Economics

An intelligent approach for predicting stock market movements in emerging markets using optimized technical indicators and neural networks

--Manuscript Draft--

Manuscript Number:	ECONJOURNAL-D-23-00153
Full Title:	An intelligent approach for predicting stock market movements in emerging markets using optimized technical indicators and neural networks
Article Type:	Research Article
Keywords:	ETF; Emerging markets; Neural networks; Feature selection; Data mining
Manuscript Region of Origin:	MEXICO
Abstract:	Big data analytic techniques associated with machine learning algorithms play an increasingly important role in various application fields, including stock market investment. This work presents a comprehensive data analytics process (CRISP-DM) applied to the selection of the best features to aid in predicting the trend direction for some ETFs based on the financial and economic features of emerging markets. We used diverse statistical methods to identify the most salient technical indicators for the time series problem, attaining a better trend prediction and reducing in 95% the computational cost of the proposed neural network model.
Manuscript Classifications:	3.7.5: Neural Networks and Related Topics; 3.8.6: Large Data Sets: Modeling and Analysis; 3.9.4: Computational Techniques • Simulation Modeling

DE GRUYTER

Editorial
Research Article

Alma Rocío Sagaceta-Mejía, Máximo Eduardo Sánchez-Gutiérrez*, and Julián Alberto Fresán-Figueroa

An intelligent approach for predicting stock market movements in emerging markets using optimized technical indicators and neural networks

Abstract: Big data analytic techniques associated with machine learning algorithms play an increasingly important role in various application fields, including stock market investment. This work presents a comprehensive data analytics process (CRISP-DM) applied to the selection of the best features to aid in predicting the trend direction for some ETFs based on the financial and economic features of emerging markets. We used diverse statistical methods to identify the most salient technical indicators for the time series problem, attaining a better trend prediction and reducing in 95% the computational cost of the proposed neural network model.

Keywords: ETF, Emerging markets, Neural networks, Feature selection, Data mining

1 Introduction

An Exchange Traded Fund (ETF) is a relatively recent financial innovation that provides an alternative method for indirectly investing in international equities. They are similar to conventional investment funds in that the market value is close to the value of the underlying assets, and they are listed on stock exchanges. This Fund allows the investors to implement different investment strategies that

Alma Rocío Sagaceta-Mejía, Departamento de Física y Matemáticas, Universidad Iberoamericana
***Corresponding author: Máximo Eduardo Sánchez-Gutiérrez,** Colegio de Ciencia y Tecnología, Universidad Autónoma de la Ciudad de México
Julián Alberto Fresán-Figueroa, Departamento de Matemáticas Aplicadas y Sistemas, Universidad Autónoma Metropolitana Unidad Cuajimalpa

incorporate diverse geographic and economic activities rather than investing in local handmade portfolios. [Deville(2008), Antoniewicz and Heinrichs(2014)].

The most popular ETFs are designed to reflect stock indices such as the S&P 500, Nasdaq and Dow Jones. The ease of administration, lower management costs and taxes, and allowing investors to enter and exit investment positions with minimal risk are benefits of ETFs. In addition, literature shows that ETFs offer greater diversification benefits than conventional local mutual funds [Miralles-Quirós et al.(2019)]. In this regard, according to [Hegde and McDermott(2004)], the success of ETFs is due to the simplicity with which investors may benefit from portfolio diversification at lower transaction costs than stock investment portfolios.

In this work, we center our attention on emerging markets ETFs, which refer to countries becoming developed markets, including countries across the Asia-Pacific region and Latin America, such as Brazil, Chile, Mexico, China, and India. Emerging economies comprise countries whose characteristics are: rapid growth, high productivity levels, increase in middle-class interest, high volatility, liquidity in local debt and equity markets, and growth potential. These ETFs are attractive for investors since emerging economies tend to grow faster than their developed counterparts, as can be seen in the information of how these markets have grown over the last decade (see Figure 1). Our focal point is the selection of features to predict the trend of two ETFs of emerging markets: iShares MSCI Chile ETF (ECH) and iShares MSCI Brazil ETF (EWZ) using technical and statistical analysis to later compare them against iShares Core S&P 500 ETF (IVV).

Fig. 1: In this plot, we show the performance of two different ETFs: the iShares MSCI World ETF (URTH) that replicates an index composed of developed market equities and the iShares MSCI Emerging Markets ETF (EMM) that replicate an index composed of large- and mid-capitalization emerging market equities.

Due to the energetic, non-linear, non-parametric, and chaotic properties of stock information, stock market prediction has been a problem for analysts and researchers

during the last decade. Some studies use regression methods [Chen and Chen(2015), Jiang et al.(2020), Zhang et al.(2020)] to forecast long-term stock costs or profits, while others use categorization techniques to predict patterns of stock cost development [Ananthi and Vijayakumar(2021), Cagliero et al.(2020)].

Several techniques for approaching the stock trend prediction problem have been developed in recent years. Originally, traditional statistical techniques were used, but with the development of artificial intelligence, machine learning algorithms help deal with intricate market data with the development of artificial intelligence [Tang et al.(2019)]. A commonly used machine learning model in the stock market prediction are neural networks, that have proven useful for discover the non-linear and non-additive data relations, attaining superior results [De Haan et al.(2016)]. Even when the conventional neural network may not be the best architecture for every problem, it can successfully capture inner data relations, providing helpful information to aid in the dimensional analysis. Nevertheless, specific sort of machine learning strategies such as Restricted Boltzmann Machine (RBM) [Liang et al.(2017)] Recurrent Neural Network (RNN) [Zhao et al.(2021)], Convolutional Neural Network (CNN) [Sezer and Ozbayoglu(2018), Barra et al.(2020)], and Long Short-Term Memory Networks (LSTM) [Nelson et al.(2017), Chen et al.(2019)] have shown remarkable performance in stock prediction.

Even though there have been studies to evaluate which models are more suitable for stock market prediction [Chen et al.(2021a)], most of them make use of a small number of stock features, such as highest and lowest price, opening and close price, and some financial indicators as rate of change, simple moving average, exponential moving average, hull moving average, relative strength index, Williams's indicator or change momentum oscillator, among others. Nevertheless, there exist many indicators of stock performance [O'Hara et al.(2000)]. For instance, the package *Pandas Technical Analysis (Pandas TA)* contains over 200 tunable technical indicators. In this article, we propose an analysis of feature selection methods based on CRISP-DM methodology to identify the most salient features, aiming to discriminate the best indicators, reduce the data to be processed, and assist investors in determining stock market behavior. The paper has been divided as follows: In Section 2, we describe the datasets, their technical indicators, the data pre-processing, the data mining methodology, the techniques for feature selection, the artificial intelligence model, and the experiments performed. Section 3 presents the resulting subsets of features for each ETF, their cross-validated accuracy, and their percentage gain from the baseline. Section 4 discusses the results obtained in the experiments, and future research lines are presented.

TOP SECTORS (%)					
ECH	**%**	**EWZ**	**%**	**IVV**	**%**
Financials	**21.53%**	**Materials**	**26.08%**	**Info. Tech**	**27.70%**
Materials	**21.28%**	**Financials**	**23.96%**	**Health Care**	**13.36%**
Utilities	**18.73%**	**Energy**	**12.83%**	**Cons. Disc.**	**12.00%**
Cons. Stap.	**13.92%**	**Cons. Stap.**	**10.14%**	**Communication**	**11.19%**
Energy	**8.34%**	**Cons. Disc.**	**8.44%**	**Financials**	**10.89%**
TOTAL	**83.8%**	**TOTAL**	**81.45%**	**TOTAL**	**75.14%**

Tab. 1: Market exposure of ETFs (ECH, EWZ and IVV)

2 Materials and methods

In this section, we describe the datasets and how they were treated before performing the experimentation. We centered our study on three datasets obtained from Yahoo Finance, calculated several technical features, and processed all the data obtained with CRISP-DM methodology. After that, we carried out a feature analysis based on diverse techniques to identify and rank the most salient features, which were finally fed to a multi-layer perceptron (MLP) to evaluate the performance of selected features.

2.1 Stocks analyzed

To prevent biases in the results, the analyzed datasets span from 12/01/2009 to 01/01/2020. This made it possible not to consider the economic alterations caused by the atypical phenomena during the pandemics, which may almost certainly influence the results. We studied three Exchange Traded Funds (ETF) from the BlackRock company: iShares MSCI Chile ETF (ECH), iShares MSCI Brazil ETF (EWZ), and iShares Core S&P 500 ETF (IVV), the former two are from emerging markets, while the last one replicates the S&P 500 index.

Table 1 shows these ETFs' main market exposure areas, while Figure 2 depicts their Opening price during the considered period.

	ECH	EWZ	IVV
Minimum	29.30	17.49	103.5
1st Quartile	40.35	36.69	137.9
Median	46.48	43.63	199.3
Mean	50.10	47.25	196.7
3rd Quartile	59.84	56.30	244.7
Maximum	80.25	81.41	325.2

Tab. 2: Summarized data for the opening price *Open*.

Fig. 2: Behavior of Open values for ECH, EWZ and IVV

The data that can be obtained from Yahoo Finance for each period are: the opening price *Open*, the highest price *High*, the lowest price *Low*, the closing price *Close*, the number of transactions *Volume*, the adjusted close price for splits, the dividends yield, and capital gain distributions *Adjusted close*. The data summarized for the opening price is presented in Table 2.

We use a classification strategy to predict a qualitative variable Γ such that

$$\Gamma(t) = \begin{cases} 1 & \text{if } Open(t) - Open(t-1) > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{1}$$

The response variable Γ corresponds to the class label of the day. Figure 3 portrays the sum of variable Γ during the considered period for each ETF.

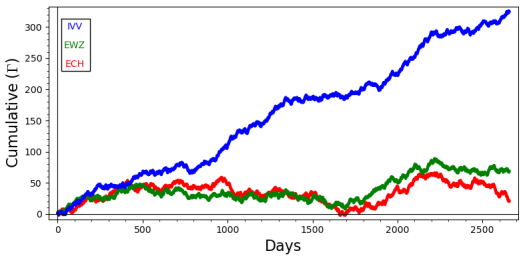


Fig. 3: The Γ Cumulative Movement for ECH, EWZ and IVV

2.2 Methodological approach based on data mining

The Cross-Industry Standard Process for Data Mining (CRISP-DM) specifies the stages in which the datasets are processed [Shearer(2000)]. CRISP-DM covers six stages: understanding the data and the domain of the problem, the data preparation, the construction of the model, the evaluation of the proposed model, and finally, the presentation of the results.

Data preparation is the main stage in any data mining methodological approach. In this stage, the raw data is cleaned, merged, scaled, and feature engineered to enhance the quality of the information, improving the machine learning algorithm's behavior [Sun et al.(2017), Jamshed et al.(2019)]. The data preparation stage can be seen as converting inconsistent and incomplete real-world data into a readable format. In this work, the data preparation stage was primarily based on the next activities:

Technical indicators

We used the Pandas Technical Analysis Library (Pandas TA) to calculate the technical indicators, which leverages the Pandas package with several customizable technical indicators, utility functions, and candlestick patterns. We obtained 210 more features using that library (including the previously obtained. Thus, we have a dataset containing 216 features for each day, including the ones obtained from the Yahoo Finance database.

Class assignment

For each day t, we obtained the Γ function previously defined. Observe that $Gamma(t)$ indicates the delta sign of the ETF Open price. Hence, we employ a binary classification strategy to identify whether this happens, so each day's class corresponds to its Γ evaluation. In order to apply the class assignment, the order of the dataset gets decreased by one.

Data normalization

Since the scales of the features computed earlier fluctuate significantly, we use a min-max approach to convert the data linearly. Every feature's minimum values are converted to 0; afterward, the maximum values are transformed to a 1, and finally, all other values are adjusted to a decimal between 0 and 1. The formula, for each value, is given by

$$\frac{\text{value}[i] - \min}{\max - \min}.$$

If the normalization is not performed, we risk diluting the effectiveness of an equally important feature (on a lower scale) because of other features having values on a larger scale.

Data cleaning

The process of preparing raw data for analysis by removing incomplete data is known as data cleaning, and this prepares the data for the data mining process, which needs valuable information. When technical features are calculated, as in the previous step, missing or non-available data is unavoidable. For example, when calculating Simple Moving Average (SMA), it is necessary to select a range of days before the day we are calculating, hence for the initial days of the dataset, the SMA cannot be obtained; thus, the data for those will be missing. To handle this issue, we proceeded to drop each day that contained unavailable data, so the database afterward spanned from 01/01/2010 to 01/01/2020.

2.3 Multi-layer perceptron for predictive analysis

A multi-layer perceptron (MLP) was used to predict the qualitative variable Γ. An MLP can be implemented as a linear and binary classifier since it finds the most appropriate boundary between the two classes. Hence, it may discern the structural

differences between two given classes, identify the linear space separating each one, and determine the likelihood of a given data point belonging to a class.

An MLP is a neural network connecting multiple neurons or perceptrons, partitioned into the input layer, the hidden layer, and the output layer. The nodes form a directed acyclic graph, meaning that the paths connect nodes in layers from one layer to the next, as shown in Figure 4. Apart from the input ones, each neuron has a nonlinear activation function, a bias, and connecting weights which the MLP train by back-propagation in a supervised learning fashion [Ecer et al.(2020)] so that the error value can be updated in a much more successful way. The MLP used in this work is depicted in Figure 4. We use this MLP to evaluate the performance of the different subsets of technical features obtained by several feature selection approaches described below.

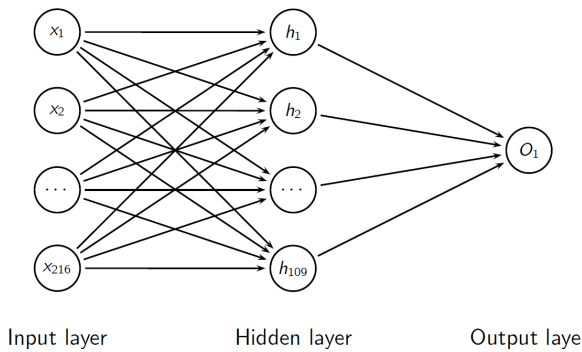


Fig. 4: Diagram of the MLP used in this work. The input layer cardinality corresponds to the number of input features; there would be 216 nodes if all features are considered. The hidden layer contains $(\text{input features} + \text{classes})/2$ nodes; if all features are considered, it would have 109 nodes. Finally, the output layer contains only one node.

2.4 Statistical measures for feature selection

As mentioned earlier, in this work, we use an approach based on exploratory analysis and reduction of the input space to improve the accuracy of multiclass classification in machine learning. This feature selection can be seen as a dimensional reduction limiting the number of variables that characterize the data. This reduction is performed by selecting the subset of features that provide more information, dropping the redundant ones to obtain a significant quantity of information from a lower-dimensional space [Reddy et al.(2020)]. In the machine learning paradigm, a

narrower input space implies a computationally efficient modeled structure since it is desirable to have input data with few variables that produce small models that generalize well. The above is especially valid for linear models where the degrees of freedom and the number of inputs are related.

To reduce the input space, we selected features after the data cleaning and scaling stages and before the predictive model's training phase. The techniques for selecting the most relevant characteristics used in this work are described below.

2.4.1 Low Variance

Low Variance feature removal is a basic and straightforward approach to feature selection. The Low Variance technique removes all characteristics whose variance does not reach the established threshold. This feature selection algorithm only works with features and not with class outputs, making it an unsupervised technique. If a feature's variance is close to zero, then a feature is approximately constant and will not improve the model's performance. In that case, it should be removed. The expression defines the variance as:

$$Var[X] = p(1-p), \tag{2}$$

where p is the probability of $X = P(X)$.

2.4.2 Chi-Squared

In this method, each feature is evaluated against the classes. For each pair of features, the Chi-Squared values are calculated; a larger Chi-Squared value commonly indicates a greater interdependence between the two attributes, this method identifies the features that are more likely to be independent of the class and thus unrelated to the classification. Since this approach is commonly used with categorical attributes, it is needed first to discretize the numeric attributes at different intervals. The expression (3) is used to calculate the Chi-Squared statistic:

$$x^2 = \sum \frac{(f_0 - f_e)^2}{f_e} \tag{3}$$

where f_0 is the observed frequency (the observed counts in the cells), and f_e is the expected frequency if no relationship existed between the variables.

2.4.3 LASSO

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization technique. Regularization is one technique to address the issue of overfitting by providing new information and, as a result, modifying the model's parameter values to induce a penalty, as can be seen in expressions:

$$LASSO = \sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{m} |w_j| \tag{4}$$

$$= RSS + \lambda \sum_{j=1}^{m} |w_j| \tag{5}$$

where

$$\hat{y_i} = y_0 + \sum_{j=1}^{m} X_{ij} w_j \tag{6}$$

The residual sum of squares (RSS) and an additional penalty for feature weights minimize the above loss function. The greater the chosen value for λ, the greater the penalty on feature weights, the more they get removed.

2.4.4 Tree-Based Feature selection

The Extra Trees Classifier is a meta-method that fits several randomized decision trees on various data set sub-samples and averages their results to enhance the prediction accuracy and counter over-fitting. Those randomized decision trees are Extremely Randomized Trees created by heavily randomizing both cut-point and attribute selection while splitting a tree node.

2.4.5 Pearson's Correlation

Pearson's correlation coefficient calculates the linear relationship between two random variables. Its values range between -1 and 1; when the coefficient value is 0, it means that the two random variables do not have a linear relationship. If the coefficient value is negative, the correlation is negative, and when the value is positive, the correlation between the two random variables is positive. This coefficient is given by:

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)\sigma_x \sigma_y} \tag{7}$$

where σ is the standard deviation of the sample.

2.4.6 Principal Feature Analysis

The Principal Feature Analysis method selects the principal features by adopting the structure of the principal components of the feature set, which retain the majority of the information, both in terms of maximum variability of the features in lower-dimensional space and minimization of the reconstruction error. This technique constructs the covariance matrix of all features and computes the eigenvectors of the matrix by applying a Principal Component Analysis (PCA). Afterward, a vector is associated with each feature, and a k-means algorithm clusters the set of vectors. The vectors closest to the centroids are identified as the principal vectors; hence the features associated with those vectors are deemed the principal features. The principal features can be considered the most dominant characteristics in each cluster and retain the least redundant information in other clusters. An in-depth description of this technique can be found in [Lu et al.(2007)].

2.4.7 Mean Absolute Difference

The mean absolute difference (MAD), as can be seen in equation (8), as the variance, is also a scale variant. This implies that the greater the MAD, the greater the discriminating power.

$$\frac{\sum_{i=1}^{N} \mid \text{value}[i] - \text{mean} \mid}{N} \tag{8}$$

where N corresponds to the total of data. With this technique, we rank the features from more discriminant to less discriminant. The value of MAD is not affected by extremely high or shallow values and non-normality.

2.4.8 Dispersion ratio

For a given (positive) feature X_i on N patterns, the arithmetic mean $\bar{x}$ and the geometric mean $\bar{x}_g$ are given by:

$$\bar{x} = \frac{1}{N}\sum_{i=0}^{N} \text{value}[i], \qquad \bar{x}_g = \sqrt[N]{\prod_{i=0}^{N} \text{value}[i]}.$$

Since $\bar{x} \geq \bar{x}_g$, with equality holding if and only if $X_{i,1} = X_{i,2} = \ldots = X_{i,N}$, then the dispersion ratio (DR) is:

$$DR = \frac{\bar{x}}{\bar{x}_g}.$$

The dispersion ratio can be used as a dispersion measure. When all the feature samples have (roughly) the same value, DR is close to 1, indicating a low relevance feature. Conversely, a higher value of DR implies a higher dispersion, thus a more relevant feature [Ferreira and Figueiredo(2012)].

2.5 Cross-validation as sampling method

Cross-validation is a model validation approach for determining how well a model's results will generalize to a new data set. Cross-validation aims to reduce issues like overfitting and underfitting by defining a data set to evaluate the model in the training phase.

This procedure split the dataset in k partitions alternating them to train and test the model, as described in Algorithm 1.

Algorithm 1 Cross-validation algorithm

1: Randomly shuffle the dataset.
2: Make a partition of the dataset into k groups

3: **for each** unique group **do**
4: Take one part as the test dataset
5: Take the remaining groups as the training dataset
6: Fit the model with the training dataset and evaluate it on the test dataset
7: **end for**

8: Average the model's accuracies

2.6 Description of the experiment's methodology

The experiments performed in this work are described in Algorithm 2.

Algorithm 2 Selection and evaluation of features

Input: ETF raw datasets
1: **for each** ETF **do**
2: Calculate each technical indicator
3: Class assignment
4: Normalization of the data
5: Cleaning of the data
6: **end for**

Input: Preprocessed ETF datasets
7: **for each** ETF **do**
8: **for each** Statistical measure **do**
9: Obtain the first quartile of the top salient features.
10: **end for**
11: Calculate the sets Selected (n)= { $f \in$ Features | f appears in at least n subsets defined in (5)}
12: **end for**

Input: Subsets of selected features
Output: $\bar{a_i}$
13: **for** $i \leftarrow 0$ **to** 8 **do**
14: **for each** K-fold cross validation partition **do**
15: Feed the MLP with the features in Selected(i).
16: Obtain the model's accuracy, a_i
17: **end for**
18: Obtain a central tendency measure, $\bar{a_i}$
19: **end for**

3 Results

The results obtained with the methodology previously described are condensed in Tables 3 and 4. Table 3 describes the results of step 11 from Algorithm 2

	Selected (n)							
ETFs	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
ECH	**216**	**101**	**72**	**39**	**20**	**10**	**4**	**0**
EWZ	**216**	**107**	**67**	**29**	**17**	**10**	**2**	**0**
IVV	**216**	**104**	**71**	**35**	**19**	**9**	**3**	**0**

Tab. 3: Number of features selected by the Algorithm

	Accuracy (%)						
ETFs	0	1	2	3	4	5	6
ECH	78.01	77.82	78.76	79.33	80.46	80.27	59.03
EWZ	76.46	75.19	75.38	75.33	76.51	77.82	51.41
IVV	77.26	77.64	77.63	77.63	77.78	78.54	71.05

Tab. 4: Median Accuracy after cross-validation

As shown in Table 4, employing the strategy described in Algorithm 2, it is possible to select a subset of features that can provide better results using less computational resources. The gain obtained by feature selection is shown in Figure 5. The features in the subset *Selected(5)* are shown in Table 5; A brief description of these features can be found in A.

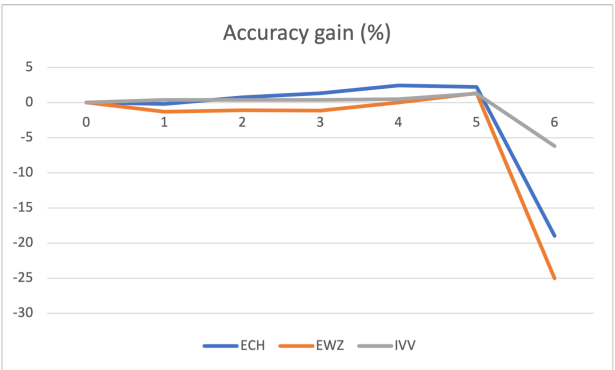


Fig. 5: Percentage gain obtained by the selection of features

ETFs	Selected(5)
ECH	AOBV_LR_2, BBP_5_2.0, BOP, CTI_12, DEC_1, EBSW_40_10, INC_1, K_9_3, J_9_3, ZS_30
EWZ	AOBV_LR_2, BBP_5_2.0, BOP, CTI_12, DEC_1, EBSW_40_10, INC_1, J_9_3, STOCHk_14_3_3, WILLR_14
IVV	BBP_5_2.0, BOP, DEC_1, INC_1, J_9_3, PVR, STOCHk_14_3_3, TTM_TRND_6, WILLR_14

Tab. 5: Features in the respective *Selected(5)* sets

4 Conclusions and future directions

Our experiments show that in each ETF analyzed, the set *Selected(5)* attain a better accuracy prediction, between 77.82% and 80.27%, than using the complete set of features while using only between 4.16% and 5.09% of the features. This may imply that a good selection of features can improve the efficiency of the computational resources while attaining similar or even better prediction results. Moreover, when we reduce the dimension of the dataset, we can construct a model with a reduced topology, with fewer freedom degrees and redundant features, which could reduce the training and prediction time. In high dimensional problems, it is always important to ask ourselves whether every input is relevant and to which extent the features contribute to determining the trend the model is attaining. The methodology used in this work may help to improve the prediction used by other machine learning techniques or in other related problems, as the one proposed by [Chen et al.(2021b)].

The indicators calculated by the *Pandas-TA* package [Johnson(2021)] belong to the following categories: Candles, Cycles, Momentum, Overlap, Performance, Statistics, Trend, Utility, Volatility, and Volume. The *Selected (5)* feature subset used for the prediction task shares similar characteristics across different ETFs. The distribution of the categories for the *Selected (5)* feature subset is shown in Table 6. Even though the distribution is not the same for each ETF, it is important to note that there are similarities. This may imply that the methodology proposed in this work enables the analysis of emerging markets and non-emerging markets ETFs.

	ECH	EWZ	IVV
Cycles	1	1	0
Momentum	4	5	4
Statistics	1	0	0
Volatility	1	1	1
Volume	1	1	1
Trend	2	2	3

Tab. 6: Number of features selected by each ETF in accordance with the feature's categories

As shown in Figure 5, there is an improving tendency of the prediction results as we approach $n = 5$, but on $n = 6$, the prediction accuracy drops to 26.46%,

30.08%, and 9.53% for each ETF, respectively. We believe this reduction is due to the lack of information of the *Selected (6)* subset as it only has between 0.92% and 1.85% of the features. This leads us to think that the categories described in 6 are not comprehensively represented *(see Table 3)*, therefore omitting relevant information that determines the trend. Hence when reducing the dimensions of this problem, it is important to determine when the reduction is detrimental for the classification problem. Since the problem of analyzing each subset is intractable as there are 2^{216} subsets, it is essential to use approaches like the statistical measures previously defined to select advantageous subsets of features.

After analyzing the selected features for each ETF, we can observe that emerging markets depend on the cyclic behaviors of the prices while developed markets do not. We believe this is due to the market exposure of these instruments, as can be seen in Table 1. Additionally, the selected feature in the Volume category for the emerging markets ETFs is AOBV (Archer's On Balance Volume), while on the developed market ETF is PVR (Price Volume Rank). These features are essentially different given that PVR classifies the day according to a relation between volume and price, while AOBV provides a quantity according to a similar relation. Finally, TTM_TRND (TTM Trend) is selected only in IVV; this supports the general idea that emerging markets' predictions use quantitative features while developed markets rely more on qualitative features. A brief description of the salient features can be found in A.

Herein we center our attention on ETFs from emerging markets with similar market exposure and market value percentages; however, it is possible to choose other kinds of ETFs where the percentage of the markets exposure distribution is essentially different. Furthermore, there are ETFs specialized in some sectors like energy, financial, commodities, or technology, where the performances may differ (Figure 6), we hypothesize that the selected features obtained by the algorithm proposed in this work will be similar in those other ETFs even if the market exposure, region or topic are different, but further research is needed.

Another interesting subject may be to determine if the methodology described in this work can provide a good selection of features that improves the performance of other neural network models such as Long-Short Therm Memory (LSTM), Restricted Boltzmann Machine (RBM), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) or even another techniques such as decision trees and random forest, self-organizing maps (SOM), time series or other regression methods. Finally, this approach may be used to select features in many other interesting related topics like investment portfolios containing various stocks.

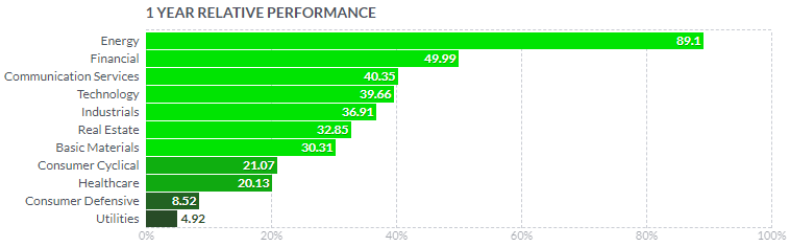


Fig. 6: One year growth performance per sector from 2020 to 2021, Source: October 2021. Year Relative Performance, https://finviz.com/groups.ashx

Acknowledgment: This work was partially funded by the authors' universities and the *Consejo Nacional de Ciencia y Tecnología (CONACyT)* from México. Its contents are the responsibility of the authors and do not reflect the views of the research granting bodies. The authors were responsible for the data analysis after the extraction and linkage.

Julián Alberto Fresán Figueroa would like to thank the support of Universidad Autónoma Metropolitana, Unidad Cuajimalpa. Máximo Eduardo Sánchez Gutiérrez would like to thank the support of Universidad Autónoma de la Ciudad de México, Unidad Cuautepec. Alma Rocío Sagaceta Mejía would like to thank the support of Universidad Iberoamericana, Ciudad de México.

References

[Ananthi and Vijayakumar(2021)] M Ananthi and K Vijayakumar. 2021. Stock market analysis using candlestick regression and market trend prediction (CKRM). *Journal of Ambient Intelligence and Humanized Computing* 12, 5 (2021), 4819–4826.

[Antoniewicz and Heinrichs(2014)] Rochelle Shelly Antoniewicz and Jane Heinrichs. 2014. Understanding exchange-traded funds: How ETFs work. *Jane, Understanding Exchange-Traded Funds: How ETFs Work (September 30, 2014)* (2014).

[Barra et al.(2020)] Silvio Barra, Salvatore Mario Carta, Andrea Corriga, Alessandro Sebastian Podda, and Diego Reforgiato Recupero. 2020. Deep learning and time series-to-image encoding for financial forecasting. *IEEE/CAA Journal of Automatica Sinica* 7, 3 (2020), 683–692.

[Cagliero et al.(2020)] Luca Cagliero, Paolo Garza, Giuseppe Attanasio, and Elena

Baralis. 2020. Training ensembles of faceted classification models for quantitative stock trading. *Computing* 102 (2020), 1213–1225.

[Chen and Chen(2015)] Mu-Yen Chen and Bo-Tsuen Chen. 2015. A hybrid fuzzy time series model based on granular computing for stock price forecasting. *Information Sciences* 294 (2015), 227–241.

[Chen et al.(2019)] Mu-Yen Chen, Chien-Hsiang Liao, and Ren-Pao Hsieh. 2019. Modeling public mood and emotion: Stock market trend prediction with anticipatory computing approach. *Computers in Human Behavior* 101 (2019), 402–408.

[Chen et al.(2021a)] Wei Chen, Manrui Jiang, Wei-Guo Zhang, and Zhensong Chen. 2021a. A novel graph convolutional feature based convolutional neural network for stock trend prediction. *Information Sciences* 556 (2021), 67–94.

[Chen et al.(2021b)] Wei Chen, Manrui Jiang, Wei-Guo Zhang, and Zhensong Chen. 2021b. A novel graph convolutional feature based convolutional neural network for stock trend prediction. *INFORMATION SCIENCES* 556 (MAY 2021), 67–94. https://doi.org/10.1016/j.ins.2020.12.068

[De Haan et al.(2016)] Laurens De Haan, Cécile Mercadier, and Chen Zhou. 2016. Adapting extreme value statistics to financial time series: dealing with bias and serial dependence. *Finance and Stochastics* 20, 2 (2016), 321–354.

[Deville(2008)] Laurent Deville. 2008. Exchange traded funds: History, trading, and research. *Handbook of financial engineering* (2008), 67–98.

[Ecer et al.(2020)] Fatih Ecer, Sina Ardabili, Shahab S Band, and Amir Mosavi. 2020. Training multilayer perceptron with genetic algorithms and particle swarm optimization for modeling stock price index prediction. *Entropy* 22, 11 (2020), 1239.

[Ehlers(2013)] John F Ehlers. 2013. *Cycle Analytics for Traders,+ Downloadable Software: Advanced Technical Trading Concepts*. John Wiley & Sons.

[Ferreira and Figueiredo(2012)] Artur J. Ferreira and Mário A.T. Figueiredo. 2012. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters* 33, 13 (2012), 1794–1804. https://doi.org/10.1016/j.patrec.2012.05.019

[Hegde and McDermott(2004)] Shantaram P Hegde and John B McDermott. 2004. The market liquidity of DIAMONDS, Q's, and their underlying stocks. *Journal of Banking & Finance* 28, 5 (2004), 1043–1067.

[Jamshed et al.(2019)] Huma Jamshed, M Khan, Muhammad Khurram, Syed Inayatullah, and Sameen Athar. 2019. Data Preprocessing: A preliminary step for web data mining. *3c Tecnología: glosas de innovación aplicadas a la pyme* 8, 29 (2019), 206–221.

[Jiang et al.(2020)] Manrui Jiang, Lifen Jia, Zhensong Chen, and Wei Chen. 2020. The two-stage machine learning ensemble models for stock price prediction

by combining mode decomposition, extreme learning machine and improved harmony search algorithm. *Annals of Operations Research* (2020), 1–33.

[Johnson(2021)] Kevin Johnson. 2021. Pandas - Technical Analysis. https://github.com/twopirllc/pandas-ta.

[Liang et al.(2017)] Qiubin Liang, Wenge Rong, Jiayi Zhang, Jingshuang Liu, and Zhang Xiong. 2017. Restricted Boltzmann machine based stock market trend prediction. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1380–1387.

[Lu et al.(2007)] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian. 2007. Feature Selection Using Principal Feature Analysis. In *Proceedings of the 15th ACM International Conference on Multimedia* (Augsburg, Germany) *(MM '07)*. Association for Computing Machinery, New York, NY, USA, 301–304. https://doi.org/10.1145/1291233.1291297

[Miralles-Quirós et al.(2019)] José Luis Miralles-Quirós, María Mar Miralles-Quirós, and José Manuel Nogueira. 2019. Diversification benefits of using exchange-traded funds in compliance to the sustainable development goals. *Business Strategy and the Environment* 28, 1 (2019), 244–255.

[Nelson et al.(2017)] David MQ Nelson, Adriano CM Pereira, and Renato A de Oliveira. 2017. Stock market's price movement prediction with LSTM neural networks. In *2017 International joint conference on neural networks (IJCNN)*. IEEE, 1419–1426.

[O'Hara et al.(2000)] H Thomas O'Hara, Cathy Lazdowski, Calin Moldovean, and Shawn T Samuelson. 2000. Financial indicators of stock price performance. *American Business Review* 18, 1 (2000), 90.

[Reddy et al.(2020)] G. Thippa Reddy, M. Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. 2020. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* 8 (2020), 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942

[Sezer and Ozbayoglu(2018)] Omer Berat Sezer and Ahmet Murat Ozbayoglu. 2018. Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach. *Applied Soft Computing* 70 (2018), 525–538.

[Shearer(2000)] Colin Shearer. 2000. The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing* 5, 4 (2000), 13–22.

[Spearman(1961)] Charles Spearman. 1961. The proof and measurement of association between two things. Appleton-Century-Crofts.

[Sun et al.(2017)] Wencheng Sun, Zhiping Cai, Fang Liu, Shengqun Fang, and Guoyan Wang. 2017. A survey of data mining technology on electronic medical records. In *2017 IEEE 19th International Conference on e-Health*

Networking, Applications and Services (Healthcom). 1–6. https://doi.org/10.1109/HealthCom.2017.8210774

[Tang et al.(2019)] Huimin Tang, Peiwu Dong, and Yong Shi. 2019. A new approach of integrating piecewise linear representation and weighted support vector machine for forecasting stock turning points. *Applied Soft Computing* 78 (2019), 685–696.

[Zhang et al.(2020)] Jun Zhang, Lan Li, and Wei Chen. 2020. Predicting stock price using two-stage machine learning techniques. *Computational Economics* (2020), 1–25.

[Zhao et al.(2021)] Jinghua Zhao, Dalin Zeng, Shuang Liang, Huilin Kang, and Qinming Liu. 2021. Prediction model for stock price trend based on recurrent neural network. *Journal of Ambient Intelligence and Humanized Computing* 12 (2021), 745–753.

A Selected Features

In this appendix, we give a summary of the best indicators selected in our methodology. We denote by $\mathfrak{h}_t$ and ℓ_t are the highest price and lowest price of the period t. Also, we denote by $\mathfrak{o}_t$ and $\mathfrak{c}_t$ as the opening and closing price of the period t, and v_t as the volume in the period t.

Cycles

A.1 Even Better SineWave (EBSW)

This indicator measures market cycles and uses a low pass filter to remove noise. Its output is a bound signal between -1 and 1, and the maximum length of a detected trend is limited by its length input. The formula can be found in [Ehlers(2013)].

Momentum

A.2 Balance of Power (BOP)

Balance of Power tells whether the underlying action in trading stock is characterized by systematic buying (accumulation) or systematic selling (distribution). The calculation of BOP is expressed as $(\mathfrak{c}_t - \mathfrak{o}_t)/(\mathfrak{h}_t - \ell_t)$.

A.3 Stochastic Relative Strength Index (StochRSI)

The Stochastic RSI technical indicator applies Stochastic Oscillator to values of the Relative Strength Index (RSI). The indicator thus produces two main plots, Full-K and Full-D oscillating between oversold and overbought levels. It is calculated as follows:

$$RSI = 100 - 10/1 + RS, \quad RS = \frac{\text{Total Gains}/n}{\text{Total Losses}/n}$$

where n is the number of RSI periods. The expressions for Full-K and Full-D are given by:

$$\%K = 100 \times \left(\frac{\text{Recent Close} - \text{Lowest } Low(n)}{\text{Highest } High(n) - \text{Lowest } Low(n)} \right),$$

$$\%D = 3 - \text{period moving average of } \%K$$

where n = number of periods used in the calculation. The formula for Stochastic Relative Strength is given by:

$$StochRSI = RSI(n) - \frac{\text{RSI Lowest } Low(n)}{\text{RSI Highest } High(n) - \text{RSI Lowest } Low(n)}$$

A.4 Correlation Trend Indicator (CTI)

This indicator represents the correlation of the price with the trend line. The correlation is measured with the Spearman algorithm [Spearman(1961)].

A.5 KDJ

KDJ indicator is a technical indicator used to analyze and predict changes in stock trends and price patterns in a traded asset. KDJ indicator is otherwise known as the random index. It is a practical technical indicator that is most commonly used in market trend analysis of a short-term stock. The indicators are obtained as follows:

$$K_t = \frac{2K_{t-1} + \frac{\mathfrak{c}_t - \ell_t}{\mathfrak{h}_t - \ell_t}}{3}$$

$$D_t = \frac{2D_{t-1} + K_t}{3}$$

$$J_t = 3K_t + 2D_t$$

A.6 Williams % R (WILLR)

The indicator Williams % R normalises the price as a percentage between 0 and 100. The formula is given by:

$$\%R = -100 * \frac{\text{Highest High} - \mathfrak{c}_t}{\text{Highest High} - \text{Lowest Low}}$$

where Highest High corresponds to the highest high in the past n periods, and Lowest Low corresponds to the lowest low in the past n periods.

Statistics

A.7 Z-score

Z-score (Zs) use the Simple Mobile Average (SMA) and the deviation of the Close values (σ) for an n period. The formulas used for this indicator of SMA are given by:

$$SMA = \frac{\sum_{i=1}^{n} \mathfrak{c}_{ti}}{n}, \qquad Zs = \frac{\mathfrak{c}_t - SMA}{\sigma}.$$

Trend

A.8 Decreasing (DEC)

The indicator Decreasing computes the difference between Close values for t and $t-1$. It is a Boolean value and is given by:

$$DEC = \begin{cases} 1, & \text{if } \mathfrak{c}_t - \mathfrak{c}_{t-1} < 0, \\ 0, & \text{other case} \end{cases}$$

A.9 Increasing (INC)

Increasing is the opposite of the Decreasing indicator, computes the difference between Close values for t and $t-1$, and it returns a Boolean value given by:

$$INC = \begin{cases} 1, & \text{if } \mathfrak{c}_t - \mathfrak{c}_{t-1} > 0, \\ 0, & \text{other case} \end{cases}$$

A.10 TTM Trend (TTM)

The TTM Trend indicator colours the price bars in red or blue if the last five prices are above or under the average price for the last five price bars. Two bars of opposite colours are the signal to buy or sell.

Volatility

A.12 Bollinger Bands %B (BBP)

Bollinger Band Percent (B_t) quantifies a symbol's price relative to the upper and lower Bollinger Band.

$$B_t = \frac{\mathfrak{o}_t - \text{Lower Band}}{\text{Upper Band} - \text{Lower Band}}$$

Volume

A.13 Archer On Balance Volume (AOBV)

The indicator corresponds to the average of the On Balance Volume for a given time. The formula for OBV reads:

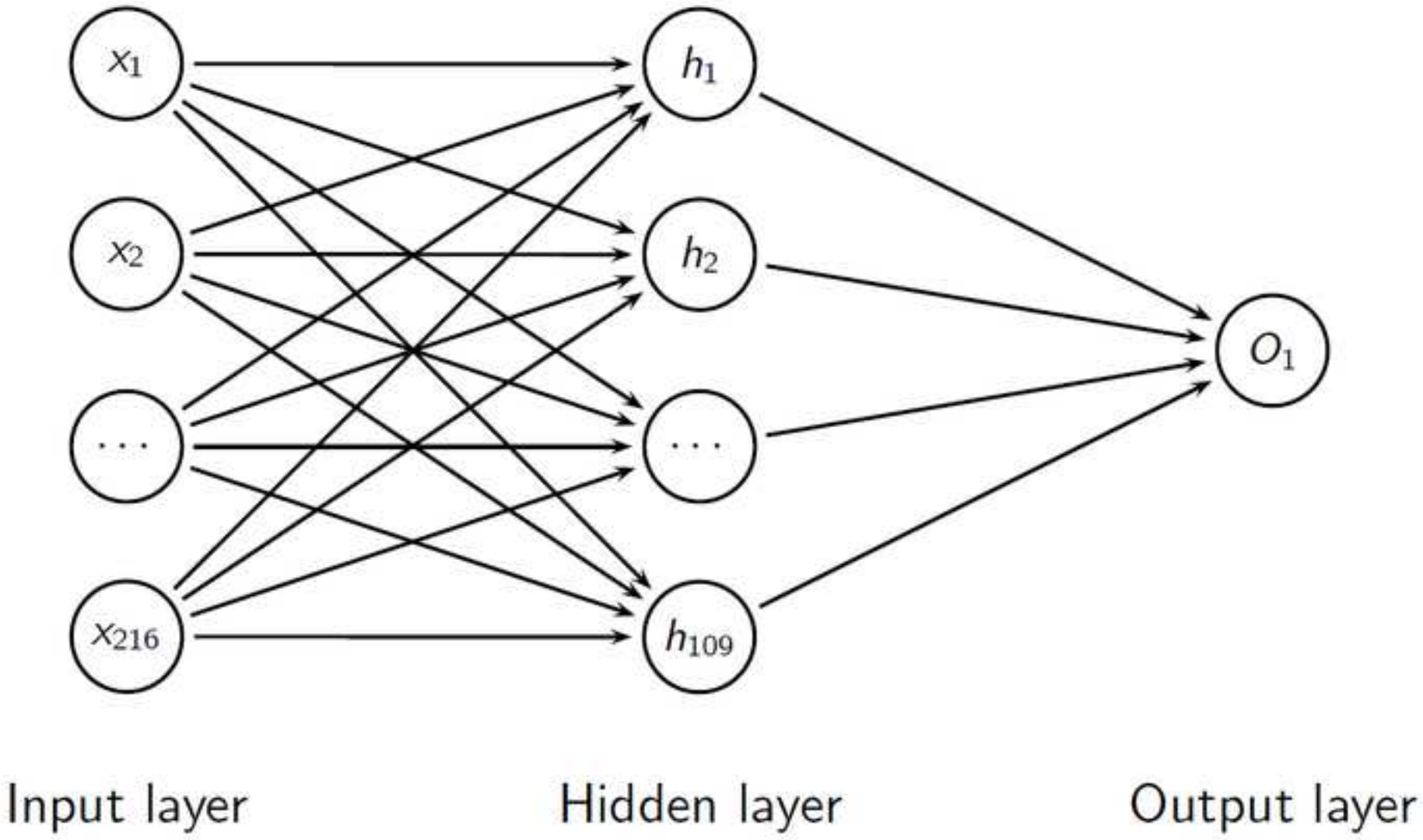
$$OBV = \begin{cases} \text{OBV}_{t-1} + \text{Volume} & \text{if} \quad \mathfrak{c}_t > \mathfrak{c}_{t-1} \\ \text{OBV}_{t-1} - \text{Volume} & \text{if} \quad \mathfrak{c}_t < \mathfrak{c}_{t-1} \\ \text{OBV}_{t-1} & \text{if} \quad \mathfrak{c}_t = \mathfrak{c}_{t-1} \end{cases}$$

A.14 Price-Volume Rank (PVR)

Price-Volume Rank compares the direction of the price change to the change in volume and assigns a number to that specific relationship. By quantifying price-volume interaction, P-V rank seeks to determine the position within a typical market cycle.

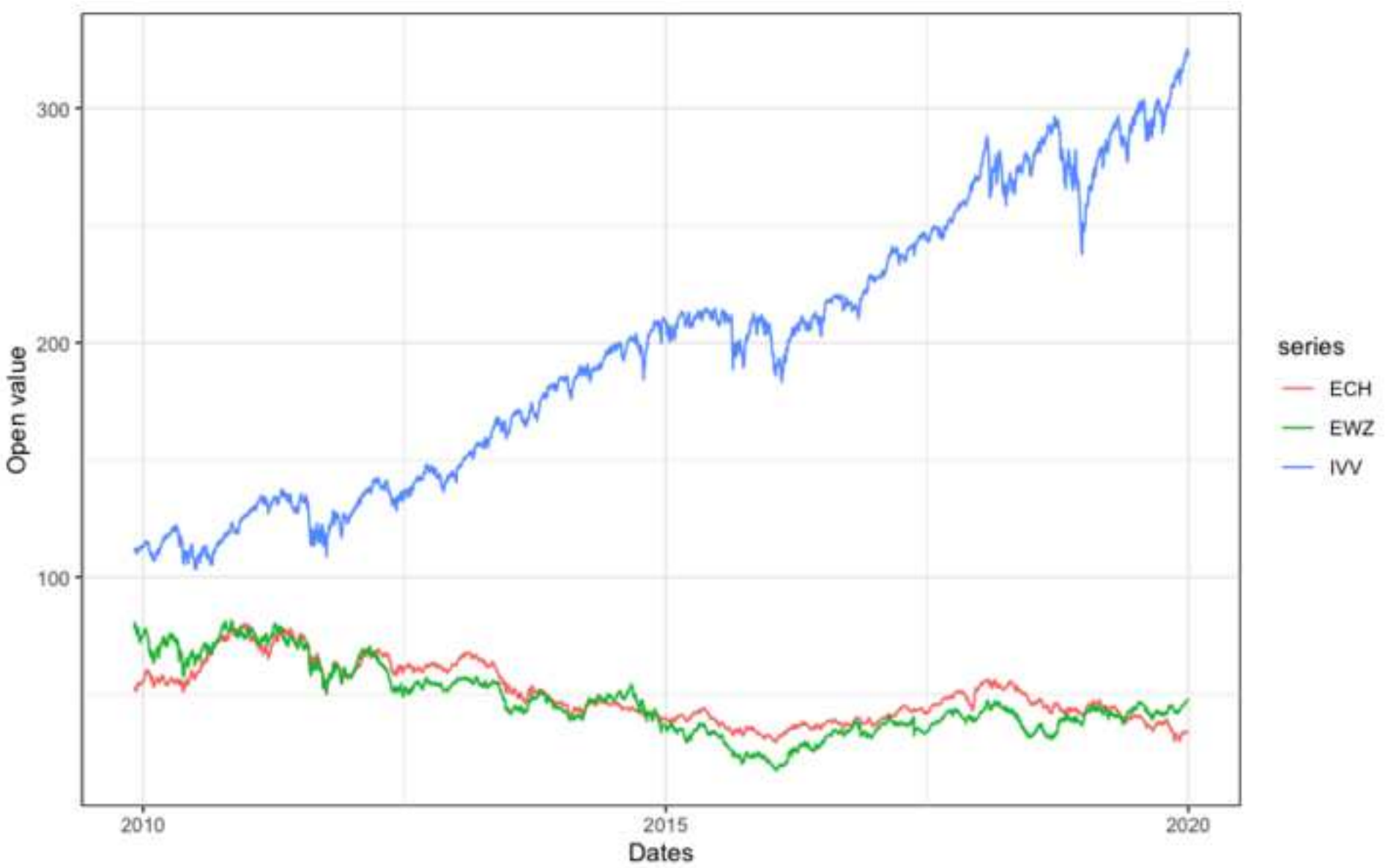
$$PVR_t = \begin{cases} 1 & \text{if} \quad \mathfrak{o}_t > \mathfrak{c}_{t-1} \quad \text{and} \quad v_t > v_{t-1} \\ 2 & \text{if} \quad \mathfrak{o}_t > \mathfrak{c}_{t-1} \quad \text{and} \quad v_t < v_{t-1} \\ 3 & \text{if} \quad \mathfrak{o}_t < \mathfrak{c}_{t-1} \quad \text{and} \quad v_t > v_{t-1} \\ 4 & \text{if} \quad \mathfrak{o}_t < \mathfrak{c}_{t-1} \quad \text{and} \quad v_t < v_{t-1} \end{cases}$$

Figure



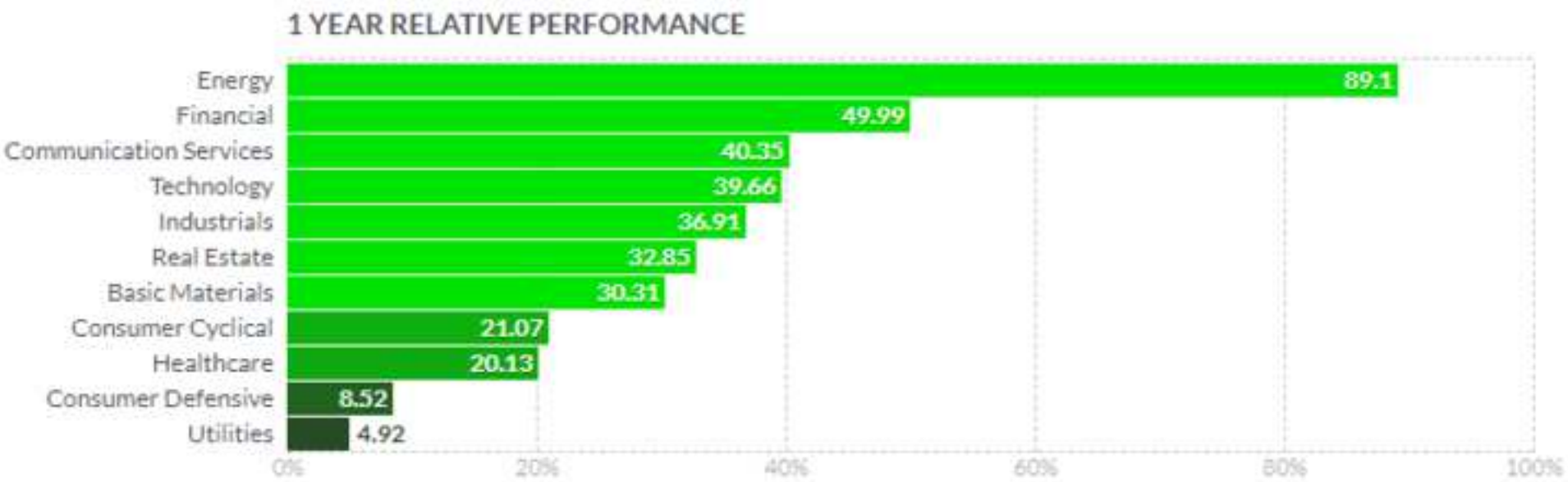
Figure

Click here to access/download;Figure;Rplot.png



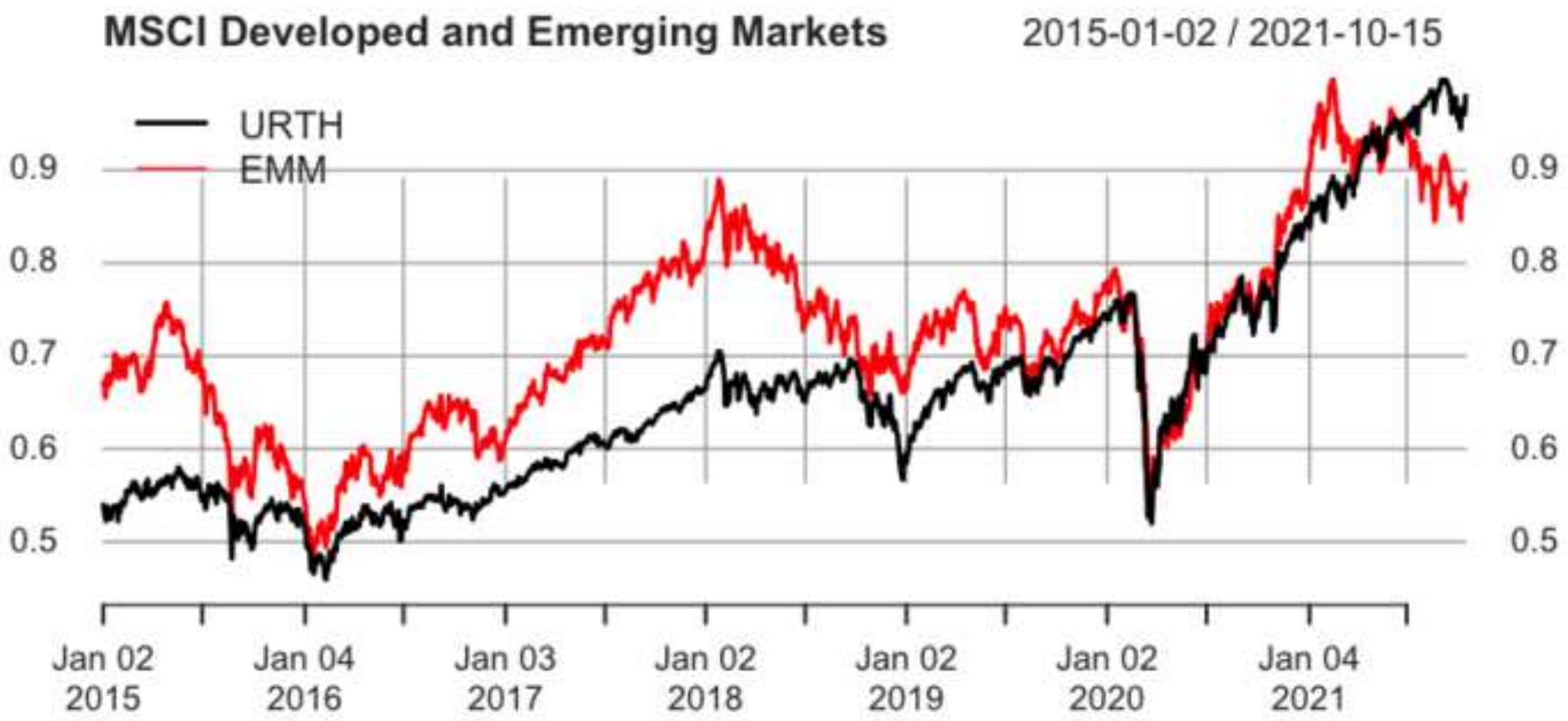
Figure

Click here to access/download;Figure;year.png



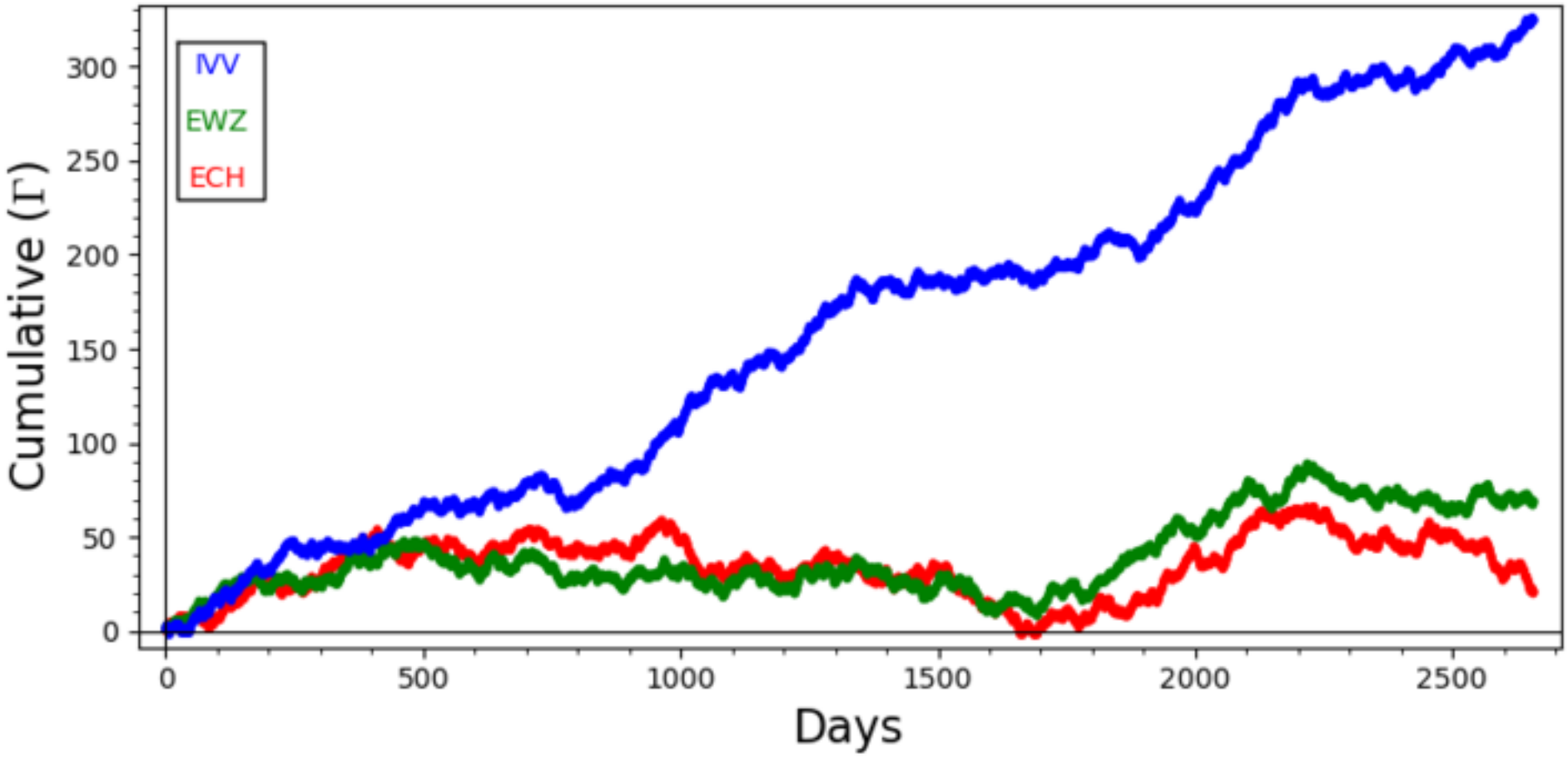
Figure

Click here to access/download;Figure;emergingm.png



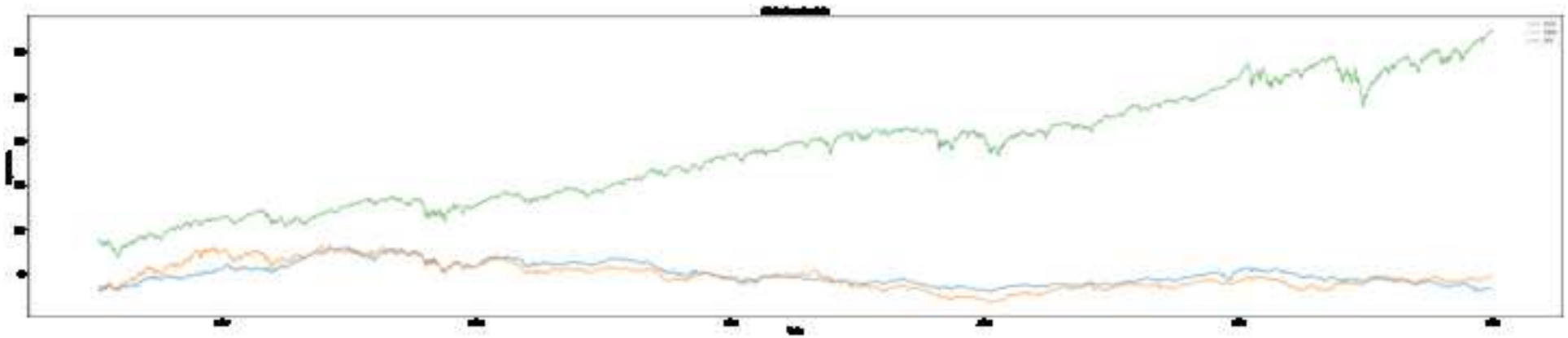
Figure

Click here to access/download;Figure;GraficaETF.png



Figure

Click here to access/download;Figure;graficas_etfs.png



Figure

Click here to access/download;Figure;accuracygain.png