Stock Price Forecasting with Artificial Neural Networks Long Short-Term Memory: A Bibliometric Analysis and Systematic Literature Review

Cristiane Orquisa Fantin, Eli Hadad

Mackenzie Presbyterian University, São Paulo, Brazil
Email: cris.orquisa@hotmail.com, eli.hadad@mackenzie.br

Abstract

This study maps the academic literature on Stock Price Forecasting with Long-Term Memory Artificial Neural Networks—RNA LSTM. The objective is to know if it is suitable for time series studies, especially for stock price projection. Through bibliometric analysis and systematic literature review, it is observed that 333 authors wrote on the topic between 2018 and March 2022, and the journals Expert Systems with Applications, IEEE Access, Big Data Journal and Neural Computing and Applications, published the most relevant articles. Of the 99 articles published in this period, 43 are associated with Chinese institutions, the most cited being that of Kim and Won, who studies the volatility of returns and the market capitalization of South Korean stocks. The basis of 65% of the studies is the comparison between the RNN LSTM and other artificial neural networks. The daily closing price of shares is the most analyzed type of data, and the American (21%) and Chinese (20%) stock exchanges are the most studied. 57% of the studies include improvements to existing neural network models and 42% new projection models.

Keywords

Stock Price Forecasting, Long-Term Memory, Backpropagation, Bibliometric Analysis, Systematic Review

1. Introduction

The financial market is characterized by being a dynamic, complex and non-linear system, characterized by data intensity, noise, non-stationary nature, unstructured and with a high degree of uncertainty [1]. As so many factors inte-
ract simultaneously, such as political events, macro and microeconomic conditions and investor expectations, predicting these movements is a very challenging task.

The growing role that the stock market plays in the world economy stimulates the development of research aimed at building theories, involving the topic of stock price prediction, and accurate methods are crucial for the management of investor portfolios.

Assessing expected returns relative to total exposure assumes that portfolio managers understand the distribution of the portfolio. Specialists can model the influence of tangible assets in relation to market value, but not of intangible assets such as rights, experiences or brand equity.

An important ally in the quest to minimize risk in relation to exposure, artificial intelligence and machine learning with their neural networks have provided a great balance of quality in recent decades, improving detection, diagnosis, prediction and problem solving [2]; this is because in this market, future events are at least partially dependent on past events and data [3], and not entirely random.

The aim of this study is to map and analyze the published academic literature on stock price prediction using artificial neural networks Long-Term Memory Artificial Neural Networks—RNA LSTM. To this end, a bibliometric analysis and a systematic review of the literature on the subject are carried out, during the period from January 1, 2000 to March 31, 2022, with a final sample of 99 articles. Bibliometric analysis refers to quantitative analysis, which is developed by counting frequencies and co-citations. The systematic review, a qualitative analysis, considers the correlation between the most significant themes, but still little studied by the academy. The research base used comes from the Web of Science—WoS database, and both the bibliometric analysis and the systematic review do not dispense with the use of R, RStudio, Biblioshiny and VOSViewer software. In the bibliometric analysis, the verification of the main laws is adopted: Lotka [4] and Bradford [5].

The literature review is presented in item 2, with the identification of theories and methods of forecasting stock prices with multilayer perceptron artificial neural networks mentioned in the articles of the final sample. The bibliometric analysis and systematic review methodologies are described in item 3, and in item 4 the results of both methodologies are reported, with descriptive statistics of the most relevant characteristics of the articles in the final sample and the knowledge gaps on the topic. Item 5 presents the conclusions, paths for future studies and limitations of this research.

2. Literature Review

The current stock price of a publicly traded company reflects the company’s past operation, current timing and future profitability prospects. To obtain a more accurate projection of the stock price, several types of studies have already been carried out. The most classic ones focus on the financial data of the target companies, added to micro and macroeconomic aspects. However, the non-linear

With the development of computer technology, new prediction methods have been proposed, such as artificial neural networks [12], whose objective is to replicate the way the human brain works. Recurrent neural networks [13] are a powerful set of artificial neural network algorithms, especially useful for processing sequential data such as sound, temporal or language. Some recurrent neural networks—RNN performed better in predicting financial data and became popular, such as Long Short-Term Memory [14].

Thus, LSTM network is a specific type of RNN that has been widely applied to solve supervised learning issues [15]. It has non-linear memory cells and gate units [16], capable of processing non-stationary long-term sequences. In addition, it can extract the characteristics of financial data and reflect the characteristics of the network [17]. It offers good performance in predicting prices in the stock market, as it is an algorithm capable of identifying non-linear and hidden relationships in the data, that is, it is a supervised learning algorithm, capable of learning from a set of data in training (given a dataset, LSTM can learn a non-linear function for regression).

Maknickiene and Maknickas [18] improved the performance of measurements in the foreign exchange market or using TM; Chen, Zhou, and Dai [19] used LSTM for Chinese design market returns and performed well. After that, several experiments with the modern LSTM network in a literary way in isolation, hybrid or combined methods with classics and financial series studies were completed, A indicates that LSTM is quite suitable for time series financial models [20].

Concretely, the LSTM network consists of three parts, including an input layer, an output layer and several hidden layers between them. The hidden layers have memory modules. The core of the memory module is the self-connecting memory cell with three ports, input, output, and forgetting. The value of each of these ports controls the flow of information in the memory module.

Information is retained by cells and memory manipulations are done by gates. Gateway: where useful information is added. Oblivion Gate: where information that is no longer useful is removed. Output Port: the task of extracting useful information from the current cell state to be displayed as a result. A vector is gen-
erated, and the information is regulated using the function that filters the values to be remembered. Vector values and regulated values are multiplied to be sent as output and input to the next cell [21].

However, designing a good network architecture for the problem studied is not a simplistic task. The model’s architecture directly interferes with its performance. The refinement process can be time consuming, as there is no formal method to perform this classification task, this is necessarily through the performance of iterative tests with several parameters, in which only the structure of greater assertiveness is maintained.

The quality of the information that the network is fed, as much or more, is reflected in the accuracy of the output layer’s response. The selection of information that will be provided to the input layer is a key factor in the design of an intelligent decision system, because even if the model is the best, it will perform poorly if the features are not well chosen. Specific methods must be used in the selection of relevant information.

3. Methodology

The aim of this study is to answer the question—using LSTM artificial neural networks, can we get reliable predictions of stock prices? For this, the 7 steps described in Figure 1, detailed below, are implemented.

Step 1—Choosing the database. Sample articles come from WoS, the world’s leading citation database.

Step 2—Using WoS Initial Search Parameters for the period from January 1, 2000 to March 31, 2022. Initially, 276 articles are identified based on variations of the keywords stock, market, LSTM, forecast, stock, predictive, regression, supervised, learn, backpropagation, supervised and backpropagation. Subsequently, exclusions are performed by applying filters in WoS itself, resulting in an intermediate sample of 127 articles, as shown in Table 1.

Step 3 - Selection of publications by using exclusion & inclusion criteria = 99 papers

Step 4 - Creation of a database and collection of articles. The 99 articles in the final sample are obtained from the following academic research databases: Web of Science

Step 5 - Bibliometric analysis. Through the R, RStudio, Biblioshiny and VOSviewer software, objective data from the articles are analyzed

Step 6 - Reading and coding the articles. Identification of the objectives, sample, methods and contributions of the articles.

Step 7 - Systematic review.

Figure 1. Steps of methodology.
Table 1. Evolution of the sample using WoS’ filters.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Description</th>
<th>Number of papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+)</td>
<td>Keywords: equal to &quot;stock market*&quot; and &quot;LSTM&quot; or &quot;stock market*&quot; and &quot;long short-term memory&quot; or &quot;forecas* stock*&quot; and &quot;LSTM&quot; or &quot;forecas* stock*&quot; and &quot;long short-term memory&quot; or &quot;predictive regression*&quot; and &quot;LSTM&quot; or &quot;predictive regression*&quot; and &quot;long short-term memory&quot; or &quot;supervision &quot;learning*&quot; and &quot;long short-term memory&quot; and &quot;backpropagation&quot;.</td>
<td>276</td>
</tr>
<tr>
<td>(−)</td>
<td>Document type: other than &quot;article&quot; or &quot;early access&quot; or &quot;data document&quot;.</td>
<td>106</td>
</tr>
<tr>
<td>(−)</td>
<td>Web of Science categories: equal to &quot;computer science artificial intelligence&quot;, &quot;computer science theory methods&quot;, &quot;computer science information systems&quot;, “interdisciplinary applications of computer science”, “hardware architecture of computer science”, “computer science software engineering”, “economics”, “business”, “business finance”, “operations research management science”, “management”.</td>
<td>43</td>
</tr>
<tr>
<td>(−)</td>
<td>Research area: equal to “computer science”, “business economics”, “operations research management science”.</td>
<td>0</td>
</tr>
<tr>
<td>(−)</td>
<td>Language: different from &quot;English&quot;.</td>
<td>0</td>
</tr>
<tr>
<td>(=)</td>
<td>Intermediate sample.</td>
<td>127</td>
</tr>
</tbody>
</table>

Step 3—Exclusion of 19 of the 127 articles for not being available in the researched sources (Google Scholar, Science Direct and Web of Science) and another 09 for not being directly related to the topic of our research, namely: 01 e-commerce, 02 cryptocurrencies, 01 real estate price bubble detection, 01 hierarchical temporal memory, 01 neuromorphic vision datasets, 01 Gray Wolf-Elman optimization, 01 stock movement during the Covid-19 pandemic and 01 stock price prediction based on in morphological similarity clustering and hierarchical temporal memory. Thus, the final sample is composed of 99 articles [15] [20]-[122].

Step 4—Creation of a database and collection of articles. The 99 articles in the final sample are obtained from the following academic research databases: Web of Science, Science Direct, and Google Scholar. From its analysis, the following information is collected to capture the general data of the article: title, author name, affiliated institution and country of origin of authors/researchers, journal name, volume and issue number, homepage and page final, year of publication, country of origin of data and number of years of sample data, keywords, Digital Object Identifier (DOI), Journal of Economic Literature (JEL) and number of citations of articles in the WoS database.

Step 5—Bibliometric analysis. Through the R, RStudio, Biblioshiny and VOS-viewer software, objective data from the articles are analyzed—countries, authors, keywords, institutions, etc., for the preparation and analysis of relationship/co-citation tables and maps. The analyzes carried out by both tools are
complemented by the verification of the main laws of bibliometrics, namely: 1) Bradford’s Law [5]—verification of journals that produce many articles in contrast to those that produce few on a given topic, and 2) Law de Lotka [4]—identification of researchers with a higher frequency of production in a given area of knowledge.

Step 6—Reading and coding the articles. Identification of the objectives, sample, methods and contributions of the articles. In addition, they are classified and coded into categories and subcategories structured according to Table 2. Each of the 08 categories has non-exclusive subcategories. This means that the same article can be classified in more than one subcategory. Thus, the sum of the frequency count of the subcategories—for each category—is what adds up to 100%. In the coding process, as many subcategories as necessary per article are assigned.

Step 7—Systematic review. After coding the (sub)categorization matrix in Table 2—for the final sample—a frequency count of the subcategories is performed to enable the identification of knowledge gaps. Such gaps are then compared with the subcategories of category 08—paths for future studies, in order to obtain aspects that can be the object of further studies on the subject.

Table 2. Matrix of (sub)categorization.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Subcategories</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Neural networks/ algorithms used in research</td>
<td>A-LSTM</td>
<td>Stock price projection with the—RNN LSTM.</td>
</tr>
<tr>
<td></td>
<td>B-Compared to LSTM</td>
<td>Stock price prediction with other artificial neural networks and results compared to RNN LSTM.</td>
</tr>
<tr>
<td></td>
<td>C-Combined with LSTM</td>
<td>Predict stock prices with blended neural networks including LSTM.</td>
</tr>
<tr>
<td></td>
<td>D-Others</td>
<td>Other topics unrelated to subcategories 1A to 1C.</td>
</tr>
<tr>
<td>2. Types of data analyzed</td>
<td>A-Closing prices</td>
<td>Daily stock closing prices.</td>
</tr>
<tr>
<td></td>
<td>B-Opening prices</td>
<td>Daily stock opening prices.</td>
</tr>
<tr>
<td></td>
<td>C-Highest and lowest prices</td>
<td>Daily highest and lowest stock prices.</td>
</tr>
<tr>
<td></td>
<td>D-Volumes</td>
<td>Stock trading volumes.</td>
</tr>
<tr>
<td></td>
<td>E-Index</td>
<td>Daily closing of the Stock Price Index.</td>
</tr>
<tr>
<td></td>
<td>F-Others</td>
<td>Others not related to subcategories 2A to 2E.</td>
</tr>
<tr>
<td>3. Analysis period</td>
<td>A-Up to 5 years</td>
<td>Data from 0 to 5 years.</td>
</tr>
<tr>
<td></td>
<td>B-More than 5 to 10 years</td>
<td>Data from 5.1 to 10 years.</td>
</tr>
<tr>
<td></td>
<td>C-More than 10 years</td>
<td>More than 10 years.</td>
</tr>
<tr>
<td></td>
<td>D-Not applicable/not informed</td>
<td>Studies that do not inform the period of analysis.</td>
</tr>
<tr>
<td>4. Objectives</td>
<td>A-Tests with new neural networks models</td>
<td>Improved share price accuracy tested with other neural networks algorithms and/or hybrid models.</td>
</tr>
<tr>
<td></td>
<td>B-Tests with other assets</td>
<td>Check whether using price and volatility indices of other assets (except stocks) can help predict stock prices.</td>
</tr>
<tr>
<td></td>
<td>C-Sentiment Analysis</td>
<td>Improved accuracy in stock price projection with sentiment analysis.</td>
</tr>
<tr>
<td></td>
<td>D-Others</td>
<td>Other topics unrelated to subcategories 7A to 7C.</td>
</tr>
</tbody>
</table>
5. Data origin

<table>
<thead>
<tr>
<th>Region</th>
<th>Exchanges</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>- CSI, SSE, NSE, HS, SH, SZSE China and Hong Kong Stock Exchanges.</td>
</tr>
<tr>
<td>C</td>
<td>- B3 Brazil Stock Exchange.</td>
</tr>
<tr>
<td>D</td>
<td>- TWSE Thailand Stock Exchange.</td>
</tr>
<tr>
<td>E</td>
<td>- IMKB Turkey Stock Exchange.</td>
</tr>
<tr>
<td>F</td>
<td>- TSE Tehran Stock Exchange.</td>
</tr>
<tr>
<td>G</td>
<td>- GSE Ghana Stock Exchange.</td>
</tr>
<tr>
<td>H</td>
<td>- ASX Australia Stock Exchange.</td>
</tr>
<tr>
<td>I</td>
<td>- DAX Germany Stock Exchange.</td>
</tr>
<tr>
<td>J</td>
<td>- KOSPI, KOSDAQ Korea Stock Exchanges.</td>
</tr>
<tr>
<td>K</td>
<td>- NSE India Stock Exchange.</td>
</tr>
<tr>
<td>N</td>
<td>- NIKKEI Japan Stock Exchange.</td>
</tr>
<tr>
<td>O</td>
<td>- IDX Indonesia Stock Exchange.</td>
</tr>
<tr>
<td>P</td>
<td>- FTSE UK Stock Exchange.</td>
</tr>
<tr>
<td>L</td>
<td>- Texts News agencies/websites, for sentiment analysis.</td>
</tr>
<tr>
<td>M</td>
<td>- No information or other There is no identification of information that can be considered as inputs for the evaluation models.</td>
</tr>
</tbody>
</table>

6. Results

| A        | Outperforms compared methods The results of the proposed model surpass the results of the compared model(s). |
| B        | Promising model The results of the proposed model are promising. |
| C        | Others Other results unrelated to subcategories 8A to 8B. |

7. Conclusions

| A        | New conclusions Presentation of new findings—adjustment to already tested neural networks models, improvement in the quality of input information, and other innovations to existing models. |
| B        | New perspectives Presentation of a new theory, new models of projections, with models of isolated, hybrid or combined neural networks. |
| C        | Conclusions similar to works presented previously Studies that do not present new perspectives or new conclusions. |
| D        | Others Other results unrelated to subcategories 9A to 9C. |

8. Pathways for future studies

| A        | Hybrid models with LSTM Studies with other hybrid models using LSMT. |
| B        | Other ANN Studies with other ANN, pure or hybrid. |
| C        | Other types of data In addition to opening, closing, high, low and trading volume data, sentiment analysis tests with other types of news and stock data, in periods such as intraday. |
| D        | Data from other sources Study the model’s performance on other Stock Exchanges. |
| E        | Other analysis periods Study and test data from different periods. |
| F        | No path commented by the author(s) No future path detailed by author(s). |
4. Analysis of Results

Item 4.1 presents the results of the bibliometric analysis, mentioned in Step 5 of the Methodology. In turn, item 5.2 contains the results of the systematic review, whose steps are described in Steps 6 and 7 of item 3 of this study.

4.1. Bibliometric Analysis

The final sample consists of 99 articles, distributed between the years 2000 and 2021, obtained from the WoS database—see Figure 2. In this period, up to 5 articles on stock price forecasting using LSTM per year are identified.

Figure 3 shows the co-occurrence map of the most used keywords in the articles.

Again, the words model and neural networks stand out, in addition to prediction and time series.

Table 3 presents the frequency of the 151 main keywords of the study, highlighting model (18 occurrences), prediction (14 occurrences), neural networks (15 occurrences), time series (13 occurrences), index (11 occurrences), machine (10 occurrences), LSTM and neural network (with 9 occurrences each).

As for the authorship of the works, 333 authors were identified. Figure 4 shows the ranking in descending order of the 26 host countries of the institutions to which these authors are associated.

![Figure 2. Annual distribution of papers.](image)

![Figure 3. Keyword co-occurrence map. Source: VOSviewer. Note: The size of the nodes represents the relevance of terms in the articles. The thickness of the lines means the strength of connection between them. Finally, the colors indicate the number of groups.](image)
Table 3. Plus keywords.

<table>
<thead>
<tr>
<th>Key words</th>
<th>The amount</th>
<th>Frequency %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>18</td>
<td>12%</td>
</tr>
<tr>
<td>Neural networks</td>
<td>15</td>
<td>10%</td>
</tr>
<tr>
<td>Prediction</td>
<td>14</td>
<td>9%</td>
</tr>
<tr>
<td>Time series</td>
<td>13</td>
<td>9%</td>
</tr>
<tr>
<td>Index</td>
<td>11</td>
<td>7%</td>
</tr>
<tr>
<td>Machine</td>
<td>10</td>
<td>7%</td>
</tr>
<tr>
<td>LSTM</td>
<td>9</td>
<td>6%</td>
</tr>
<tr>
<td>Neural network</td>
<td>9</td>
<td>6%</td>
</tr>
<tr>
<td>Volatility</td>
<td>7</td>
<td>5%</td>
</tr>
<tr>
<td>Classification</td>
<td>6</td>
<td>4%</td>
</tr>
<tr>
<td>Others</td>
<td>39</td>
<td>26%</td>
</tr>
<tr>
<td>Total</td>
<td>151</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 4. Publication of articles by country to which authors are associated.

According to the RStudio software, of the 99 articles, 77 (78%) are classified as articles written by authors associated with institutions in the same country (SCP), and 22 (22%) are articles written by authors associated with institutions in different countries. countries (MCP).

Figure 5 indicates that 653 citations are related to articles written by authors associated with institutions located in China. The other citations are from authors linked to institutions of the following in Korea (380), USA (298), Pakistan (120), India (69), and the other citations, scattered among 21 other countries.

Figure 6 shows the co-citation network among journals in the final sample of 99 articles. The most cited, according to the VOSviewer software, are Expert Systems with Applications, Knowledge-Based Systems (503 co-citations), IEEE Access (235 co-citations), Neural Computing and Applications (105 co-citations) and Soft Computing (42 co-citations).

Of these, only Expert Systems with Applications stands out below, indicating that the journals that publish the most on a given topic are not necessarily the most co-cited; this fact is actually due to the relevance of each published article.

Table 4 indicates the journals in which the 99 articles of the final sample are
published, through the application of Bradford’s Law [5]. The law states that there are few journals producing many articles and many journals producing few articles on a given topic. For Brookes [35], this law estimates the degree of relevance of certain academic journals that work in specific areas of knowledge. Thus, if the journals are classified in decreasing order of productivity, they can be distributed in zones with variation in the proportion 1:n:n^2, and so on.

Zone A is identified as the core of the disciplines, being composed of journals with 5 references or more, highlighting Expert Systems with Applications, IEEE Access, Big Data Journal and Neural Computing and Applications. Zone B presents periodicals with 2 to 4 publications, and Zone C, periodicals with a single publication.

Figure 5. Frequency of article citations in the countries of the institutions with which the authors are associated.

Figure 6. Map of co-citations between journals. Source: VOSviewer. Note: The size of the nodes represents the relevance of terms in the articles. The thickness of the lines means the strength of connection between them. The colors indicate the number of groups.
Table 4. Bradford’s law on journals.

<table>
<thead>
<tr>
<th>Zone</th>
<th>Daily</th>
<th>Individual quantity</th>
<th>Accumulated amount</th>
<th>Accumulation percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zone A</td>
<td>Expert Systems with Applications</td>
<td>12</td>
<td>12</td>
<td>12.1%</td>
</tr>
<tr>
<td></td>
<td>IEEE Access</td>
<td>11</td>
<td>23</td>
<td>23.2%</td>
</tr>
<tr>
<td></td>
<td>Big Data Journal</td>
<td>5</td>
<td>28</td>
<td>28.3%</td>
</tr>
<tr>
<td></td>
<td>Neural Computing and Applications</td>
<td>5</td>
<td>33</td>
<td>33.3%</td>
</tr>
<tr>
<td>Zone B</td>
<td>Applied Smooth Computing</td>
<td>4</td>
<td>37</td>
<td>37.4%</td>
</tr>
<tr>
<td></td>
<td>Multimedia Tools and Applications</td>
<td>4</td>
<td>41</td>
<td>41.4%</td>
</tr>
<tr>
<td></td>
<td>Scientific Programming</td>
<td>4</td>
<td>45</td>
<td>45.5%</td>
</tr>
<tr>
<td></td>
<td>Smooth Computing</td>
<td>4</td>
<td>49</td>
<td>49.5%</td>
</tr>
<tr>
<td></td>
<td>Forecast Diary</td>
<td>3</td>
<td>52</td>
<td>52.5%</td>
</tr>
<tr>
<td></td>
<td>Neurocomputing</td>
<td>3</td>
<td>55</td>
<td>55.6%</td>
</tr>
<tr>
<td></td>
<td>PEERJ Computer Science</td>
<td>3</td>
<td>58</td>
<td>58.6%</td>
</tr>
<tr>
<td></td>
<td>Algorithms</td>
<td>2</td>
<td>60</td>
<td>60.6%</td>
</tr>
<tr>
<td></td>
<td>Big Data</td>
<td>2</td>
<td>62</td>
<td>62.6%</td>
</tr>
<tr>
<td></td>
<td>Computational Economics</td>
<td>2</td>
<td>64</td>
<td>64.6%</td>
</tr>
<tr>
<td></td>
<td>Studies and Research in Economic Computing and Economic Cybernetics</td>
<td>2</td>
<td>66</td>
<td>66.7%</td>
</tr>
<tr>
<td></td>
<td>Quantitative finance</td>
<td>2</td>
<td>68</td>
<td>68.7%</td>
</tr>
</tbody>
</table>

Table 5 presents the ten most cited works on the RNN LSTM topic. Among them, the work by Kim and Won [20] stands out, with 235 (20.8%) of the citations and an annual average of 47.0. The article focuses on predicting the volatility of the stock price index, using a model that integrates the LSTM with several General Autoregressive models conditional Heteroskedasticity—GARCH. The second and third places are the articles by Long et al. [22] with 133 citations, an average of 33.3 per year, and Kudugunta and Ferrara [24] with 132 citations, an average of 26.4 per year.

In turn, Lotka [4] states that a small number of authors produce many works and that the production obtained by this small number of researchers is equal in quantity to the performance of the others. This law is called the inverse square law—see Equation (1).

\[ \text{an} = \frac{a_1}{n^2}, \quad n = 1, 2, 3 \]  

In which:
- \( \text{an} \) = number of authors who published \( n \) articles;
- \( a_1 \) = number of authors who published an article;
- \( n \) = number of articles published by author.

For Equation (2), Chung and Cox [23] clarify that the number of authors with a single published article, according to Lotka’s Law, would be:

\[ a_1 = \frac{6}{\pi^2} = 0.6079 = 60.8\% \]
Table 5. The ten most cited articles.

<table>
<thead>
<tr>
<th>References</th>
<th>Number of citations</th>
<th>Frequency of citations %</th>
<th>Total citations per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim and Won [20]</td>
<td>235</td>
<td>20.8%</td>
<td>47.0</td>
</tr>
<tr>
<td>Long et al. [22]</td>
<td>133</td>
<td>11.7%</td>
<td>33.3</td>
</tr>
<tr>
<td>Kudugunta and Ferrara [24]</td>
<td>132</td>
<td>11.7%</td>
<td>26.4</td>
</tr>
<tr>
<td>Baek and Kim [25]</td>
<td>123</td>
<td>10.9%</td>
<td>24.6</td>
</tr>
<tr>
<td>Bukhari et al. [26]</td>
<td>120</td>
<td>10.6%</td>
<td>40.0</td>
</tr>
<tr>
<td>Pang et al. [27]</td>
<td>120</td>
<td>10.6%</td>
<td>40.0</td>
</tr>
<tr>
<td>Sohangir et al. [28]</td>
<td>93</td>
<td>8.2%</td>
<td>19.6</td>
</tr>
<tr>
<td>Jin et al. [29]</td>
<td>70</td>
<td>6.2%</td>
<td>23.3</td>
</tr>
<tr>
<td>Borovkova and Tsiamas [30]</td>
<td>57</td>
<td>5.0%</td>
<td>14.3</td>
</tr>
<tr>
<td>Xing et al. [31]</td>
<td>49</td>
<td>4.3%</td>
<td>9.8</td>
</tr>
<tr>
<td>Total</td>
<td>1,132</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Thus, an author with two published articles must have a frequency of 15.2% \((0.6079/2^2)\). For an author with three published articles it would be 6.8% \((0.6079/3^2)\) and an author with four published articles would be 6.8% \((0.6079/4^2)\).

It appears that the 99 articles in the final sample are produced by 333 authors, with one author publishing 4 articles, six authors publishing 3 articles, twenty authors publishing 2 articles and three hundred and thirty-three authors publishing a single article. 27 authors (08%), including only those who publish the most, are responsible for 62 (18.2%) publications. That said, there are not a smaller number of researchers matching the performance of the others, making it impossible to confirm Lotka’s Law.

4.2. Systematic Review

A systematic literature review seeks to identify knowledge gaps related to the topic of this study. For this, in Step 6 of Item 3—Methodology, a (sub)categorization matrix is defined—see Table 2. Categories and subcategories are identified for each of the 99 articles in the final sample. In this way, the frequency count is made in relation to the total of the subcategories and not to the total of the 99 articles.

The subcategories that have the potential to be prioritized in future research are highlighted. In category 1, Neural Networks/Algorithms Used in the Research, the theme “Forecasting stock prices with other artificial neural networks and results compared to RNN LSTM” is the most relevant (65%), followed by “Forecasting stock prices with blended neural networks, including LSTM” (22%) and Share Price Projection with RNN LSTM (23%).

Regarding category 2 “Type of Data Analyzed”, the closing price was the most used data in the surveys, alone (22%) or associated with other data (61%).

Regarding the period of analysis, category 3, 37% of the articles were concen-
trated in periods of up to 05 years, 37% in periods of 05 to 10 years, 16% in peri-
dods of more than 10 years, and 09% did not inform.

In category 4, the objectives of the papers are highlighted. 74% of them project
the price of shares with LSTM alone or associated with the most varied RNN
(without relevant concentrations); 19% test hybrid neural networks and senti-
ment analysis models.

Category 5 indicates that 21% of the papers are exclusively based on data from
US stock exchanges, 20% exclusively use data from Chinese stock exchanges,
08% jointly use data from American and Chinese stock exchanges and other
associations. 38% of articles do not use data from the US or Chinese stock ex-
changes, but from the stock exchanges of Brazil, Thailand, Turkey, Tehran,
Ghana, Australia, Germany, Korea, India, Japan, Indonesia or the United King-
dom, in association or in isolation. Thus, it can be seen that there is plenty of
room for studies based on Stock Exchanges in other developed and/or develop-
ing countries. 08% perform sentiment analysis, that is, they use texts exposed in
the media, and not data from Stock Exchanges.

In turn, category 6 presents the results of the studies carried out. In 57% of
them, the proposed models outperform the results of the models to which they
were compared; in 42% the results are defined as promising, indicating that they
can be improved.

According to category 7, 57% of the studies present adjustments in already
tested neural network models, with improvement in the quality of input infor-
iation and/or other innovations in existing models. 42% of the articles present
new theories or new models of projections, using simple, hybrid or combined
neural networks.

Finally, category 8 indicates paths for future studies, that is, knowledge gaps
according to the authors of the 99 papers in the final sample. In 27% of the pa-
ers, the authors suggest the use of innovative data, such as other news sources
and/or stock data, in different time periods, such as intraday. 22% suggest stu-
dies with hybrid models of LSTM not yet tested, associated or not with other
neural networks, other types of data, other sources and other periods. 10% sug-
gest studies with other types of neural networks, pure or hybrid.

5. Conclusion

Publications on this topic are concentrated from 2020 onwards. The keywords
most associated with these studies are model, neural networks, prediction and
time series. 333 authors wrote on the subject between 2018 and March 2022; 43
of the 99 articles published in this period are associated with Chinese institu-
tions. The journals that publish the most significant articles on the topic are Ex-
pert Systems with Applications, IEEE Access, Big Data Journal and Neural
Computing and Applications. The most cited article is by Kim and Won [20],
Stock Price Index Volatility Prediction: A Hybrid Model Integrating LSTM with
Various GARCH-Type Models, which studies the volatility of Kospi 200 stock
index returns and capitalization of the stock market in South Korea, cited 235 times. The daily closing price of shares is the most analyzed type of data, and studies are still concentrated on American (21%) and Chinese (20%) stock exchanges. 57% of the studies present adjustments to already tested neural network models and 42% present new theories or new projection models. In 27% of the articles, the authors suggest future studies with news sources, other stock data, or the use of different time series. 22% suggest studies with hybrid models of LSTM not yet tested, associated or not with other neural networks, other data, other sources and periods.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


