Expanding a Machine Learning Class Towards its Application to the Stock Market Forecast

Expanding a Machine Learning Class Towards its Application to the Stock Market Forecast

Enrique González-Núñez. [0000-0002-5410-0483][1*],
Luis A. Trejo [0000-0001-9741-4581][1†] and
Michael Kampouridis [0000-0003-0047-7565][2†]

[1]School of Engineering and Science, Tecnologico de Monterrey, Carretera al Lago de Guadalupe Km. 3.5, Atizapán, 52926, Edo. de México, México.
[2]School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, Essex, United Kingdom.

*Corresponding author(s). E-mail(s): enrique.gonzalez@tec.mx;
Contributing authors: ltrejo@tec.mx; mkampo@essex.ac.uk;
†These authors contributed equally to this work.

Abstract

In this work, we present a new and efficient algorithm to perform a short-term market trend forecast, based on the Artificial Organic Networks (AON) metaheuristic machine learning framework. Regarding this goal, we present the concept of Artificial Halocarbon Compounds (AHC) or AHC-algorithm as a bio-inspired supervised machine learning algorithm based on the AON framework. Through our research, we contrast the forecast acquired with the proposed AHC model, to previously reported outcomes using the Artificial Hydrocarbon Networks (AHN) in similar tasks. The AHN algorithm is the first formally defined topology based on the AON, making the AHN algorithm a vital benchmark to contemplate. After comparing the AHC-algorithm to the original AHN-algorithm, we found out that due to the high computational complexity of the latter, the new topology is more convenient when modeling more complex systems; being this characteristic the main contribution of the AHC-algorithm, allowing it to be a more adaptable, dynamic, and reconfigurable topology. Likewise, we compared the results of the AHC-algorithm against the outcomes derived from an ARIMA model; we also made a cross-reference contrast against results concerning the prediction of other stock market indices using former state-of-the-art machine learning methods. The proficiency of the AHC-algorithm is assessed by doing a forecast of the IPC Mexico index obtaining good results, achieving a computed R-square of 0.9919, and an $\mathbf{8 \times 10^{-4}}$ mean relative error for the forecast.

Keywords: Artificial Intelligence, Machine learning, Bio-inspired, Metaheuristic, Stock market index, Financial Forecasting

1 Introduction

The Index Tracking Problem (ITP) or stock market prediction as more commonly known, is a complex process affected by many factors [1, 2]. As remarked in [3], stock market forecasts, despite being a recurrent subject of many investigation groups, remain an essential financial research topic within other aspects due to their economic impact. As an update of the example provided in the previously referred article, the New York Stock Exchange had a $40.5 trillion market capitalization (market value of all shares traded from public companies listed in its market) as of December 2022 and had a $52.2 trillion market capitalization as of December 2021. The ITP is a trading strategy based on the buy-and-hold of assets [4, 5], that uses an index tracker to replicate the performance of a stock market index or any other security found in the capital markets, that considers the risk factor of investing, anticipating the appealed potential profits that perhaps can be obtained from these markets. Its behavior is reproduced through different techniques, including different state-of-the-art artificial intelligence methods [3], capable of developing models that provide reliable forecasts.

This paper returns to the objective previously defined in [6], aiming to apply the Artificial Organic Networks (AON) metaheuristic machine learning framework, to develop a new efficient algorithm, enable to perform a short-term market trend forecast. Regarding this goal, we present the concept of Artificial Halocarbon Compounds (AHC) or AHC-algorithm as a bio-inspired supervised machine learning algorithm, and as a new topology based on the AON framework. Through our research, we contrast the forecast acquired with the proposed AHC model, to previously reported outcomes using the Artificial Hydrocarbon Networks (AHN) in similar tasks. The AHN algorithm is the first formally defined topology based on the AON, making the AHN algorithm a vital benchmark to contemplate. After comparing the AHC-algorithm to the original AHN-algorithm, we found out that due to the high computational complexity of the latter, the new topology is more convenient when modeling more complex systems; being this characteristic the main contribution of the AHC-algorithm, allowing it to be a more adaptable, dynamic, and reconfigurable topology. Likewise, we compared the results of the AHC-algorithm against the outcomes derived from an ARIMA model; we also made a cross-reference contrast against results concerning the prediction of other stock market indices using former state-of-the-art machine learning methods.

The rest of the article is structured as follows. In Section 2, we offer a literature review related to the forecast of stock market indices using machine learning methods. Furthermore, we introduce some main notions of the AON framework. Afterward, in Section 3, we provide elements about the used dataset and its preprocessing. Likewise, we describe the methodology followed to perform the experiments. Then, Section 4 explains how the AHC-algorithm is designed, as well as its main characteristics, implementation, and computational complexity. Through the discussion, we review some of the disadvantages that the original Artificial Hydrocarbon Networks (AHN) topology has; these disadvantages are considered regarding the definition of the new AHC-algorithm. Section 5 presents the results obtained from the experiments, including a cross-reference evaluation. Ultimately, Section 6 affords the conclusions of this work, complementing with possible future lines for research.

2 Background

In this section, we offer a theoretical framework. In this respect, Section 2.1 presents a literature review. Later, Section 2.2 illustrates some main notions of the AON framework. Next, Section 2.3 delivers some fundamental concepts of the AHN-algorithm.

2.1 Literature Review

Numerous articles analyze the employment of machine learning techniques as predicting tools for stock market indices, contemplating the characteristics around their complex behavior, such as their noisy, unpredictable dynamics, making this a difficult task. In this respect, Ye [7] offered the forecast of the Google (GOOG) and Tesla (TSLA) stock prices using methods such as Support Vector Regression (SVR), Gated recurrent units (GRUs), Long Short-Term Memory (LSTM), and Extreme Gradient Boosting (XGBoost). Sunki et al. [8] delivered the forecast of the Netflix (NFLX) stock price applying ARIMA, LSTM, and FBProphet methods. In contrast, Shi et al. [9] predicted the Standard and Poor's 500 (S&P 500) implementing XGBoost. Correspondingly, Aliyev et al. [10], employed the ARIMA-GARCH and the LSTM methods to produce the forecast of the Russian Stock Exchange (RTS). In their work, Singh [11] presented the forecast of the Indian Stock Market Index (NIFTY 50) utilizing different models, including amongst them Artificial Neural Networks (ANNs), Adaptive Boost (AdaBoost), and k-Nearest Neighbors (KNN). Similarly, Harahap et al. [12] predict the Nikkei 225 (N225) using SVR, Back Propagation Neural Networks (BPNNs), and Deep Neural Networks (DNNs). Finally, González-Núñez et al. [13] used Genetic Algorithms (GA) to predict different indices like the S&P 500, the N225, the Dow Jones Industrial Average (DJIA), the Financial Times Stock Exchange (FTSE), and the Cotation Assistée en Continu index (CAC), amongst others.

2.2 AON Framework

As stated above, the main goal of this research is to define a new algorithm following the notions and main characteristics of AON as a machine learning class [14, 15]; considering that the AON technique allows modeling systems as a gray box, yet conceding to partially understand the behavior of the system. As depicted in Table 1, the organization of the framework comprehends the following aspects:

- It defines the set of components and interactions required to build a structure.
- Heuristic rules that state its organization, inspired by basic chemical rules and observations.
- The artificial organic network is expressed mathematically.

– It is used through an implementation.

Table 1 Artificial Organic Networks Framework.

AON Framework[1]	
Level	**Description**
Implementation	Training and inference
Mathematical model	Structure and functionality
Heuristics	Rules of organization: three-level energy scheme
Interactions	Relationships: covalent bonds, chemical balance interaction
Components	Units: atoms, molecules, compounds, mixtures

[1] Source [15].

Hence, an AON is a set of graphs built based on heuristic rules, as per the framework's guidelines; these organization rules are inspired by chemistry to form organic compounds and define their interactions. Each graph represents a molecule with atoms as vertices and chemical bonds as edges. These molecules interact via the chemical balance to form a mixture of compounds; therefore, an AON is a combination of compounds formed by different components. Thus, following the designation of the framework, four components have been defined: atomic units, molecular units, compounds, and mixtures; further, two interactions are characterized among the components: covalent bonds and chemical balance interaction.

The molecules can be seen as packages of information; the bonds between the structures of the model describe its complexity. The seven defined characteristics of the AON are: structural and behavioral properties, encapsulation, inheritance, organization, mixing properties, stability, and robustness. Accordingly, AON has a structure and behavior that states its two main characteristics: modeling non-linear systems and partial interpretation of unknown information. In consistency with the framework organization, AON as a learning method needs a two-step implantation: a training process to build its structure and estimate all the parameter values inspired by organic chemistry rules, and an inference process that consists of using the obtained structure to find an output considering a specific input value.

The AON framework has been classified as enclosing three types of algorithms: a) chemically inspired algorithms with defined heuristic rules, based on functional groups and molecular structures of chemical organic compounds, b) artificial basis algorithms that define specific functional groups and molecular structures independently of chemical organic compounds, and c) hybrid algorithms that define structures based on a mixture of chemically inspired and artificial basis algorithms.

2.3 Topological Structure and Prevalence of AHN

Circumscribing the components and interactions disposed on the framework's main characteristics as established by Ponce et al. [15], the AON requires the usage of a functional group; these functional groups are the type of molecules that determine a topological configuration of the AON for its implementation. Consequently, AON was implemented through one existing topology: Artificial Hydrocarbon Networks

(AHN); thus, the AHN model is AON's first and only formal topology defined up to now. The AHN algorithm is defined as a chemically bio-inspired on the way chemical hydrocarbon compounds are formed; it was proposed to perform an optimization of a cost-energy function, based on two mechanisms to form organic compounds and produce an efficient number of molecules to build the structures:

i It uses least-squares regression (LSR) to define the structure of each molecule.
ii It uses gradient descent (GD) to optimize the position and number of molecules in the feature space.

AHN has shown improvements in predictive power and interpretability compared to other well-known machine learning models, such as neural networks and random forests. However, as explained in [16] big data are mainly characterized by the amount of information that can be processed, the speed of data generation, and the variety of data involved; existing machine learning algorithms need to be adapted to profit the advantages of big data and process more information efficiently. Thus, AHN has the disadvantage of being very time-consuming and is unable to deal with big data, since the model uses GD that, due to its complexity, hinders the scalability of the AHN model.

To find more examples of applications of the AHN-algorithm, we recommend the work presented in [13]. For the interested readers seeking more in-depth information on the AHN model implementation and the AON framework, we suggest to explore the lecture by Ponce et al. reported in [15].

3 Methodology

Contrasting experiments have been carried out to confirm that the AHC-Algorithm, as a proposed supervised machine learning algorithm, can effectively perform a short-term market price trend forecast, as initially defined in [6]. In this sense, Section 3.1 provides details of the dataset we used and its preprocessing. After that, Section 3.2 illustrates the steps followed along the methodology of the experiments.

3.1 Data

The experiments on this work were performed using existing data from Mexico. The variables included in the dataset are:

- The daily reported closing price of the IPC Mexico stock market index.
- The daily reported MXN-USD foreign exchange rate (FX).
- The quarterly reported gross domestic product (GDP).
- The monthly reported consumer price index (CPI).
- The monthly risk-free rate (RFR).
- The monthly unemployment rate (UR).
- The monthly reported current account to GDP rate (BOP).
- The monthly reported investment rate (GFCF).

All data collected covers a period from the 1st of June 2006 to the 31st of May 2023. The IPC and the FX data were retrieved from Yahoo Finance, and the rest of the variables obtained from the OECD. Table 2 shows some descriptive statistics for data of the closing price of the IPC used in this research.

Table 2 Descriptive statistics of the IPC index.

Descriptive Statistics						
Mean	SD	Min	25%	50%	75%	Max
39,899.97	8,813.83	16,653.15	33,262.48	41,960.44	46,190.08	56,609.53

The data were preprocessed as follows: a) the macroeconomic variables (MEVs) are treated as "continuous signals" instead of discrete information, so for each input, an independent approximation is made using least-squares polynomial regression (LSP), b) the data are scaled, for this purpose, the dataset is standardized by removing the mean, c) the dimensionality of the data is reduced using principal component analysis (PCA), it is done considering three principal components (PC).

At this point, it is essential to remark that, even though eight features can be pondered as a small amount, the employment of PCA is pivotal in the implementation of AHC, since as explained later, the computational complexity of the AHN-algorithm depends on the number of features. Additionally, as a measure to prevent overfitting, besides preprocessing, the data were split into subsets, the initial section for training, and the remaining portion for testing. Finally, the model is assessed by using an out-of-sample forecast, where the performance of the financial model is tested on data not used for building the model; this implementation was applied instead of a one-day-ahead forecast due to the progress achieved in our research by the moment the present results were recovered. For the interested reader, the dataset used on our research is available at [17].

3.2 Forecast of the Index

The methodology observed to achieve the forecast of the index, using the AHC-algorithm, consists of the following steps (see Figure 1):

1 Data are preprocessed as explained above in Section 3.1; afterward, the dataset is passed into the AHC-algorithm.
2 The training parameters are established by doing a grid search, as described in Section 5.1.
3 The AHC model is fitted using the training set.
4 The model performs a forecast using the testing set.
5 The results of the forecast and the performance of the AHC-algorithm are compared in Section 5.3 against the reported results using the AHN-algorithm in a similar financial task; likewise, Section 5.5 also includes a comparison against some of the results found across the literature review.

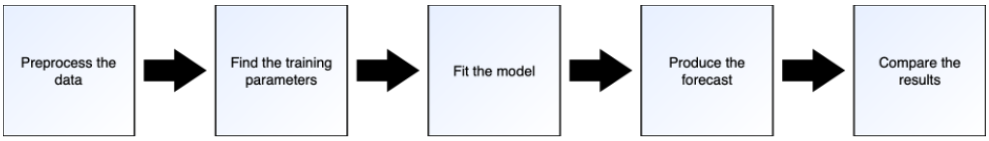

Fig. 1 Methodology to produce the forecast of stock market applying the AHC-algorithm.

All the experiments have been implemented in Python, employing in some cases the SciPy [18], Scikit-learn [19], and statsmodels [20] libraries. For the interested reader, the code used in this research is accessible at [21] and the dataset at [17].

4 Artificial Halocarbon Compounds

We explain here how the AHC-algorithm is designed, as well as some of its primary notions. In this respect, Section 4.1 exposes the main drivers that inspired the design. Section 4.2 presents the new attributes defined for the AHC-algorithm as a method inspired by organic chemistry. Section 4.3 describes how the AHC-algorithm forms compounds based on a different functional group. Section 4.4 portrays how the algorithm is implemented. Finally, Section 4.5 analyzes the computational complexity of the AHC-algorithm.

4.1 A New Conception: Diversifying the AON Topologies

Since the forecast of a stock market index as a time series phenomenon [22, 23] is a dynamic, non-linear process, that at the same time has largely non-linear dependencies with other factors such as the MEVs, it can become very complex and difficult. As Chacon [24] explains, among the challenges found when doing a prediction model for a time series, exists the necessity of detecting non-linear dependencies across time, excluding the noise and behavior of the time series. Correspondingly, Hou [25] and Sheta [4] state that precise stock return forecasting as a financial problem remains a particularly demanding task due to its complex, dynamic, non-linear, and chaotic nature. In contrast, Ordoñez [26] exemplifies the importance and complexity of analyzing the forecast of stock returns with the existence of financial investment companies, with specialized business areas with teams of employees focused on studying these kinds of projections contemplating their complexities. Consequently, continuing with the work started in [6], the problem was narrowed down following two main constraints:

1 The forecast is done using the historical prices of the concerning index rate and at least five additional MEVs.
2 The MEVs are selected based on their correlated coefficient (CC) to the analyzed index.

In the most straightforward notion, as explained in the literature [15], the AHN process, as a supervised learning algorithm, where an AON structure is built through the algorithm, is identified as f; along the process, a structure of an organic compound is produced by segmenting the dataset received to create a model of the system. Each section is analyzed by a molecule that fits second- or third-degree polynomial terms using LSR, the position and number of molecules are optimized using GD; a set of molecules produces a compound.

By virtue that AHN is the only existing arrangement for the AON machine learning class, the postulation of the new algorithm adept at performing the stated objectives, centers its attention on the conception of a topology distinct from the AHN, attending some of its disadvantages. For example, the AHN-algorithm is very time-consuming, hence it uses GD to optimize the position of the molecules in the feature space. Indeed, this strategy of providing a better capability to deal with big data, while reducing algorithmic time consumption, must be designed pondering the requirement to avoid losing either predictive power and/or interpretability. We consider that these properties are two of the main characteristics that the original AHN topology initially possessed. Consequently, the approach for establishing a different topology is driven by the subsequent main goals:

1 To find an alternative to GD to define the position and/or number of molecules, so the algorithmic time consumption while creating an AON arrangement can be reduced.

2 To circumscribe a new topology that must be able to handle time series, so it can be capable of producing a financial forecast.

3 The new organic structure will be produced based on a functional group different from the hydrocarbons; consequently, the standard LSR method will be discarded. Additionally, new types of curves to be fitted will be explored.

4.2 Artificial Halocarbon Compounds: A Hybrid Bio-inspirational Algorithm

Artificial Halocarbon Compounds (AHC) or the AHC-algorithm, is a proposed supervised machine learning algorithm based on the AON framework inspired by chemical halocarbon compounds. As a new AON arrangement, one of its principal considerations is focused on relinquishing the GD mechanism to optimize the position and/or number of molecules, hence the time consumption can be reduced while creating an AON structure. The laydown of GD has been analyzed before [6, 27]. Furthermore, the idea of artificial aromatic compounds as a recursive network was roughly explored by Ponce et al. [15]; regardless, these attempts did not finally peak in an AON arrangement formalizing a new topology. Therefore, the strategy of using K-means introduced in [6] is recalled toward generating a new type of organic structure based on the AON framework.

4.2.1 Components and Interactions

In the proposed scheme, since the GD method is substituted by K-means as part of the mechanism to form organic compounds, and the hydrocarbons functional group is replaced as a topology for the new structure arrangements, intrinsically some of the main changes of the algorithm take place through the components and interactions level of the AON framework. Accordingly, the new topology is based on the third type of algorithm from the AON framework: hybrid algorithms that define structures based on a mixture of chemically inspired and artificial basis algorithms.

In this hybrid approach, the feature space is segmented (clustered) using K-means per the number of molecules required, in this way its position is defined. Therefore, each time an iteration occurs, the data are segmented as many times as the same number of molecules to be created; afterward, the structure of each molecule is computed for the corresponding segment. Although this appeal is performed based on a hybrid model, it still complies with the seven characteristics defined for an AON: structural and behavioral properties, encapsulation, inheritance, organization, mixing properties, stability, and robustness.

As stated before, the structure of each molecule is not defined directly by using a standard LSR method; instead, as a significant feature that will characterize the new AON arrangement, a dynamic topology is offered. Hence, the new topology is qualified for choosing a wider variety of options to build the organic structures for a compound, according to the cost-energy function, and therefore maintaining an overall low error of the models being produced. These dynamic options consider whether to substitute the type of curve to be fitted for a respective segment, such as sine or cosine, among others, instead of just fitting a second- or third-degree polynomic term, or even choosing amid different fitting methods, such as multiple non-linear regressive (MNLR) model [6], autoregressive (AR) within others, in this sense replacing the method employed to characterize each molecule. All these replacements are analyzed while the computation of the algorithm is done, just as if an inspirational chemical reaction was taking place, and at the end of the reaction, the arrangement with the best final substitution from the structures compared is provided.

Halocarbons functional groups

The halocarbons functional group [28], or halocarbons molecules, are hydrocarbons where a halogenation reaction has taken place; specifically, organic compound structures in which one or more carbon atoms are covalently bonded to one or more halide atoms that have substituted/replaced hydrogen atoms. Halides or halogens are the six elements of Group 17 or VIIA of the periodic table; these elements are: fluorine (F), chlorine (Cl), bromine (Br), iodine (I), astatine (At), and tennessine (Ts); the most common substitution is made by chlorine halocarbons, known as organochlorine compounds. The halocarbons have four valence electrons. In consequence, with the introduction of the halocarbons functional group, the set of atomic units originally defined by Ponce et al. [15] expands from two to eight different atomic units for the AHC-algorithm; therefore, F is called fluorine atom, Cl is called chlorine, Br is called bromine, I is called iodine, At is called astatine, and Ts is called tennessine, being these the new six atomic units added to the existent H and C.

The halocarbons are classified into three types: haloalkanes, haloalkenes, and aryl halides. Haloalkanes are saturated compound structures formed by halogenated hydrocarbons, where all the carbon atoms are linked by single bonds and at least one hydrogen atom has been replaced by a halide; in opposition, haloalkenes are unsaturated compounds, which means that they contain one or more double bonds between carbon atoms, making them more likely to undergo addition reactions with other compounds. Aryl halides, also known as haloaromatics or haloarenes, are organic compounds that contain a halogen atom bonded to one or more aromatic rings. An aromatic ring or arene is formed by carbon atoms linked in a cyclic arrangement, like the benzene ring, which is the most common example, and is formed by six carbon atoms organized in a hexagonal shape.

4.3 Forming of Compounds

The AHC-algorithm bio-inspiration is particularly based on the aryl halides subset alluding to their distinctive properties over the other halogenated organic compounds; the presence of the arene in aryl halides makes them more stable. Thus, due to their unique physical and chemical characteristics are widely used as building blocks in the synthesis of various organic compounds, such as pharmaceuticals, agrochemicals, and polymers.

Concerning the molecule's structure, since the halocarbons are formed by the substitution of hydrogen atoms in hydrocarbons, thus, the initially defined set of CH-primitive molecules with carbon as the central atom, and their corresponding covalent bonds between elements, constitute the initial bases for a cyclic recursive CH molecule in the new AHC-algorithm. In addition, since the benzene ring is not only the most common but a fundamental structure in aromatic compounds, it will constitute an essential part of the new structure. As mentioned before, the benzene arrangement consists of six-membered carbon atoms in a hexagonal ring with alternating single and double bonds (see Figure 2). Each carbon atom in the benzene ring is also bonded to one hydrogen atom. The benzene ring is highly stable due to its particularly aromatic bonding characteristics.

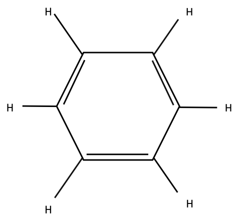

Fig. 2 Diagram of the benzene ring.

Another crucial attribute considered is the capacity of arenes to bond or join together (fused) in two or more aromatic rings. These kinds of more extensive arrangements are known as polycyclic aromatic compounds; some examples of these polycyclic aromatic compounds include, amongst others:

- Naphthalene: It consists of two benzene rings fused; it is a typical aromatic compound found in coal tar.
- Anthracene: It consists of three fused benzene rings; it is also found in coal tar and is used as a starting material for the synthesis of dyes, as colored substances that bond chemically are widely used to fabric textiles and other substrates.
- Phenanthrene: It consists of three fused benzene rings; it is found in fossil fuels.
- Pyrene: It consists of four fused benzene rings; it is commonly found in coal tar.

Since the mathematical formulation of the target function *f*, identified in an earlier stage of this work, was defined through the MNLR model [6], it is essential to remark that the stated math expression possesses ten coefficients. Hence, at least ten hydrogens will be needed in the proposed scheme to represent all the terms of the multivariate polynomial expression accounted to handle the dynamic system. As it can be observed in the skeletal formulas represented in Figure 3, different polycyclic aromatic compounds have the presence of ten hydrogens atoms; however, due to their easier representation in the skeletal formula, the new polycyclic aromatic network topology will mainly remain based on the anthracene disposition.

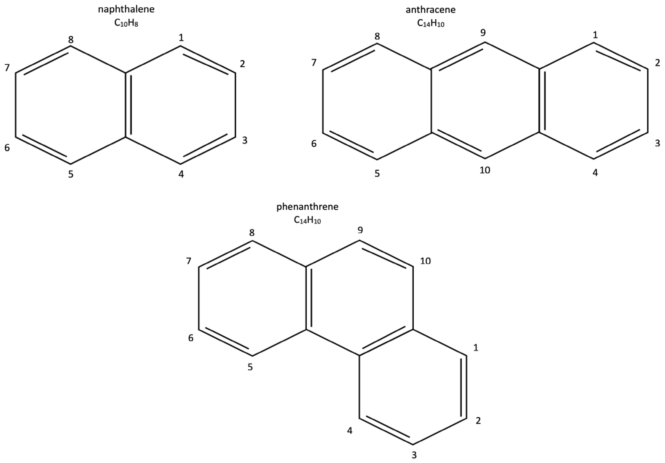

Fig. 3 Skeletal formula representation of some polycyclic aromatic compounds.

Afterward, the inspirational halogenation where atoms of H are replaced by halides is characterized in the AHC-algorithm by chemically reacting the anthracene-based molecule with all the halogens, but at the end substituting the H atoms just by the halide which offers the lowest-cost energy function. Thus, as a dynamic topology, AHC-algorithm compares (computes) the result of reacting with all the different halogens and choosing the one with the lowest error; Figure 4 shows an example of an AHC molecule where all H atoms are replaced by Cl atoms. Through the halogenation mechanism, instantiated in Table 3, the replacement of the original LSR mechanism implemented in the AHN-algorithm that defined the structure of each molecule, takes place.

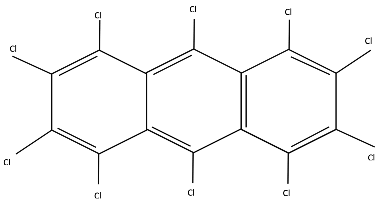

Fig. 4 Example of the AHC molecule where all H atoms are replaced by Cl atoms.

Table 3 AHC Halogenations, with different substitutions of the original LSR mechanism that define the structure of a molecule in the AHN-algorithm.

AHC Halogenations		
Substitution	**Reaction**	**Description**
H → F	Fluorination	Substitution with a second-degree MNLR expression.
H → Cl	Chlorination	Substitution with a two-lag AR model.
H → Br	Bromination	Substitution with a cube-root MNLR expression.
H → I	Iodination	Substitution with a hyperbolic sine MNLR expression.
H → At	Astatination	Substitution with a hyperbolic cosine MNLR expression.
H → Ts	Tennessination	Substitution with an element-wise 2x MNLR expression.

The motives to choose these types of polynomial expressions are based on empirical reasons after running different experiments scrutinized in [6] and in the next section; nevertheless, depending on the specific type of system, criteria, or application aimed to be modeled, it is encouraged to explore with these and/or other different kinds of polynomial expressions or methods used in the substitution mechanisms for the AHC halogenations. In addition, based again on the same empirical reason, the attribute of creating mixtures considered in the AHN-algorithm, was discontinued for the AHC-algorithm since this feature was no further required.

Finally, the enthalpy property [15] applied in the AHN-algorithm for optimizing compounds was also ceased. This property was implemented as a method for building an optimal compound using the minimum number of CH-primitive molecules, resembling that enthalpy in thermodynamics measures the heat energy (transferred or exchanged), along the chemical reactions. Regardless, it is still desired to find the lowest-cost energy function, again, this is achieved by the inspirational halogenation described earlier; moreover, this method allows the introduction of the entropy property, since along the process a set of possible combinations to form a compound are computed.

In thermodynamics [29], entropy is a property that describes the distribution of energy by measuring the disorder in chemical reactions. It can be understood as the molecular disorder in a system or the degree of randomness of its particles; as the randomness or disorder goes higher, the entropy increases as well. In essence, entropy describes the energy within a system by quantifying the number of possible energy distributions, specific arrangements (microstates), or configurations of particles and energy that are available within a system, contemplating that in a system, energy can be distributed among particles in different ways. It is essential to notice that entropy describes the distribution of energy, not the total energy of the system. Consequently, the concept of entropy property or entropy-rule is formally introduced to the AHC-algorithm, and is measured while the dynamic topology compares the level of energy of the different reactions from the available halogens; the aim is to keep the lowest entropy of the model system, based on the output error.

4.4 Phases of the AHC-algorithm and implementation

The artificial halocarbon compounds is implemented using the AHC-algorithm illustrated in Algorithm 1, which is the main routine. Algorithm 1 iteratively calls the FORM-COMPOUND() routine, in charge of forming structures of organic compounds based on halogenations of anthracene, this routine is illustrated in Algorithm 2.

Figure 5 illustrates the phases of the AHC-algorithm. After the initialization steps, in its core, the AHC-algorithm creates the structure of an artificial halocarbon compound based on the number of polycyclic aromatic molecules needed. This phase corresponds to line 5 in Algorithm 1.

Next, the AHC-algorithm defines the structure of each polycyclic molecule in the compound depending on the best halide selected during the halogenation reaction, this step corresponds to line 6 in Algorithm 1. As stated earlier, the attribute of creating a set of compounds to combine them in a mixture, initially defined in the AHN-algorithm, was discarded as a feature since it was not found further required based on empirical experience.

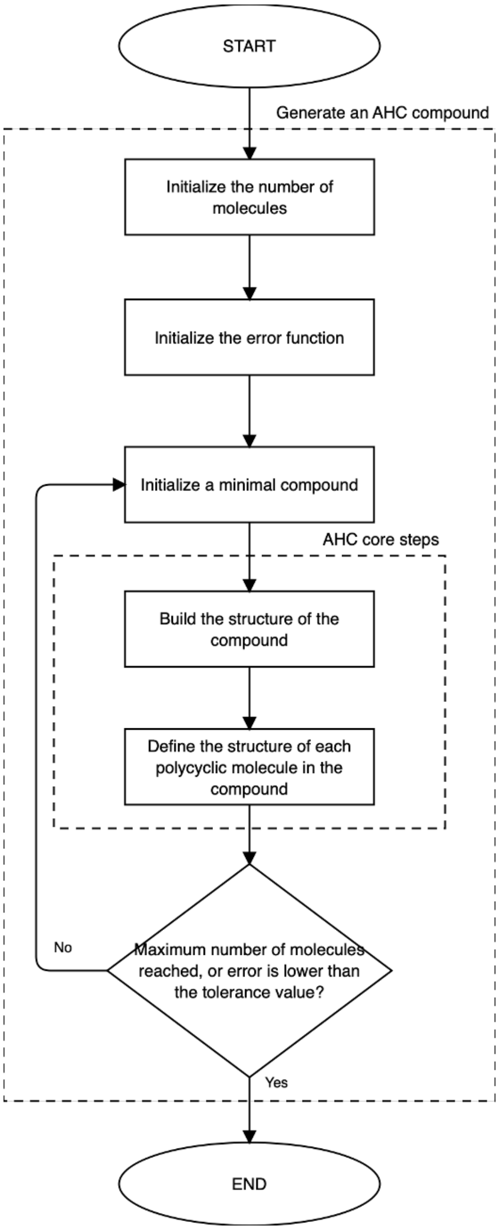
START
Generate an AHC compound
Initialize the number of molecules
Initialize the error function
Initialize a minimal compound
AHC core steps
Build the structure of the compound
Define the structure of each polycyclic molecule in the compound
No
Maximum number of molecules reached, or error is lower than the tolerance value?
Yes
END

Fig. 5 Flow diagram depicting the phases of the AHC-algorithm.

Algorithm 1 AHC-ALGORITHM $(\Sigma, n_{max}, \epsilon, \lambda)$**:** Implementation of the artificial halocarbon compounds using AHC-algorithm.

Input: the system $\Sigma = (x, y)$, the maximum number of molecules n_{max}, the tolerance value $\epsilon > 0$, and the regularization factor λ.

Output: the structure of the compound C, and the type of halogenation τ for each molecule in C. The coefficients A are included within the structure C.

```
Initialize the number of molecules, $n \leftarrow 2$.
Initialize the error function, $\varepsilon \leftarrow \infty$.
while $(n \leq n_{max})$ and $(\varepsilon > \epsilon)$ do
    Initialize a minimal compound $C$.
    Split $\Sigma$ into $n$ subsets $\Sigma_i$ with their centers $Q_i$, using K-means.
    Obtain a new structure of $C$, find the halogenations $\tau$, and update the error
      function $\varepsilon$ with FORM-COMPOUND($\sum_{i=2}^{n}, Q, C, \lambda$).
end while
return $C$, and $\tau$
```

Algorithm 2 FORM-COMPOUND $(\sum_{i=2}^{n}, Q, C, \lambda)$**:** Routine to form structures of organic compounds based on halogenations.

Input: the n splits of the system $\sum_{i=2}^{n}$, the centers of the splits Q, the initial compound C, and the regularization factor λ.

Output: the final structure of the compound C, the type of halogenation τ for each molecule in C, and the updated error function ε.

```
Initialize $\tau$ with all the types of halogenation.
for each partition $\sum_i$ do
    for each type of halogenation $\tau_j$ do
        Find the energy level of the subset $\Sigma_i$ with each halogen $\tau_j$, considering $\lambda$.
    end for
    Update the final behavior of the molecule in $\Sigma_i$, by selecting the best
      halogenation $\tau_j$, following the ENTROPY-RULE.
end for
Update the error function $\varepsilon$ using the true fractional relative error defined in [6].
return $C$, $\tau$ , and $\varepsilon$
```

4.5 Computational Complexity

Considering Algorithm 1, the worst-case assumption of the computational time complexity for the AHC-algorithm is $O(5n_{max} * 10^3 + n_{max} * 15^3)$, measured as follows:

Locating at the while loop inside Algorithm 2, the time complexity for the first assignment can be assumed $\sim O(1)$, since the initialization of a minimal compound consists of defining a DataFrame that holds the structure of a default CH_2-primitive molecule. The second step splits Σ into i partitions Σ_i for $i = 1, \ldots, n_{max}$. To create the partitions, we used K-means (from the Scikit-learn library); the computational complexity for this method is $O(kqTD)$, where T is the number of iterations, q is the number of data points, k is the number of clusters, and D is the dimensionality of the data. Next, the third step considers the computation of a new compound structure based on the halogenation reactions for each molecule, as explained in Section 4.3; the target function f of each reaction is characterized through the MNLR model. For each halogenation, the MNLR model is solved using an inverse matrix; the time complexity for solving an inverse matrix is $O(n^3)$. Since six halogenations are computed, then six inverse matrix operations are performed; moreover, as explained in Section 5.1 five halogenations consider an MNLR expression defined with ten coefficients, and one with 15. Thus, the worst-case complexity for the third step is $O(5n_{max} * 10^3 + n_{max} * 15^3)$. In consequence, the overall time complexity of one iteration in the while-loop is $O(5n_{max} * 10^3 + n_{max} * 15^3)$, because $O(5n_{max} * 10^3 + n_{max} * 15^3) > O(kqTD) > O(1)$ if $q \geq n_{max} \geq 2$.

5 Results & Analysis

In this section, we present the results obtained from the conducted experiments. In this respect, Section 5.1 describes how the hyperparameters are tuned by doing a grid search in the training phase. Next, Section 5.2 illustrates the forecast results using the testing set. Subsequently, Section 5.3 compares the prediction obtained with the AHC-algorithm and its performance against the reported results using the AHN-algorithm in a similar financial task. Later, Section 5.4 compares our results obtained with the AHC-algorithm against the outcomes derived from an ARIMA model. Finally, Section 5.5 includes a benchmark of the forecast achieved against some of the results found across the literature review.

5.1 Training phase

Following the AON framework, the AHC is trained with an initial set of different parameters, so later a hyperparameter tuning could be conducted. Considering that the AHC-algorithm requires four inputs (the system and three parameters) as defined in the previous section, Table 4 shows the set of parameters used for a combination of setups that are trained along this phase.

Table 4 Set of initial parameters.

	AHC Parameters			
ϵ	6×10^{-4}		9×10^{-4}	
n_{max}	2	4	8	12
λ	0	1×10^{-10}	0.95	1

From Table 4, it can be observed that the set of parameters allowed for a total combination of 32 setups. However, the experiments were repeated using seven different subset sizes for the training and testing, resulting in a total combination of 224 models trained. Table 5 shows the seven split sizes used during the experiments.

Table 5 Split sizes applied for the training and testing sets.

Subset sizes Train/Test (%)						
98/2	95/5	92/8	90/10	85/15	80/20	75/25

For each case, all the input data (system and set of parameters) are provided to the AHC-algorithm for the training process. Earlier, in Section 4.3 we described the inspirational halogenation mechanism that occurs along the algorithm's execution. According to Table 3, we described the type of polynomial expressions employed for the substitutions. These polynomial expressions are based on the MNLR model first presented in [6] and outlined here in the subsequent definition, considering ten coefficients for three independent variables (since three PCs are used):

Definition 1 (MNLR model, with three independent variables)**.** *Let* $\hat{a}_0, \hat{a}_1, ..., \hat{a}_9$ *be the coefficients of the mathematical expression to model* f*, where* x_{t1}*,* x_{t2} *and* x_{t3} *are the inputs,* λ *is the scalar value of the regularization term,* $\hat{y}$ *is the value obtained from the model and* e_t *is the true error between both; then, the true value* y_t *is said to be:*

$$\begin{gathered} y_t = \hat{a}_0 + \hat{a}_1 x_{t1} + \hat{a}_2 x_{t2} + \hat{a}_3 x_{t3} + \hat{a}_4 \lambda x_{t1}^2 + \\ \hat{a}_5 \lambda x_{t2}^2 + \hat{a}_6 \lambda x_{t3}^2 + \hat{a}_7 \lambda x_{t1} x_{t2} + \hat{a}_8 \lambda x_{t1} x_{t3} + \hat{a}_9 \lambda x_{t2} x_{t3} + e_t \end{gathered} \tag{1}$$

Recapitulating the AHC Halogenations, the Fluorination reaction was implemented by applying the MNLR model as stated in Eq. 1. In the case of Bromination, Iodination, Astatination, and Tennessination reactions the, lambda term was not considered, and the quadratic terms were respectively replaced by other kinds of mathematical functions; an Iodination example with hyperbolic sine substitution is now illustrated in Eq. 2:

$$\begin{gathered} y_t = \hat{a}_0 + \hat{a}_1 x_{t1} + \hat{a}_2 x_{t2} + \hat{a}_3 x_{t3} + \hat{a}_4 \sinh(x_{t1}) + \\ \hat{a}_5 \sinh(x_{t2}) + \hat{a}_6 \sinh(x_{t3}) + \hat{a}_7 x_{t1} x_{t2} + \hat{a}_8 x_{t1} x_{t3} + \hat{a}_9 x_{t2} x_{t3} + e_t \end{gathered} \tag{2}$$

In the case of the Chlorination, the reaction was implemented with a two-lag AR model. Nevertheless, the AR expression used is based on the original MNLR model; in this regard, instead of using three PCs, the new equation includes two-lags of the observed (true) value and two PCs as variables, redefining the consideration of Eq. 1 from ten to 15 coefficients as observed in Eq. 3:

$$\begin{gathered} y_t = \hat{a}_0 + \hat{a}_1 y_{t-1} + \hat{a}_2 y_{t-2} + \hat{a}_3 x_{t1} + \hat{a}_4 x_{t2} + \hat{a}_5 \lambda y_{t-1}^2 + \\ \hat{a}_6 \lambda y_{t-2}^2 + \hat{a}_7 \lambda x_{t1}^2 + \hat{a}_8 \lambda x_{t2}^2 + \hat{a}_9 \lambda y_{t-1} y_{t-2} + \hat{a}_{10} \lambda y_{t-1} x_{t1} + \\ \hat{a}_{11} \lambda y_{t-1} x_{t2} + \hat{a}_{12} \lambda y_{t-2} x_{t1} + \hat{a}_{13} \lambda y_{t-2} x_{t2} + \hat{a}_{14} \lambda x_{t1} x_{t2} + e_t \end{gathered} \tag{3}$$

Again, it was stated before in Section 4.3 that the AHC topology will particularly remain based on the anthracene disposition because it has ten hydrogens and as a polycyclic aromatic compound it has a simpler representation in the skeletal formula. However, for this particular case where Eq. 3 has 15 coefficients, the AHC topology takes form based on the skeletal formula of the Hexacene (see Figure 6). From the skeletal formula, it can be observed that the Hexacene has 16 hydrogens; the discrepancy is eliminated by assuming a 16^{th} coefficient with a constant value of cero.

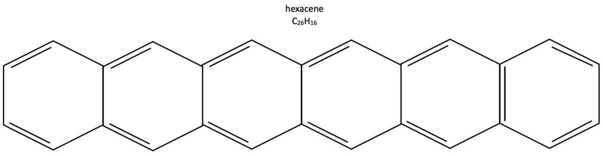

Fig. 6 Skeletal formula representation of the Hexacene.

5.2 Forecasting phase

As stated above, earlier in the training phase a set of parameters and split sizes are defined, allowing for 224 different models trained. We then conducted hyperparameter tuning on the models to produce a forecast with the best parameters identified from all results. On this basis, in the forecasting phase (inference phase as defined in the AON framework), all 224 models are employed to produce an out-of-sample forecast using the testing set. Figure 7 illustrates a heatmap from these results; afterward, the results are ranked using the mean of the error from the testing set of each model. Next, we applied a Friedman test to the ranked top five models to determine any significant difference among them. The results from the Friedman test provided a statistic equal to 5.9999, and a p-value of 0.4231, showing no statistical difference among the results. Hence, for the sake of simplicity, we chose the first parameters from the top. The best parameters we found are $\epsilon = 9 \times 10^{-4}$, $n_{max} = 12$, and $\lambda = 1 \times 10^{-10}$. However, to conduct the final forecast with these parameters in this phase, we chose a data split size of 75% for training and the remaining 25% for testing. The reason to choose this split size is that the results from this forecast are compared in Section 5.3 to the results presented in [27], where data from the examples included therein are split into 70% for training and 30% for testing (the 5% difference is not considered relevant).

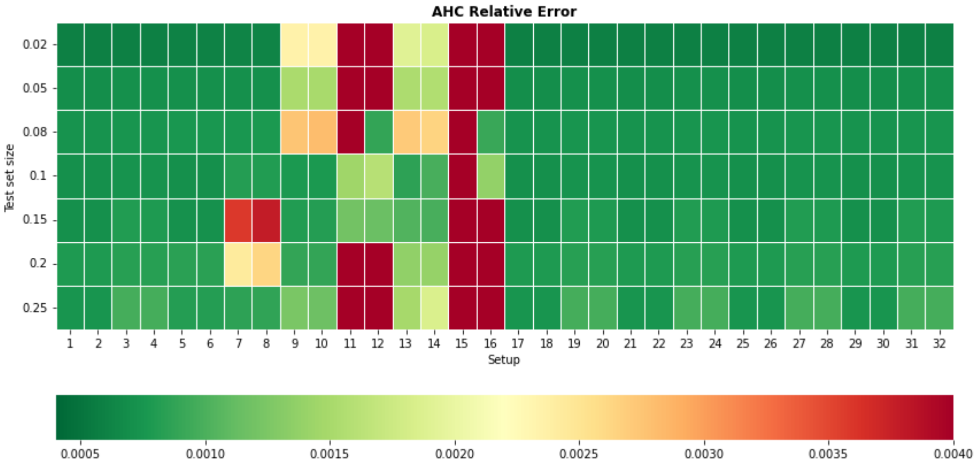

Fig. 7 Results heatmap.

The resulting organic structure computed from these parameters consists of a compound formed by two chlorinated polycyclic aryl halides (two hexacenes) molecules; the respective computed coefficients for each molecule are shown in Table 6, being awarded that a 16^{th} coefficient with value zero is considered to complete each hexacene, as previously mentioned. The outcome of the forecast made with the computed AHC model provided very encouraging results:

1 Figure 8 illustrates the graph with a blue curve corresponding to the original IPC Mexico values y_t from the testing set, and a red curve depicting the estimated values $\hat{y}$, merely no difference can be appreciated between them.
2 Figure 10 presents the graph of Residuals; a good homogeneous distribution can be observed along the Predictor axis. It can be noticed that at the left-hand side of the values of the Predictor, there are some residual values outside the homogeneous distribution strip (outliers), mostly when the Predictor is < 8.4. Nonetheless, as can be appreciated in Figure 8, the values of the Index are close to 8.3 at the beginning of 2020; in this sense, it is significant to remark that this period corresponds to the beginning of the COVID-19 pandemic when the economy of the world was severely affected, so it is natural to expect in this period more variance that can introduce more errors in the forecast. For a more detailed view of the period, we provide Figure 9, which is a zoom-in of Figure 8, for the period 2020-12 to 2021-09. The observed behavior in Figure 9 of the forecast of the IPC computed with the AHC model can be compared to an AR method since, as explained above, from Table 6 we have noticed that the compound formed has two chlorinated molecules, and this reaction was implemented with a two-lag Auto Regression (AR) model as explained in Section 5.1.
3 To evaluate the model, Table 7 shows within other computed statistical measures for the forecast, a very high R-square and Adjusted R-square with values of 0.9919 and 0.9918 respectively. Additionally, Table 7 provides the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), noticing that both have very

low values (big negative numbers). These metrics yield reliance on how well the model fits and explains the original system.

4. Table 8 displays a few measures of the error from the forecast produced.
5. Table 9 shows some descriptive statistic of the relative error with an 8×10^{-4} mean and a 6×10^{-4} median, the behavior of the error is depicted in Figure 11.

Table 6 Structure of the computed AHC model: two molecules, and 16 coefficients.

	Computed AHC model	
Molecule	1	2
τ	Cl	Cl
$\hat{a}_0$	2.727210×10^{-2}	9.269068×10^{-2}
$\hat{a}_1$	1.092679	1.004407
$\hat{a}_2$	-9.594364×10^{-2}	-1.519266×10^{-2}
$\hat{a}_3$	1.104696×10^{-4}	4.278086×10^{-4}
$\hat{a}_4$	-3.755683×10^{-4}	-5.639691×10^{-4}
$\hat{a}_5$	1.608554×10^{-09}	1.029201×10^{-09}
$\hat{a}_6$	$-3.398093 \times 10^{-10}$	$-6.563887 \times 10^{-10}$
$\hat{a}_7$	6.812609×10^{-10}	9.461997×10^{-10}
$\hat{a}_8$	$-1.551042 \times 10^{-09}$	$-3.771424 \times 10^{-10}$
$\hat{a}_9$	6.335330×10^{-10}	1.835423×10^{-10}
$\hat{a}_{10}$	$-1.204230 \times 10^{-11}$	$-6.824526 \times 10^{-11}$
$\hat{a}_{11}$	2.267883×10^{-10}	1.791191×10^{-11}
$\hat{a}_{12}$	$-2.745103 \times 10^{-10}$	3.667814×10^{-10}
$\hat{a}_{13}$	5.987430×10^{-08}	2.903713×10^{-08}
$\hat{a}_{14}$	7.366559×10^{-10}	$-1.010653 \times 10^{-09}$
$\hat{a}_{15}$	0	0

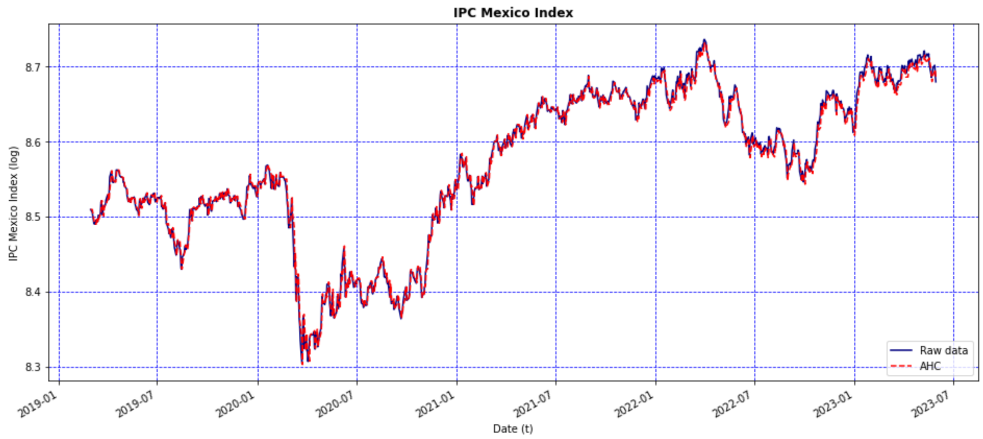

Fig. 8 Graph of the forecast of the IPC computed with the AHC model using a test set size of 25%.

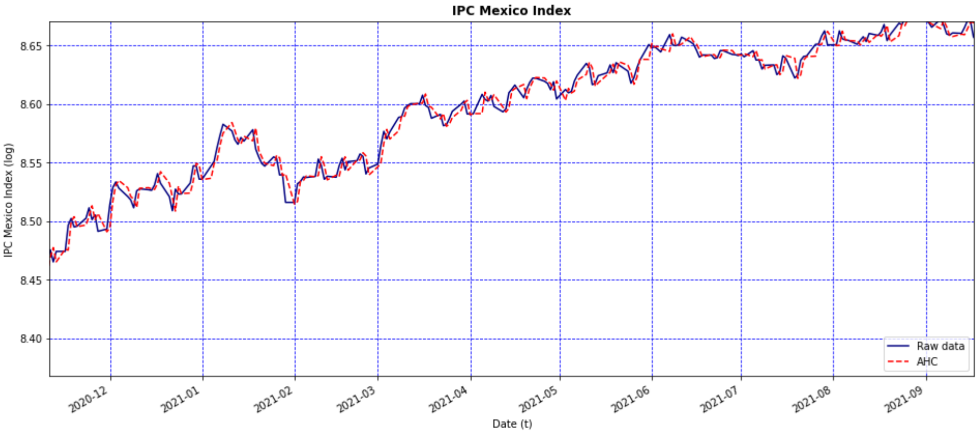

Fig. 9 Zoom of the forecast of the IPC computed with the AHC model.

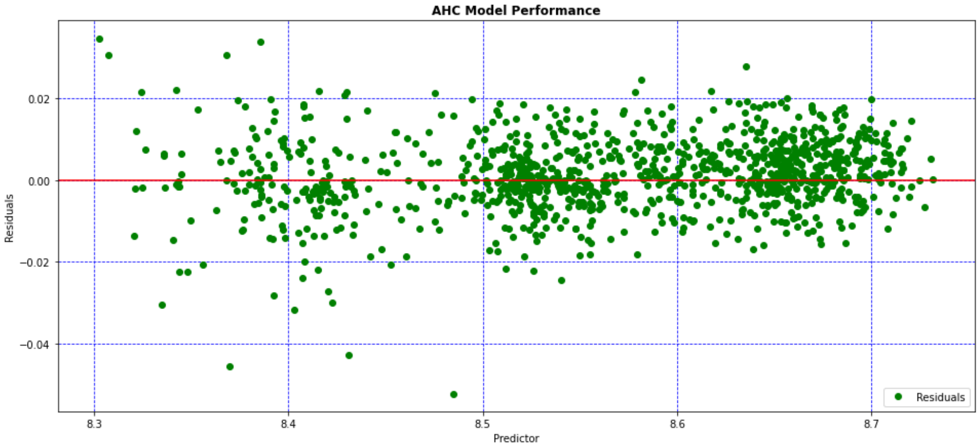

Fig. 10 Residuals of the AHC model.

Table 7 Statistical measures of the sum of squares, the R-square, the Adjusted R-square, the AIC, and the BIC of the AHC model.

Model Performance						
RSS	SSR	TSS	R-square	Adj. R-square	AIC	BIC
0.0903	11.0074	11.0976	0.9919	0.9918	-10,431.5535	-10,411.5095

Table 8 Error measures of the AHC model.

Forecast Error		
MSE	RMSE	MAPE (Relative Error)
2.03×10^{-05}	4.5122×10^{-03}	0.0008

Table 9 Descriptive statistics of the relative error.

Relative Error (MAPE)						
Mean	Median	SD	MAD	Max	Min	Range
0.0008	0.0006	0.0007	0.0005	0.0062	1.2754×10^{-07}	0.0062

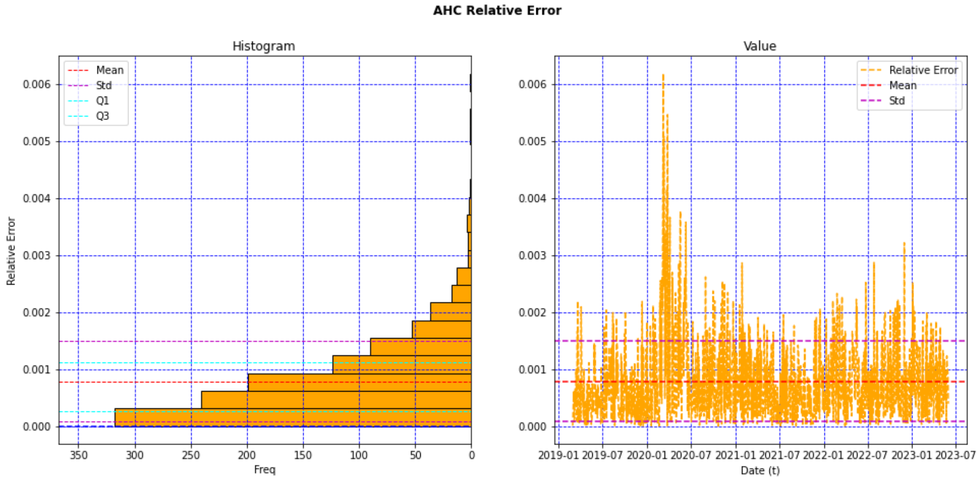

Fig. 11 Behavior of the relative error of the AHC model.

5.3 Benchmark AHC vs. AHN performance

Ayala et al. in [27] give an example of the employment of the AHN-algorithm to produce the forecast of the exchange rate BRIC currencies to USD; particularly, the Brazil Real to USD (BRL/USD) is illustrated. In this regard, the results presented by the previous authors are used as a benchmark, since the FX prediction can be considered a very similar problem to the one addressed in this research. The data used in [27] correspond to the historical data of monthly frequencies from July 1997 to December 2015 for a total of 221 samples. The values were normalized, split in sizes of 70% for training and 30% for the testing sets, and then all the input data were provided to the AHN-algorithm for the respective training process, where an organic structure is computed to produce the model f, that corresponds to the value Δy_t between y_t and y_{t-1}.

Consequently, by comparing the results from both cases, it can be stated that, for this instance, the model f of the system obtained from the AHC-algorithm is more complex and robust in contrast to the model f computed by the AHN-algorithm, based on the following elements:

1 It is well known that in the everyday operation of stock markets, the behavior of a market index has far more volatility (in a riskier financial market), than the behavior of exchange rate markets; thus, a stock market index is more difficult to predict.

2 The model built by the AHC-algorithm contemplated two-lags of the historic data of the stock market index, and at least two PCs of the seven additional MEVs that are considered; in contrast, for the prediction of the FX, the AHN-algorithm only used two-lags of the historic value.

3 The data used for the AHC model are based on a daily frequency reported value of the index and the other seven MEVs from June 2006 to May 2023, which constitutes a dataset of 4,435 samples for eight variables, making a total of 35,480 values; in contrast, the AHN model is fed with the historical data of the BRL/USD on monthly frequencies from July 1997 to December 2015, which constitutes a dataset with a total of 221 values (221 samples for one variable).

4 As a crucial fact, the AHC-algorithm has been designed to allow for a more adaptable, dynamic, and reconfigurable topology, when computing model f, making it more convenient for more complex systems. These properties can be appreciated in the following example: to compute a model -taking into account a less volatile variable-, the AHN-algorithm performed at least ten iterations to build a compound of ten molecules to produce a forecast of the BRL/USD exchange rate; in contrast, the AHC-algorithm used two iterations to create a compound of two molecules to produce a forecast of the IPC index. This is a 5:1 ratio in the number of iterations.

5 Furthermore, we compared the computational complexity of both algorithms. In this regard, the reported time complexity of the AHN-algorithm [15] is $O(c_{max} n_{max} q \ln 1/\epsilon)$ with $q \geq n_{max} \geq c_{max} \geq 2$, and a small value $\epsilon > 0$, where q is the number of samples, c_{max} is the number of compounds, and ϵ is the tolerance value. Nonetheless, this computational complexity is measured based on the time complexity of $O(C^2 N)$ for least squares estimates, where C is the dimensionality of the data, and N is the number of training samples. However, as reported in [27], the AHN-algorithm finds the parameters of each molecule using a Vandermonde matrix (polynomial interpolation), which generally requires a time complexity proportional to $O(n^3)$. In consequence, considering the number of features and the number of molecules, the time complexity of the AHN-algorithm for building the structure of the compound is $O(2(3k)^3 + (n_{max} - 2)(2k)^3)$, where k is the number features. The time complexity of the AHC-algorithm is $O(5n_{max} * 10^3 + n_{max} * 15^3)$, as reported in Section 4.5. Table 10 shows the evaluation of the time complexity for the AHN-algorithm in terms of total number of steps; in contrast, Table 11 shows the evaluation of the time complexity for the AHC-algorithm. Based on these results, we can claim that the AHC algorithm is more convenient to model more complex systems. In the AHC-algorithm case, PCA is applied during the data preparation step; hence, the MNLR model is computed considering a fixed number

of coefficients. As a result, the time complexity does not depend on k, keeping its value constant regardless of the number of features used to build the model.

6 The AHC-algorithm yields better results (again, using fewer iterations to predict a more volatile variable). The graph of the forecast provided by the AHC model behaves much better, the error is smaller (0.0102 vs. 0.0008) and the model has a very high R-square (0.99), no R-square value is provided for the AHN model.

.
Table 10 AHN-algorithm computational complexity $O(2(3k)^3 + (n_{max} - 2)(2k)^3)$ in terms of the total number of steps, k corresponds to the number of features.

AHN-algorithm computational cost				
n_{max} vs k	1	2	3	4
2	54	3,760	813,186	416,353,664
3	62	4,272	923,778	472,976,768
4	70	4,784	1,034,370	529,599,872
5	78	5,296	1,144,962	586,222,976
6	86	5,808	1,255,554	642,846,080
7	94	6,320	1,366,146	699,469,184
8	102	6,832	1,476,738	756,092,288
9	110	7,344	1,587,330	812,715,392
10	118	7,856	1,697,922	869,338,496

.
Table 11 AHC-algorithm computational complexity $O(5n_{max} * 10^3 + n_{max} * 15^3)$ in terms of the total number of steps, k corresponds to the number of features.

AHC-algorithm computational cost				
n_{max} vs k	1	2	3	4
2	16,750	16,750	16,750	16,750
3	25,125	25,125	25,125	25,125
4	33,500	33,500	33,500	33,500
5	41,875	41,875	41,875	41,875
6	50,250	50,250	50,250	50,250
7	58,625	58,625	58,625	58,625
8	67,000	67,000	67,000	67,000
9	75,375	75,375	75,375	75,375
10	83,750	83,750	83,750	83,750

5.4 Model Comparison vs. ARIMA performance

Besides comparing the AHC-algorithm performance against the AHN-algorithm (the original topology defined for the AON class) as we did in the previous subsection, here we compare the outcomes of our proposed model against the results obtained from a classical statistic method like ARIMA; for this objective, the ARIMA model was implemented through the statsmodels library. To have a valid framework for contrasting the results, we kept consistency with the experiments illustrated in Section 5.2; in this respect, we preprocessed the data using the same steps described in Section 3.1 and split the dataset into two subsets based on the DateTime, employing the first 75% of the observations for training, and remaining 25% for testing.

The definition of the parameters (p, d, q) of the ARIMA model is done as follows: the value of the component p used for the lag order from the mathematical term of the AR is set to 2 using heuristics. The value of the parameter d used for the degree of differencing the data to make it stationary, is computed with the aid of the Augmented Dickey-Fuller (ADF) test included in the statsmodel library; the data are differentiated as many times as necessary until it reached an ADF statistic equal to or below -1 (see Table 12). Finally, the value of the component q used for the size window or order of the mathematical term of the Moving Average (MA), is also computed with the aid of the autocorrelation and partial autocorrelation test included in the statsmodels library.

Table 12 Values of the ADF Test after data were differentiated.

ADF Test	
ADF Statistic:	-26.0689
p-value:	0.0000
Critical Values:	
1%:	-3.432
5%:	-2.862
10%:	-2.567

Once the components (p, d, q) are quantified, the ARIMA method is applied using these parameters; the reliability of the model is assessed via the relative error (MAPE). Next we evaluate the performance of the model, just as we did for the AHC model:

1 Figure 12 illustrates the graph with a blue curve corresponding to the original IPC Mexico values y_t from the testing set, and a red curve depicting the estimated values $\hat{y}$ using ARIMA.

2 To evaluate the model, Table 13 shows within other computed statistical measures for the forecast, the R-square and Adjusted R-square, both with a value of 0.9979 respectively. Additionally, Table 13 provides the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), noticing that both have very low values (big negative numbers). These metrics yield reliance on how well the model fits and explains the original system.

3 Table 14 displays a few measures of the error from the forecast produced.

4 Table 15 shows some descriptive statistic of the relative error with an 8×10^{-4} mean and a 5×10^{-4} median.

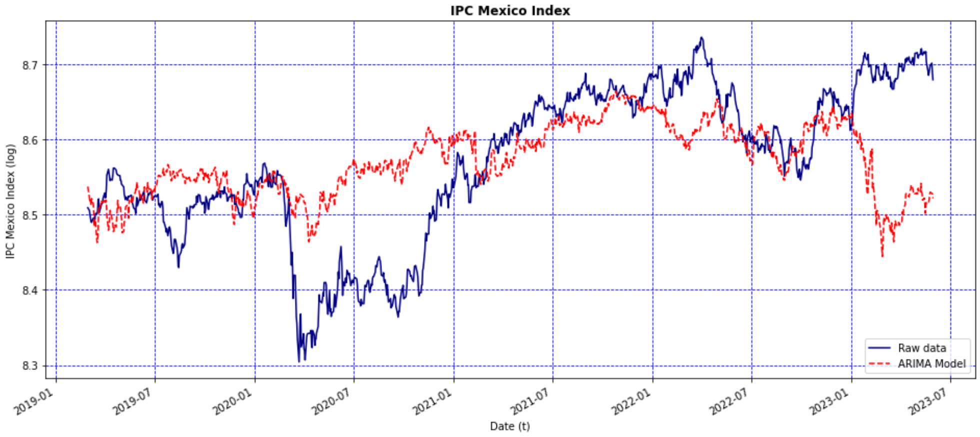

Fig. 12 Graph of the forecast of the IPC computed with the ARIMA model using a test set size of 25%.

Table 13 Statistical measures of the sum of squares, the R-square, the Adjusted R-square, the AIC, and the BIC of the ARIMA model.

Model Performance						
RSS	SSR	TSS	R-square	Adj. R-square	AIC	BIC
0.3015	140.0247	140.3262	0.9979	0.9979	-30,927.6973	-30,952.1354

Table 14 Error measures of the ARIMA model.

Forecast Error		
MSE	RMSE	MAPE (Relative Error)
0.0001	0.0095	0.0008

Contrasting the outcomes acquired from the AHC and the ARIMA models can be attended that both methods had provided similar results. A MAPE comparison is provided in Table 16 (columns extracted from Tables 9 and 15). Moreover, it can be noticed that the AIC and BIC coefficients of the ARIMA model are smaller, denoting that the model could be more adequate or that explains better the original system; regardless, the AHC-algorithm has the subsequent advantages:

Table 15 Descriptive statistics of the relative error.

Relative Error (MAPE)						
Mean	Median	SD	MAD	Max	Min	Range
0.0008	0.0005	0.0009	0.0006	0.0105	0.0	0.0105

- As a classic statistical model, the ARIMA method can provide high-average performance solutions; however, as a classic analytic method its implementation can be complex with inherent elevated computational costs.
- The development of the forecast graph from the test set delivered by the AHC-algorithm behaves much better than the graph provided by the ARIMA model since, as it can observed in Figure 12, the gaps between the raw data and the model are more noticeable.

Considering these benefits, and the fact that the difference between the error from each method is negligible, we conclude that the AHC-algorithm can be an alternative to classic statistical models.

Table 16 Comparison of the statistics of the MAPE.

Statistics Comparison of the Relative Error (MAPE)							
	Mean	Median	SD	MAD	Max	Min	Range
AHC-algorithm	0.0008	0.0006	0.0007	0.0005	0.0062	1.2754×10^{-07}	0.0062
ARIMA	0.0008	0.0005	0.0009	0.0006	0.0105	0.0	0.0105

5.5 Cross-Reference Comparison

We present here a cross-reference comparison against the results we obtained using the AHC-algorithm delivered in Section 5.2. Table 17 reviews some of the outcomes found within the literature mentioned earlier in Section 2.1 and compares them to our results.

From our assessment, we identified that some articles like Ye [7] and Sunki et al. [8], use a specific stock price (security) such as Google or Netflix; in contrast, in works like Shi et al. [9], Aliyev et al. [10], Singh [11], and Harahap et al. [12], they report the usage of an index. Anyhow, these works only consider the historical data of the analyzed variable as the only input to produce the forecast. Respectively, to yield the forecast of the stock market, we pass to the AHC-algorithm the historical data of the index with seven macroeconomic variables as part of the parameters; moreover, in [13] the authors report the employment of the AHC-algorithm for the forecast of further indices with the same considerations. Likewise, it can be observed from Table 17 that the R-square of 0.9919 obtained in the forecast of the IPC using the AHC model in this research, is comparable to other state-of-the-art methods like the ARIMA-GARCH model with an R-square of 0.977 employed for the forecast of the RTS, or the LSTM method applied for the prediction of the Google stock with an R-square of 0.9965.

Table 17 Cross-reference comparison vs. the AHC model's results.

Results from the testing sets					
Index	Method	Error	R-square	Period	Testing Set Size
IPC	AHC	$4.5122 \times 10^{-03*}$	0.9919	2006 - 2023	25%
GOOG[1]	SVR	1.10335*	0.9979	2014 - 2023	20%
GOOG[1]	GRUs	2.4863*	0.9811	2014 - 2023	20%
GOOG[1]	LSTM	0.0074*	0.9965	2014 - 2023	20%
GOOG[1]	XGBoost	2.4301*	0.9820	2014 - 2023	20%
TSLA[1]	SVR	23.9910*	0.9953	2020 - 2021	20%
TSLA[1]	GRUs	41.7746*	0.8685	2020 - 2021	20%
TSLA[1]	LSTM	0.0055*	0.9952	2020 - 2021	20%
TSLA[1]	XGBoost	44.122*	0.8533	2020 - 2021	20%
NFLX[2]	ARIMA	7.8919*	NA	2019 - 2021	15%
NFLX[2]	LSTM	10.3376*	NA	2019 - 2021	15%
NFLX[2]	FBProphet	9.1186*	NA	2003 - 2021	NA
S&P 500[3]	XGBoost	$5.5235 \times 10^{08*}$	0.316	2017 - 2023	15%
RTS[4]	ARIMA-GARCH	35.93*	0.977	2000 - 2022	10%
RTS[4]	LSTM	14.91*	0.996	2000 - 2022	10%
NIFTY 50[5]	ANN	36.865*	0.999	1996 - 2021	20%
NIFTY 50[5]	SGD	42.456*	0.999	1996 - 2021	20%
NIFTY 50[5]	SVM	68.327*	0.998	1996 - 2021	20%
NIFTY 50[5]	AdaBoost	2277.710*	-0.930	1996 - 2021	20%
NIFTY 50[5]	RF	2290.890*	-0.952	1996 - 2021	20%
NIFTY 50[5]	KNN	2314.720*	-0.993	1996 - 2021	20%
N225[6]	SVR	NA	0.81	2016 - 2019	10%
N225[6]	DNN	NA	0.79	2016 - 2019	10%
N225[6]	BPNN	NA	0.82	2016 - 2019	10%
N225[6]	SVR	NA	0.58	2016 - 2019	20%
N225[6]	DNN	NA	0.58	2016 - 2019	20%
N225[6]	BPNN	NA	0.56	2016 - 2019	20%
S&P 500[7]	GA	0.1347†	0.5027	2006 - 2023	15%
DJIA[7]	GA	0.0263†	0.5228	2006 - 2023	15%
FTSE[7]	GA	0.0531†	0.4959	2006 - 2023	15%
N225[7]	GA	0.0064†	0.6111	2006 - 2023	15%
CAC[7]	GA	0.0462†	0.5149	2006 - 2023	15%
S&P 500[7]	AHC	0.0049†	0.729	2006 - 2023	15%
DJIA[7]	AHC	0.0007†	0.9777	2006 - 2023	15%
FTSE[7]	AHC	0.0063†	0.7074	2006 - 2023	15%
N225[7]	AHC	0.0064†	0.6111	2006 - 2023	15%
CAC[7]	AHC	0.0038†	0.8317	2006 - 2023	15%

[1]Results reported in [7]. [2]Results reported in [8]. [3]Results reported in [9]. [4]Results reported in [10]. [5]Results reported in [11]. [6]Results reported in [12]. [7]Results reported in [13]. *Reported as RMSE. †Reported as Relative Error. (NA) Value not provided.

6 Conclusions & Future Work

In this work, the main aspects of the AON framework are studied toward the definition of a new algorithm based on this machine learning class; this analysis led to the review of the five levels of the AON framework: implementation, mathematical model, heuristics, interactions, and components. Furthermore, we point out the necessity of AON for using a functional group to determine its topological configuration along the implementation. Also, we recalled how AON has been implemented -before this work- using AHN as the unique formal topology defined so far. Along the review, we mentioned the two mechanisms that AHN employs to optimize a cost-energy function while forming the structures of organic compounds, as part of the process to model a given system. These mechanisms are: the utilization of LSR to define the structure of each molecule, and the employment of GD to optimize the position and number of molecules in the feature space.

The AHN algorithm has a very high computational cost due to the use of GD and is not capable of dealing well with big data. In this work, we aimed at defining a new topology to overcome these disadvantages. We proposed an alternate functional group different from hydrocarbons to define a new topology. To meet our goal, the concept of Artificial Halocarbon Compounds or AHC-algorithm is introduced as a supervised machine learning algorithm based on the AON framework.

Within the considerations behind its design, the AHC-algorithm complies with the following properties:

- It is based on the third type of AON framework: a hybrid algorithm that defines a structure based on a mixture of chemically inspired and artificial basis algorithms.
- The conception of the new arrangement was focused on changes in the layer of components and interactions from the AON framework.
- For the new algorithm, GD was arrogated and substituted by the K-means method, as part of the mechanism to form organic compounds.
- The halocarbons functional group was selected to define a new topology, specifically from halogenations of anthracene, from the haloaromatic type.
- As a dynamic topology, the algorithm can conduct reactions using the variety of atoms from the halogens to build an organic structure.
- The halogenation process is ruled by the entropy-property, to assure the computation of the lowest cost-energy function.
- The new topology is capable of producing time series forecasting.

The AHC-algorithm was tested by modeling and producing a forecast of the IPC Mexico stock market index, using 17 years of data, and with a set of parameters that allowed a combination of 224 models. The final forecast model provided very encouraging results on the testing set, R-square equal to 0.9919, and a mean relative error of 8×10^{-4}. When comparing the AHC-algorithm to the original AHN-algorithm, we claim that due to the high computational complexity of the latter, the new topology is more convenient when modeling more complex systems; being this characteristic the main contribution of the AHC-algorithm, allowing it to be a more adaptable, dynamic, and reconfigurable topology, when computing model *f*. In future work, we plan to test the algorithm using other stock market indices and compare it with other state-of-the-art methods employed for financial analysis.

Acknowledgments. The authors would like to acknowledge the financial support of Tecnólogico de Monterrey, Mexico, in the production of this work.

Declarations

- Funding - The authors declare the publication of this article is funded by Tecnologico de Monterrey.
- Conflicts of interest/Competing interests - Not applicable, the authors have no relevant financial or non-financial interests to disclose.
- Ethics approval - Not applicable, the authors declare that our research does not involve humans and/or animals.
- Consent to participate - The authors declare that all authors gave their consent to participate.
- Consent for publication - The authors declare that all authors gave their consent for publication.
- Availability of data and material - The authors declare that the data used in this work is available at:
  https://ieee-dataport.org/documents/datasets-stock-market-indices
  https://dx.doi.org/10.21227/yvfx-n484
- Code availability - The authors declare that the code used in this work is available at:
  https://github.com/egonzaleznez/ahc
- Authors' contributions - The authors declare that all authors contributed equally to this work.

References

[1] Soler-Dominguez, A., et al.: A survey on financial applications of metaheuristics. ACM Computing Surveys (Volume: 50) (2017)

[2] Elliott, G., *et al.*: Handbook of Economic Forecasting, 1st edn. Elsevier Ltd., UK (2013)

[3] Christodoulaki, E., *et al.*: Technical and sentiment analysis in financial forecasting with genetic programming. In: 2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr) (2022). https://doi.org/10.1109/CIFEr52523.2022.9776186

[4] Sheta, A.F., *et al.*: Evolving stock market prediction models using multi-gene symbolic regression genetic programming. ICGST, USA (2015)

[5] Stoean, C., *et al.*: Deep architectures for long-term stock price prediction with a heuristic-based strategy for trading simulations. PLOS ONE (Volume: 14) (2019) https://doi.org/10.1371/journal.pone.0223593

[6] González-Núñez, E., Trejo, L.A.: Artificial organic networks approach applied to the index tracking problem. In: MICAI 2021. LNCS (LNAI). Springer, Cham (2021)

[7] Ye, S.: Applying ensemble learning to multiple stock pricepredictions: A comparative study (2024) https://doi.org/10.54254/2755-2721/50/20241501

[8] Sunki, A., SatyaKumar, C., . Surya Narayana, G., Koppera, V., Hakeem, M.: Time series forecasting of stock market usingarima, lstm and fb prophet (2024) https://doi.org/10.1051/matecconf/202439201163

[9] Shi, B., Tan, C., Yu, Y.: Predicting the s&p 500 stock market with machine learningmodels (2024) https://doi.org/10.54254/2755-2721/48/20241621

[10] Aliyev, F., et al.: Applying deep learning in forecasting stock index: Evidence from rts index. (2023) https://doi.org/10.1109/AICT55583.2022.10013496

[11] Singh, G.: Machine learning models in stock market prediction (2022) https://doi.org/10.35940/ijitee.C9733.0111322

[12] Harahap, L.A., et al.: Nikkei stock market price index prediction using machine learning. (2020) https://doi.org/10.1088/1742-6596/1566/1/012043

[13] González-Núñez, E., Trejo, L.A., Kampouridis, M.: A comparative study for stock market forecast based on a new machine learning model. Big Data and Cognitive Computing **8**(4) (2024) https://doi.org/10.3390/bdcc8040034

[14] Ponce, H., et al.: Artificial Organic Networks: A New Algorithm Bio-Inspired on Organic Chemistry. Accent Social and Welfare Society (2012)

[15] Ponce, H., *et al.*: Artificial Organic Networks: Artificial Intelligence Based on Carbon Networks, 1st edn. Springer, Cham (2014)

[16] Ponce, H., *et al.*: Development of Fast and Reliable Nature-Inspired Computing for Supervised Learning in High-Dimensional Data. Nature Inspired Computing for Data Science, pp. 109–138. Springer, Cham (2019)

[17] González-Núñez, E., Trejo, L.A.: Datasets of Stock Market Indices. https://dx.doi.org/10.21227/yvfx-n484 (2024)

[18] Virtanen, P., et al.: Scipy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods (Volume: 17) (2020)

[19] Pedregosa, F., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research (Volume: 12) (2011)

[20] Seabold, S., *et al.*: statsmodels: Econometric and statistical modeling with Python. In: 9th Python in Science Conference (2010)

[21] González-Núñez, E.: AHC related code. Retrieved from: https://github.com/egonzaleznez/ahc

[22] Chiu, D.Y., *et al.*: U.s.a. s&p 500 stock market dynamism exploration with moving window and artificial intelligence approach. In: The 7th International Conference on Networked Computing and Advanced Information Management. IEEE, Gyeongju, South Korea (2011)

[23] Chiu, D.Y., *et al.*: Applying artificial immune algorithm to explore the seasonal effect in the stock market. In: International Conference on Software Intelligence Technologies and Applications & International Conference on Frontiers of Internet of Things 2014. IET, Hsinchu, Taiwan (2014)

[24] Chacón, H., et al.: Improving financial time series prediction accuracy using ensemble empirical mode decomposition and recurrent neural networks (2020) https://doi.org/10.1109/ACCESS.2020.2996981

[25] Hou, X., et al.: An enriched time-series forecasting framework for long-short portfolio strategy (2020) https://doi.org/10.1109/ACCESS.2020.2973037

[26] Ordóñez, J.M.: Predicción del comportamiento de los mercados bursátiles usando redes neuronales. Technical report, Universidad de Sevilla, Sevilla, España (2017)

[27] Ayala-Solares, J.R., et al.: Supervised learning with artificial hydrocarbon networks: An open source implementation and its applications (2020) https://doi.org/10.48550/arXiv.2005.10348

[28] Meislich, H., *et al.*: Química Orgánica, 3rd edn. McGraw-Hill, Colombia (2001)

[29] Greiner, W., *et al.*: Thermodynamics and Statistical Mechanics. Springer, USA (1995)