J., Keir, S., Lewis, S., & Sandercock, P. (2004). Immediate computed tomography scanning of acute stroke is cost-effective and improves quality of life. *Stroke*, *35*, 2477–2483. http://dx.doi.org/10.1161/01.str.0000143453.78005.44.

Zhao, H. (2007). A multi-objective genetic programming approach to developing Pareto optimal decision trees. *Decision Support Systems*, *43*, 809–826. http://dx.doi.org/10.1016/j.dss.2006.12.011.

Zhou, J., Li, X., & Mitri, H. S. (2015). Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. *Natural Hazards*, *79*, 291–316. http://dx.doi.org/10.1007/s11069-015-1842-3.