Journal of Biomedical Informatics 54 (2015) 305–314

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A novel neural-inspired learning algorithm with application to clinical risk prediction

Darwin Tay [a,b], Chueh Loo Poh [b,*], Richard I. Kitney [a]

[a] Department of Bioengineering, Imperial College London, UK
[b] Division of Bioengineering, Nanyang Technological University, Singapore

ARTICLE INFO

Article history:
Received 11 August 2014
Accepted 22 December 2014
Available online 6 January 2015

Keywords:
Cardiovascular disease
Classification
Clinical risk prediction
Neural-inspired algorithms

ABSTRACT

Clinical risk prediction – the estimation of the likelihood an individual is at risk of a disease – is a coveted and exigent clinical task, and a cornerstone to the recommendation of life saving management strategies. This is especially important for individuals at risk of cardiovascular disease (CVD) given the fact that it is the leading causes of death in many developed counties. To this end, we introduce a novel learning algorithm – a key factor that influences the performance of machine learning-based prediction models – and utilities it to develop CVD risk prediction tool. This novel neural-inspired algorithm, called the Artificial Neural Cell System for classification (ANCSc), is inspired by mechanisms that develop the brain and empowering it with capabilities such as information processing/storage and recall, decision making and initiating actions on external environment. Specifically, we exploit on 3 natural neural mechanisms responsible for developing and enriching the brain – namely neurogenesis, neuroplasticity via nurturing and apoptosis – when implementing ANCSc algorithm. Benchmark testing was conducted using the Honolulu Heart Program (HHP) dataset and results are juxtaposed with 2 other algorithms – i.e. Support Vector Machine (SVM) and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS). Empirical experiments indicate that ANCSc algorithm (statistically) outperforms both SVM and EDC-AIRS algorithms. Key clinical markers identified by ANCSc algorithm include risk factors related to diet/lifestyle, pulmonary function, personal/family/medical history, blood data, blood pressure, and electrocardiography. These clinical markers, in general, are also found to be clinically significant – providing a promising avenue for identifying potential cardiovascular risk factors to be evaluated in clinical trials.

1. Introduction

Cardiovascular disease (CVD) is an epidemic and major health concern in today's world. It is the leading causes of mortality in many developed countries, such as the United States (U.S.) and the United Kingdom (U.K.) [1,2]. To this end, an exigent clinical task is the ability to accurately predict whether an individual is likely to experience CVD in the near future so that appropriate and personalized preventive/life saving strategies can be recommended to the patient; an approach to reduce avoidable mortality. This prediction of health related patient outcomes has received increased recognition as an essential activity in clinical practice, research and assessment [3]. In response to this escalating demand for accurate, proactive and personalized prediction of disease risk, machine learning techniques have been recognized as a possible avenue to address this challenge. However, the development of versatile and reliable machine learning-based prediction models that allow clinicians to use in clinics/hospitals to instantly classify patients' risk remains a major medical screening conundrum [4]. One key factor that influences the performance of clinical risk prediction models is the robustness to learn and generalizability of the machine learning algorithm used. This dovetails with a phenomenon known as the selective superiority problem [5] where every learning algorithm has an inductive bias that would work reasonably well for some, but not all, datasets or application domains. Therefore, to alleviate the aforementioned challenges, we introduce a novel neural-inspired learning algorithm that has its own

* Corresponding author at: 70 Nanyang Drive, N1.3-B2-09, Singapore 637457, Singapore. Fax: +65 6791 1761.
E-mail addresses: darwintay@imperial.ac.uk (D. Tay), CLPoh@ntu.edu.sg (C.L. Poh), r.kitney@imperial.ac.uk (R.I. Kitney).

http://dx.doi.org/10.1016/j.jbi.2014.12.014
1532-0464/

unique inductive bias and use it to predict whether an individual is at risk of CVD.

The proposed machine learning algorithm – called the Artificial Neural Cell System for classification (ANCSc) – is inspired by the characteristics exhibited by 3 natural phenomena responsible for developing and enriching the brain function – namely (1) neurogenesis, (2) neuroplasticity as a result of the dynamic interplay between nature and nurture, and (3) apoptosis. These mechanisms (among others) enable human to learn, identify, differentiate and organize objects, patterns, sounds, concepts, etc. This model of neural operations has many features in common, generally in the field of machine learning, to the task of classification – the problem of identifying which category an observation belongs to, on the basis of a pre-specified set of data containing observations with known category membership. Hence, these neural processes – which to our knowledge have not been exploited for the development of machine learning algorithms – become an ideal candidate for the study and modeling of learning systems.

Neurogenesis, in neuroscience, is the process by which new neurons are generated in the nervous system from neural stem/progenitor cells [6]. The generated neurons are not stagnant throughout the life of a species and can be stimulated by behavioral and environmental factors [7]. This is vital and necessary for adapting the brain to any changing elements it encounters; refining the neural pathways and synapses essential for learning and adapting to changes, and circumvent any undesirable side effects. This process of molding and reshaping the brain in face of changes in behavior, environment and neural processes is often referred to as neuroplasticity [8]. Intentional exposure to new environments and (supervised/guided) inculcation of desirable information/behavior to a human (e.g. taught by an instructor) may trigger neuroplastic changes as well. This process, considered as nurturing, capitalizes on what the nature can provide (i.e. the individuals' innate qualities), enriches and leverages on the individuals' ability so that they can perform at their greatest potential.

Motivated by the profound significance of the aforementioned mechanisms in human learning process, and the ability to autonomously trim off non-essential cells during human development (commonly known as apoptosis), we implemented ANCSc algorithm. In a nutshell, this algorithm bio-mimics the mechanisms underlying the neuronal behavior associated with the process of learning and interaction with the external environment. It allows artificial neurons (i.e. candidate solutions) to (1) proliferate in the solution space (bio-mimicking neurogenesis), (2) progressively and independently refine and adapt to the (data) environment presented (bio-mimicking neuroplasticity as a result of nurturing), and (3) survive or undergo programmed cell death as part of an effort to construct a concise and efficacious classification model (bio-mimicking apoptosis). The utilization of these learning mechanisms is a novel contribution towards the development of neural-inspired learning algorithms, and in our opinion, would promote the development of robust classification models. Through this paper, we aim to suggest new approaches that might be of value to the construction of learning systems.

The predictive ability of ANCSc algorithm was evaluated by constructing CVD predictive models using the Honolulu Heart Program (HHP) dataset [9–11] – a prospective study of environmental and biological causes of CVD. The performance of ANCSc algorithm was juxtaposed against Support Vector Machine (SVM) [12–14] and Evolutionary Data-Conscious Artificial Immune Recognition System (EDC-AIRS) [15] algorithms in order to corroborate its robustness to learn and generalize. SVM and EDC-AIRS algorithms were selected in this assessment because they have been demonstrated to yield competitive performance when tested with a widely benchmarked heart disease dataset (i.e. Statlog Heart) [15]. Experimental results indicate that ANCSc algorithm achieve a reasonable classification performance and outperforms both SVM and EDC-AIRS algorithms in this clinical risk prediction task. Further, we have analyzed the key clinical markers – identified by the postulated risk prediction models – that are deemed to be associated with CVD; an important aspect in clinical prediction research.

The body of the paper is organized as follows. Section 2 provides a brief overview of neural processes that motivated the implementation of ANCSc algorithm. A detailed description of the proposed ANCSc algorithm is presented in Section 3. Materials and methods used in this study are delineated in Section 4. Performance of ANCSc, EDC-AIRS and SVM algorithms are offered in Section 5. Section 6 discusses the key results and properties associated with the algorithm. Finally, conclusions are drawn in Section 7.

2. Overview of neural processes

Neurons, a group of specialized impulse-conducting cells that process and transmit information through electrical and chemical signals, form the core components of the nervous system (e.g. the brain). The human brain contains on average 86.1 billion neurons [16], connected to each other to form neural networks. Communication among the neurons occurs via synapses – specialized connections between neurons that allow electrical and chemical signals to be transmitted. This interaction among neurons is the cellular basis for tasks like thinking and decision making. In particular, neurons are interconnected in smaller groups – called neuronal pools – defined on the basis of function (i.e. each neuronal pool is responsible for enabling a specific function to be carried out) [17].

New neurons are generated in the human brain from neural stem/progenitor cells – a process called neurogenesis. It is most active during prenatal development and declines sharply over the adolescence period [6]. Neurogenesis in the adult brain occurs primarily in two discrete areas – namely the dentate gyrus of the hippocampus and the subventricular zone, along the lateral ventricles. The number of new neurons added to an adult brain is dependent on the rate of cell generation and the probability of cell survival (i.e. generated cells might undergo programmed cell death after a period of time – a phenomenon known as apoptosis) [6]. As demonstrated in several studies, the rate at which neurogenesis occurs is modulated by several intrinsic and environmental stimuli. Intrinsic regulators include age [18], gender [19] and genetic factors [20] while environmental stimuli comprise of environmental enrichment [21], physical [22] and social [23] activities, stress [24], smell [25] and diet [26]. It is noteworthy that adult neurogenesis, in any cases, occurs (during most part of the life) at a very low rate [6,8]. Further, there is also growing evidence suggesting an association between adult hippocampal neurogenesis to several processes like neuro-inflammation, learning and memory. It has been demonstrated that neuro-inflammation inhibits neurogenesis in adult hippocampus [27] while increased hippocampal neurogenesis is potentially involved in ameliorated learning and memory [28–30]. Long-surviving neurons in the brain have been postulated to be more stable and preserve the encoding of the learned environment, whereas newly generated neurons are more plastic – which allows the brain to adapt itself to the new environment (i.e. occurrence of neuroplasticity as a consequence of learning) [6].

In the parlance of literature, neuroplasticity refers to the malleability of the brain – usually observable as changes in neuronal structure (e.g. changes in the position of the neurons) and connectivity, functional changes in the brain and neurogenesis. These changes typically occurs as a result of learning (e.g. taught/nurtured by an instructor), training (e.g. practicing to improve the ability to perform a task) and experience (e.g. exposure to certain event or environment); rendering the brain capable of adapting to environmental dynamics [8]. It is noteworthy that it has become increasingly evident that both neurogenesis and neuroplasticity occur in the human brain throughout life; instead of during prenatal development or juvenile period only [7,31].

Apoptosis, the process of controlled cell death, is an important feature that offers significant advantages during an organism's life-cycle. It promotes healthy (e.g. nervous system) development where defective apoptotic processes would be detrimental – leading to diseases like cancer (as a result of inadequate apoptosis) or atrophy (as a consequence of excesses apoptosis).

3. Artificial Neural Cell System for classification (ANCSc) algorithm

The details of ANCSc algorithm will be presented in this section. It is a supervised classification algorithm that bio-mimics how new neurons are populated, refined and maintained in the mammalian brain. Through this process, it aims to "educate" the ANCSc classification model to learn (in an incremental manner) the key patterns that underlie the training data.

To provide a comprehensive description of ANCSc algorithm, Section 3.1 describes the key terms and parameters vital for the understanding of the algorithm, while Section 3.2 provides a tour of the training routine associated with the algorithm.

3.1. Key concepts and parameters

This subsection describes the definitions for the key terms and parameters used in relation to the ANCSc algorithm.

3.1.1. Key terms

- *Affinity:* The Euclidean distance between two artificial neurons (feature vectors). In this implementation, this distance is between 0 and 1 (where 0 represents high affinity while 1 indicates low affinity).
- *Apoptosis:* The removal of artificial neurons from the artificial cognitive system that mimics the naturally occurring and genetically determined process of self-destruction of unwanted cells. It is a regulated process that offers the advantage of producing a parsimonious, yet accurate, artificial cognitive system for performing classification at the end of the training routine.
- *Artificial cognitive system:* A collection of representative artificial neurons (which evolve during the training process of ANCSc) capable of describing the training data presented. Given that communication among neurons (e.g. within a neuronal pool) enables human to think or recognize objects, we propose the use of KNN algorithm [32] to perform classification (at the end of each training cycle) due to their metaphorical similarity – i.e. both defines a pool of elements for conducting a task of interest.
- *Artificial neuron:* In neuroscience, neuroplastic changes (i.e. slight changes in the position of the neurons) have been proposed as the consequence of learning and memory formation in species like human [31]. To bio-mimic this phenomenon, we metaphorically relate feature vectors with artificial neurons developed in ANCSc algorithm which contribute to the formation of the artificial cognitive system. This metaphoric establishment was made as both feature vector and artificial neuron – in a sense – represent an entity that provides some form of information. Synaptic connections between neurons are not considered in order to simplify the construction of the learning model. Artificial neurons can be added, modified or removed from the postulated artificial cognitive system during ANCSc training cycle.
- *Artificial neuronal pool:* A group of proximal artificial neurons that describe a specific pattern determined within the data problem presented. Its formation is regulated by the associated classification performance and defined on the basis of cell proliferation, adaptation and survival. The size of the artificial neuronal pool determines the number of artificial neurons (i.e. k value) to be used for classification by KNN.
- *Class:* The category assigned to a given feature vector. For binary classification problems, each feature vector is assigned to one of the 2 pre-defined categories.
- *Feature vector:* An n-dimensional vector of categorical/numerical features that describe the characteristics of an object/observation.
- *Neuroplasticity:* The adaptation of artificial neurons (i.e. modification of the feature vectors) in the artificial cognitive system triggered by the process of learning and generalization. This procedure aims to promote the generation of highly representative artificial neurons capable of describing the given data environment.
- *Testing data:* A collection of data items – that represent observations/measurements of a subject of interests – used to estimate the performance of the classification model trained with the training data. It is a distinct set of data that is used in an iterative process to evaluate and improve the performance of the trained model.
- *Training data:* A collection of data, similar to the testing data, used to develop a classification model. Training data are commonly used in various areas of information science for the discovery of predictive relationship between the feature vector and the class. In this particular context, they serve as the data environment that promotes proliferation, adaptation and survival of neural cells.

3.1.2. Key parameters

- *Learning plateau threshold (LPT):* A termination criterion which defines the number of learning cycles that the ANCSc algorithm would iterate for before termination. Improvement in classification accuracy (during a learning cycle) would reset this (integer) parameter.
- *Neural density (ND):* This value, which ranges between 0 and 1, aims to spread artificial neurons with high affinity. This offers the potential advantage of generating a set of representative artificial neurons.
- *Neurogenic space (NS):* A parameter, used during the neurogenesis phase, which determines the size of the region at which artificial neurons would develop in the fetal artificial cognitive system. The value of this parameter ranges between 0 and 1.
- *Neurogenic rate (NR):* The rate at which artificial neurons are generated during neurogenesis phase. The value of this parameter ranges between 0 and 1.

Algorithm 1: Overview of ANCSc Algorithm

Input: **D** (training data)
T (testing data)
Output: **O** (class label prediction)

Initialization
Step 1: Set $t = 1$. Normalize **D** and **T** to the range [0,1].

Neurogenesis Phase
Step 2: Populate a pool of artificial neurons $\mathbf{P_1}$ to form the initial cognitive system **C**.
$\mathbf{P_1}$ is generated by searching for representative data items in **D**, $\mathbf{P_1} \subseteq \mathbf{D}$.
Step 3: A_1 = accuracy of classification model $\mathbf{P_1}$ when evaluated with **D**.
Set $t = t + 1$.

Neuroplasticity via nurturing Phase
Step 4: Identify $p_i \in \mathbf{P_t}$ that resulted in largest number of misclassification.
If class label of p_i contradicts with NPS artificial neurons at its neighborhood, removed p_i from $\mathbf{P_t}$.
Otherwise, generate centroid artificial neuron p_j among the NPS artificial neurons (with same class label).
Add p_j to $\mathbf{P_t}$.
Step 5: A_t = accuracy of classification model $\mathbf{P_t}$ when evaluated with **D**.
If A_t is greater or equals to best accuracy achieved thus far, update C to $\mathbf{P_t}$. Otherwise, discard $\mathbf{P_t}$.
Set $t = t + 1$.
Step 6: Scatter closely clustered $p_i \in \mathbf{C}$. Resulting model forms $\mathbf{P_t}$.
Step 7: A_t = accuracy of classification model $\mathbf{P_t}$ when evaluated with **D**.
If A_t is greater or equals to best accuracy achieved thus far, update **C** to $\mathbf{P_t}$. Otherwise, discard $\mathbf{P_t}$.
Set $t = t + 1$.
Step 8: If termination criteria are satisfied, proceed to Step 9. Otherwise, go to Step 4.

Apoptosis Phase
Step 9: If eradication of $p_i \in \mathbf{C}$ does not deteriorate classification performance when evaluated with **D**, remove p_i from **C**.
Otherwise, keep p_i.

Evaluation
Step 10: Evaluate performance of classification model **C** on **T**. Generated class labels of **T** are assigned to **O**.

- *Neuronal pool size (NPS):* The number of artificial neurons that should be used to determine the classification of a given test data item. This integer value is used as the k value parameter required in KNN algorithm.
- *Neuroplastic coefficient (NPC):* This parameter specifies the degree to which the generated artificial neurons migrate in the artificial cognitive system. This offers an opportunity for the artificial neurons to generalize and circumvent situation like overfitting. The value of this parameter ranges between 0 and 1.
- *Neuroplastic threshold (NPT):* The number of cycles allowed for ANCSc algorithm to generalize the artificial cognitive system before termination. This (integer) parameter resets if there is an improvement to the classification accuracy.

3.2. Training routine of ANCSc

This subsection provides a detailed description of the key routines, methods and equations proposed in ANCSc algorithm. The canonical flow of the algorithm is illustrated in Fig. 1 while Algorithm 1 provides the corresponding pseudocode. In this implementation, all data are normalized using the equation below so that the Euclidean distance between any 2 feature vectors is between 0 and 1.

$$\bar{f}_i = \frac{f_i - f_i^{min}}{f_i^{max} - f_i^{min}}$$

where $\bar{f}_i$ represents the normalized value for feature i, f_i refers to the measured value in feature i, and f_i^{min} (f_i^{max}) denotes the minimum (maximum) value in feature i.

The ANCSc algorithm consists of 3 key development phases – namely neurogenesis, neuroplasticity via nurturing and apoptosis phases. All steps proposed in ANCSc algorithm to develop the classification methodology – aiming to construct a reduced set of representative artificial neurons for classification – are explained independently below.

3.2.1. Neurogenesis phase

The primary objective of this phase is to generate a reduced set of representative artificial neurons (or data items) from the training dataset. This establishes the fetal artificial cognitive system that would be refined and enhanced in the later phases. It begins the process of populating new artificial neurons by requiring the specification of 2 parameters – namely neurogenic space (NS) and neurogenic rate (NR). It proceeds by searching for the region (radius defined by NS) that is most populated with data items (within the training dataset). Upon finding it, a uniformly distributed subset of data items from that region is selected. This selection technique of uniformly distributed data item is similar to the Kennard-Stone (KS) algorithm [33]. However, unlike KS algorithm, we proposed that the number of data items (numNeurons) to be selected be dynamically determined by the following equation:

$$\text{numNeurons} = \|\text{NS}\| * \text{NR} \quad (1)$$

where $\|\text{NS}\|$ is the number of data items found within the region defined by NS, and NR is a user-defined probability parameter that determines the proportion of data items that would be selected as artificial neurons for the development of the fetal artificial cognitive system. This NR parameter is tantamount to the intrinsic and environmental stimuli (described in Section 2) that regulate the rate of neurogenesis in human brain.

Subsequently, all data items within the previously defined region are removed and the aforementioned process repeats to create the fetal artificial cognitive system. At the end of this phase, a set of representative artificial neurons would form the artificial cognitive system. An illustration of this process is given in Fig. 2. Finally, the classification performance of the constructed fetal artificial cognitive system is evaluated (using KNN – see evaluation phase below) with the initial training data.

3.2.2. Neuroplasticity via nurturing phase

Neuroplasticity, as a consequence of nurturing, plays a significant role in promoting the construction of a robust classification model that promises enhanced performance over one that regurgitates memorized patterns learned during the neurogenesis phase. This phase was inspired by observation of how neuronal structures change (i.e. change in the position of the neurons) in tandem with healthy brain development, learning and memory formation. Changes in connectivity among the neurons (i.e. synaptic connec-

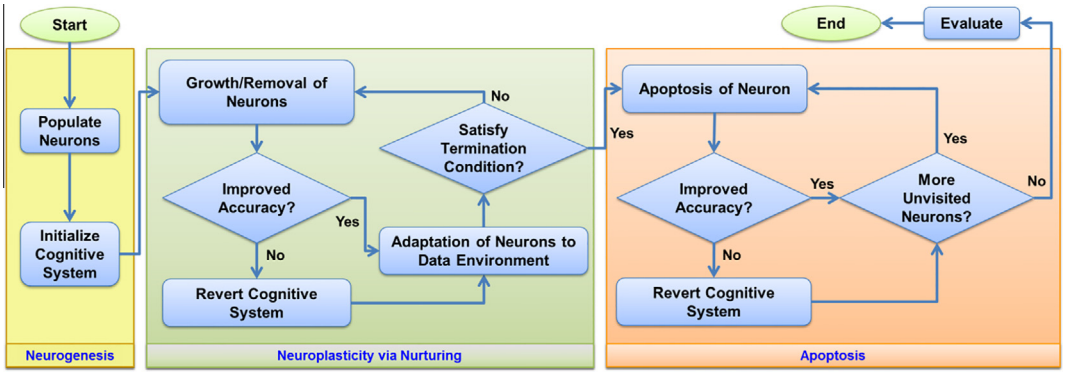

Fig. 1. Canonical Flow of ANCSc Algorithm. The ANCSc algorithm consists of 3 key phases: neurogenesis, neuroplasticity via nurturing and apoptosis. During the neurogenesis phase, the initial set of artificial neurons is created. These artificial neurons then evolve (through cell proliferation, adaptation and survival) in the subsequent 2 phase, generating a set of representative artificial neurons capable of describing the underlying patterns of the data presented.

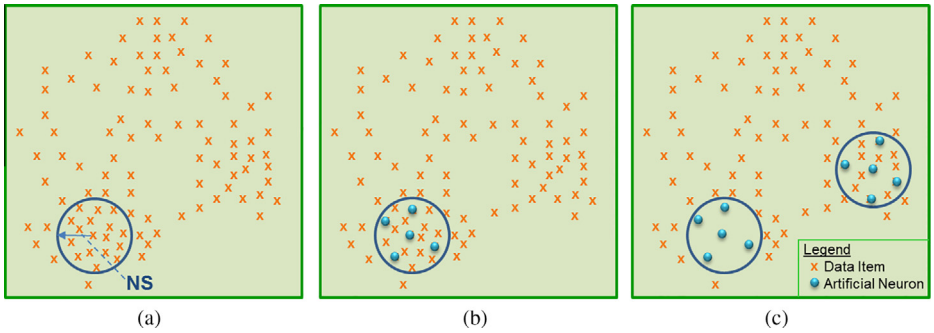

Fig. 2. Graphical Illustration of Neurogenesis Phase. (a) From the initial training data, the region (defined by NS) that is most populated with data items is identified. (b) From the identified region, artificial neurons are created using KS algorithm (in this example, numNeurons = 5). (c) Upon selection, data items within that region are removed. This process is repeated.

tion) are not considered in order to postulate a simple and efficient learning model.

The primary objective of this phase is to (1) grow artificial neurons at locations that would contribute to better classification performance, (2) remove existing artificial neurons that exacerbate the classification performance, and (3) adapt engendered artificial neurons to the input data environment to promote better classification performance. This phase begins by identifying the artificial neuron that resulted in the largest number of misclassifications. If the class of this artificial neuron (for example, it is class 1) contradicts with most of the other artificial neurons (i.e. they are of class 0) at its proximity, it is removed from the artificial cognitive system. Otherwise, a new artificial neuron with the same class (as those at its proximity) is generated at the centroid of those artificial neurons, and added to the artificial cognitive system. A condition that must be satisfied for this addition is that the class of the artificial neuron to be added must belong to the minority data class. This is to encourage a balanced number of artificial neurons (i.e. similar number of artificial neurons with class 0 and 1 labels) to thrive in the developed artificial cognitive system. We hypothesize that this would potentially deliver a solution that could generalize better.

An aging mechanism is implemented, "aging" the newly added artificial neurons. This is to allow "younger" artificial neurons to have an opportunity to be involved in the learning process (i.e. mimicking the concept – in neuroscience – that younger neurons in human brain are more plastic [6]). If this phase resulted in an artificial cognitive system that shows improved performance, it would be kept for future development. Otherwise, it would be discarded.

To better adapt the engendered artificial neurons to the input data environment, closely clustered artificial neurons are scattered apart if it does not compromise the resulting classification performance. This adaptation step begins by searching for the artificial neuron (dNeuron) – within a region whose radius is defined by the neural density (ND) parameter – that is most populated with other artificial neurons. Upon finding this artificial neuron, the closest artificial neuron (cNeuron) affiliated to it (i.e. with highest affinity to dNeuron) is modified so that they are more distributed apart. The degree of spread is determined by the neuroplastic coefficient (NPC) parameter and defined with the following equation:

$$\mathrm{cNeuron}_i = \mathrm{cNeuron}_i + \mathrm{NPC} * (\mathrm{cNeuron}_i - \mathrm{dNeuron}_i) \quad (2)$$

where $\mathrm{cNeuron}_i$ and $\mathrm{dNeuron}_i$ are the ith attribute of cNeuron and dNeuron, respectively. Through modicum adjustment of the artificial neurons in the artificial cognitive system, we aim to promote the construction of a more diverse set of representative artificial neurons; sequella for mitigating the risk of overfitting. Similar to the previous step, a (separate) aging mechanism is implemented. This is to ensure that different artificial neurons that are densely clustered together have a chance to deviate and generalize. Likewise, if this newly developed artificial cognitive system constructed in this phase demonstrates ameliorated performance, it would be saved. Otherwise, it would be removed from further consideration.

3.2.3. Termination of neuroplasticity via nurturing phase

The stopping criterion for neuroplasticity via nurturing phase is reached if there is no improvement in the classification performance after LTP (a user-defined value) iterations or the same classification performance is achieved consecutively after NPT (a user-defined value) iterations. Otherwise, neuroplasticity via nurturing phase repeats, inculcating the artificial cognitive system with key patterns that underlie the training data.

3.2.4. Apoptosis phase

Naturally occurring apoptotic processes are very important in healthy development of organism. For example, apoptosis occurs between the fingers and toes of a human during the embryonic stage (which initially appears like duck's webbed feet), giving them the freedom to maneuver individually. Metaphorically, this feature may offer ANCSc algorithm the ability to trim away redundant neurons, delivering a concise and efficacious classification model.

This process of removing redundant artificial neurons is carried out upon termination of the neuroplasticity via nurturing phase. It aims to eradicate redundant artificial neurons that do not contribute to the construction of an efficacious and concise artificial cognitive system, but instead exacerbate the overall performance. The determination of which artificial neuron to apoptosize is governed by 2 questions. First, whether the Euclidean distance of the artificial neuron under examination and another artificial neuron in the postulated artificial cognitive system is smaller than the product of NS and NR? Second, whether removal of the neuron under examination would contribute to an improved artificial cognitive system? If the answer is 'yes' to both these questions then that artificial neuron is removed. Otherwise, it remains in the artificial cognitive system. This process would lead to a reduced set of representative artificial neurons that is used for classification (by KNN).

3.2.5. Evaluation phase

At the end of each training phase described above, KNN algorithm is used to predict the class value of unseen data items. It works by determining the k (defined by NPS parameter) artificial neurons closest to an unseen data item and adopting a majority vote scheme to suggest the class value. This is similar to activating the neurons in the corresponding neuronal pool – in human brain – when one recall an event or object. Accuracy is used as the measurement metric to determine the performance of the artificial cognitive system constructed.

3.3. Data class-specific ANCSc parameters

Neurogenesis has been shown to occur in 2 distinct areas of the brain, namely the dentate gyrus of the hippocampus and the anterior part of the subventricular zone. Each area harbors a population of neural stem/progenitor cells that divide and proliferate independently. Moreover, each area is responsible for different function – the hippocampus is claimed to be the putative area for information storage while the subventricular zone is associated with the development of the olfactory bulb. The occurrence of autonomous neurogenesis in areas of the brain responsible for different function suggests that decentralized development may be the strategy that nature adopts.

These observations underscore the importance of locality and task specific regulation. One approach to bio-mimic this computationally is to independently analyze and model each data class (i.e. having an independent parameter set for each data class). This, when applied to an immune-inspired algorithm [15], demonstrated improved performance. Therefore a similar technique was implemented in ANCSc algorithm. The parameters that orchestrate the proliferation, adaptation and survival of the neural cells include NS, NR, ND and NPC. Hence, these parameters were duplicated and optimized independently for each data class. Genetic algorithm (GA) [34] – a search heuristic inspired by natural evolution – was employed to optimize these parameters.

Fig. 3 illustrates the canonical flow of the strategy used to solve binary classification problems. Two sets of parameters were initialized and optimized in parallel – namely a common set (i.e. a single set of parameters used to model both data classes) consisting of 7 parameters and an independent set comprising 11 parameters. Upon termination of the optimization process carried out by GA,

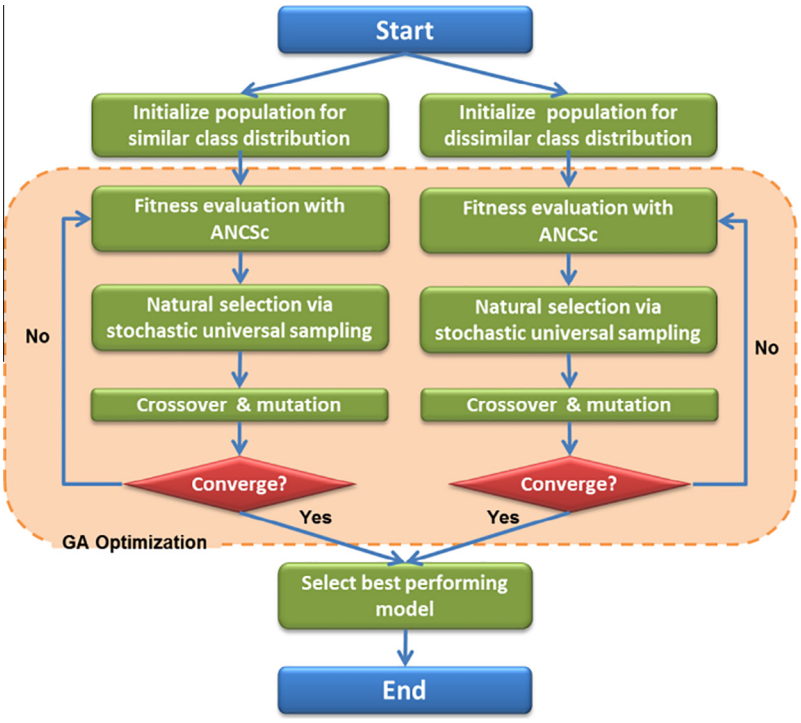

Fig. 3. Proposed strategy for optimization of parameter set for binary class classification problems. Optimization using GA was conducted on ANCSc algorithm that has a common and independent set of parameters. The best performing model computed from these two experiments was selected as the resulting classification model.

the best performing classification model obtained is used to predict future unseen data items.

4. Materials and methods

4.1. Performance evaluation of ANCSc algorithm

The assessment of ANCSc algorithm was conducted by employing it (together with EDC-AIRS and SVM) as a technique to develop predictive models for CVD risk prediction. Examination data of subjects – from the HHP dataset – age between 46 and 55 were utilized by the respective algorithms to perform 2-year CVD risk prediction (i.e. estimation of whether an individual is likely to experience CVD within their next 2 years). In order to fortify ceteris paribus experimental design among the algorithms used in this evaluation (i.e. ANCSc, EDC-AIRS and SVM algorithms), 3 consecutive optimization steps – namely model selection, feature selection and feature construction – were performed during the training phase for each algorithm. This methodology is illustrated in Fig. 4. Genetic algorithm (GA), unless otherwise stated, was utilized in this study to optimize the parameters. The parameters of GA were determined experimentally to work well for this clinical prediction problem and the details are as follow: population size: 100; maximum generation: 100; fitness function: accuracy obtained from classifier; natural selection: stochastic universal sampling; crossover type: discrete recombination; crossover probability: 0.8; mutation rate: $1/P$, where P is the number of parameters.

The first step, model selection, aims to select the most potent prediction model for this clinical prediction problem. GA was used to optimize the parameters of ANCSc and EDC-AIRS algorithms while uniform design (UD) [35] method was used to determine the cost and gamma parameters required by SVM kernel (i.e. radial basis function). UD approach was adopted for SVM as it has been shown to produce promising results, and at the same time alleviate the computational loads associated with the search for the optimal cost-gamma pair [36]. The parameter details for SVM are kernel function: radial basis function (RBF); cost: $[2^{-5},2^{13}]$; gamma: $[2^{-15},2^{3}]$; for EDC-AIRS are seed: 1; clonal rate: 10; hyper-mutation rate: 2; stimulation threshold: 0.9; initial memory pool size: [0,200]; KNN value: [1,15]; affinity threshold scalar: [0,1]; total resource: [150,300]; $\text{Radius}_{\text{density}} = [0,3]$; $\text{Radius}_{\text{max}} = [0,3]$; and for ANCSc are seed: 1; NPS: [1,15]; LPT: [0,10]; NPT: [0,100]; NS: [0,0.5]; NR: [0,1]; ND = [0,0.5]; NPC = [0,0.5].

The feature selection step aims to identify informative clinical markers that would contribute to the development of an accurate and parsimonious prediction model. GA was used to carry out this task and the set of clinical markers identified was passed to the feature construction step. Feature construction is the process of discovering unknown relationship between features and augments the existing feature space with new composite features [37]. Cartesian Genetic Programming (CGP) [38], a highly effective form of genetic programming that has demonstrated success in garnering parsimony (i.e. more human-comprehensible) [39], was employed to construct new features. The parameter details for CGP are #inputs: feature dimension; #output: 1; #rows: 1; #columns: 10; arity: 2; levels back: 10; functions: {addition, subtraction, multiplication, division}.

Finally, the respective trained models were assessed using the validation dataset (i.e. new data sample not used to train the model). This validation phase is very important because reporting results based on the training dataset may be overly optimistic and prone to over-fitting.

The performance yielded by ANCSc algorithm was (statistically) compared to those achieved by SVM and EDC-AIRS algorithms. We have chosen McNemar's test to perform this statistical analysis as it has been demonstrated to have low type 1 error [40]. To perform this test, the algorithms were first trained with the training data and tested with the validation data. The predicted outcome for each data item in the validation dataset was recorded and used to construct the contingency table shown in Fig. 5. Referring to the figure, if the sum of 'b' and 'c' is greater than 25, chi-square test with 1 degree of freedom is used to perform the McNemar's test. Otherwise, to provide a better estimation of the small sample (i.e. $b + c \leqslant 25$), binomial distribution is used for (exact) McNemar's test. The 2 algorithms are considered to be statistically different if the p-value computed with McNemar's test is smaller than 0.05.

4.2. Datasets

The Honolulu Heart Program (HHP) [9–11], initiated in 1965 by the National Heart, Lung and Blood Institute (NHLBI) as a prospective study of environmental and biological causes of CVD among Japanese Americans living in Hawaii, was analyzed in this study. Subjects, followed for the development of CVD, collected between 1965 and 1968 (exam 1) were utilized as the baseline data. It consists of 8006 Japanese-American men living on the island of Oahu, Hawaii. At the time of study, participants received a comprehensive examination (e.g. physical measures, medical history/lifestyle, dietary, anthropometric measures, etc.). This resulted in 412 clinical features being collected. Out of these participants, only individuals (a total of 7383) who were free from angina pectoris (AP), coronary insufficiency (CI) and myocardial infarction (MI) were considered.

Cardiovascular events that occurred after the baseline examination (i.e. exam 1) were monitored through surveillance of hospital discharges, subsequent examinations, death certificates and

		Algorithm 2	
		Misclassification	Correct Classification
Algorithm 1	Misclassification	a	b
	Correct Classification	c	d

Fig. 5. Contingency Table for McNemar's Test. 'a' indicates the number of data items misclassified by both algorithm 1 and algorithm 2; 'b' represents the number of data items misclassified by algorithm 2 but correctly classified by algorithm 1; 'c' denotes the number of data items misclassified by algorithm 2 but correctly classified by algorithm 1; 'd' dictates the number of data items correctly classified by both algorithm 1 and algorithm 2.

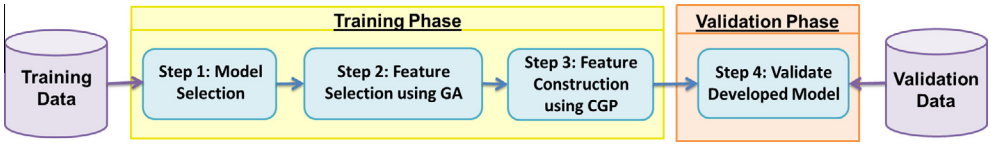

Fig. 4. Methodology employed for the development of clinical risk prediction model (using ANCSc, EDC-AIRS and SVM algorithms). The training phase, which uses the training data, is responsible for developing the prediction models. It consists of 3 distinct steps which include model selection, feature selection and feature construction. The developed prediction model is then validated for generalizability, during the validation phase, using the validation data.

autopsy records. A total of 392 individuals were found to experience cardiovascular diseases between exam 1 and exam 2 (which occurred between 1968 and 1970). Cardiovascular diseases, in this study, include AP, CI, MI, transient ischemic attack (TIA), stroke and congestive heart failure (CHF). To establish a 2-year risk prediction model, participants' record was matched between exam 1 and exam 2 (i.e. follow-up examination of participants not conducted in exam 2 were removed). Finally, to mitigate class imbalance data problem (i.e. the tendency of the algorithm overwhelmed by the major class and ignores the minor one) [41,42], a balanced number of cases and controls were randomly selected. In addition, uninformative features (i.e. features with constant value for all participants) were removed, resulting in a total of 370 clinical features and 172 instances.

In order to perform an accurate assessment of the performance of each algorithm, 70% of the baseline data (120 instances) was used to develop/train the model (commonly referred to as the training instances) while the remaining (common known as the validation instances – 52 instances) was used to validate the developed model.

5. Experimental results

To evaluate the performance of ANCSc algorithm, we conducted experiments related to CVD risk prediction using the HHP dataset. Each algorithm was executed 3 times and consistent results were achieved. During the training phase, model selection was conducted first and results – measured using performance metrics like sensitivity (SN), specificity (SP) and balanced accuracy (BA) (i.e. average between sensitivity and specificity) – were compared against those achieved by SVM and EDC-AIRS algorithms (see Fig. 6a). Empirical analysis indicates that ANCSc algorithm (SN: 72.1%; SP: 74.4%; BA: 73.3%) outperforms EDC-AIRS (SN: 60.5%; SP: 70.9%; BA: 65.7%) and SVM (SN: 40.7%; SP: 60.5%; BA: 50.6%) algorithms for all examined performance metrics. Next, feature selection was conducted to identify informative clinical markers. A total of 179, 174 and 168 features were identified to be informative to ANCSc, EDC-AIRS and SVM algorithms respectively. As illustrated in Fig. 6b, prediction models developed using ANCSc (SN: 75.6%; SP: 80.2%; BA: 77.9%) and EDC-AIRS (SN: 72.1%; SP: 83.7%; BA: 77.9%) algorithms in general performed comparably but are more competitive than SVM algorithm (SN: 57.0%; SP: 60.5%; BA: 58.7%). Finally, feature construction was conducted to generate informative features that would enhance the performance of the predictive model constructed. A total of 0, 1 and 8 features were created for ANCSc, EDC-AIRS and SVM algorithms respectively. In this case, EDC-AIRS algorithm (SN: 70.9%; SP: 89.5%; BA: 80.2%) performs slightly better than ANCSc algorithm (SN: 75.6%; SP: 80.2%; BA: 77.9%) while both ANCSc and EDC-AIRS algorithms outperform SVM algorithm (SN: 61.6%; SP: 69.8%; BA: 65.7%) (see Fig. 6c). A summary of these experimental results is provided in Table 1.

To assess the generalizability of the respective trained models, evaluation was conducted using the validation dataset. Results, as shown in Fig. 7, suggest that ANCSc algorithm (SN: 61.6%; SP: 86.1%; BA: 73.6%) outperforms both EDC-AIRS (SN: 66.7%; SP: 50.0%; BA: 58.3%) and SVM (SN: 47.2%; SP: 66.7%; BA: 56.9%) algorithms. To corroborate this observation, McNemar's test was conducted. Statistical findings indicate that ANCSc algorithm outperforms both EDC-AIRS (p-value: 0.022) and SVM (p-value: 0.019) algorithms. A summary of these results is provided in Table 2.

6. Discussion

We have developed a novel algorithm called ANCSc and applied it to perform CVD risk prediction. ANCSc algorithm is a supervised classification algorithm inspired by the importance and robustness of several neural mechanisms (i.e. neurogenesis, neuroplasticity, nurturing and apoptosis) that occur during the development of the brain. These mechanisms have motivated us to capitalize on an ensemble of learning/optimization techniques (like clustering, under-sampling, evolutionary algorithm and instance-based learning) that – combined in accordance to the neural activities associated with brain development and learning – provides a good classification approach to the field of machine learning.

The first step of ANCSc algorithm (i.e. neurogenesis phase) generates the fetal artificial cognitive system by developing an initial reduced set of representative artificial neurons. This is an impor-

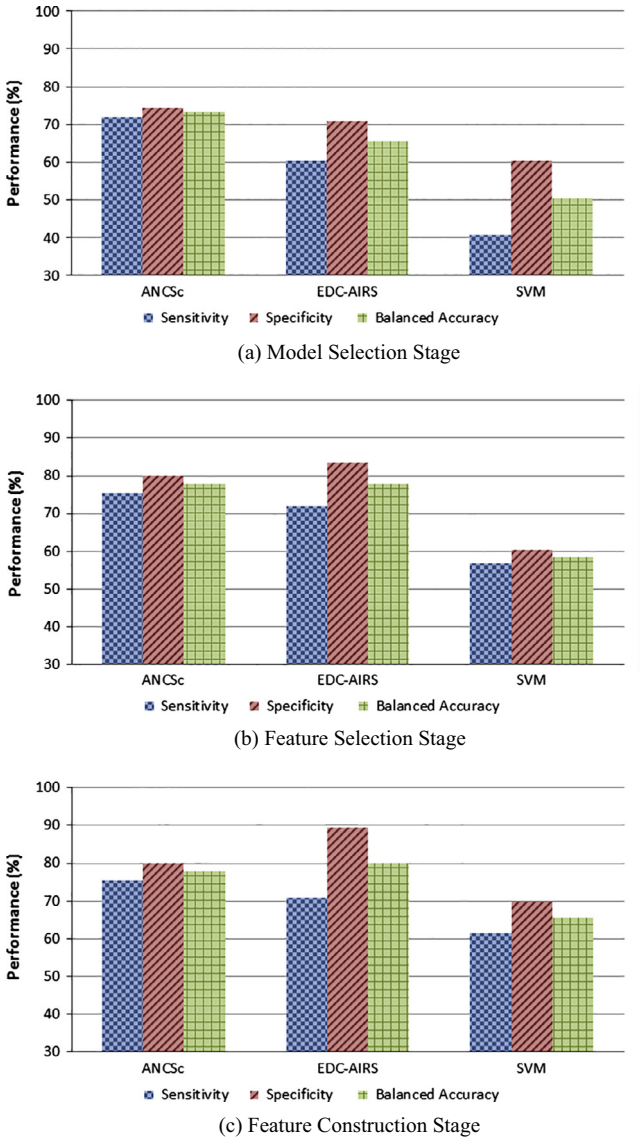

Fig. 6. Progressive performance of ANCSc, EDC-AIRS and SVM algorithms at different optimization steps. During the model selection stage, the parameters associated with each algorithm were optimized independently. The complete dataset is used in this stage and the performance obtained from 10-fold cross-validation is presented in (a). During the feature selection stage, the respective algorithm used the optimized parameters' value computed (during the model selection stage) to identify predictive features (using GA as the feature selection technique). The performance of the most predictive subset of features obtained is shown in (b). The respective reduced set of features obtained was subsequently used in the feature construction stage to generate new predictive features (using CGP as the feature construction technique). Classification performance of each algorithm using this newly discovered set of features is illustrated in (c).

Table 1
Performance summary of ANCSc, EDC-AIRS and SVM algorithms (training phase).

Stage	Algorithm	#Features	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)
Model selection	ANCSc	370	0.721	0.744	0.733
	EDC-AIRS	370	0.605	0.709	0.657
	SVM	370	0.407	0.605	0.506
Feature selection	ANCSc	179	0.756	0.802	0.779
	EDC-AIRS	174	0.721	0.837	0.779
	SVM	168	0.570	0.605	0.587
Feature construction	ANCSc	179	0.756	0.802	0.779
	EDC-AIRS	175	0.709	0.895	0.802
	SVM	176	0.616	0.698	0.657

The performance measurements, obtained after different optimization stage combinations during the training phase (using 10-fold cross validation), are shown. The number of features used by each algorithm during different optimization stage combinations is presented in the table as well.

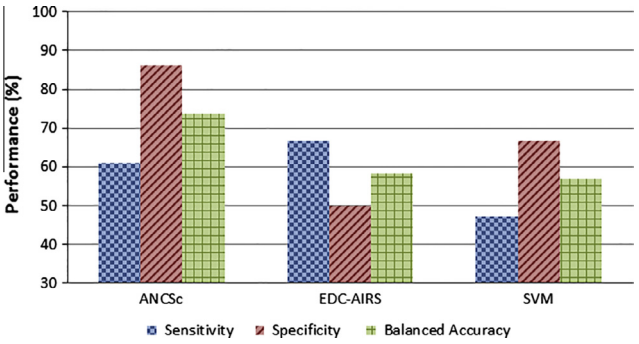

Fig. 7. Performance of ANCSc, EDC-AIRS and SVM algorithms during validation phase. These performance measurements were obtained by evaluating each trained model (obtained after all optimization stages – i.e. model selection, feature selection and feature construction – have been performed) using the validation dataset.

tant process as it alleviates the complexity and computational loads required to tune and "nurture" the postulated set of artificial neurons carried out during the 'neuroplasticity via nurturing' phase. In this regard, the 'neuroplasticity via nurturing' phase optimizes the artificial neurons in the fetal artificial cognitive system by evolving them during each learning cycle through the growth of new artificial neurons, performance of niche refinement to existing ones and/or eradication of existing artificial neurons that hinder the inculcation process. Through this repeated learning process, the aim is to "educate" the artificial cognitive system with key patterns found within the training data; enabling the artificial neurons to develop further and collectively realize their full potential. Termination of this learning algorithm proceeds with the removal of redundant artificial neurons (i.e. apoptosis phase) that potentially exacerbate the resulting classification performance. This is an important phase as it attempts to alleviate over-fitting, and fortify an effective and parsimonious prediction model.

To assess the robustness of ANCSc algorithm to learn and generalize, it was used to develop prediction model that aims to predict the risk of an individual experiencing CVD in the near future – a highly coveted but elusive clinical task. Prediction performance was compared with EDC-AIRS and SVM algorithms. Broadly, both feature selection and feature construction improve the performance of the prediction models during training phase. Specifically with feature selection, ANCSc, EDC-AIRS and SVM algorithms improve their cross-validation balanced accuracy by 6.28%, 18.6% and 16.0% respectively while reducing the number of features needed to develop the prediction model by 51.6%, 53.0% and 54.6% respectively. This reduction in the number of features required is highly desirable as each clinical feature is often associated with a financial cost, measurement time and/or risk for obtaining it. Feature construction is yet another useful technique for the development of clinical prediction model as it capitalizes on features that are already obtained and generates new ones that are capable of ameliorating the performance of the prediction models. This inevitably saves cost and time while eradicating extra burdens on the patients.

Referring to the experimental results, SVM performs poorly on the HHP dataset during both the training and validation phases. We believe that this is because of the presence of exceptional cases in medical data that resulted in the poor performance of SVM. ANCSc algorithm, on other hand, performs reasonably well during both the training and validation phases. In particular, ANCSc algorithm – in terms of balanced accuracy – is 20.8% and 22.7% better than EDC-AIRS and SVM algorithms when evaluated using the validation dataset. We hypothesize that the success of ANCSc algorithm is because of its ability to handle exceptional cases and optimize the neuronal cultures based on evolution, cooperation and altruism proposed by the algorithm. Further, since ANCSc algorithm uses KNN to perform prediction, it offers the advantage of providing reasoning for the decision it has made for the new cases the prediction model is inquired with – a highly desirable attribute in medicine – by presenting (to the clinicians) representative cases (from the developed cognitive system) that are most similar to the new cases that need to be explained [43].

Key clinical markers identified by ANCSc algorithm include risk factors related to diet/lifestyle, pulmonary function, personal/family/medical history, blood data, blood pressure, and electrocardiography. All these clinical markers in general – except for personal history (e.g. age left parent's home, wife present job, number of older brothers/younger sisters, etc) to our knowledge – are also identified as clinically significant in the literature [44–48]. We believe that personal history, a currently understudied factor, could be viewed as an intricate element that contributes to the

Table 2
Performance summary of ANCSc, EDC-AIRS and SVM algorithms (validation phase).

Algorithm	#Features	Sensitivity (%)	Specificity (%)	Balanced accuracy (%)	McNemar's test[a] (*p*-value)
ANCSc	179	0.611	0.861	0.736	–
EDC-AIRS	175	0.667	0.500	0.583	**0.022**
SVM	176	0.472	0.667	0.569	**0.019**

[a] McNemar's test was conducted between ANCSc algorithm and the algorithm the *p*-value is associated with in the table.

stress level individuals are experiencing as part of their lives. This chronic stress factor, on other hand, is a risk factor for CVD [49]. The observation that these statistically significant clinical markers can also be clinically significant provides a promising avenue for identifying potential cardiovascular risk factors to be evaluated in clinical trials.

One limitation of this work is the use of a single clinical risk prediction task to evaluate ANCSc algorithm. This limits our power to conclusively state the potential of ANCSc algorithm. However, it does provide some insights into the robustness of the algorithm. As part of our future work, we aim to apply ANCSc algorithm to solve other challenging tasks.

7. Conclusions

We have presented a novel supervised learning algorithm inspired by natural phenomena related to neurogenesis, neuroplasticity, nurturing and apoptosis. Leveraging on the fetal artificial cognitive system developed from the input data environment, ANCSc algorithm "nurture" it in an attempt to unleash its greatest potential. Application of ANCSc algorithm to clinical risk prediction has been carried out with promising results.

The learning approach postulated by ANCSc algorithm, in our opinion, has potential for learning profound data structures and producing a concise model capable of descripting the problem. Additionally, it offers a novel learning methodology in which classification problems can be solved by approaching them from a different perspective.

Acknowledgments

This work was supported by Nanyang Technological University-Imperial College London Joint PhD Scholarship, The Engineering and Physical Science Research Council (EPSRC) and the Ministry of Education (Singapore).

References

[1] Go AS, Mozaffarian D, Roger VL, Benjamin EJ, Berry JD, Borden WB, et al. Heart disease and stroke statistics-2013 update: a report from the American Heart Association. Circulation 2013;127:6–245.
[2] British Heart Foundation Statistics Database. Coronary Heart Disease; 2010. Internet: <http://www.bhf.org.uk/publications/view-publication.aspx?ps=1001546 [08.08.13].
[3] Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, Why, and How? BMJ 2009;338(b375).
[4] Tay D, Poh CL, Van Reeth E, Kitney R. The effect of sample age and prediction resolution on myocardial infarction risk prediction. IEEE J Biomed Health Inform 2014.
[5] Brodley C. Addressing the selective superiority problem: automatic algorithm/model class selection. In: Proc. 10th machine learning conf.; 1993.
[6] Wiskott L, Rasch MJ, Kempermann G. A functional hypothesis for adult hippocampal neurogenesis: avoidance of catastrophic interference in the dentate gyrus. Hippocampus 2006;16(3):329–43.
[7] Lillard AS, Erisir A. Old dogs learning new tricks: neuroplasticity beyond the juvenile period. Dev Rev 2011;31(4):207–39.
[8] Taupin P. Adult neurogenesis and neural stem cells in mammals. Nova Publishers; 2006. p. 208.
[9] Syme S, Marmot M, Kagan A, Kato H, Rhoads G. Epidemiologic studies of coronary heart disease and stroke in Japanese Men Living in Japan, Hawaii and California: introduction. Am J Epidemiol 1975;102(6):477–80.
[10] Marmot M, Syme S, Kagan A, Kato H, Cohen J, Belsky J. Epidemiologic studies of coronary heart disease and stroke in Japanese Men Living in Japan, Hawaii and California: prevalence of coronary and hypertensive heart disease and associated risk factors. Am J Epidemiol 1975;102(6):514–25.
[11] Robertson TL, Kato H, Rhoads GG, Kagan A, Marmot M, Syme SL, et al. Epidemiologic studies of coronary heart disease and stroke in Japanese Men Living in Japan, Hawaii and California: incidence of myocardial infarction and death from coronary heart disease. Am J Cardiol 1977;39(2):239–43.
[12] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, Pittsburgh, Pennsylvania, United States. ACM New York, NY, USA; 1992.
[13] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[14] Vapnik V. An overview of statistical learning theory. IEEE Trans Neural Networks 1999;10(5):988–99.
[15] Tay D, Poh C, Kitney R. An evolutionary data-conscious artificial immune recognition system. In: Genetic and evolutionary computation conference (GECCO), Amsterdam, The Netherland; 2013.
[16] Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J Compar Neurol 2009;513(5):532–41.
[17] Martini FH, Nath JL, Bartholomew EF. Fundamentals of anatomy & physiology. Pearson; 2011, p. 1264.
[18] Kuhn H, Dickinson-Anson H, Gage F. Neurogenesis in the dentate gyrus of the adult rat: age-related decrease of neuronal progenitor proliferation. J Neurosci 1996;16(6):2027–33.
[19] Tanapat P, Hastings N, Reeves A, Gould E. Estrogen stimulates a transient increase in the number of new neurons in the dentate gyrus of the adult female rat. J Neurosci 1999;19(14):5792–801.
[20] Kempermann G, Kuhn HG, Gage FH. Genetic influence on neurogenesis in the dentate gyrus of adult mice. Proc Natl Acad Sci 1997.
[21] Nilsson M, Perfilieva E, Johansson U, Orwar O, Eriksson PS. Enriched environment increases neurogenesis in the adult rat dentate gyrus and improves spatial memory. J Neurobiol 1999;39(4):569–78.
[22] Praag HV, Kempermann G, Gage FH. Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. Nat Neurosci 1999;2:266–70.
[23] Fowler CD, Liu Y, Ouimet C, Wang Z. The effects of social environment on adult neurogenesis in the female prairie vole. J Neurobiol 2002;51(2):115–28.
[24] Gould E, Tanapat P. Stress and hippocampal neurogenesis. Biol Psychiatry 1999;46(11):1472–9.
[25] Tanapat P, Hastings NB, Rydel TA, Galea LA, Gould E. Exposure to fox odor inhibits cell proliferation in the hippocampus of adult rats via an adrenal hormone-dependent mechanism. J Compar Neurol 2001;437(4):496–504.
[26] Stangl D, Thuret S. Impact of diet on adult hippocampal neurogenesis. Genes Nutr 2009;4(4):271–82.
[27] Ekdahl CT, Claasen JH, Bonde S, Kokaia Z, Lindvall O. Inflammation is detrimental for neurogenesis in adult brain. Proc Natl Acad Sci 2003.
[28] Neves G, Cooke S, Bliss T. Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. Nat Rev Neurosci 2008;9(1):65–75.
[29] Gould E, Beylin A, Tanapat P, Reeves A, Shors T. Learning enhances adult neurogenesis in the hippocampal formation. Nat Neurosci 1999;2:260–5.
[30] Shors TJ, Miesegaes G, Beylin A, Zhao M, Rydel T, Gould E. Neurogenesis in the adult is involved in the formation of trace memories. Nature 2001;410:372–6.
[31] Gage F. Neurogenesis in the adult brain. J Neurosci 2002;22(3):612–3.
[32] Cover T, Hart P. Nearest neighbor pattern classification. Inform Theory, IEEE Trans 1967(1):21–7.
[33] Kennard R, Stone L. Computer aided design of experiments. Technometrics 1969;11(1):137–48.
[34] Holland J. Genetic algorithms. Sci Am 1992:66–72.
[35] Fang KT, Lin DKJ, Winker P, Zhang Y. Uniform design: theory and application. Technometrics 2000;42(3):237–48.
[36] Tay D, Poh C, Goh C, Kitney R. A biological continuum based approach for efficient clinical classification. J Biomed Inform 2014;47:28–38.
[37] Liu H, Motoda H. Feature extraction, construction and selection: a data mining perspective. Kluwer; 1998. p. 410.
[38] Miller J, Thomson P. Cartesian genetic programming, EuroGP. Springer-Verlag; 2000.
[39] Kowaliw T, Banzhaf W. The unconstrained automated generation of cell image features for medical diagnosis. GECCO; 2012.
[40] Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 1998;10(7):1895–924.
[41] Japkowicz N. Learning from imbalanced data sets: a comparison of various strategies. AAAI Workshop on learning from imbalanced data sets; 2000.
[42] Li D, Liu C, Hu S. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010;40(5):509–18.
[43] Caruana R, Kangarloo H, Dionisio JD, Sinha U, Johnson D. Case-based explanation of non-case-based learning methods. AMIA annual symposium proceedings archive; 1999.
[44] Kris-Etherton P, Eckel RH, Howard BV, St Jeor S, Bazzarre TL. Lyon diet heart study: benefits of a mediterranean-style, national cholesterol education program/american heart association Step I dietary pattern on cardiovascular disease. Circulation 2001;103:1823–5.
[45] Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, et al. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the Interheart Study): case-control study. Lancet 2004;364(9438):937–52.
[46] Ebi-Kryston KL. Respiratory symptoms and pulmonary function as predictors of 10-year mortality from respiratory disease, cardiovascular disease, and all causes in the whitehall study. J Clin Epidemiol 1988;41(3):251–60.
[47] Barrett-Connor E, Khaw KT. Family history of heart attack as an independent predictor of death due to cardiovascular disease. Circulation 1984;69(6):1065–9.
[48] Kannel William B, Gordon T, Castelli William P, Margolis James R. Electrocardiographic left ventricular hypertrophy and risk of coronary heart disease: the Framingham study. Ann Intern Med 1970;72(6):813–22.
[49] Dimsdale JE. Psychological stress and cardiovascular disease. J Am Coll Cardiol 2008;51(13):1237–46.