ISSN : 2327-7211

Journal of Data Analysis and Information Processing

Volume No. 12

Issue No. 1

January - April 2024

ENRICHED PUBLICATIONS PVT. LTD

S-9, IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR-5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-47026006

ISSN : 2327-7211

Journal of Data Analysis and Information Processing

Aims & Scope

Journal of Data Analysis and Information Processing (JDAIP) is an international journal dedicated to the latest advancement of data analysis and information processing methods. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas of data analysis and information processing.

All manuscripts must be prepared in English and are subject to a rigorous and fair peer-review process. Generally, accepted papers will appear online within 3 weeks followed by printed hard copy. The journal publishes original papers including but not limited to the following fields:

· Application of Technology in Policy and Decision Making
· Application of Technology in Teaching and Learning Process
· Artificial Intelligence
· Artificial Neural Networks
· Big Data Analytic
· Business Intelligence
· Computational Intelligence
· Data Mining
· Data Modeling
· Deep Learning
· Digital Signal Processing
· Dimension Reducing
· Empirical Mode Decomposition
· Fourier Analysis
· Graph Mining
· Image Processing, Video Processing
· Information Extraction
· Information Hiding
· Information Retrieval
· Information Theory
· Knowledge Discovery

ISSN : 2327-7211

· Machine Learning
· Multivariate data analysis, Matrix Analysis
· Pattern Recognition
· Semantic Technology
· Social Network Mining
· Spatial and Temporal Data Analysis
· Statistics
· Statistics and Applied Probability
· Stochastic Process
· Stream Mining
· Time Series Analysis
· Wavelet Analysis

We are also interested in: 1) Short Reports – 2-5 page papers where an author can present either an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book Reviews – Comments and critiques.

ISSN : 2327-7211

Journal of Data Analysis and Information Processing

ISSN : 2327-7211

Journal of Data Analysis and Information Processing

(Volume No. 12, Issue No. 1, January - April 2024)

Contents

Dixit Player with Open CLIP

Ryan Wei
Syosset High School, New York, USA

ABSTRACT

A computer vision approach through Open AI's CLIP, a model capable of predicting text-image pairs, is used to create an AI agent for Dixit, a game which requires creative linking between images and text. This paper calcu lates baseline accuracies for both the ability to match the correct image to a hint and the ability to match up with human preferences. A dataset created by previous work on Dixit is used for testing. CLIP is utilized through the com parison of a hint to multiple images, and previous hints, achieving a final accuracy of 0.5011 which surpasses previous results.

Keywords : Computer Vision, AI, CLIP, Dixit, Open AI, Creative Gameplay, Open CLIP, Natural Language Processing, Visual Models, Game AI, Image-Text Pairing

Introduction

In recent years, various board games such as chess have served as benchmarks for progress in AI. However, this research has focused primarily on logical, de terministic games, creating a void in AI research centered on creative and social gameplay [1]. We attempt to begin filling this void by creating an AI which can play the game Dixit [2].

Dixit is a complex game which demands logical, creative, and social ability. It is a challenging benchmark for the creative capabilities of AI and serves as a platform to improve models which connect images and text. In each episode of the game, a storyteller must carefully choose a card and a corresponding de scription for other players to base their card selections on. Each player then votes for which card they believe is the storytellers.

We attempt to build an AI agent which can accomplish the task of guessing the card which either successfully matches up to the storytellers' or matches up to the human choice. Previous work on Dixit [3] has used basic machine learn ing algorithms, achieving slightly better results than human counterparts on identifying the storyteller's card (which is one of the key tasks for a Dixit player). An important way in which we capitalize on this prior work is the use of the Dixit play data shared by the authors of [3].

The method achieving the best results in [3] is based on the well-established TF-IDF features. In contrast, we propose a new and more modern approach, based on computer vision and natural language processing models, namely CLIP [4] [5].In our experiments, we consider two key tasks facing a Dixit player: identifying the storyteller's card and predicting which card in a lineup would garner most votes from other players (note that the latter may not be the same as the storyteller's card)! We show that our proposed method, based on evaluating card-to-hint relevance using "historical" play data in the training set, does better on both tasks than a number of previously proposed baselines.

2. Our Approach

2.1. Dixit

Dixit was chosen due to our belief that it was a good test of Open AI's zero-shot capabilities. The 84 cards which Dixit uses are abstract, artistically provocative paintings which often result in less literal

descriptions. Hints describe an emo tion evoked by the respective cards or are explained through a cultural reference.

This is magnified by the scoring system. In the game, a storyteller is encour aged to generate a description which is not too obvious, but not too vague since the best outcome for their score occurs when some, not all, players guess the storyteller card. Two examples of a matching hint-card pair are displayed in Figure 1. Due to the nature of the game, a player must excel at the task of asso ciating abstract/creative descriptions with the correct image.

2.2. The CLIP Model(s)

Open AI's CLIP [4] is a model which attempts to align image and text. It is trained on a large dataset (originally of 400 million text-image pairs, although subsequent efforts trained CLIP on even larger datasets) acquired from the internet, utilizing contrastive representation learning to maximize the cosine scores of the correct image-text pairs.

CLIP simultaneously trains an image encoder (mapping images to a vector in a 512-dimensional embedding space) and a text encoder (mapping text to a vec tor in the same embedding space). The training objective is to project the matching pairs close to each other and farther away from others. The similarity is measured by the cosine between two vectors. Note that while the training objective focuses on matching images and text, the cosine similarity can also be used to judge association between two images or two text strings. A concise summary is displayed in Figure 2.

Due to the dataset and its training, CLIP has displayed impressive zero shot performance achieving state of the art image recognition abilities [4] [5].

(a) (b)

Figure 1. (a) "Gotham City's sidekick" referring to Robin, Batman's, sidekick. Painting interpreted as robin eggs; (b) "Finally!" A look of many emotions is displayed on the girl's face as she comes across a flower.

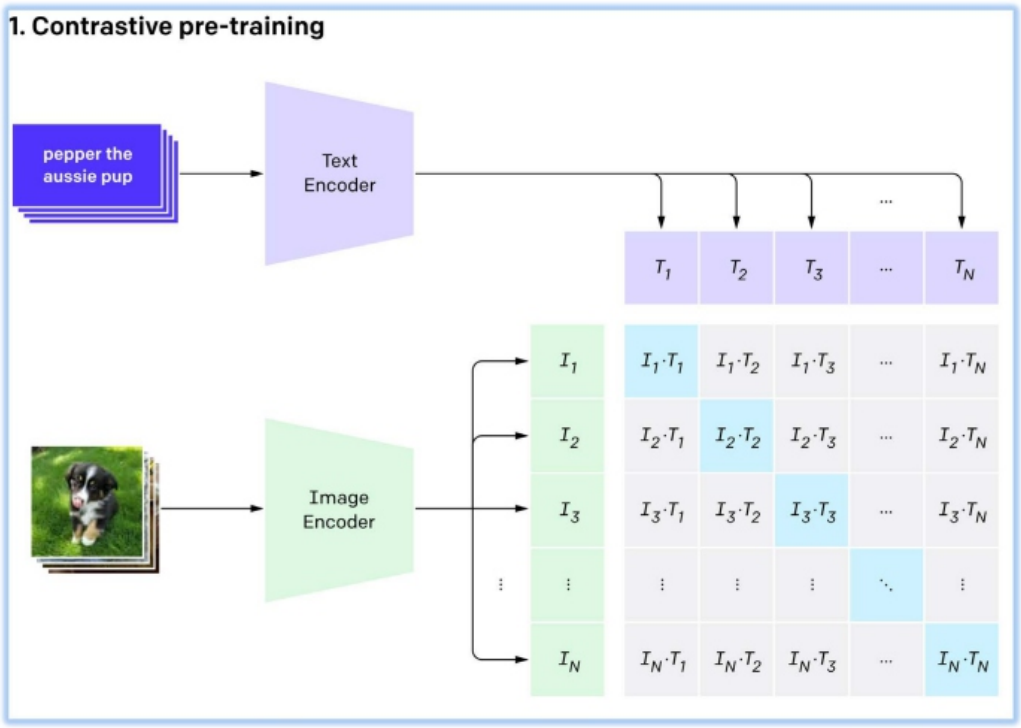


Figure 2. Simple summary of CLIP training, taken from [4].

However, CLIP has primarily been tested and trained on standard datasets of text-image descriptions, meaning that its ability for image recognition with more abstract/creative descriptions such as in Dixit has not been extensively evaluated.

2.3. Dataset

Our proposed approach uses the data set collected by the authors of [3]. Each"episode" provides information from a single turn of Dixit played by human players on an online platform. The data include: the storyteller's card ID, storyteller's hint, IDs of other cards played in response to the hint, and the votes re-ceived by each card. Each episode comes from a game played by 4, 5 or 6 players.

The dataset includes a partition into train, validation, and test sets, with 92,981, 11,624 and 11,624 episodes respectively; we maintain this partition in our experiments.

Formally, let the set of cards played in an episode i be { 1 , , i} I ip c c , where piis the number of players in the game (4, 5 or 6), and each ij c is an index into the 84 Dixit cards. The storyteller's hint is hi (a text string), for the storyteller's card 1i c . For each ij c we have ij v the fraction of the players who voted for ij c based on hint hi. The episode data then includes {pi, ci.....cpi, hi ,, , ,,, ,vi, pi1}

2.4. Hint History

For the final CLIP model, we use the hint history for each card, meaning that the hints of each episode in the training data are encoded and stored in the corres ponding storyteller card list. Formally, in each episode i of the training set, the encoded hint, hi is stored in the list corresponding to card 1i c . After our compu tation, we end with a dictionary of 84 terms, each containing the embedded ver sions of all associated hints.

For each episode in the validation test set, containing pi cards, hint hi, and storyteller card 1i c , hi's similarity score is compared with the list of previous hints used for card 1i c in the training set. We test a few different evaluations on the resulting scores.

Min Similarity between the card c and the hint h is determined as

$$\min_{i:c_1^i=c} \cos\left(CLIP\left(h^i\right), CLIP(h)\right) \quad (1)$$

Intuition: High minimum value indicates that no training play episode in which c was the storyteller's card had a hint that was very different from h, thus suggesting that h may be a good match to c.

Max Similarity between the card c and the hint h is determined as

$$\max_{i:c_1^i=c} \cos\left(CLIP\left(h^i\right), CLIP(h)\right) \quad (2)$$

Intuition: High maximum value indicates a training play episode in which c was the storyteller's card had a hint that closely resembled h, thus suggesting that h might be a good match to c.

Top 5 Similarity between the card c and the hint h is determined as

$$\max \sum_{\substack{i=0 \\ rsorted\left(i:c_1^i=c\right)}}^{4} \cos\left(CLIP\left(h^i\right), CLIP(h)\right) \quad (3)$$

Intuition: A high sum of the top 5 values indicates multiple training play epi sodes in which c was the storyteller's card which had a hint that closely resembled h, thus suggesting that h might be a good match to c.

Average Similarity between the card c and the hint h is determined as

$$\sum_{i:c_1^i=i} \cos\left(CLIP\left(h^i\right), CLIP(h)\right) \Big/ \left|i : c_1^i = c\right| \quad (4)$$

Intuition: High average value indicates a good overall similarity between h and all hints associated with c in training, thus suggesting that training episodes in which c was the storyteller's card had generally high similarity scores to h.

Range Similarity between the card c and the hint h is determined as

$$\max_{i:c_1^i=c} \cos\left(CLIP\left(h^i\right), CLIP(h)\right) - \min_{i:c_1^i=c} \cos\left(CLIP\left(h^i\right), CLIP(h)\right) \quad (5)$$

Intuition: A high range indicates higher maximum scores, but a lower minimum score could cause range to perform poorly. This is not an evaluation which is expected to improve the overall accuracies, it is meant to depict how influential maximum and minimum are.

3. Experiments and Results

Our approach yields successful results in comparison to baselines and other methods for both accuracy in selecting the storyteller's card and for matching up to human preferences. Card selection accuracy is displayed in Table 1 while the accuracy of the model matching up to human behavior is displayed in Table 2. We also calculate the KL divergence between the distributions of card choices made in each episode of the dataset and the distributions created by the AI, displayed in Table 3.

3.1. Baseline Approaches

Before testing how capable open CLIP is with Dixit data, we first implement two basic strategies of selection to serve as a baseline. The first is to randomly select a card. The second is a recreation of the baseline described in Vatsakis et al.—the selection of the card which was most frequently chosen as the storytellers in the dataset.

Naïve CLIP

Our third baseline is a naive use of CLIP. We embedded the hint and the 4, 5, or 6 images before calculating the cosine similarities between the hint and images for each episode. This gives us 4, 5, or 6 values which we normalize into a softmax distribution. The card with the corresponding max probability in the softmax is the AI's choice.

3.2. Ability to Choose the Correct Card

The AI performs the task well, surpassing the numbers achieved by the baselines, humans, and the Vatsakis model. All accuracies are displayed, split into 5091, 1947, and 4585 episodes of 4, 5, and 6 players respectively, as well as the overall mean accuracy. The 5 different evaluations of our resultant matrix are displayed, with top 5 achieving the best accuracy. To cut down significantly on computation time, we precompute the embeddings of all 84 cards and corresponding hint histories, combined into a 512 × ni matrix, where ni represents the number of hints chosen for card i. This precomputation took about one hour to process,

Table 1. Accuracies of guessing the storyteller's card selection.

Method	Total	4 players	5 players	6 players
Random baseline	0.2086	0.2509	0.2025	0.1657
Frequency baseline	0.2090	0.2931	0.2332	0.2076
Naive CLIP	0.2767	0.3239	0.2661	0.2236
Human, from [3]	0.4782	0.542	0.472	0.410
Keyword model [3]	0.4042	0.443	0.407	0.360
Full model [3]	0.4793	0.523	0.488	0.427
Hint history (max)	0.4590	0.4970	0.4504	0.4207
Hint history (min)	0.2117	0.2497	0.2013	0.1740
Hint history (avg)	0.3180	0.3571	0.3063	0.2604
Hint history (top 5)	**0.4995**	0.5343	0.4972	0.4619
Hint history (max range)	0.4223	0.4614	0.4304	0.3754

Table 2. Accuracies of selecting card(s) preferred by human players.

Method	Total	4 players	5 players	6 players
Random baseline	0.2086	0.2509	0.2025	0.1657
Frequency baseline	0.2090	0.2931	0.2332	0.2076
Naive CLIP	0.3969	0.4585	0.3878	0.3324
Hint history (max)	0.5045	0.5608	0.4992	0.4443
Hint history (min)	0.2893	0.3518	0.2717	0.2275
Hint history (avg)	0.4137	0.4722	0.4299	0.3418
Hint history (top5)	**0.5647**	0.6154	0.5547	0.5125
Hint history (max range)	0.4632	0.5258	0.4520	0.3754

Table 3. KL divergence numbers.

Method	Total	4 players	5 players	6 players
Random baseline	2.299	3.302	2.196	1.229
Frequency baseline	0.2090	0.2931	0.2332	0.2076
Naive CLIP	1.361	2.019	1.282	0.6653
Hint history (max)	1.750	2.514	1.698	0.9227
Hint history (min)	2.422	3.507	2.306	1.266
Hint history (avg)	2.142	3.083	2.052	1.134
Hint history (top 5)	**1.634**	2.359	1.570	0.8561
Hint history (max range)	1.887	2.717	1.826	0.9914

cutting down the run time of testing to 10 seconds. The results for the different evaluations are displayed in Table 1.

While running on the validation test set, each hint is encoded (into a 512 × 1 vector), transposed, and multiplied with the hint history matrices of $\left\{c_1^i,\cdots,c_{p^i}^i\right\}$.

3.3. Comparison to Human Behavior

The second metric is to judge the AI's ability to match up with human behavior. Recall from the Introduction that this is related to the strategic goals of Task B. For each episode, the cards that were chosen with the highest frequency in the data were stored. If multiple cards shared the highest probability, they would all be viable choices for the AI. There is no benchmark to compare to, but more than half of the AI's choices matching up with the preferences of human players is an impressive start. Additionally, the ranking of the methods is the same as shown in Table 2.

3.3.1. KL Divergence

If we consider the goal of "replicating human judgment" by the AI player, we need to look beyond selecting the winner of the vote. The vote data available to us from [3] is indeed more detailed. For instance, if 3 players out of 6 vote for card 1, 2 vote for card 4 and 1 for card 3, then we have information beyond "card 1 is the winner"—we can also aim to estimate the full scope of human preferences. We can treat the vote distribution in a game episode (that sums to 1 over the cards) as a "true distribution" of the human vote; intuitively, if the vote fraction for a card is $x \in [0, 1]$ we treat it as "the probability of a human player voting for this card is x". We would like the AI player to predict this probability. To this end, we compare the probability distributions generated by the AI, converted into a vote distribution summing to 1 through applying the softmax distribution, and the "human probability distributions" in the dataset through KL divergence[6]. We use KL divergence as opposed to other calculation types to tell us how much information is lost, giving us a quantifiable number. Again, the method rankings are the same, as indicated by Table 3.

3.3.2. Combining Naïve CLIP with Hint History

In order to get the best possible outcome, we combined the two methods tested with CLIP: comparing the hint to each image and comparing the hint to the hint history for each card. This would also more closely resemble a human's thought process when playing Dixit. We added the two probabilities, weighing their influences, before converting the sums into a softmax distribution again. The differences weren't significant, but there was still a small increase. We adjusted the weightings by 0.01 for 100 iterations. The best results were 0.75 and 0.25 for the naive CLIP and the hint history, respectively, raising the accuracy to 0.5020 from 0.4995.

3.3.3. Restricting Hint History

We also wanted to investigate how big of an effect the amount of training data the AI could see would have on the overall accuracy. We randomly selected 500, 250, 100, 25, 10, and 5 hints from the hint history and calculated the scores using the algorithm described in section 3.3.2. Results are displayed in Table 4.

3.4. Final Results

Table 5 displays the result of our best model, combining naive CLIP and hint history in a 0.75 and 0.25 weighing. On the validation, the overall accuracy is 0.5020 and on the test dataset, it is 0.5011, surpassing the overall accuracy of humans and Vatsakis as displayed in Table 5.

4. Discussion

In our experiments there were some surprising results. The most notable of these was that averaging the scores does relatively poorly, while range does de cently well, a result which we didn't expect. To understand the accuracies, we closely examined a few specific episodes, finding one particularly illustrative of the trends in the data. In the 48th episode of the training dataset, we are given a hint of "Don't trust Fibonacci." along with 6 cards [11, 48, 12, 25, 19, 79] (displayed in Figure 3) where card 11 is the storyteller's card. The image/text match ing gives us a softmax distribution of [0.2673, 0.1433, 0.2467, 0.1455, 0.1957,0.2184] respectively. The softmax for the history matching is [1, 0, 0, 0, 0, 0]. The top 5 terms for card 11 and the 5 other cards are displayed in Table 7 and Table 8. Combining this through addition and weighing the two probabilities in a 0.75, 0.25 split, we get a final softmax distribution of [0.4505, 0.1075, 0.1850, 0.1091,0.1468, 0.1638]. The first probability is the largest and that is the AI's answer, the correct answer.

From this data, we can draw a few conclusions. Fibonacci is a common term associated with explaining card 11 and other cards. Card 19 has 4 descriptions of "Fibonacci," but its last hint is "Golden ratio". These differences illustrate the effectiveness of the top 5 method over taking the max. For card 79 and 25, the hints don't bear much resemblance to the storyteller's hint. Other cards don't have the same volume of use of the term, "Fibonacci," while card 11 does. Extending the history evaluation to the top 10 scores gives us an even closer look at the AI's process.

The corresponding similarity scores in Table 6 show that after there are no longer any hints which contain the word fibonacci for card 11, the scores begin to drop off. One of the first hints below this batch is "Golden ratio." Other hints include "Freebonacci", "mathematics", and "Either way could be interesting," which had similarity scores between 0.6 and 0.8. There was a noticeable differ ence for these hints, but they still maintain some sense of similarity with the original hint.

This episode, along with others, confirmed our observation that for each card, there are multiple different elements and emotions evoked, and each one could be described in multiple different ways. The fibonacci example showed that de scriptions tend to come in batches, each batch describing a specific element of the card in a certain way.

Table 4. Effect of restriction of hint history data on accuracy. Random selections of the number of restrictions for each of the 84 card dictionaries were made

Restriction	Overall	4 players	5 players	6 players
500	0.4600	0.4946	0.4612	0.4209
250	0.4236	0.4569	0.4201	0.388
100	0.3775	0.4109	0.3780	0.3402
25	0.2937	0.3229	0.2964	0.2602
10	0.2512	0.2942	0.2450	0.2061
5	0.2309	0.2754	0.2142	0.1887

Table 5. Final accuracy on validatoin and test datasets for our best method along with Vatsakis numbers.

Data set	Overall	4 players	5 players	6 players
Val	0.5020	0.5343	0.5039	0.4654
Test	0.5011	0.4516	0.4893	0.4592
Val, from [3]	0.4793	0.523	0.488	0.427
Test, from [3]	0.4731	0.516	0.472	0.424

Figure 3. 6 cards [2] in a specific episode of the dataset provided by [3].

Table 6. Top 5 similiarity scores calculated by the AI player for each card in this episode. Corresponding hints are displayed in Table 7 and Table 8.

Card	1st	2nd	3rd	4th	5th
11	0.9165	0.9165	0.9120	0.9120	0.9120
48	0.9120	0.6062	0.6004	0.6004	0.5913
12	0.9120	0.8873	0.8148	0.6622	0.6081
25	0.6227	0.6040	0.6016	0.5966	0.5944
19	0.9120	0.9120	0.9120	0.9120	0.6786
79	0.6384	0.6106	0.6077	0.6071	0.6059

Due to the wide scope of the game, many different cards can result in similar descriptions, meaning that taking the max only was not as effective. Top 5 appeared to be the sweet spot from our testing, as adding more tends to take away from the influence of the most similar hints. This also explains why averaging doesn't work well, as each card may have multiple batches of descriptions which are completely different to the given hint, pulling the score of the correct card down. In our example, Fibonacci was a relatively common description, taking up 10 hints for the drop off from the first batch to the second batch to occur. The more unique descriptions would have much sharper drop-offs.

This reasoning also explains why taking the largest minimum performed poorly. Due to the diversity of descriptions, a very small or large minimum did not mean much. The minimum's ineffectiveness is further highlighted by the relative success of the range. High range indicates a high maximum value, while the minimum doesn't have much of an effect and would likely be similar across the cards. Perhaps unsurprisingly, we observe the same trends for the AI-human evaluation.Limiting hint history showed a drastic drop off, until it began to hold back the naive CLIP method, as the accuracies show in Table 4. On average, each card contained 1100 hints, while the minimum number was 767. The accuracy dropped at a steady rate, and it generally took at least 15 hints to improve on naive CLIP. It took a limitation of 500 to drop the accuracy by 0.042.

Overall, however, we were able to improve on the numbers achieved by Vatsakis et al., showing that the recent advancements in computer vision are more effective than traditional machine learning algorithms for image recognition tasks. It is likely that the consideration of both text and image creates a more ba lanced judgement of each episode. Additionally, analyzing the images means there is always an impartial aspect to the calculation—the image is always the same, unlike the descriptions which can occasionally be difficult to understand for even humans. The amount of training data is another key aspect in the suc cess of the agent, as many obscure hints that are difficult for CLIP to understand are covered due to similar hints in training being recognized by CLIP's zero-shot ability.

Table 7. Top 5 hints for cards 11, 48, and 12.

Card 25	Card 19	Card 79
It's not much but it's honest work	Fibonacci	I don't think this is correct
I didn't expect this…	Fibonacci	This doesn't make ANY sense
Not my problem anymore	Fibonacci	It's supposed to do that
Every 6th year, this doesn't happen	Fibonacci	Not correct
I'm not sure how long this will last	Golden ratio	This just doesn't make any sense

Table 8. Top 5 hints for cards 25, 19, and 79

Card 25	Card 19	Card 79
It's not much but it's honest work	Fibonacci	I don't think this is correct
I didn't expect this…	Fibonacci	This doesn't make ANY sense
Not my problem anymore	Fibonacci	It's supposed to do that
Every 6th year, this doesn't happen	Fibonacci	Not correct
I'm not sure how long this will last	Golden ratio	This just doesn't make any sense

Limitations

We improve the accuracy of our AI agent to surpass that of Vatsakis and humans. However, the data is from a casual website, requiring only registration to play. Many rounds are played by casuals and first-timers, meaning that the human accuracy score may not be the best benchmark. Although Dixit is a complex game which requires a specific skill set that test understudied elements of current AI, it is not well known. The quality of the data and our benchmarks/baselines may be questioned. However, our numbers are still impressive for a task that is not easy for both humans and AI.

There are more improvements which can be made for AI Dixit players. The most obvious one is that of fine-tuning CLIP for Dixit. CLIP's zero-shot abilities are great, but it can still improve through optimization of its parameters. This is especially true for image-text pairs in Dixit. Additionally, we did not address the topic of generating descriptions, a task that is much more difficult. It would likely require extensive training and optimizations in order to create a model which works well, along with more time in order to test it.

5. Conclusions

In this paper, we developed a Dixit AI agent, utilizing the capabilities of CLIP in order to obtain the best accuracy possible for choosing the correct card or matching human behavior. We obtain a 0.5003 accuracy rate on the test data, surpassing that of humans (0.4782) and the Vatsakis model, (0.4793). With extensive training and fine-tuning, this number can likely be improved.

Creating an AI agent which can guess the correct card correctly at a greater rate would be an impressive step of advancement for computer vision, widening the scope of its ability to identify abstract and creative image-text pairings.

Another task would consist of being able to generate the Dixit descriptions effectively, a challenging, but interesting task. CLIP could be used as a function to calculate how effective certain words and certain strings of words are and trained to prefer the ideal types of descriptions in Dixit. This is a task that would be time-intensive and challenging, but still interesting.

Acknowledgements

We thank Greg Shakhnarovich for guidance and many helpful discussions throughout the project.

Conflicts of Interest

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

References

[1] Singh, A.K., Ding, D., Saxe, A., et al. (2022). Know Your Audience: Specializing Grounded Language Models with the Game of Dixit. arXiv: 2206.08349. https://doi.org/10.48550/arXiv.2206.08349

[2] Roubira, J.-L. and Carpuat, M. (2008) Dixit. [Board game].

[3] Vatsakis, D., Mavromoustakos-Blom, P. and Spronck, P. (2022). An Internet-Assisted Dixit-Playing AI. Proceedings of the 17th International Conference on the Foundations of Digital Games, Athens, 5-8 September 2022, 1-7. https://doi.org/10.1145/3555858.3555863

[4] Radford, A., Kim, J.W., Hallacy, C., et al. (2021) Learning Transferable Visual Models from Natural Language Supervision. arXiv: 2103.00020.https://doi.org/10.48550/arXiv.2103.00020

[5] Cherti, M., Beaumont, R., Wightman, R., et al. (2022) Reproducible Scaling Laws for Contrastive Language-Image Learning. arXiv: 2212.07143.https://doi.org/10.48550/arXiv.2212.07143

[6] Han, J. and Kamber, M. (2008). Introduction to Data Warehousing and Mining. https://hanj.cs.illinois.edu/cs412/bk3/KL-divergence.pdf.

Correlation Analysis between Ethnic Diversity and Success Rate on a Massive Repository of Movies Data Set and the Board of Directors of Fortune 500 in Terms of Net Sales and Gross

Sarah Bamatraf
Strategic Research Section UAE Space Agency Abu Dhabi, Abu Dhabi, UAE

ABSTRACT

In contemporary workplace, organizations are emphasizing on individual's diversity and inclusion initiatives in order to reinforce managerial adaptability, increase competitive advantage and decrease legal risks. Nonetheless, in recent times, there has arisen a debate on whether diversity is a variable that has an immediate effect on success or not. This study focused on determining if diversity in terms of ethnicity, gender, age, etc., has effects on success, by investigating two different data sets; the first one is a massive repository of movies data set and actors to determine if there is a correlation between multiple movie related variables and box office earnings. While the second data focused on Fortunes top 500 companies in the United States (US) vs. 500 less profitable companies in the US. Moreover, the study explores how diversity among Board of Directors (BOD) of fortune 500 companies affects the net sales and gross profits. The movie data set was collected from two main websites; Internet Movie Database (IMDB) and Rotten Tomatoes (RT), the imdb data set contained 107,645 records, while as the rotten tomatoes contained 13,904 records. In addition, information about Fortunes 500 companies was obtained from various websites manually, as immediate data sets were hard to find since it's the first study that focuses on diversity and success of fortune companies. The data set contained data of fortunes top 500 companies with information of all of its BOD about 5358 records, and less profitable companies of 4434 records. The reason in which these data sets were chosen was to study the ethnic diversity factor and its impact on success rate, and also due to the fact that IMDB and Rotten Tomatoes are the most recognized websites that provide access to a massive repository of movie data sets. While the fortune company's data set was chosen to demonstrate diversity in the chosen dataset where one was for movies and the other was enterprise based. Furthermore, the data was analyzed in python to establish the relationship between the various variables. In all of the correlation analysis, the Pearson's coefficient was less than 0.1. Therefore, it was concluded that ethnic diversity has an insignificant effect on the success of movies and the Fortune 500 companies.

Keywords Diversity, Ethnicity, Repository

1. Introduction

1.1. Problem Definition

Recently, organizations are focusing more on corporates diversity to reinforce organizational adaptability and encourage competitive advantage. However, a debate has arisen that tackles whether diversity is a variable that has an immediate effect on success or not. The advancement in technology and the easiness of travel in the 21st century have caused a fundamental change in workplace dynamics. Cross-border trade and investment barriers have been eradicated by the advancement in telecommunication and transportation.

The breaking of the transportation and telecommunication barriers has consequently diversified the

workforce in various organizations. According to [1], workforce diversity refers to similarities and differences that occur among a company's employees or a group of people working together to achieve a common goal. The similarities and difference are in terms of race, cultural background, age, physical abilities, physical disabilities, gender, sexual orientation, and religion. Diversity is known to make a working environment heterogeneous.
Various researches have been done to establish the effect of diversity on the success of a company or a group of people working together towards a common goal.

1.2. Motivation

One of the ways that organizations seek to improve their performance is through diversification of workforce. In recent years, many organizations and companies have implemented diversity initiatives that are beyond the traditional monolithic structure, compliance standards and affirmative action. Employers seek to employ competitive workforce regardless of age, gender, race, ethnicity, religion, language, perception, and attitude. As the global economy continues to expand, business leaders have learned the value of a multi-cultural workforce with regard to different aspects of their businesses. There is a greater understanding of diversity as a competitive edge when leveraged. Thus, the need for it to be accounted for in the workforce equation for continued development. Advantages of diversity in work place [2]:

- Improves approachability to new and diverse customer marketplaces.
- Increases innovation and productivity.
- Increases revenue.
- Leads to the development of new products and services.
- Allows greater flexibility and adaptability in a more globalized environment.
- And improves social cohesion.

1.3. Objectives

1) To investigate impact of ethnic diversity on the success of a movie in terms of movie ratings and Box Office earnings.
2) To investigate the impact of ethnic diversity of the Board of Directors of Fortune 500 companies on the success of the company in terms of net sales and gross profit.

1.4. Hypothesis

Since the research is carried out on two independent data set (movie dataset and Fortune 500 company's dataset), there are to null hypothesis and two alternative hypotheses for each dataset.

• Null hypotheses

1) H01 Ethnic diversity of movie writers and directors has no impact on the success of the movie.
2) H02 Ethnic diversity of Board of Directors of Fortune 500 Company has no impact on the net sales and gross profit.

• Alternative Hypotheses

1) Ethnic diversity directly affects the success of a movie.
2) Ethnic diversity of the board of directors has an impact on the net sales and gross profit.

1.5. Research Contribution

The study focused on the effect of ethnic diversity to the success of a movie in terms of movie ratings and

box office earnings and, the effect of ethnic diversity on the BOD of a company and the success of the company in terms of net sales and net profits. In the research, movie data from The Internet Movie Database (IMDB) and Rotten Tomatoes database were used. Also, data of Fortune 500 companies were used. The Internet Movie Database is a movie repository that store comprehensive data about movies, scriptwriters, movie directors, movie release dates among other relevant details. The IMDB repositories stores more than 900,000 movie titles. Furthermore, the site data contained more than 2.3 million individuals. Also, data contains details of scriptwriters, movie directors, movie producers, movie reviewers among others. The in-depth storage of data in the Internet Movie Database makes it a rich source of information for various analyses. Figure 1 displays the most common genres provided by IMDB Website and the number of movies for each genre within the dataset.

Rotten Tomatoes is a repository for movie data as well as movie review aggre gation. The Rotten Tomatoes provides a platform whereby movies can be revised by professional movie reviews, movie critics, and the audience. The website uses a special algorithm to combine the reviews into a single aggregation referred to as the Tomatometer. The Tomatometer is trusted by millions of people around the world in determining whether to watch a movie not based on the review.
The research further utilizes data from Fortune 500 companies, Fortune 500 companies refer to a list that is compiled annually by the Fortune magazine. The list ranks the top 500 companies in the United States of America based on the total revenue per given fiscal year. The data provided by the Fortune magazine includes details about the Chief Executive officers of the company, the list of the
Members of Board of Directors, Net profits, and Net sales among other impor tant information. The dataset from the repository mentioned above is collected and analyzed in python using important python libraries.

2. Related Work

Social media is considered a massive source for sharing contents, thus, giving the liberty for millions of users to comment on all type of subjects on a daily Furthermore, it is evident that businesses consider these massive repositories as a rich source for valuable data as they have a strong interest in tapping in that world in order to gather information that improves their decision-making process.
As an example, using social media for creating predictive models that helps filmmakers make more profitable decisions [3].

The Internet Movie Database (IMDb) is an online comprehensive database that contains information related to movies, actors, television shows, production, etc. Furthermore, imdb features 963,309 movie titles, around 2,297,335 ac tor's data [4]. And has a separate web page dedicated for each of the actor's his tory and information. In addition, it offers ratings for each movie by aggregating the results of the overall rating given by the audience [5].

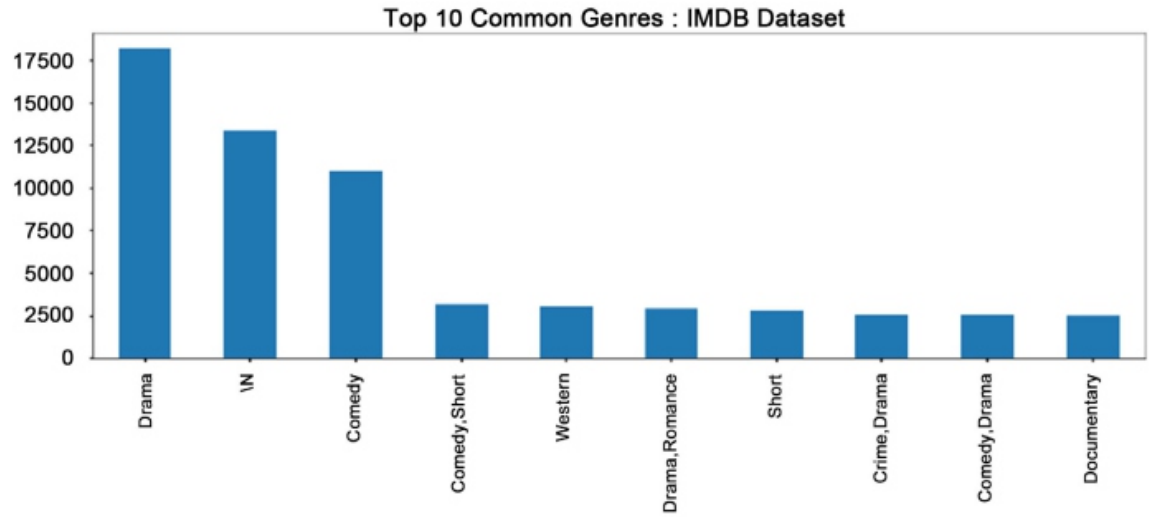


Figure 1. Top 10 common genres provided by IMDB.

Rotten Tomatoes is a website that is considered a rich source for movies data.
It computes rating by a measurement known as the Tomatometer; which is basically a measurement of quality entertainment, representing the percentage of positive expert reviews for films and TV shows to help users with their entertainment viewing decisions as it displays a comprehensive guide to what to watch for the audience [6]. Oghina, Breuss, Tsagkias, and de Rijke [7] examined Information Retrieval (IR) system; a system that uses various signals from various sources for ranking objects efficiently. Moreover, [7] have focused on predicting movie ratings from numerous social media signals. In order to efficiently improve racial and ethnic diversity in a workplace environment, organizations are required understand some of the principle terms and definitions that includes the following [8]:

• Racial Discrimination: racial discrimination in a workplace is defined as any act of exclusion, restriction or preference that is based on race, color, descent or national or ethnic origin, which prevents an employee's ability to exercise their rights to be equally treated in a workplace [8].

• Ethnic Group: defined as a group of individuals whom members are identified through factors such as common heritage, culture, ancestry, language, dialect, history, identity and geographic origin [8].

• Ethnic Minority: is defined as any ethnic group that is not dominant socially, economically or politically [8].

• Implicit Bias: negative associations that people unknowingly hold. They are articulated automatically, without conscious awareness [8].

• Inclusion: authentically incorporating traditionally excluded individuals and/or groups into processes, activities and decision/policy making in a way that shares power [8].

2.1. Variables That Measure Workplace Diversity in Organizations

There are several variables that can be used to measure workplace diversity. The most important ones include the following:

1) Age Diversity

Age is a generational difference between employees of an organization. Growing age diversity is increasingly becoming part of quite a number of business organizations. According to Kunze [9], the social identity and categorization theory may be used to understand this relationship. According to this theory, it is suggested that individuals tent to classify themselves based on dimensions that seem

relevant to them. As a result, individuals tend to favor employees of their own group, and discriminate employees from other age groups. Thus, the employees generational belonging is an important criterion for distinction that may stir emotional conflict at workplace.

2) Gender Diversity

Ali defines gender diversity as psychological and experienced disparities that are culturally or socially attached to being of a male or a female.

3) Ethnic Diversity

An ethnic group refers to a group of people with a sense of common origin, and most often, a sense of common destiny. Pitts [10] argue that as firms impress ethnic diversification, there is need to pay more attention on how different ethnic groups interact with each other at workplace. Hoogendoorn Van Praag [11] defines ethnic diversity as the heterogeneity in races, languages, and cultures among employees of an organization.

2.2. The Impact of Workplace Diversity on Employees and Organizational Performance

Workforce diversity has an influence on both employee and organizational per formance, and consequently, on the organizational performance. This means that a positive effect of workplace diversity at employee level will have a positive effect at organizational level, and vice versa.

2.3. Conceptual Framework

1) Age Diversity

Most often, organizations avoid utilizing the expertise of old employees because of stereotypes and false assumptions that they are slow in adopting to changes and new technologies, prone to health problems, poor performance, and expensive compared to the older generation [12]. According to a study done by Hamilton Nickerson [13], on simple production technology, it was found that work-forces with age diversity were less productive. A similar finding is reported by Leonard Levine [14], where they indicated that retail store businesses that had age diversity among employees were slow in making profits. In another study by Ilmarinen [15], it was reported that there was no relationship between em ployees age and work performance. Many studies have shown that older em ployees perform work-related tasks as effective and efficient as younger em ployees. According to Williams OReilly [16], having a heterogenous age em ployee team is more productive than having an employee team with a homogenous age.

2) Gender Diversity

Mixed gender team of employees performs better compared to a team of employees of the same gender [17]. In studies carried out by [18], it was evident that there was a positive relationship between organizational performance and gender diversity, based on the organization resources. Many other studies have found a negative effect of gender diversity on team performance in male domi nated samples, and insignificant effects in female dominated samples [19]. Ac cording to Samuel [20], the organizations competitive advantage increases when gender diversity is at a moderate level, while a greater level of gender diversity reduces organizational performance. The study results obtained by [21], showed an inverted U-shaped connection between employees' gender and organizational performance. In a similar study conducted [21], it was found that there was an inverted U-shaped relationship between gender mix and employee productivity.

The research also found that moderately heterogenous teams demonstrated bet ter performance compared to gender homogenous teams. Gender diversity con tributed positively to the services industry, while in the manufacturing industry, it had negative effects. Therefore, gender diversity in service industry workforce might bring a positive impact compared to companies in the manufacturing industry.

3) Ethnic Diversity

Jackson, et al. [14] studied the effect of gender diversity on performance. In his studies, he found that ethnically diverse teams of employees exhibited poor performance compared to homogenous teams. A close study by Jones [22], also demonstrated that ethnic groups were less cohesive compared to teams. Thus, ethnic diversity is likely to have a less positive impact on group performance when compared to team performance. An ethnically diverse team of employees possess high creativity and innovation that comes with learning opportunities.

According to Sander Hoogendoorn [12], ethnic diversity at moderate level has no impact on organizational performance. Samuel [20], also reported a positive effect of ethnic diversity on innovation, productivity, market share, and sales.

Jones [22] investigated the effect of ethnic diversity in the Oil Gas Industry. The researchers reported a positive relation between ethnic diversity and team per formance. In another study conducted by Jones [22], it was found that there was no relationship between ethnic diversity and sales productivity, revenue, and customer satisfaction.

3. Data Analysis

3.1. Movies Data Analysis; an Analysis of IMDB Movies Dataset and Rotten Tomatoes Dataset

This study was conducted by choosing two major film websites (Rotten Tomatoes), the reasons in which these websites were chosen due to the fact that these are the essential websites that provides up to date movies ratings and have massive amount of data. Other websites provide datasets that are not up to date or include missing data.

Two main movie websites were chosen as a source for data gathering; which were imdb and rotten tomatoes websites. The objective of the study was to ana lyze the rotten tomatoes data to determine the relationship between diversity in terms of movie writers, movie directors and the success of a movie in terms of earnings. To carry out the analysis, 14,235 movie sets were scrapped from Rotten Tomatoes website. The data contained 29 variables each describing a varied aspect of the movie. The analysis was done using Python programming language that required use of various libraries to ease the analysis process. The list of libraries used in the analysis includes Numpy, pandas, Scikit-lean, OS, and SYS.

The analysis was carried out in Jupyter Notebook. Table 1 demonstrates va riables of the imdb dataset and description for each of the variable studied while Figure 2 shows distribution of average rating.

Table 1. Demonstrates IMDB dataset.

Variable	Description of the variable
Movie ID	Unique ID identifying the movie
Year	The year the movie was release
New Distribution	Data about distribution of the movie
Votes	Number of votes on the movie
Ranks	Raking of the movie
Title Comp	Title of the movie
Budget	Amount of money used in movie production
Box Office	Box office income
Currency Budget	Currency of the budget
Currency Box Office	Currency of the box office income
rtriticRatings	Movie ratings given by movie critics
rtAllCriticsNumReviews	Number of reviews by all the critics
rtAllCriticsNumFresh	Number of fresh movie critics
rtAllCriticsNumRotten	Number of Rotten Tomatoes Critics
rtAllCriticsScore	Scores by the critics
rtTopCriticsRating	Rating provided by top critics
rtTopCriticsNumReviews	Number of reviews by top rotten tomatoes top critics
rtAudienceRating	Ratings provided by the audience
rtAudienceNumRatings	Number of ratings by the audience
rtAudienceScore	Scores given by the audience
BudgetUSD	Budget of the movie in United Stated Dollars
BoxOfficeUSD	Box office income in United States Dollars

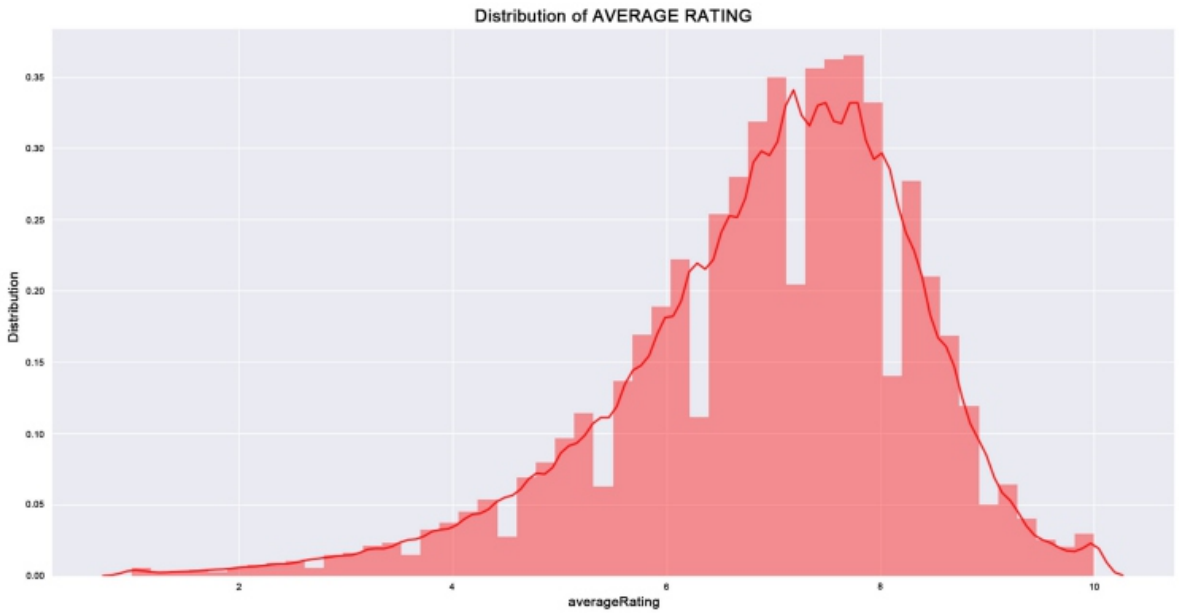


Figure 2. Shows distribution of average rating.

3.2. Methodology

Before starting the analysis, it was paramount that the data were cleaned for the analysis. The preparation of the data entailed carrying out the following process;

1) Import the required libraries (Pandas, Numpy, OS, Scikit-learn and OS).
2) Loading all the files using the imported libraries.
3) Merging the data in the files into a single dataset.

Implementation.

After setting up the environment, firstly, the dataset was loaded, the Rotten
Tomatoes movies dataset and IMDB movies dataset in the Jupyter environment.
The next task was merging the datasets to find the number of movies which were common in both datasets, an inner join on Title column was used to merge the two datasets. It was found that the number of movies which were common in both datasets were 8553. It was also concluded that in IMDB dataset, Title and Original Title column have same number of unique observations. There were 107,645 distinct movies in IMDB Dataset and 13,904 movies in Rotten Tomatoes Dataset.

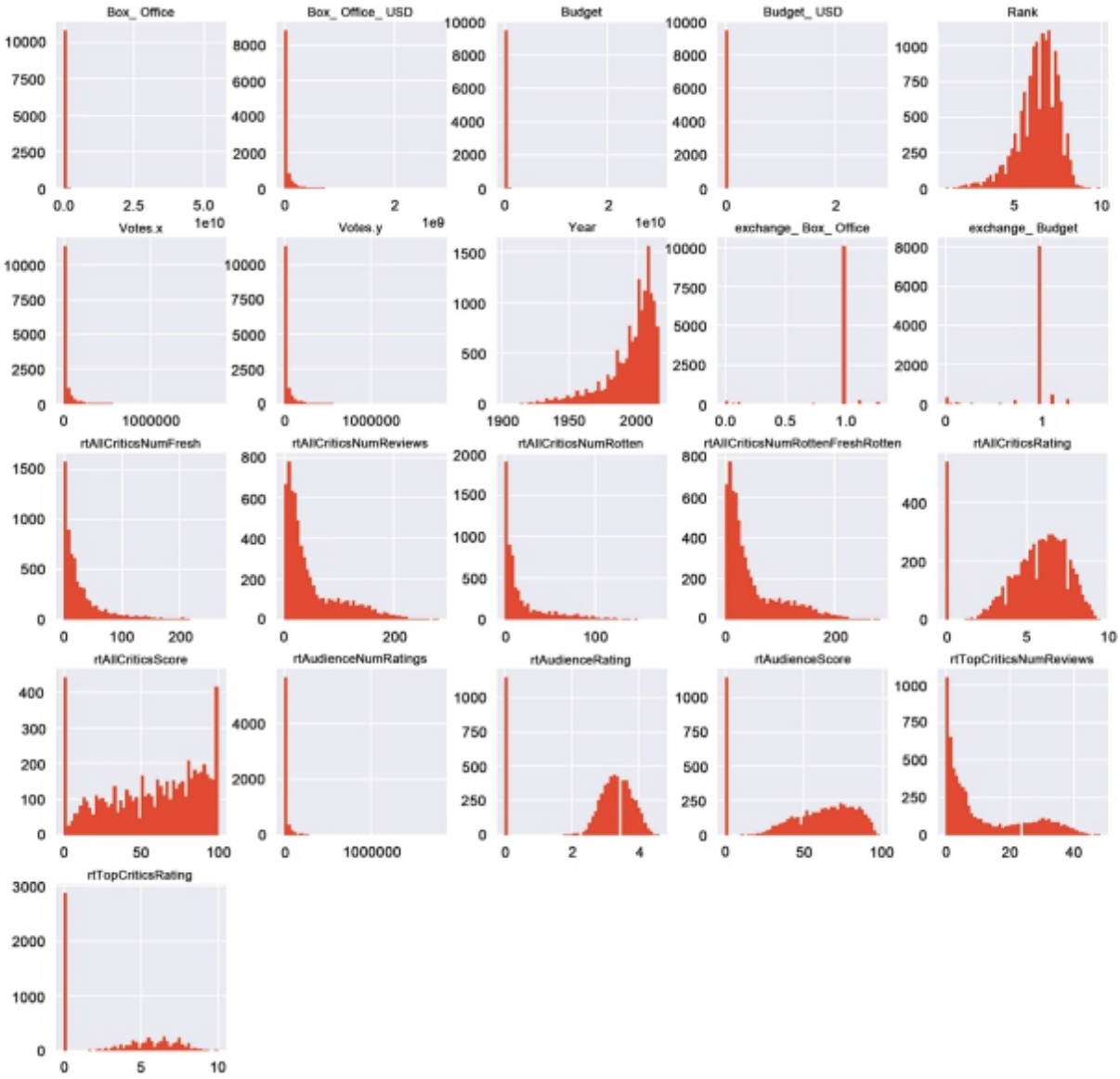
After finding the number of movies which are similar in both datasets, in terms of critics score and audience score and revenues. The Tomatometer rating which was discussed above is based on the published opinions of hundreds of film and television critics—is a trusted measurement of movie and TV programming quality for millions of moviegoers and it can also be observed that Tomatometer rating is greater than the audience ratings which shows that critics rated the movie in more positive manner as compared to the audiences. It can be observed that most of the movies are drama and comedy. Also, for 13,367 movies the genre is not recognized. Next, the number of movies or programmes of various title types present in the dataset was analyzed. And finally, for the IMDB Dataset, the distribution was also analyzed in terms of runtimes in minutes and found that most of the movies were of 90 minutes in runtime.

3.3. Rank Versus Ratings (Critics/Audience)

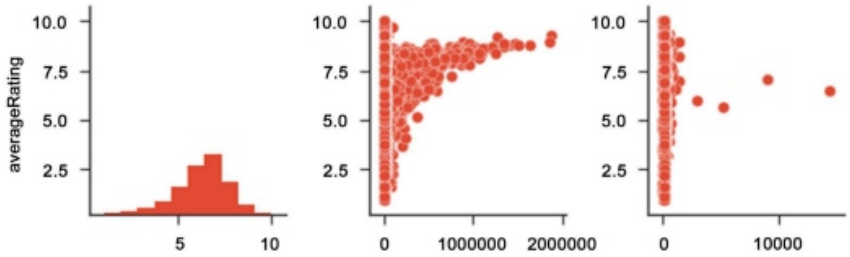
In a similar manner, what was also analyzed was the scatter plots among the audience/critics score and the overall rank of the movie in the two scatterplots. But, here in this case, we can observe that Audience Score is highly correlated to the Rank of the movie as compared to the Critic Score. Figure 3 demonstrates the various numeric variables. And Figure 4 exhibits a correlation matrix heat map.

3.4. Pearson Method of Correlation

Statistical tests were also performed to understand how these variables are dependent on each other and how significantly they differ from each other. The results shown below are the Pearson's coefficient of Correlation which has been already discussed above in the scatter plots. One of the measures used to es tablish if there is a relationship between two variables is the Pearson prod uct-moment correlation coefficient normally referred to as Pearson's coefficient.


Box_ Office
Box_ Office_ USD
Budget
Budget_ USD
Rank
Votes.x
Votes.y
Year
exchange_ Box_ Office
exchange_ Budget
rtAllCriticsNumFresh
rtAllCriticsNumReviews
rtAllCriticsNumRotten
rtAllCriticsNumRottenFreshRotten
rtAllCriticsRating
rtAllCriticsScore
rtAudienceNumRatings
rtAudienceRating
rtAudienceScore
rtTopCriticsNumReviews
rtTopCriticsRating

(a)


averageRating

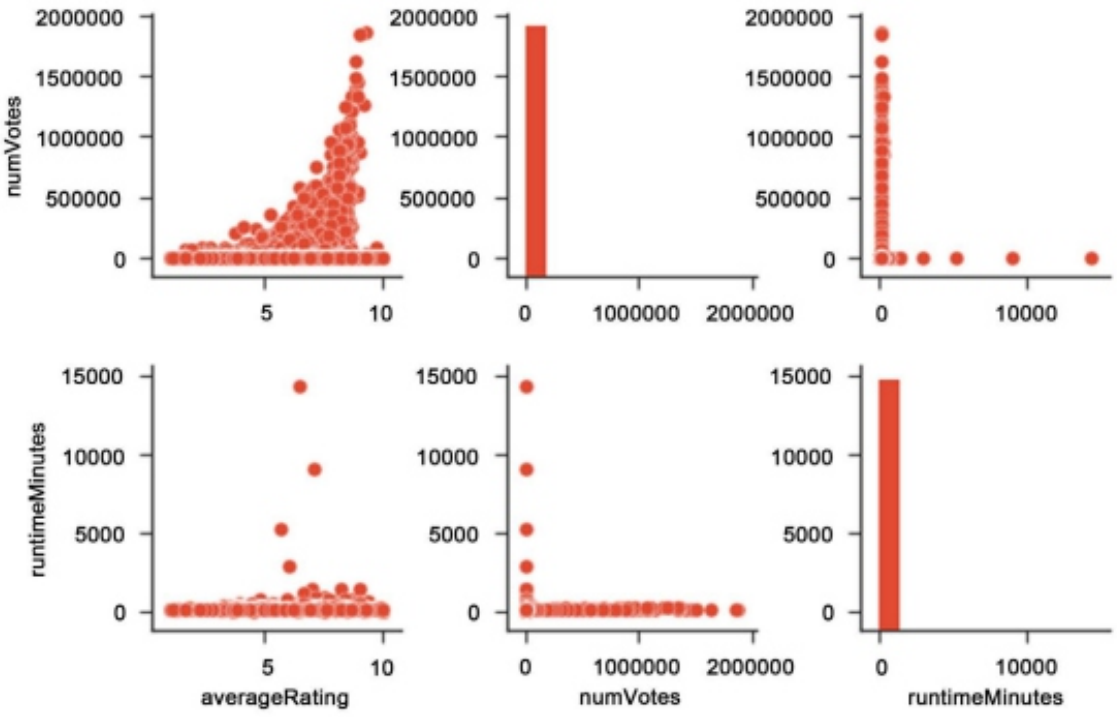


Figure 3. Distribution of various numeric variables in the dataset.

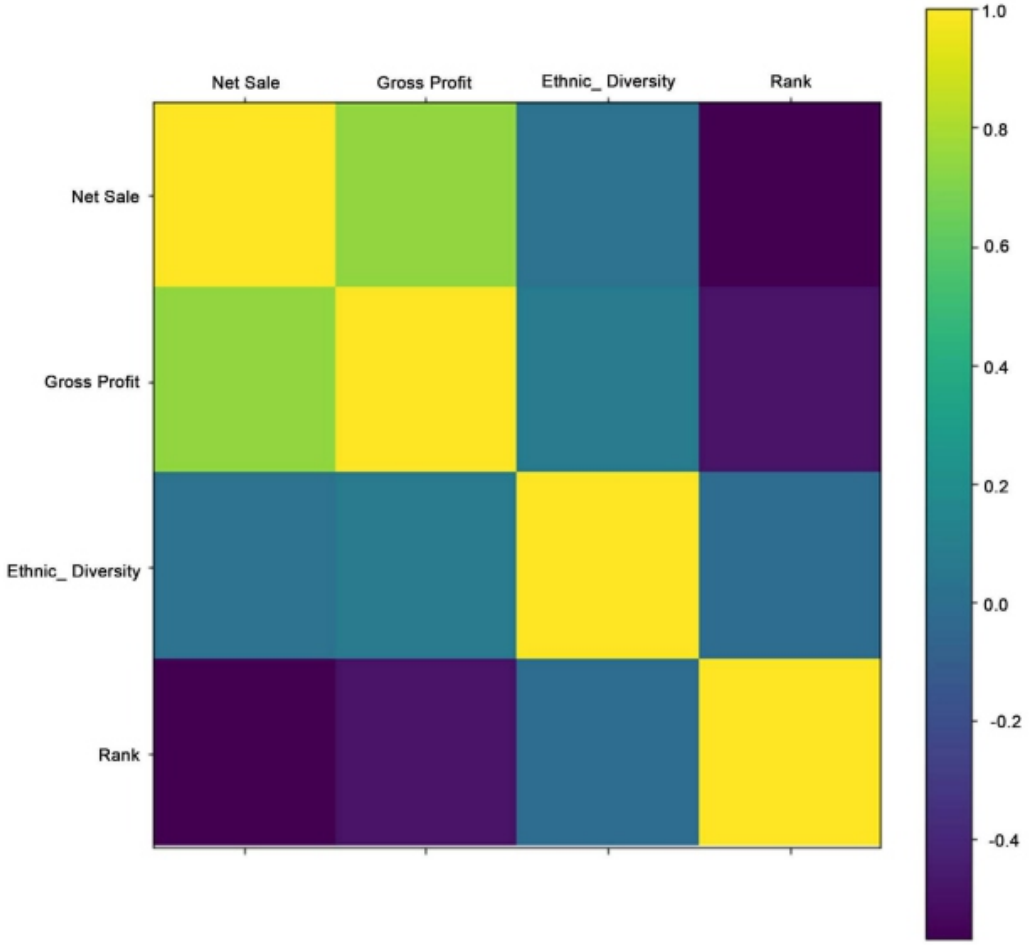


Figure 4. Correlation matrix heat map.

The Pearson's coefficient measures the strength of the linear relationship between two variables. The formula for calculating the Pearson's coefficient is as shown below:

$$1-\lambda=1-\sum_{i=1}^{R}\rho_i^2=1-\frac{1}{2D}$$

The value of Pearson product-moment correlation coefficient ranges between −1 and 1. A strong positive correlation tends to lean towards positive one while a strong negative correlation tends to lean towards negative one. If the value of Pearson product-moment correlation coefficient is zero, there is an indication that there is no correlation between the variables under study. There are various ways in which the value of the Pearson product-moment correlation can be calculated. One of the most common method entails plotting a scatter plot and then establishing a line of best fit through the plotted data points. Pear son product-moment correlation coefficient is then used to indicate how far the plotted points are from the best line of fit. In this project, the Pearson product-moment correlation coefficient was determined using the Numpy library using the corref() function.

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

3.5. Fortune's Companies Data Analysis

Data about the Board of Directors of Fortune 500 companies was collected. The data was contained in two CSV files. One file contained data about the details of the Board of directors, while the other file contained information about net sales and gross profit. To have an overview of the data, the head () fiction was used. Furthermore, the data about the company was grouped as either top gainers or top losers. The following table shows a snippet of the Board of directors' top gainers data and top loser data. Table 2 shows Fortune's Less Profitable Companies Diversity Score.

The data of the top losers and the top gainers company were merged with the data of the net sales and gross sales of the respective companies. Using the data of the board of directors of each of the fortune 500 companies, the ethnic diversity score was calculated. After the calculation, a histogram was plotted to visualize how diverse the board of directors are in the company basing on the earlier calculated ethnic diversity score. Figure 5 shows a plot for Ratings versus Diversity for Writers.

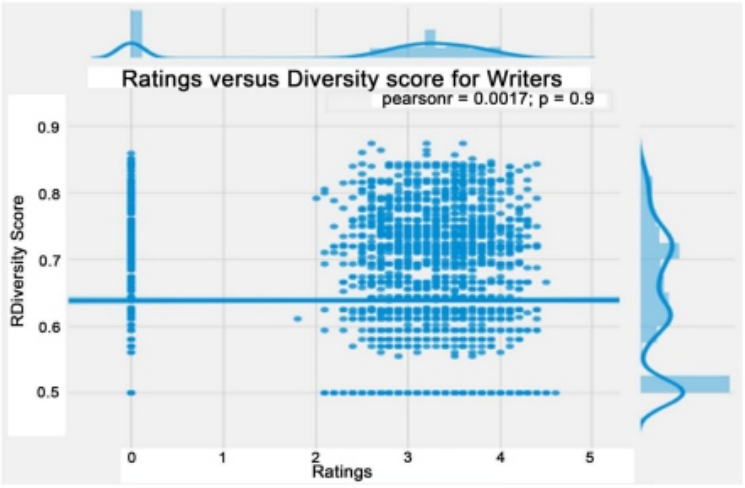
4. Results

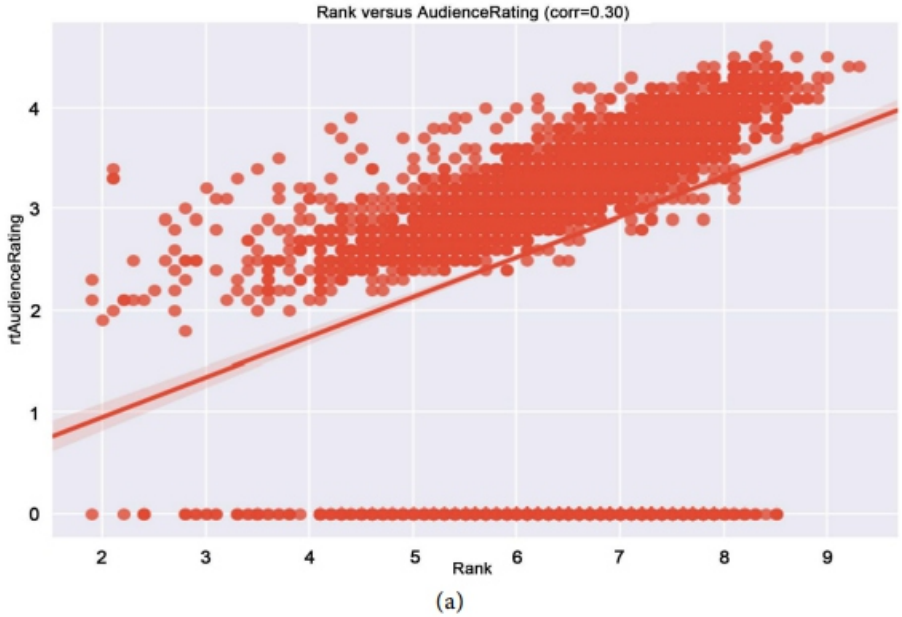
4.1. Distribution of Ethnic Diversity among the Writers

Figure 6 shows Box Office Versus Diversity Score for Writers. The movie writers have varied diversity, consisting of Hispanics, Blacks, and whites, among others. It is thus important to establish how the diverse the movie writers are according to the Rotten Tomatoes data, it can be seen that the diversity among the writers is normally distributed save the maximum number of observations. Furthermore, the chart shows that most of the movies have writer ethnic diversity score that is greater than 0.6 and less than 0.8.

Table 2. Fortune's less profitable companies diversity score.

Company Name	Net Sale	Gross Profit	African	American	Asian	British	Indian	total ethnicities	diversity
Big Lots Inc.	5.2 B	2.1B	0	9	0	0	0	9	0
Momentive Performance Materials Inc	544 M	146 m	0	10	1	0	0	11	0.16
Markel Corporation	5.61 B	2.06 B	0	14	0	0	0	14	0
Noble Energy, Inc.	3.49 B	2.41 B	0	11	0	0	0	11	0
Leidos Holding, Inc.	7.04 B	852 M	0	12	1	0	0	13	0.14
Rockwell Collins, Inc.	6.82 B	1.95 B	0	10	0	0	0	10	0
Sprague	2.39 B	210.9 M	0	8	0	0	0	8	0
YRC World Wide Inc.	4.7 B	3.34 B	0	10	0	0	0	10	0
The Hanover Insurance Group, Inc.	4.95 B	1.98 B	0	11	0	0	0	11	0
Fiserv, Inc.	5.51 B	2.55 B	0	9	0	0	0	9	0





(a)

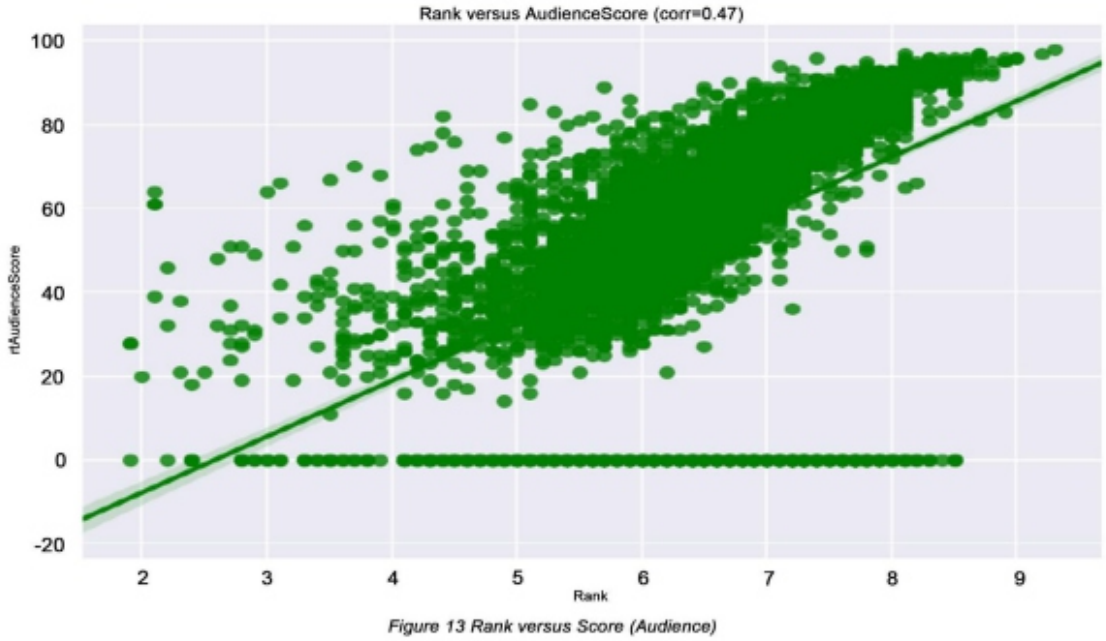


Figure 6. Rank versus audience rating and audience score.

4.2. Distribution of Ethnic Diversity among the Directors

It can be seen that the diversity score with the highest distribution is between 0.68 and 0.70. Thus, it can be concluded that the distribution of diversity among the movie directors is relatively normal.

4.3. Correlation between Writers Ethnic Diversity and Movie Ratings

Figure 7 shows the Rank Versus Audience Rating and Audience Score. To vi sualize how diversity among the writers relates the diversity movie ratings, a re gression joint plot was drawn. From the regression chart, it can be seen that the Pearson product moment correlation coefficient is 0.00017 (the Pearson correla tion product moment coefficient is approximately equal to zero). Since the value of r is almost close to zero, thus be concluded that there is no relationship be tween the diversity among the movie writers and he movie rating.

4.4. Correlation between Box Office Earnings and Diversity among Writers

Analysis was also done to establish if there is a relationship between writer's di versity score and the movies box office earnings. From the regression chart above, it can be seen that the Pearson's correlation coefficient, r, is equal to 0.11.

The value of r is significant and thus, it can be concluded that there is a weak
positive relationship between movie diversity and box office earrings. The rela tionship is such that as the diversity score increases, the box office earnings increase.

5. Discussion

The analysis in the previous section has presented how ethnic diversity score of movie writers and directors effects on the success of a movie. The section has also analyzed how the ethnic diversity score of Board of Directors of fortune 500 companies affects the net sales and net profits. The general conclusion from the analysis is that ethnic diversity has an impact on success of movies and the fortune 500 companies. In this context, the success of a movie is measured in terms of ratings and box office

earnings while the success of fortune 500 companies is measured in terms of net sales and the profits. Despite the established correlation, the impact of ethnic diversity is very small and may be considered to be insignificant. The impact of ethnic diversity is attributed to various factors. Jackson [14] studied the effect of ethnic diversity on performance. In his studies, he found that ethnically diverse teams of employees exhibited poor

Performance compared to homogenous teams. A close study by Jones [22], also demonstrated that ethnic groups were less cohesive compared to teams. Thus, ethnic diversity is likely to have a less positive impact on group performance when compared to team performance. Despite the fact that an ethnically diverse team of employees possess high creativity and innovation that comes

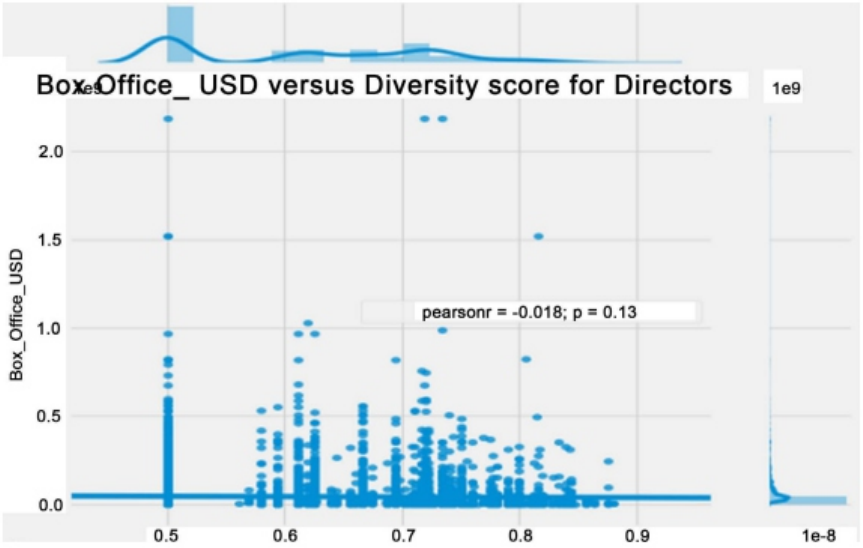


Figure 7. Box office versus diversity score for writers.

with learning opportunities and complementary, Hoogendoorn [12], established that ethnic diversity at moderate level has no impact on organizational performance. Samuel [20], also reported a very weak positive effect of ethnic diversity on innovation, productivity, market share, and sales in a company. Also, Jones [22] investigated the effect of ethnic diversity in the Oil Gas Industry. The research reported a weak positive relation between ethnic diversity and team performance. In another study conducted by Jones [22], it was found that there was no relationship between ethnic diversity and sales productivity, revenue, and customer satisfaction. It is evident that the outcomes of this research are in tandem with what other researches had established earlier. All researches agree that ethnic diversity has a weak impact on success of a company in terms of net profits and sales. The same trend extends to the effect of ethnic diversity in terms of writers and movies on the success of a movie in terms of movie rating and box office earnings.

6. Future Work

Future work can focus on obtaining the ethnicities of actors from Wikipedia as the ethnicities that were obtained for this study was by the use of a name classifier. Website such as Wikipedia and ethnicelebs were examined for ethnicities but it does not contain all the ethnicity of the actors. Also, Wikipedia was scraped for data regarding ethnicity, nonetheless, it only had information regarding actor's nationality not ethnicity.

7. Conclusions

According to the analysis done, the Pearson's coefficient has been found to be less than 0.1 in all the regression analysis. This implies a weak relationship between ethnic diversity and the other variables. Based on this, the null hypothesis is accepted. Therefore, it can be concluded that ethnic diversity of movie writers and directors has no significant impact on the success of a movie and Ethnic diversity of Board of Directors of Fortune 500 Company has no impact on the net sales and gross profit.

Moreover, the analysis in has presented how ethnic diversity score of movie writers and directors affects the success of a movie. The paper has also analyzed how the ethnic diversity score of board of directors of fortune 500 companies is associated with net sales and net profits.

The general conclusion from the analysis is that ethnic diversity has an impact on success of movies and the Fortune 500 companies. In this context, the success of a movie is measured in terms of ratings and box office earnings while the success of Fortune 500 companies is measured in terms of net sales and the profits. Despite the established correlation, the association of ethnic diversity is very small and may be considered to be insignificant.

It is evident that the outcomes of this research are in tandem with what other researches had established earlier. All the researchers agree with this study findings, which showed that ethnic diversity has a weak impact on success of a company in terms of net profits and sales.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Saxena, A. (2014) Workforce Diversity: A Key to Improve Productivity. Procedia Economics and Finance, 11, 76-85. https://doi.org/10.1016/S2212-5671(14)00178-6

[2] Siciliano, J.I. (1996) The Relationship of Board Member Diversity to Organizational Performance. Journal of Business Ethics, 15, 1313-1320.https://doi.org/10.1007/BF00411816

[3] Apala, K.R., Jose, M., Motnam, S., Chan, C.C., Liszka, K.J. and de Gregorio, F. (2013) Prediction of Movies Box Office Performance Using Social Media. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, 25-29 August 2013, 1209-1214.

[4] Augustine, A. and Pathak, M. (2008) User Rating Prediction for Movies. Technical report, Citeseer.

[5] Amazon.com, Inc. (1990) Internet Movies Database. https://www.imdb.com/

[6] Fandango (2016) Rotten Tomatoes. https://www.rottentomatoes.com/about/

[7] Oghina, A., Breuss, M., Tsagkias, M. and de Rijke, M. (2012) Predicting IMDB Movie Ratings Using Social Media. In: European Conference on Information Retrieval, Springer, Berlin, 503-507. https://doi.org/10.1007/978-3-642-28997-2_51

[8] 2017 PeopleScout, a TrueBlue Company—All Rights Reserved (2017) Improving Racial and Ethnic Diversity in the Workplace.http://www.peoplescout.com/racial-and-ethnic-diversity-in-the-workplace/

[9] Kunze, F., Boehm, S. and Bruch. H. (2009) Age Diversity, Age Discrimination, and Performance Consequences—A Cross Organizational Study. Academy of Manage-ment Proceedings, 2009, 1-6. https://doi.org/10.5465/ambpp.2009.44249873

[10] Ali, M., Kulik, C.T. and Metz, I. (2011) The Gender Diversity—Performance Rela tionship in Services and Manufacturing Organizations. The International Journal of Human Resource Management, 22, 1464-1485.https://doi.org/10.1080/09585192.2011.561961

[11] Pitts, D.W. and Wise, L.R. (2010) Workforce Diversity in the New Millennium: Pros pects for Research. Review of Public Personnel Administration, 30, 44-
69.https://doi.org/10.1177/0734371X09351823

[12] Hoogendoorn, S. and Van Praag, M. (2012) Ethnic Diversity and Team Performance: A Field Experiment. IZA Discussion Paper No. 6731. https://doi.org/10.2139/ssrn.2114911

[13] Selvaraj, P.C. and Darwin, J.R. (2015) The Effects of Work Force Diversity on Em ployee Performance in Singapore Organisations. International Journal of Business Administration, 6, 17-29. https://doi.org/10.5430/ijba.v6n2p17

[14] Hamilton, B.H., Nickerson, J.A. and Owan, H. (2004) Diversity and Productivity in Production Teams. https://doi.org/10.2139/ssrn.547963

[15] Leonard, J. and Levine, D.I. (2003) Diversity, Discrimination, and Performance. https://doi.org/10.2139/ssrn.420564

[16] Ilmarinen, J. (2005) Towards a Longer Worklife: Ageing and the Quality of Wor klife in the European Union. Finnish Institute of Occupational Health, Ministry of Social Affairs and Health.

[17] Williams, K.Y. and O'Reilly, C.A. (1998) Demography and Diversity in Organizations: A Review of 40 Years of Research. Research in Organizational Behavior, 20, 77-140.

[18] Wood, R. and Bandura, A. (1989) Impact of Conceptions of Ability on Self-Regulatory Mechanisms and Complex Decision Making. Journal of Personality and Social Psy chology, 56, 407-415. https://doi.org/10.1037/0022-3514.56.3.407

[19] Frink, D.D., Robinson, R.K., Reithel, B., Arthur, M.M., Ammeter, A.P., Ferris, G.R.,Kaplan, D.M. and Morrisette, H.S. (2003) Gender Demography and Organization Performance: A Two-Study Investigation with Convergence. Group & Organization Management, 28, 127-147. https://doi.org/10.1177/1059601102250025

[20] Idowu, S.O., Capaldi, N., Zu, L.R. and Gupta, A.D. (2013) Encyclopedia of Corpo rate Social Responsibility, Springer, New York.https://doi.org/10.1007/978-3-642-28036-8

[21] Van der Vegt, G.S. and Bunderson, J.S. (2005) Learning and Performance in Multi disciplinary Teams: The Importance of Collective Team Identification. Academy of Management Journal, 48, 532-547. https://doi.org/10.5465/amj.2005.17407918

[22] Jones, M., Qazi, M. and Young, K.D. (2005) Ethnic Differences in Parent Preference to Be Present for Painful Medical Procedures. Pediatrics, 116, e191-e197.

Use of a Neural Network to Measure the Impact of Social Distribution and Access to Infrastructure on the HDI of the Municipalities of Mexico

Fernando I. Becerra López, Ricardo Pérez Ramírez
Department of Mathematics, Universidad de Guadalajara, Guadalajara, Mexico

ABSTRACT

The Human Development Index (HDI) was created by the United Nations (UN) and is the basis for many other indicators, as well as being the origin of many public policies worldwide. It is a summary measure of life expectancy, education, and per capita income. These components, in addition to being global measures, show difficulty in being impacted and, with this, advancing in the level of human development. This work shows a model that relates va riables of social distribution and access to infrastructure in Mexico, with the HDI. These variables were chosen through a statistical analysis based on a set of indicators measured by the National Institute of Statistics and Geography (INEGI) periodically at the municipal level. The statistical analysis shows that there is no simple correlation between these variables and the HDI, so that a supervised learning model based on a neural network was used, therefore proposing a classification technique based on the distribution of data in the underlying metric space. In addition, an attempt was made to find the simplest possible model to reduce the computational cost and in turn obtain information on the variables with the greatest impact on the HDI, with the aim of facilitating the creation of public policies that impact it.

Keywords: Multilayer Perceptron, Human Development Index, K-Means, Non-Linear Correlation

1. Introduction

In all countries there are indices and indicators that help governments monitor the performance of their policies, these can refer to education, health, infra structure, and social distribution, among others; although these are only methodological proposals and are likely to receive comments to improve their use fulness, for example, as in [1] where a different way of evaluating marginaliza tion in Mexico is proposed. An advantage of having diverse types of indicators is that with these it is possible to make analyses between different indicators on an objective indicator, as in [2] where development is taken as a variable that is influenced by distinct factors, features such as social, and economic, among others.

With the above in mind, the Human Development Index, HDI, is selected for this work as an index that reflects the quality of life of a population and taking the view that the indices are impacted not only by the methodology with which they are created but also by other features, a selection of other indices are pro posed, which have no appreciable direct relationship but at the same time it is inferred that modifications to these have an impact on the quality of life of a population.

HDI was published for the first time in 1990 by the United Nations Develop ment Program, UNDP (1990). This index was introduced due to the need to have a measure of development in the countries and its fundamental objective is to measure the development of the human being, unlike, for example, the Gross Domestic Product, GDP, of a country, which reflects development, but based only on its economic activity. Therefore, three components were chosen to cal culate the HDI, which focuses on health, education, and wealth, which represent the fundamental axes in the development of a person. HDI has become a very important tool for governments, including organizations such as the Economic

Committee for Latin America and the Caribbean (CEPAL) [3].

The health indicator, calculated by the longevity of a population, is determined by the life expectancy at birth of a person. It is of special relevance since it indirectly reflects a population's access to health services, as well as adequate nutrition, since, without these two indirect characteristics, it would be difficult to increase life expectancy.

Regarding the education component, this is calculated with the literacy rate of a population, something of significant importance since it provides the opportunity to access knowledge. In fact, it is currently desired that the population have access to higher levels of knowledge for a better performance in their productive lives.

Lastly, the index of the wealth of a population tries to reflect the capacity that this must face the basic needs that an individual may have for its development.

This index is calculated with the GDP per capita of each country or region together with a correction, purchasing power parity, to homogenize the level of said purchasing power between different regions.As can be seen, the three components of the HDI really aim to reflect what an individual's life of well-being can be like, a long life with access to education to develop the desired and well-paid economic activity. Unfortunately, these three components are averages in the population and, therefore, a global measure, which can hide the reality and the dispersion that the population experiences in each of these factors.

In addition to the, designing public policies that help increase these three components and, consequently, the HDI does not turn out to be intuitive.

Therefore, the search for other indices with a more sensitive impact on the decisions made by the government can be helpful in the design and implementation of public policies that help improve the HDI of the regions.

2. Variables Proposed to Influence the HDI Level

Considering what was expressed in the introduction, indices were selected that are believed to be more local and easily obtained (all are provided by the National Institute of Statistics and Geography, INEGI, from the year 2010, and at the municipal level, where the methodological manuals are in [4] and [5]). These indices or variables are the following.

The percentage of the population that lives in communities of less than 5,000 inhabitants in a municipality (PL < 5000), is a variable that is used in the reports of marginalization prepared by the Government of Mexico, which is important, since in more than 50% of the municipalities in the country have 100% in this index, in addition populations of this type tend to have less access to services and less economic development.

The Labor Force Participation Rate (LFPR) of a municipality, which refers to the quotient of economically active people who are working or looking for a job (a person can conduct an economic activity from the age of fifteen) between the entire population.

$$LFPR = \frac{LF_{(15 \text{ or more})}}{P_{(15 \text{ or more})}} \times 100 \qquad (1)$$

where: (15 mas y) LF is the labor force aged fifteen or over and (15 mas y) P is the total population greater than or equal to fifteen years.

The degree of accessibility to paved roads (AccesInfra), which is obtained thanks to the work of the National Council for the Evaluation of Social Development Policy, CONEVAL, and which reflects the ease that different communities have in using paved roads. The AccesInfra grade per municipality is

is taken as the weighted sum of the AccesInfra's grades per community, which make up the municipality.

$$\frac{\sum_{i=1}^{n}(p_i \times g_i)}{\sum_{i=1}^{n} p_i} \qquad (2)$$

where i p is the population of the i-th community that makes up the municipality, i g is the degree of accessibility of the i-th community and n is the number of communities that make up the municipality.In addition, the percentage of the population of a municipality which does not native to the same federal entity (%PobMig) was chosen, considering that the phenomenon of migration is intricately linked to the search for better job op-portunities and life prospects. Due to this, municipalities that have high scores in this index could be interpreted as municipalities that offer high standards of living and that is why they attract populations from other states.

Lastly, the population density of the municipality (DENS10), which is a variable that might not seem to have as much relevance, but since many of the infra structure construction decisions are public and private, such as universities, hospitals, etc. is linked to covering the largest possible population, these constructions are aimed at municipalities with densely populated areas.

3. Relationship between the Variables and the HDI

As can be seen, several of the indices presented not only have a municipal focus, but are even obtained from a community level, so that the reality that the inhabitants may be experiencing can be better reflected. Likewise, indices such as the LFPR and the AccesInfra are sensitive to public policies, in addition to the fact that they all function as control variables for human development (although this does not mean that it is impossible to create public policies that help improve them in the short term).

A main objective of this work is to find a relationship between these five variables with the HDI of each municipality, to then indirectly relate to the three axes that support the HDI and facilitate the structuring of public policies based on these five indices and thereby improve the HDI of the municipalities. For example, with the urban planning of services and infrastructure, which better benefits its surrounding communities with a high index of PL < 5000, so that they have greater access than densely populated communities. Also, the search for private investment, national or foreign, for the creation of new jobs, to increase the attraction of population from other states or retention of its own population and that, in turn, would be reflected in the LFPR.

4. The Proposed Model Using a Neural Network

Doing a linear correlation analysis between the selected variables and the HDI, it can be seen that it does not exist for any of them, as can be seen in Table 1, so it will be necessary to use a model that can find non-linear and multivariate correlations, being Neural networks are a good tool for this type of problem, as demonstrated in [6]. For this reason, a Multilayer Perceptron (MLP) was selected as a model [7], which is a generalization of the simple Perceptron proposed by [8], which through its processing units (neurons), and their dynamic states of activation [9], processes the input data in order to find patterns in the data and thereby offer a model capable of generalizing [10].

Before explaining the architecture and the results obtained, it is important to mention that the outputs of an MLP express the probability of belonging to a certain set, therefore, it was decided to classify (cluster) the municipal HDI values into three groups with the method of K-means as in [11]. The decision to classify into three groups was based on a statistical analysis, in which three

Table 1. Correlation matrix.

	AcessInfra	DENS10	PL < 500	LFPR	%PobMig	HDI
AcessInfra	1.0000	0.1778	−0.3870	0.1019	0.2767	0.5617
DENS10	0.1778	1.0000	−0.3398	0.1901	0.3203	0.3299
PL < 5000	−0.3870	−0.3398	1.0000	−0.2044	−0.3990	−0.6309
LFPR	0.1019	0.1901	−0.2044	1.0000	0.1881	0.1942
%PobMig	0.2767	0.3203	−0.3990	0.1881	1.0000	0.4964
HDI	0.5617	0.3299	−0.6309	0.1942	0.4964	1.0000

classes offered greater separability between groups, this against a greater or lesser number of clusters. Once the limits of the classes were obtained, they were labeled, so that the class labeled with [1,0,0] represents the group of municipalities with a "high" HDI, the label [0,1,0] represents the group of municipalities with a "medium" HDI and, finally, the label [0,0,1] is the representative of the municipalities with a "low" HDI.

For the training of the MLP, it was decided to separate the data by municipalities and at the same time validated for robustness by the Student's T test, such that 70% of these were used to train the model and the remaining 30% to validate the model and how capable it is to generalize or, in other words, assess the perdition of membership of an input dataset, on which it was not trained.The architecture selected for the problem was an input layer with 5 neurons, a hidden layer with 25 neurons, and an output layer with 3 neurons, all with logistic activation function (Equation (3)), trained with the backpropagation algorithm and in an "off-line" mode because the data presented concurrency [12][13]. This architecture is selected, since a better performance (stability) was observed in the generalization of the data and the norm of its derivative reached almost zero, Figure 1. Also, it was made an analysis of convergence for several amounts of neurons at the hidden layer to guarantee avoid overfitting and keep the model simple.

$$\sigma = \frac{1}{1+e^{-x}} \tag{3}$$

For the validation of the model, precision was taken as a metric, which is defined by:

$$\text{Acc} = \frac{\text{number of correct predictions}}{\text{total number of predictions}} \times 100 \tag{4}$$

An 81% accuracy was obtained for the training data and 74% for the validation data, giving 79% in the evaluation with all the data, which confirms that there is a correlation of the variables with the HDI of the municipalities.

Already having the model, tests were conducted both to validate it and to observe the behavior of the selected indices. For example, the median of each of the indices was selected, since due to the bias of the data it provides us with a better measure of central tendency, obtaining a prediction of the average HDI, which was expected.

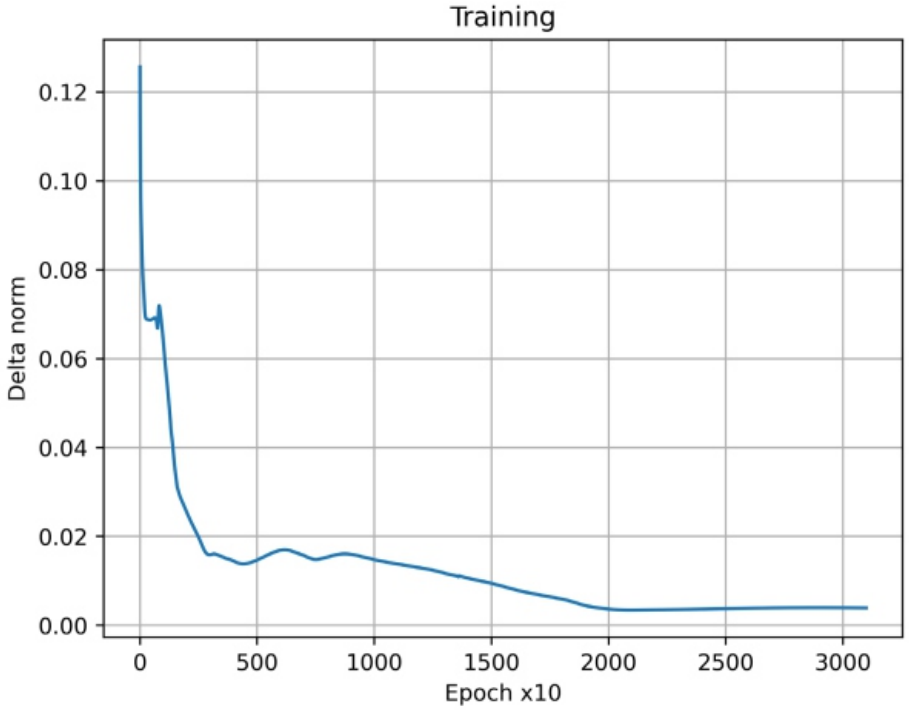


Figure 1. MLP training.

In addition, an arrangement of hypothetical municipalities was created, where each of them had the median in all the indices except one, which is evaluated with its minimum value and then with its maximum value, to observe the relationship between the indices regarding these municipalities. It was observed that the indices with the greatest impact are AccesInfra, %PobMig and PL < 5000, the latter with an inverse relationship to the HDI (that is, a higher value is reflected in a decrease in the HDI). For example, a municipality with all their indices equal to the median, but %PopMig at the minimum obtained a low HDI predic tion of 76.75%, while with %PopMig at the maximum a medium HDI prediction of 97.42% was obtained.

After this, the exercise of leaving all the indices at their lowest or highest value was conducted, with only one of them varying between the range of minimum and maximum values. With this, it was identified that only one of the indices, %PobMig, has an impact on the classification made by the model, while the others do not. This tells us, omitting the case, that the movement of a single variable has no relevance and, therefore, there is no one-to-one relationship with the HDI.

Something to confirm the above in a quantitative way and to be able to observe that all the indices are important for the model, an analysis of characteris tics was carried out taking the proposal of [14], in which the importance i S of an input to the model as seen in Equation (5) and being the criterion in [15] the one used. Equation (6), where wij is the weight of the i-th data to the j-th neuron, thus obtaining Table 2 of results and where it is observed that the value S of the %PobMig index is the highest and the AccesInfra the lowest, although the latter is still important in scale with the others.

$$S_i = \sum_{j=1}^{n} s_{ij} \qquad (5)$$

Table 2. Individual importance of features for the model.

	AcessInfra	DENS10	PL < 500	TPE	%PobMig
S	7157.90	13650.49	15769.78	13267.01	24340.26

$$s_{ij} = \left(w_{ij}\right)^2 \quad (6)$$

This leads us to conclude that the proposed model (MLP) establishes a better relationship between the proposed variables and the HDI over other models such as, for example, multiple linear regression.

5. Conclusions

After observing that the model managed to relate the selected indices and the HDI of the municipalities, a statistical analysis of the municipalities with high HDI is conducted, calculating the averages of the indices, and comparing them with those obtained globally, as well as from the municipalities that are not included in the high HDI cluster.

When conducting the, it is observed that the municipalities with a high HDI have higher average values in access to paved roads, a higher population density, a slightly higher economic participation by their population, higher percentages of migrant population from other states, as well as a smaller percentage of the population living in small communities. Something that is also observed in the relationships found with this model was that AccesInfra, %PobMig and PL <5000 have greater weight compared to DENS10 and TPE. This suggests that policies focused on improving AccessInfra, %PobMig and PL < 5000, would have a greater positive impact on the HDI of these communities. This is a special approach since public policy decisions are rarely made using machine learning models. Furthermore, the idea of investigating indirect variables and their influence on human development is also new.

An example of this type of public policy for AccesInfra is the policy of the current Federal Government of Mexico to pave access to municipal capitals for municipalities that did not have this. For %PobMig, one can take what has been done in China with its special economic zones that have attracted people from the interior of the country, where the HDI is usually lower than in said special zones. Finally, the PL$ < $5000 is an index that cannot be impacted so quickly, since it will depend on the resources and services that these communities receive to help in their development, urbanization and growth (in the worst case that these small communities disappear and are grouped into a main one in the same municipality or in another part of the country).

For all the above, it can be inferred that creating public policies that consider the selected variables of social distribution and access to paved roads, would have a positive impact on the HDI of the municipalities and, therefore, on the original variables that they calculate it: health, schooling, and GDP.It is important to note that, although the model performs well with the present data, it is difficult to interpret the real effects due to the public policies implemented. However, changes to the HDI locally may be reviewed in the future.

Conflicts of Interest

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

References

[1] Gutiérrez-Pulido, H. and Gama-Hernández, V. (2010) Limitantes de los índices de marginación de Conapo y propuesta para evaluar la marginación municipal en México. Papeles de población, 16, 227-257.

[2] Peláez-Herreros, Ó. (2012) Análisis de los indicadores de desarrollo humano, marginación, rezago social y pobreza en los municipios de Chiapas a partir de una perspectiva demográfica. Economía, sociedad y territorio, XII, 181-213. https://doi.org/10.22136/est00201290

[3] Salas-Bourgoin, M. (2014) A proposal for a modified Human Development Index. CEPAL Review, 112, 29-44.

[4] Coneval, C.N. (2018) Grado de accesibilidad a carretera pavimentada. Coneval, Ci udad de México.https://www.coneval.org.mx/Medicion/Paginas/Grado_accesibilidad_carretera.aspx

[5] Inegi, I.N. (2017) Metodología de Indicadores de la Serie Histórica Censal. Inegi, Ci udad de México.https://www.inegi.org.mx/contenidos/programas/ccpv/cpvsh/doc/serie_historica_censal_met_indicadores.pdf

[6] Abdulsalama, K.A. and Babatunde, O.M. (2019) Electrical Energy Demand Fore casting Model Using Artificial Neural Network: A Case Study of Lagos State Nige ria. International Journal of Data and Network Science, 3, 305-322. https://doi.org/10.5267/j.ijdns.2019.5.002

[7] Hilera González, J.R. and Martínez Hernando, V.J. (1995) Redes Neuronales Artifi ciales: Fundamentos, Modelos y Aplicaciones. RA-MA, Madrid.

[8] Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Sto rage and Organization in the Brain. Psychological Review, 65, 386-408. https://doi.org/10.1037/h0042519

[9] Cybenko, G. (1989) Approximation by Superpositions of a Sigmoidal Function. Ma thematics of Control, Signals, and Systems (MCSS), 2, 303-314.https://doi.org/10.1007/BF02551274

[10] Lecun, Y. (1989) Generalization and Network Design Strategies. In: Pfeifer, R., Schreter, Z., Fogelman, F. and Steels, L., Eds., Connectionism in Perspective Elsevier, Elsevier, Toronto.

[11] Fix, E. and Hodges, J.L. (1989) Discriminatory Analysis. Nonparametric Discrimi nation: Consistency Properties. International Statistical Review, 57, 238-247.https://doi.org/10.2307/1403797

[12] Møller, M.F. (1993) A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. Neural Networks, 6, 525-533 https://doi.org/10.1016/S0893-6080(05)80056-5

[13] Higham, C.F. and Higham, D.J. (2018) Deep Learning: An Introduction for Applied Mathematicians. arxiv:1801.05894. https://arxiv.org/abs/1801.05894

[14] Sohangir, S. and Gupta, B. (2014) Neuro Evolutionary Feature Selection Using NEAT. Journal of Software Engineering and Aplications, 7, 562-570.https://doi.org/10.4236/jsea.2014.77052

[15] Belue, L.M. and Bauer, K.W. (1995) Determining Input Features for Multilayer Per ceptrons. Neurocomputing, 7, 111-121.https://doi.org/10.1016/0925-2312(94)E0053-T.

Forecasting Shark Attack Risk Using AI: A Deep Learning Approach

Evan Valenti
SafeWaters.AI, Boston, USA

ABSTRACT

This study aimed to develop a predictive model utilizing available data to forecast the risk of future shark attacks, making this critical information ac cessible for everyday public use. Employing a deep learning/neural network methodology, the system was designed to produce a binary output that is sub sequently classified into categories of low, medium, or high risk. A significant challenge encountered during the study was the identification and procurement of appropriate historical and forecasted marine weather data, which is integral to the model's accuracy. Despite these challenges, the results of the study were startlingly optimistic, showcasing the model's ability to predict with impressive accuracy. In conclusion, the developed forecasting tool not only offers promise in its immediate application but also sets a robust precedent for the adoption and adaptation of similar predictive systems in various analogous use cases in the marine environment and beyond.

Keywords Machine Learning, Deep Learning, AI, Artificial Intelligence, Predictive AI, AI/ML, Shark Research, Shark Attack Research, Marine Biology, Shark Biology

1. Introduction

1.1. The Importance of Forecasting Shark Attacks

The waters surrounding our coastlines offer innumerable opportunities for recreation, livelihood, and exploration. Yet, as is the nature of the vast marine environment, it also holds certain risks. Foremost among these for many beach goers and marine enthusiasts is the potential for shark attacks. Safeguarding these waters, SafeWaters has taken the mantle to mitigate shark attack risks for the American populace. A glaring statistic to note is that in 2021, 47 Americans became victims of unprovoked shark attacks, accounting for a staggering 64% of the worldwide total of such incidents. This translates to 137 recorded shark bites across the globe, of which 11 resulted in fatalities [1]. Such numbers are not merely anomalous blips on the radar. Historical data reveals an escalating trend, with the number of unprovoked shark attacks in the U.S. escalating from a mere four in 1955 to a concerning 57 in 2015 [2]. These statistics underscore the pressing need for robust measures to predict and prevent such unfortunate encounters.

1.2. Current Methods and Their Limitations

Present-day strategies to mitigate shark attacks, while well-intentioned, often introduce more problems than they solve. A prime example is the utilization of shark nets. While designed to protect swimmers and surfers, these nets have a devastating environmental impact. Alarmingly, shark nets contribute to the death of millions of sharks every month. To put this into perspective, approximately 273 million sharks meet untimely deaths annually due to such human interventions [3]. These measures not only diminish the shark population but also disrupt marine ecosystems, where sharks play a pivotal role as apex preda tors. The repercussions of these actions resonate through the food chain, leading to imbalances that may, in the long run, prove even more detrimental to marine life and human activity in coastal areas. Yet, these

aren't the only shortcomings. Many of the prevalent methods are reactive, localized, and fail to provide real-time or future-oriented insights.

1.3. Why an AI Approach Might Offer Advantages

In the ever-evolving landscape of technology, Artificial Intelligence (AI) stands out as a beacon of potential, particularly in the domain of predictive analytics. Utilizing past data, AI models can discern patterns and trends that are often too intricate for human analysis. This retrospective analysis is pivotal, but AI's real prowess lies in its ability to forecast future events based on these patterns. By implementing an AI-based approach to shark attack prediction, there's the po tential to provide timely warnings, allowing beachgoers and authorities to take early precautions. Furthermore, such a tool can be constantly updated with real-time data, ensuring that its predictions are always grounded in the latest available information. Beyond just shark attacks, the adaptability of AI means that similar models can be employed for a myriad of other marine-related predictions, ushering in a new era of maritime safety and ecological awareness.

2. Methods

2.1. Data Collection

2.1.1. Global Shark Attack File Utilization

To establish the foundation for the study, we tapped into the comprehensive resource of the Global Shark Attack File. This repository offers a holistic dataset, capturing various instances of shark attacks, encompassing both provoked and unprovoked encounters. The primary focus was to extract information relevant to the parameters of our study.

2.1.2. Dataset Cleaning and Structuring

Upon downloading the dataset, our first objective was to curate the information to ensure clarity and relevance. All rows, save for those representing the 'date' and 'location' of the attacks, were pruned. To facilitate the machine learning process, a new column titled "attack" was appended to the dataset. In this column, a "1" was placed against each entry, serving as a binary indicator of the occurrence of a shark attack on that specific date and location.

2.1.3. Geocoding the Dataset

For the next phase, we employed Python, a versatile programming language re nowned for its data manipulation capabilities. With a script crafted in Python, the dataset was looped to geocode each location entry. This was accomplished by leveraging the robust capabilities of Google's Geocode API, which returned lati tude and longitude coordinates for each listed location. Consequently, these coordinates were appended to the dataset, associating each shark attack event with its precise geographical point.

2.1.4. Integrating Marine Weather Data

Recognizing the potential correlation between marine weather conditions and shark behavior, it was deemed essential to incorporate marine weather data. A Historic Marine Weather API was utilized to fetch relevant weather data for each row indicating an attack. The fetched data was subsequently output to a new CSV file.

2.1.5. Expanding the Dataset

To ensure a thorough representation of both shark attack and non-attack days, another Python script was

conceived. This script was responsible for retrieving marine weather data for every day of the year, for each beach globally with a documented shark attack. This was traced back until 2015, governed by the constraints of the available historical data.

2.1.6. Final Dataset Compilation

Upon retrieval, individual CSVs were concatenated, culminating in a comprehensive dataset. For each location, days devoid of any recorded shark attacks were introduced into the dataset. These entries were labeled with a "0" in the "attack" column, offering a binary distinction for machine learning algorithms to discern between days with and without shark attacks. The expansive dataset is full with detailed insights and granular data points that encompass the entire spectrum of marine conditions and their potential influences on shark behavior. The masking + indicator column method was applied to handle any missing da ta.Through this meticulous process, we ensured that our dataset not only cap tured instances of shark attacks but also provided a holistic view of marine con-ditions, geographical coordinates, and the relative frequency of these events.

This expansive dataset served as the cornerstone for our predictive modeling.

The collected data is most appropriate for this use case as marine weather condi tions have been the backbone for understanding attacks throughout attack research history, already knowing cloudy days & murky waters have been the con ditions for a magnitude of attacks. Expanding the scope of marine weather vari able an letting our advanced neural network learn the relationships and weights of the 30 different monitored marine weather conditions is essential to learn their impact on the sharks behavior and aggression.

2.2. Model Selection

2.2.1. Selection of a Fully Connected Neural Network with Binary Output

In the realms of Artificial Intelligence and machine learning, multiple models exist, each with its specific capabilities, advantages, and potential shortcomings. For the task at hand—predicting the risk of shark attacks based on historical and meteorological data—a fully connected neural network (FCNN) was selected.

2.2.2. Rationale behind Choosing the Fully Connected Neural Network

The FCNN was considered apt for several reasons:

1) Comprehensive Feature Learning: FCNNs have the ability to automatically and adaptively learn spatial hierarchies of features from input data. Given the mul tidimensional nature of our dataset—encompassing geographical coordinates, marine weather conditions, and temporal data—the ability of the FCNN to capture intricate patterns in such datasets made it a logical choice.

2) Binary Classification: Our objective was to predict the occurrence or nonoccurrence of a shark attack, which is essentially a binary classification problem. FCNNs, when combined with a sigmoid activation function in the output layer, are adept at such binary classifications.

3) Flexibility: The neural network architecture allows for easy adjustments. By tweaking the number of layers and nodes, or neurons, in each layer, the network can be adapted to handle varying complexities in data.

2.2.3. Training, Validation, and Testing Procedures

• Data Preprocessing:

• Date and time were parsed from strings to datetime objects, enabling the ex traction of meaningful features like "Month", "Day", "Hour", and "Minute".

• Categorical features like “moon_phase”, “weatherDesc”, and “swellDir16Point”were transformed using one-hot encoding to convert them into a machine-readable format without introducing ordinal relationships where none exist.
• The MinMaxScaler was applied to normalize the features to ensure that no variable overshadows another due to differences in their magnitudes.
• Dataset Splitting:
• The data was divided into training and testing sets (70% and 30%, respectively) using stratified sampling, ensuring that the proportion of positive (attack) and negative (no attack) samples remained consistent across both sets.
•Neural Network Architecture
• The network comprises an input layer, three hidden layers, and an output layer.
• To prevent over fitting and improve generalization, dropout layers were introduced between the hidden layers.
• The final layer employed a sigmoid activation function, aligning with our binary classification objective.
• Model Compilation and Training
• The model was compiled using the Adam optimizer and binary cross-entropy as the loss function—standard for binary classification tasks.
• Early stopping was incorporated into the training process, monitoring the validation loss. This halts training if the model’s performance on the validation data does not improve after ten epochs, ensuring efficient training and preventing over fitting.
• The model was then trained using the training set, setting aside 20% of it for validation.
• Model Evaluation
• Once trained, the model’s predictions on the test set were converted to binary values (1 for “attack” and 0 for “no attack”).
• Evaluation metrics like precision, recall, and F1-score were calculated for the positive class (attack instances) to gauge the model’s accuracy in predicting actual shark attacks.

2.2.4. Conclusion
The chosen FCNN was specifically designed, structured, and optimized for the nature of the data at hand and the binary classification objective. With its deep layers and data preprocessing steps, the model efficiently learned from the historical data to predict shark attack occurrences, positioning it as a promising tool for safety measures and decision-making in marine activities.

3. Implementation
3.1. Introduction to the Implemented Technology
To make forecasting shark attacks as accessible and user-friendly as possible, the AI-driven solution was integrated into a mobile application. This allows users to quickly and seamlessly determine the risk of shark attacks for their chosen location, backed by a powerful AI model trained on historical and marine weather data.

3.2. Functionality of the Mobile App
• User Interface:
• Upon launching the app, users are greeted with a simple interface prompting them to input their location of interest.
• Processing the user Input:

• Once the location is submitted, the mobile app sends this data as a POST request to an endpoint. This endpoint is hosted on a Flask application, which is set up and running on Google Cloud, ensuring high availability and scalability.
• Geocoding the Location:
• The Flask application processes the incoming location data and uses a geocoding service to convert the provided location name into its corresponding latitude and longitude coordinates. This is a vital step, as our AI model and marine weather API both require specific geographical coordinates for pre cise forecasting.
• Fetching Marine Weather Forecasts:
• With the derived latitude and longitude in hand, the Flask app then sends a request to the marine weather API. In response, the API provides marine weather forecasts for the upcoming seven days for the specified location.
• Running the AI Model for Predictions:
• The fetched seven-day marine weather forecast is input into the h5 file containing the weights of the trained model. This model processes the weather data and returns a seven-day risk prediction in the form of binary outputs (0 or 1 for each day).
• Classifying and Displaying the Risk:
• The binary outputs from the model are classified into three categories:
1) Low Risk (if the model's output is 0 to 0.33)
2) Medium Risk (if the model's output is between 0.34 and 0.66)
3) High Risk (if the model's output is between 0.67 and 1)
• These risk levels are then displayed in the mobile app, providing users with a clear and comprehensible seven-day forecast of shark attack risk for their chosen location.
• It is important to know the risk forecasts rely on the accuracy of the marine weather forecasts as the forecasted marine weather variables are used as input in the forecasting model.

3.3. Conclusion

The mobile application serves as an intuitive bridge between end-users and a sophisticated AI model. Through a streamlined process, users can quickly ascertain the shark attack risk for any location, enabling them to make informed decisions about their marine activities. The backend, hosted on Google Cloud, ensures that the app remains responsive and accurate, utilizing real-time marine weather data and the power of deep learning to provide reliable forecasts.

4. Results

4.1. Introduction

In the domain of shark attack forecasting, the primary objective is to predict attacks accurately, ensuring the safety of individuals in and around marine waters.

Evaluating the efficacy of our model is, therefore, crucial in validating its applicability and usefulness. This section will present the results from our deep learning model, emphasizing its performance metrics, and comparing its predictions to existing preventative measures.

4.2. Model's Performance Metrics

• Accuracy:
• The model's forecast accuracy for the positive class on the test set is an impressive 0.8289. This implies that in about 82.89% of instances, the model correctly predicted days with a higher risk of

Shark attack.
• Precision:
• A precision score of 1.00 indicates that every time our model predicted a high-risk day, it was correct. In other words, there were no false positives.
• Recall:
• The model achieved a recall of 0.80, suggesting that it was able to correctly identify 80% of actual high-risk days.
• F1-Score:
• With an F1-score of 0.89, the model showcases a harmonious balance between precision and recall, ensuring that the model is neither too conservative nor too liberal in its predictions.
• Confusion Matrix Visualization: (Figure 1)
• For a more granular view of the model's performance, a confusion matrix has been included. This matrix provides a visual representation of the model's true positive, false positive, true negative, and false negative predictions. By analyzing this matrix, stakeholders can gain deeper insights into the model's strengths and areas of potential improvement.

4.3. Comparison to Existing Preventative Measures

While our model utilizes advanced machine learning and real-time marine weath er data to forecast shark attack risks, the traditional preventative measures have been rather generic and are based on common-sense practices. Some of these include:
• Avoiding shiny jewelry as it can resemble fish scales and attract sharks.
• Limiting excessive splashing which can lure sharks closer.
• Refraining from swimming during early mornings or late evenings.
• Avoiding swimming during cloudy days or in murky waters as it reduces vi sibility for sharks, increasing the chances of accidental bites.

While these practices are sound advice, they are broad and do not offer specific guidance for a particular day or location. Moreover, these practices lean towards caution, potentially discouraging people from enjoying marine activities even on days with low shark attack risks.
Our AI-driven approach offers several advantages over these traditional measures:

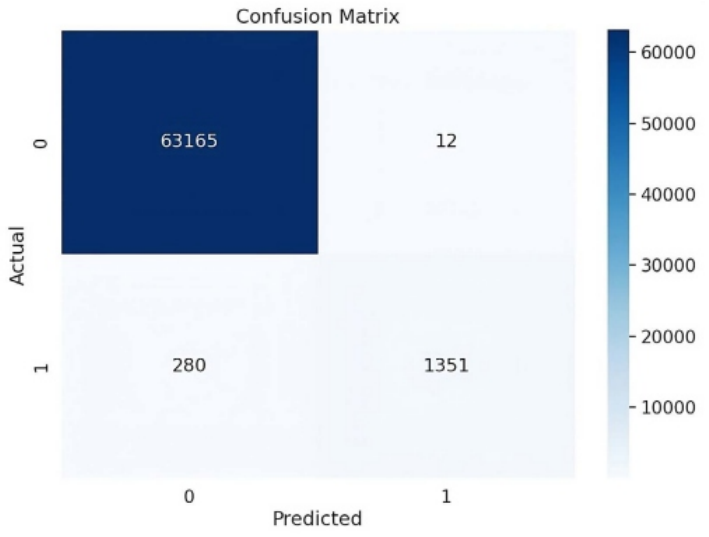


Figure 1. Confusion matrix visualization.

1) Specificity: By providing risk assessments for specific locations and days, individuals can make informed decisions about their activities.
2) Real-time Data Integration: The model's integration with marine weather data ensures that its predictions are based on current conditions, increasing its reliability.
3) Accessibility: Through a mobile app, users have instant access to risk assessments, making it more user-friendly than recalling a list of general best practices.

4.4. Conclusion

The results from our deep learning model are promising, showcasing high accuracy and precision. While traditional shark attack prevention advice is valuable, the integration of AI and real-time weather data offers a more dynamic, specific, and user-friendly approach to risk assessment. As the world continues to advance technologically, such tools can set a precedent, ensuring marine safety through data-driven insights.

5. Discussion

5.1. Introduction

The development and deployment of a predictive model for shark attack forecasting marks a pivotal moment in marine safety. As with all pioneering ventures, it's essential to evaluate its overall significance, understand how it compares to traditional methods, and determine areas that can be fine-tuned or expanded upon for future iterations. This section delves into the significance of our model, its improvements over existing measures, and potential areas of refinement.

5.2. Improvements over Existing Methods

• Dynamic Update of Attack Data:

• One of the most significant advantages of our model over traditional methods is its adaptability. As new shark attack data becomes available, the model can easily integrate this information to refine its predictive capabilities. Traditional methods remain static, but our model evolves, ensuring that it remains relevant and accurate over time.

• Variable Inclusion and Expansion:

• The ability to incorporate additional variables as they become accessible or as more research emerges means that our model can continuously grow in complexity and precision. This dynamic nature contrasts with the fixed guidelines of existing shark attack prevention advice, which can't be easily updated or expanded without broad public re-education campaigns.

• Personalized and Location-specific Risk Assessment:

• By offering location-specific risk assessments, our model provides more actionable insights than broad preventative guidelines. This tailored advice can empower individuals to make informed decisions about their safety.

5.3. Limitations and Areas for Improvement

• Data Reliability and Completeness:

• The efficacy of the model is intrinsically tied to the quality and completeness of the data it ingests. Any inaccuracies or gaps in the shark attack database could potentially skew predictions. Future iterations could focus on data verification mechanisms or diversifying data sources to create a more holistic dataset.

• Generalization across Different Coastal Ecosystems:

• Different coastal areas might have distinct marine ecosystems, which could influence shark behavior

in unique ways. Tailoring the model to account for these regional differences could improve its predictive accuracy across various geographies.

• Feedback Mechanism:

• Currently, once a prediction is made, there is no feedback loop to confirm if the forecasted risk materialized or not. Integrating a user feedback mechan ism within the mobile app could provide valuable real-world validation data.

• Model Interpretability:

• While neural networks are powerful predictors, they are often termed as "black boxes" due to their lack of interpretability. Ensuring stakeholders understand the model's decisions might be crucial for broader acceptance. Fu ture work could focus on model transparency or incorporating explainable AI techniques.

• External Factors:

• Certain external factors, such as sudden changes in local fish populations, human activities like fishing tournaments, or marine celebrations, could in fluence shark movement and behavior. Including such events as additional variables could refine the model's predictive power.

5.4. Conclusion

The launch of our AI-driven shark attack forecasting model represents a sub stantial leap forward in marine safety. Its dynamic nature, ability to incorporate new variables, and location-specific predictions offer unparalleled advantages over traditional preventative measures. However, as with all innovative solu tions, it's essential to acknowledge its limitations and continuously seek avenues for refinement. Embracing a mindset of continuous improvement ensures that the model remains at the forefront of marine safety innovations.

6. Project Conclusion

The journey from conceptualizing the need for an AI-driven shark attack fore casting model to its eventual deployment has been both challenging and enligh tening. This research sought to harness the power of machine learning to fill a gap in marine safety—forecasting the risk of shark attacks. The importance of this endeavor was underpinned by data showing the alarming rise of such inci dents, particularly in the U.S.

6.1 Main Findings

1) Feasibility: Through data collection, preprocessing, and model selection, it was established that it is indeed feasible to use available data to forecast future shark attack risks. With a focus on binary output that could be further classified into risk categories, the model displayed a promising ability to predict potential threats.

2) Model Performance: The fully connected neural network model demonstrated significant accuracy in its predictions, especially for the positive class. With a forecast accuracy of 0.8289 for positive instances, the model also show cased commendable precision, recall, and F1-score metrics. This signifies a robust prediction capability, especially given the novel nature of this venture.3) Comparison with Existing Methods: Compared to traditional measures which revolve around broad guidelines, the model stands out with its dynamic, personalized, and location-specific risk assessments. While there's no direct benchmark in the realm of shark attack prediction, the model's approach sur passes general prevention advice both in granularity and adaptability.

4) Implementation Insights: The creation of a mobile app offers a direct, us er-friendly interface to the public. By allowing users to input their locations and subsequently providing a tailored risk assessment,

the application not only serves as a testament to the practical application of AI but also as an invaluable tool for marine safety.

6.2. Future Directions

1) Data Expansion and Refinement: As more shark attack data becomes available, future iterations of the model can integrate this information, refining its predictive power, improving accuracy. Furthermore, collaboration with ma-rine biologists and shark experts can help in identifying new relevant data points.
2) Model Variants: Exploring other machine learning or deep learning archi tectures might offer even better predictive performance. Variations of neural networks, such as convolutional or recurrent architectures, might be worth investigating, especially as the dataset grows in complexity.
3) User Feedback Integration: By integrating a feedback mechanism within the mobile app, it's possible to collect real-world validation data. This feedback can be invaluable in continuously refining the model's accuracy.
4) Broadening the Application: While the current focus is on shark attacks, similar frameworks could potentially be applied to predict other marine-related incidents, setting a precedent for diverse use-cases.
In wrapping up, the successful development and deployment of the AI-driven shark attack forecasting model showcases the immense potential of artificial intelligence in pioneering new safety solutions. As the tides of technology continuously ebb and flow, this research stands as a beacon, illuminating the path to a safer marine future.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] International Shark Attack File (2023) Yearly Worldwide Shark Attack Summary. Florida Museum. https://www.floridamuseum.ufl.edu/shark-attacks/yearly-worldwide-summary/#:~:text=Global%20total%20of%20unprovoked%20shark,average%20of%2072%20incidents%20annually

[2] Quartz Staff (2019) Shark Attacks Are on the Rise in the US and Australia. Quartz. https://qz.com/1560529/shark-attacks-are-on-the-rise-in-the-us-and-australia/

[3] Stanich, A. (2023) Discover How Many Sharks Are Killed per Year and How You Can Help Them. A-Z Animals. https://a-z-animals.com/blog/discover-how-many-sharks-are-killed-per-year-and-how-you-can-help-them/.

A Hybrid Neural Network Model Based on Transfer Learning for Forecasting Forex Market

Salum Hassan Faru1, Anthony Waititu2, Lawrence Nderu3
1Department of Mathematics, The Pan African University, Institute for Basic Sciences, Technology and Innovation (PAUSTI), Nairobi, Kenya
2Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya
3Department of Computing and Information Technology, Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya

ABSTRACT

The forecasting research literature has developed greatly in recent years as a result of advances in information technology. Financial time-series tasks have made substantial use of machine learning and deep neural networks, but build ing a prediction model from scratch takes time and computational resources. Transfer learning is growing popular in tackling these constraints of training time and computational resources in several disciplines. This study proposes a hybrid base model for the financial time series prediction employing therecurrent neural network (RNN) and long-short term memory (LSTM) called RNN-LSTM. We used random search to fine-tune the hyperparameters and compared our proposed model to the RNN and LSTM base models and eva luate using the RMSE, MAE, and MAPE metrics. When forecasting Forex currency pairs GBP/USD, USD/ZAR, and AUD/NZD our proposed base model for transfer learning outperforms RNN and LSTM base model with root mean squared errors of 0.007656, 0.165250, and 0.001730 respectively.

Keywords:Deep Learning, Transfer Learning, Time Series Analysis, RNN, LSTM

1. Introduction

Foreign Exchange (dubbed Forex or FX) is a global currency trading market considered the most liquid global financial market. According to the Bank for International Settlements, trading in Forex markets averaged $5.3 trillion daily in April 2013 [1]. Due to its anonymous nature and increased instability of price rates, the Forex market is deemed very complicated and volatile, resembled with
the black box [2].

The Forex market is an example of a financial time series market where cur rency exchange rates are traded in pairs. These pairs are categorized into three groups, namely major, minor, and exotic currency pairs [3]. Predicting price time series in financial markets, which have a non-stationary nature, takes much work [4] [5]. They are dynamic, chaotic, noisy, and non-linear series that the market is prejudiced by the general economy, characteristics of the industries, politics, and even the psychology of investors [6] [7]. Researchers for forecasting
the Forex market have proposed many different methods. In the late 1990s, popular ways were statistical methods [3]. The neural network was a revolutionary discovery for time series forecasting problems, and many other methodologies have arisen from this occasion [8] [9].

Due to the limitations of classic machine learning methods, researchers and data scientists have recently embraced the concept of transfer learning. Traditional machine learning models, such as RNN, LSTM, CNN, GRU, SVM, and others, necessitate training from scratch, which is computationally expensive

and demands vast data to attain good performance [10]. They also use an isolated training strategy, in which each model is trained separately for a specific task without relying on prior information. Researchers are now using transfer learning to overcome these constraints; however, transfer learning in time series prediction problems has yet to be widespread [11].

Pre-trained models for time-series predictions still need to be improved, despite transfer learning's growing popularity in tackling these constraints of training time and computational resources in several disciplines like Natural Language Processing (NLP), image, video, and audio problems. Google's Tensor Flow Hub and NVIDIA contain several pre-trained models for various problem domains such as text, photo, video, and audio.

In recent years, the most popular Contract for Differences (CFDs) brokers have displayed standardized risk warnings, including the proportion of losses on a CFD provider's accounts held by retail investors, and the data shows that losing funds made up 54% to 83% of the total, with 76% being the average1. From this observation, despite having different prediction models, the percentage of profitable traders needs to be bigger!

This study proposes a hybrid model for the Forex market prediction employing the recurrent neural network (RNN) and long-short-term memory (LSTM) called RNN-LSTM. Based on transfer learning, the model is used to predict Forex market currency pairs as a pre-trained model for future work related to the time series problem. The study creates a dashboard to help users make the best trade selections and maximize their winning rates to reduce the number of unprofitable traders.

The rest of the paper is structured as follows; Section 2 designates the literature review, and Section 3 materials and methods. Section 4 presents the experimental study and results analysis. Section 5 discusses of results, and Section 6 concludes the work.

2. Literature Review

As a result of the advancement of information technology, the forecasting research literature has considerably increased. The results reveal that neural network models outperform statistical models like ARIMA (Autoregressive integrated moving average), indicating its applicability for forecasting foreign exchange rates. For the Turkish TL/US dollar exchange rate series, [12] used both the ARIMA time series model and neural networks. The outcomes demonstrate the superiority of ANNs over statistical models, with the ANN approach outperforming the ARIMA time series model in terms of accuracy [9].

The Recurrent Neural Network (RNN) has been employed in most studies because it has the advantage of remembering some information about the sequence through hidden states. In contrast to a standard neural network, the input and output are entirely independent. However, [13] has observed the drawbacks of RNN and stated that gradient-based learning approaches take far too long, since the error vanishes or explodes as it propagates back. In a RNN, the issue of vanishing and exploding gradients has been addressed in various ways, and the LSTM is one of the most well-known.

A C-RNN forecasting technique based on deep RNN and CNN for Forex time series data was offered by [14]. Data-driven analysis was employed by the researcher to fully harness the Spatiotemporal properties of Forex time series data.

The C-RNN foreign exchange time-series data forecast approach has increased applicability and accuracy compared to LSTM and CNN, according to an experimental comparison of the predicted strategy on the exchange rate data of nine major foreign exchange currencies.

The direction of the Forex market using LSTM was forecasted by [15]. The macroeconomic information and the technical indicator dataset were both examined by the researcher. The study created a hybrid model that integrated the two LSTMs that corresponded to the study's two datasets, and it was proven to be quite successful on actual data. The key motive for the researcher to use a hybrid was to avoid the

shortcomings of LSTMs. When ME_LSTM and TI_LSTM models were executed separately, many transactions with wrong signals were generated, reducing the accuracy.
For predicting stock prices, [16] suggested a novel deep learning-based model.
They constructed a hybrid model by merging the well-known LSTM and GRU neural network models. The model uses the LSTM output as the input to the GRU unit. They employed S&P 500 historical time series data and traditional evaluation criteria like MSE, MAPE, and so on for model validation.
Transfer learning is a method for encoding features from a model that has already been learned, preventing us from having to create a new model from scratch. A pre-trained model is often developed on a large datasets, and the weights gained from the trained model can be used with your custom neural network for any other similar application. These freshly constructed models can be used directly for task predictions or in training processes for related applications. This method reduces both the training time and the generalization error [17].
To forecast short-term stock price movement, [11] created a deep transfer with related stock information (DTRSI) model that blends transfer learning and a deep neural network. The researcher developed an LSTM base model and trained it using substantial stock data from the US and Korean markets. When the model parameters were tuned, it predicted the stock price of Hyundai Motor Co. with an average accuracy of 64.54 per cent and the stock price of Amazon Co. Inc with an average accuracy of 62.65 per cent.Also, [17] suggested a methodology for the stock market prediction that com bined transfer learning of data from industrial chains with deep learning algo rithms such as multilayer perceptron (MLP), RNN, LSTM, and GRU. These al gorithms were used to forecast the trend of 379 stock market indices in China. They discovered that RNNs are occasionally the best prediction option when dealing with detailed time-series data. For this reason, the MLP was selected for transfer learning stock market index prediction, and the maturity yield surpasses the buy-and-hold strategy.
Due to the drawbacks of RNN and limitations of LSTM, and to leverage the strength of RNN and LSTM, this study suggests the hybrid RNN-LSTM modelused as the primary model for transfer learning time-series tasks. Despite the ap plication of transfer learning, [11] studies show that the prediction performance of stock markets could be more impressive. Consequently, additional studies can be conducted to enhance the performance. To the best of our knowledge, no pa per has used the hybrid model RNN-LSTM to pre-train the source domain for time series prediction problems, especially Forex market currency prediction using transfer learning, thus filling the gap.

3. Materials and Methods

A data science process is a series of instructions outlining how a person or a team should carry out a data science project. According to [18], the three most common processes used for data science projects are Cross Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Databases (KDD), and Sample-Explore-Modify-Model-Assess (SEMMA).The most often used framework for carrying out data science initiatives is CRISP-DM, which provides a structured and systematic approach to data mining projects, which leads to increased efficiency and success. KDD highlights the high-level applications of particular Data Mining techniques and refers to the general process of discovering knowledge in massive databases2. The CRISP-DM data science process was used in this study due to the nature of our work and the objectives.

3.1. Data Description

Depending on the broker, the Forex market has more than 100 currency pairs. For instance, Exness broker3 has seven major pairs, 26 minor pairs, and 67 exotic pairings. For training our hybrid base model, we employed six major currency pairs, seven minor currency pairs, and seven exotic currency

pairs for 20 currency pairs. We used five currency pairs as our target currency pairs: one central currency pair, two minor currency pairs, and two exotic currency pairs, as given in Table 1.

Table 1. Currency pairs used

Source data		
Major	**Minor**	**Exotic**
AUD/USD	AUD/CHF	AUD/JPY
EUR/USD	CAD/JPY	CHF/JPY
NZD/USD	EUR/CHF	EUR/NOK
USD/JPY	GBP/AUD	EUR/AUD
USD/CAD	GBP/NZD	CAD/HKD
USD/CHF	NZD/JPY	AUD/SGD
	AUD/CAD	ZAR/JPY
Target data		
GBP/USD	CAD/CHF	USD/ZAR
	EUR/CAD	AUD/NZD

3.2. Technical Indicators

Price movements follow trends, according to technical analysts. Those price changes often follow recognized patterns that can be partially attributed to market psychology based on the widely-held belief that market participants act the same when faced with analogous situations. We have also used the five most popular technical indicators to forecast the closing price for the future hour, including the stochastic oscillator, Bollinger band, relative strength index (RSI), and exponential moving averages (EMA).

3.3. Modeling

3.3.1. Recurrent Neural Networks (RNN) Model

According to [19], the RNN belongs to the class of neural networks that process sequential data. They accept a series of vectors (xx x 1 2 ,,, n) as input and generate a second series (hh h 1 2 ,,, n) that contains details of the input sequence at each step. In particular, RNNs use a recurrent hidden state, whose activation at each iteration depends on the activation at the previous iteration, to address variable-length sequences.Figure 1 shows the architecture of RNN model. The implementation of up dating recurrent hidden state Rt h is as follows:

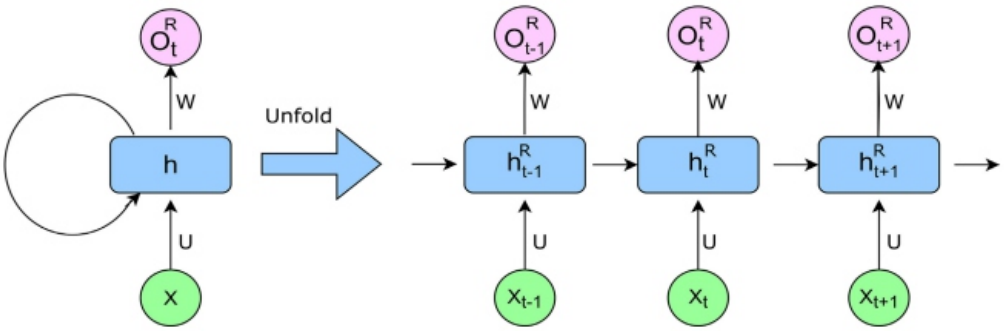


Figure 1. RNN model architecture.

$$h_t^R = g\left(W \times x_t + U \times h_{t-1}^R + b\right) \quad (1)$$

$$O_t^R = g\left(W^o \times h_t^R + b^o\right) \quad (2)$$

In this case, g is a bounded and smooth function, such as a hyperbolic tangent or logistic function and R Ot is the output/predicted value from recurrent neural network. The network's input vector t x , along with its prior hidden state 1Rt h −and the bias b, determine its recurrent hidden state Rt h per time t step. The weight matrices, W and U, are filters that choose how much significance to assign based on the input at hand and the hidden state from the past. They gener ate an error, which is returned via back propagation and used to change theweights of their parameters until the error cannot be reduced further [20].

3.3.2. Long Short-Term Memory (LSTM) Model

The LSTM, which [21] proposed, is a specific type of recurrent network that addresses the problems of vanishing and exploding gradients by includingmemory units that enable the system to comprehend when to forget the earlier hidden states and when to revamp the hidden conditions with the latest infor mation. In the literature, models with hidden units and different linkages within the memory unit have been put forth with remarkable practical success [20]

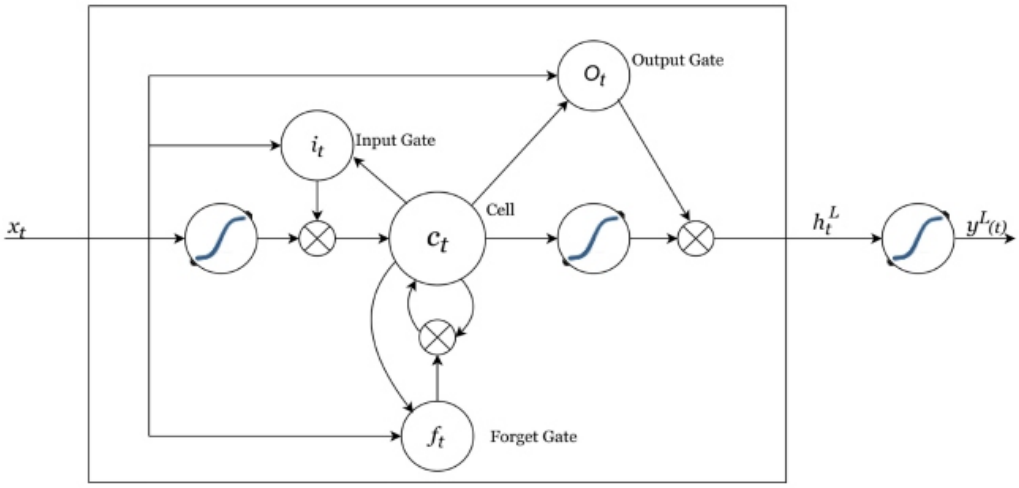


Figure 2. LSTM network.

The LSTM transition equations are as following [20]:

$$i_t = \sigma\left(W^i \times x_t + U^i \times h_{t-1}^L + b^i\right) \quad (3)$$

$$f_t = \sigma\left(W^f \times x_t + U^f \times h_{t-1}^L + b^f\right) \quad (4)$$

$$o_t = \sigma\left(W^o \times x_t + U^o \times h_{t-1}^L + b^o\right) \quad (5)$$

$$u_t = \tanh\left(W^u \times x_t + U^u \times h_{t-1}^L + b^u\right) \quad (6)$$

$$c_t = i_t \odot u_t + f_t \odot c_{t-1} \quad (7)$$

$$h_t^L = o_t \odot \tanh\left(c_t\right) \quad (8)$$

$$y^L\left(t\right) = \sigma\left(W^o \times h_t^L\right) \quad (9)$$

From Equation (3) to (9), t i , t f , and t o is the input, forget, and an output gate respectively, where t c denote a memory cell, an activation function is giv en by t u , LSTM hidden state by Lt h , and the predicted value is given by ( ) L y t .

Figure 2 illustrates the defaulting connection among these LSTM units.In contrast to a standard recurrent unit, which simply replaces its scope per time step, the LSTM can decide whether to maintain the present memory with the aid of the additional gates. Intuitively, the output gate determines the amount of internal memory state accessible, while the input gate chooses what fresh information will be updated. How much of the previous memory cell is forgotten is then determined by the forget gate.

3.3.3. Proposed Hybrid RNN-LSTM Base Model

We provide thorough information about our proposed hybrid model in this section. The hybrid is made up of RNN and LSTM, as seen in Figure 3. As [3] stated, both are sophisticated neural network models that can outperform regression-based predictions in terms of accuracy.

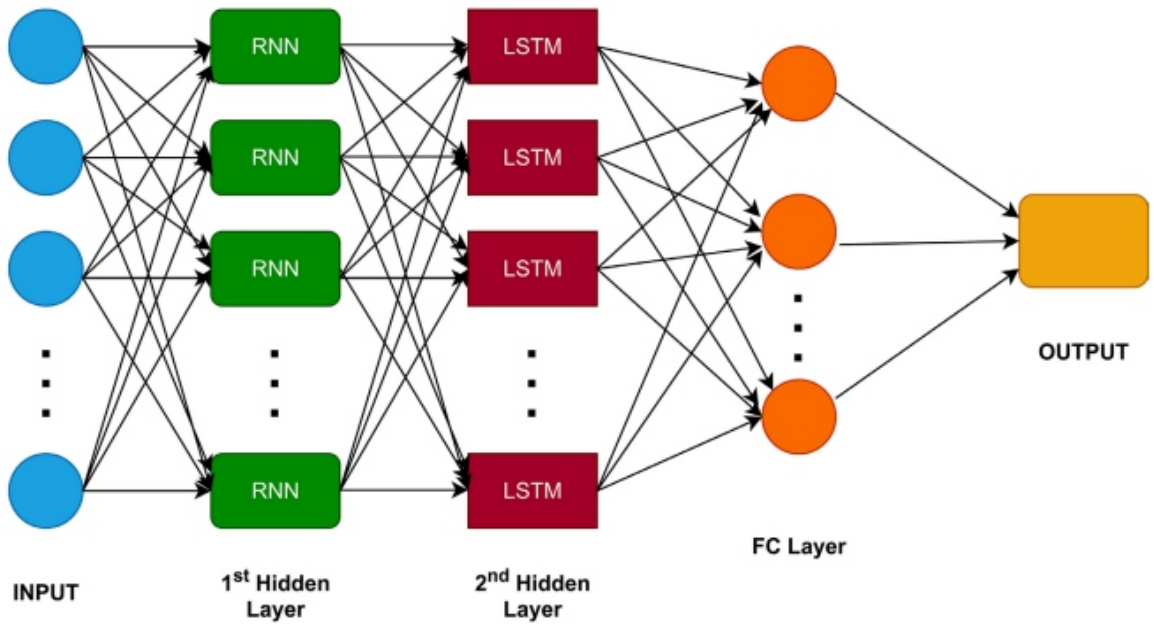


Figure 3. A hybrid RNN-LSTM base model architecture.

Figure 3 represents the architecture of hybrid base model where FC is fully connected layer. The primary logic behind employing these networks was that currency price prediction is a regression-based problem. To begin, we fed the data into RNN network, and obtain the output R Ot as shown in Equation (2).

To get the final prediction ˆt y , we pass the output of the RNN layer to the LSTM layer.Mathematically, the input in LSTM layer for our hybrid will change from t x to R Ot as shown below,

$$i_t^H = \sigma\left(W^i \times O_t^R + U^i \times h_{t-1}^L + b^i\right) \quad (10)$$

$$f_t^H = \sigma\left(W^f \times O_t^R + U^f \times h_{t-1}^L + b^f\right) \quad (11)$$

$$o_t^H = \sigma\left(W^o \times O_t^R + U^o \times h_{t-1}^L + b^o\right) \quad (12)$$

$$u_t^H = \tanh\left(W^u \times O_t^R + U^u \times h_{t-1}^L + b^u\right) \quad (13)$$

$$c_t^H = i_t^H \odot u_t^H + f_t^H \odot c_{t-1}^H \quad (14)$$

$$h_t^H = o_t^H \odot \tanh\left(c_t^H\right) \quad (15)$$

And lastly, the predicted price ˆt y is give by;

$$\hat{y}_t = \sigma\left(W^o \times h_t^H\right) \quad (16)$$

where o W is the weight matrix of the output gate and Ht h is the hidden state of Hybrid unit given by Equation (15).

3.3.4. Transfer Learning

Transfer learning can manage scenarios where domains and distributions differ, unlike typical machine learning and data mining techniques, which presume that training and testing data are from the same feature space and distribution. These features allow the model to make advantage of relevant source data and apply the underlying knowledge to the target job, resulting in increased performance [22].

Due to this benefit, in our study, we heavily rely on datasets of currency pairs in the Forex market for training the base model and learning knowledge from them before applying the knowledge to the target data (i.e., where data from only the target currency pair is used). Figure 4 depicts the transfer learning framework employing a hybrid base model for the source and destination domains.

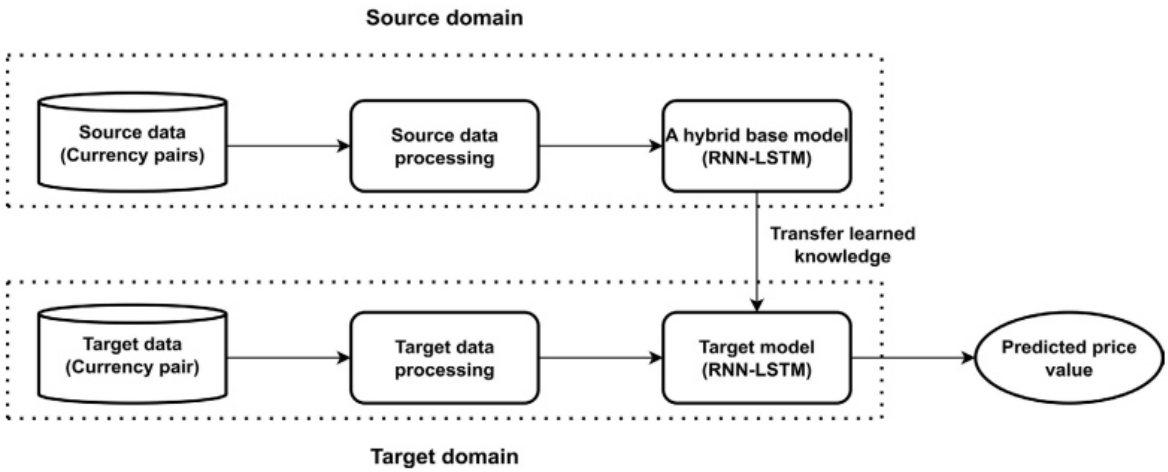


Figure 4. A framework of transfer learning.

4. Experiment and Results

The research environment's software and hardware configuration are displayed in Table 2.

Table 2. Configuration of an experimental environment.

Items	Parameter
Central processing unit (CPU)	Core(TM) i7-7600U CPU@2.80GHz
Graphics processing unit (GPU)	Intel(R) HD Graphics 620
Random-access memory (RAM)	16 GB
Programming language	Python3.8.8
Deep learning framework	TensorFlow v2.9.1

For our experiment, the historical data of Forex market where downloaded from the Tickstory website4, with the time period of 3 months and hourly time frame.

4.1. Data Preprocessing

None of the datasets we used contained missing values. To avoid bias, we transformed the data into a scale range of [0, 1], acknowledging that variables recorded at different scales do not equally contribute to model fitting and learned function. We divided the dataset into the ratio of 6:2:2, with 60% for training, 20% for validation, and 20% for testing. We created the training dataset with shape (1052, 120, 20) and the testing dataset with shape (293, 120, 20) using the sliding window method approach. The shape reflects the number of samples, time steps, and currency pairs used in each set. The 120 time steps indicate the number of prior time steps used for forecasting the subsequent one-hour closing price, and the 20 currency pairs indicate the total number of currency pairs used in the model.

4.2. Performance Metrics

Every prediction model must be evaluated in order to determine its accuracy [23]. In this study, the three basic assessment metrics for regression issues for model evaluation: root mean squared error (RMSE),

mean absolute error (MAE), and mean absolute percentage error (MAPE) are used to evaluate and compare the RNN, LSTM, and hybrid RNN-LSTM base model performance.
RMSE show an estimation of the residual between the real t y value and anticipated value, ˆt y , MAE estimates the average magnitude of the errors in forecasts without looking at their direction, and MAPE indicates an average absolute percentage error; it is preferable if the MAPE is as low as possible [16]. Equations representing these metrics are given below:

$$\text{RMSE} = \sqrt{\frac{1}{N} * \sum_{t=1}^{N} (y_t - \hat{y}_t)^2} \tag{17}$$

$$\text{MAE} = \frac{1}{N} * \sum_{t=1}^{N} |y_t - \hat{y}_t| \tag{18}$$

$$\text{MAPE} = \frac{1}{N} * \sum_{t=1}^{N} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \tag{19}$$

4.3. Results Analysis

We provide our findings for both source and target domain. We used the closing prices of 20 and 5 currency pairs to train and evaluate our hybrid base model for the source domain.

4.3.1. Case 1

We trained and validated the base model for our hybrid RNN-LSTM model using 20 currency pairings as the source domain, as shown in Table 1. After training and validating the base model, we transferred the learned weights and biases to the target domain model to anticipate currency pairings.
However, Table 3 shows unsatisfactory outcomes for our forecasts of the target domain currency pair, like GBP/USD. As a result, we opted for scenario 2.

Table 3. Model forecast evaluation case 1

Model	RMSE	MAE	MAPE
RNN	16.138934	16.138692	0.935115
LSTM	15.880864	15.880837	0.934288
RNN-LSTM	15.757492	15.754454	0.955955

4.3.2. Case 2

In this instance, we reduced the number of currency pairs in the source domain utilized for training and validating the basic model from 20 to 5, as shown in Table 4.

Table 4. Currency pairs used in case 2.

Source data		
Major	**Minor**	**Exotic**
AUD/USD	EUR/CHF	ZAR/JPY
USD/CAD	GBP/NZD	
Target data		
GBP/USD	USD/JPY	USD/ZAR
EUR/USD		AUD/NZD

We trained and validated our base model with the default parameters and the RMSE and Loss (MSE) for 5 currency pairs used are shown in Figure 5.

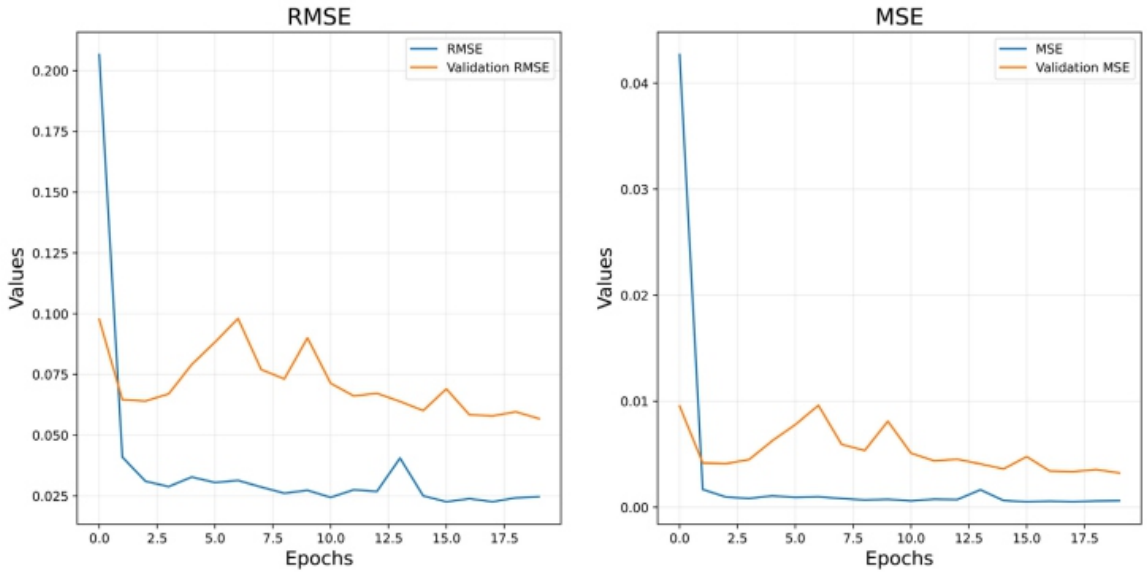


Figure 5. RMSE and loss before parameter tuning.

To reduce the training and validation RMSE and Loss, we fine-tuned the hyperparameters of our hybrid RNN-LSTM base model as there are displayed in Table 5. To navigate the vast search space, we used a random search to find the best hyperparameter optimization. The optimal set of hyperparameters with the lowest error consisted of 2 hidden layers for RNN and LSTM, 40 neurons, a dropout probability of 0.2, a batch size of 64, and 70 epochs. The improved RMSE and Loss are shown in Figure 6.

Table 5. Hyperparameters.

Hyperparameter	Interval
Hidden layers for RNN and LSTM	2 to 3
Number of neurons	14 to 70
Dropout rate	0.1 to 0.9
Batch size	16, 32, and 64
Epochs	10 to 100

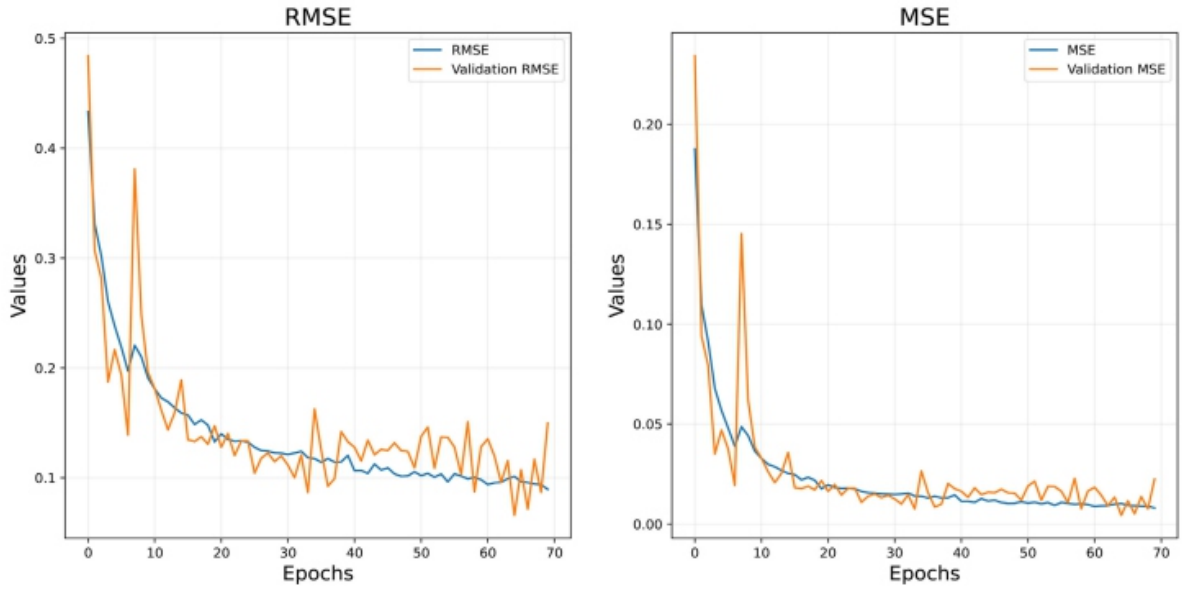


Figure 6. RMSE and loss after parameter tuning.

We utilized the weights and biases from our base hybrid model for the target domain and trained our dataset on top of it. To forecast the one-hour closing price of frequently traded Forex currency pairings like GBP/USD, EUR/USD, USD/JPY, USD/ZAR, and AUD/NZD, we used both historical data of closing prices for the past 3 months and technical indicators.

To train the target model for prediction, we froze the top third of the base model's layers, attached our dense layer, trained it with 64 batches, and ran 70 epochs. We made predictions for the currency pairs using both the RNN base model and the LSTM base model, with the same structure as the proposed RNN-LSTM hy brid model, for comparison purposes. The results of the predicted closing price are displayed in Table 6.

Table 6. Model forecast evaluation.

Model	Pairs	RMSE	MAE	MAPE
RNN	GBP/USD	0.022116	0.020246	0.017770
	EUR/USD	0.008276	0.007461	0.007584
	USD/JPY	2.091997	1.764091	0.012051
	USD/ZAR	0.275423	0.263631	0.014701
	AUD/NZD	0.005693	0.004293	0.003826
LSTM	GBP/USD	0.017378	0.015948	0.014063
	EUR/USD	**0.003796**	**0.003054**	**0.003117**
	USD/JPY	**1.072977**	**0.879307**	**0.005954**
	USD/ZAR	0.322514	0.312474	0.017475
	AUD/NZD	0.003056	0.002676	0.002392

	GBP/USD	**0.007656**	**0.006473**	**0.005756**
	EUR/USD	0.016708	0.015403	0.015498
RNN-LSTM	USD/JPY	3.443044	3.260152	0.022532
	USD/ZAR	**0.16525**	**0.13648**	**0.007774**
	AUD/NZD	**0.001730**	**0.001355**	**0.001210**

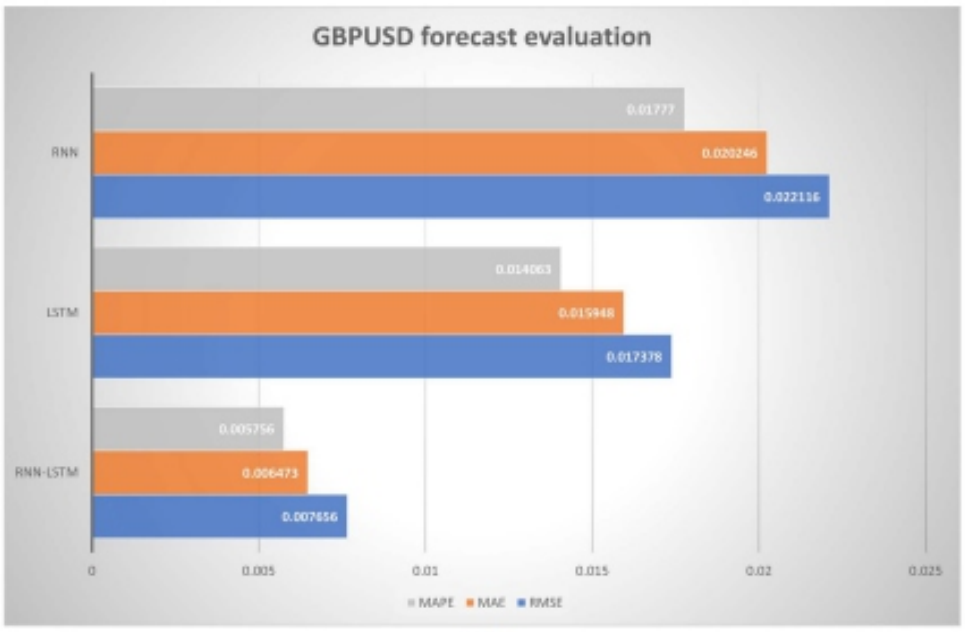
Our hybrid RNN-LSTM model achieved low RMSEs for GBP/USD, USD/ZAR, and AUD/NZD, with values of 0.007656, 0.16525, and 0.001730 respectively,according to Table 6. The LSTM base model had small RMSEs for EUR/USD and USD/JPY, with values of 0.003796 and 1.072977. The RNN-LSTM model produced the smallest MAE and MAPE values for GBP/USD, USD/ZAR, and AUD/NZD, while the LSTM model produced the smallest values for EUR/USD and USD/JPY.

5. Discussion

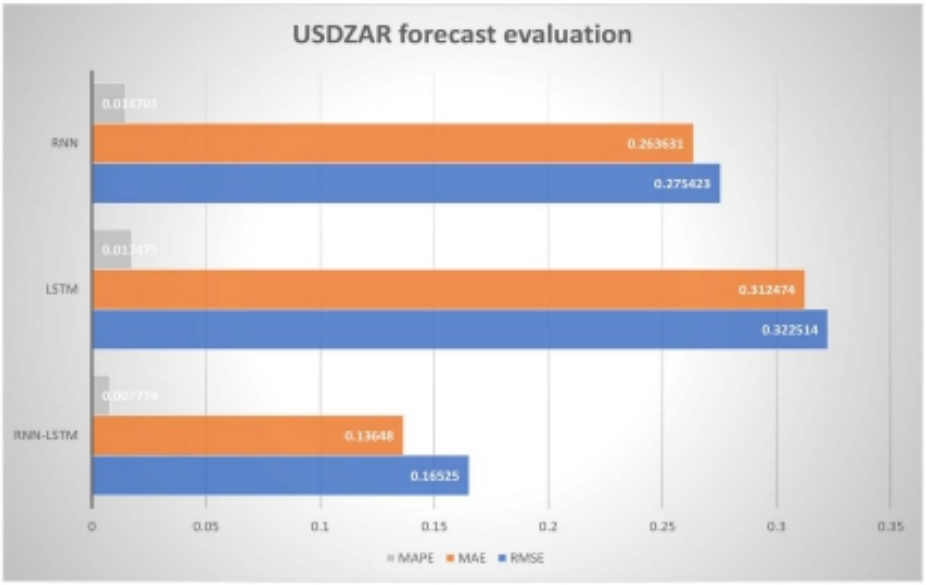
In this study, we initially proposed to train and validate the hybrid RNN-LSTM base model using 20 different currency pairs in the source domain (Table 1).

However, when forecasting the target domain currency pair, the results were not satisfactory (Table 3). One of the factors that might have contributed to this was the use of a large number of datasets with different scales before applying trans fer learning.To address this issue, we retrained the base model using only 5 closely related currency pairs that have similar scales. The results showed improvement, as seen in Table 6, suggesting that the proposed hybrid RNN-LSTM base model works better with small datasets.

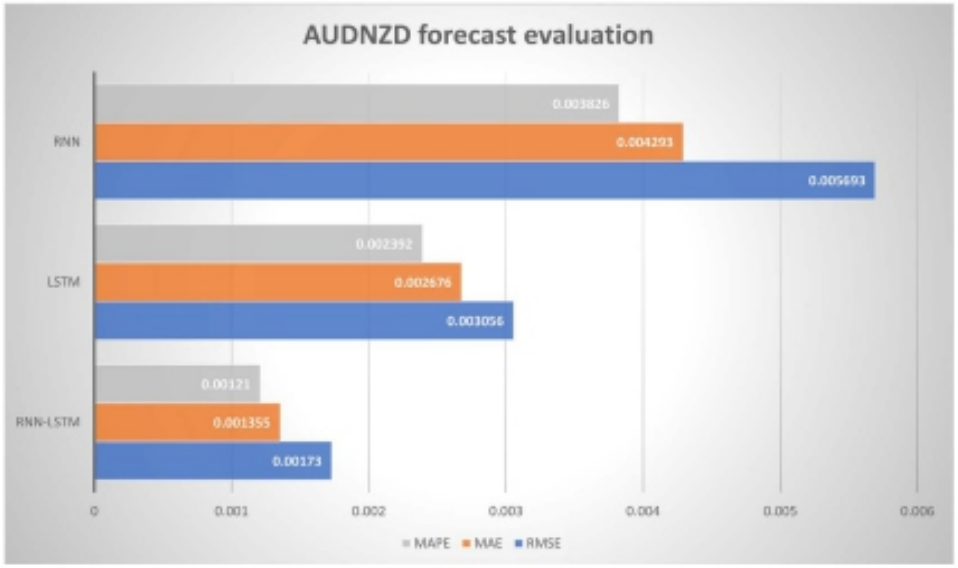
When forecasting GBP/USD, USD/ZAR, and AUD/NZD as shown in Figure 7, our proposed hybrid RNN-LSTM base model for transfer learning outper forms RNN and LSTM base model with both RMSE, MAE and MAPE. The LSTM basic model provided superior forecasts for the EUR/USD and USD/JPY as depicted in Figure 8.



(a)

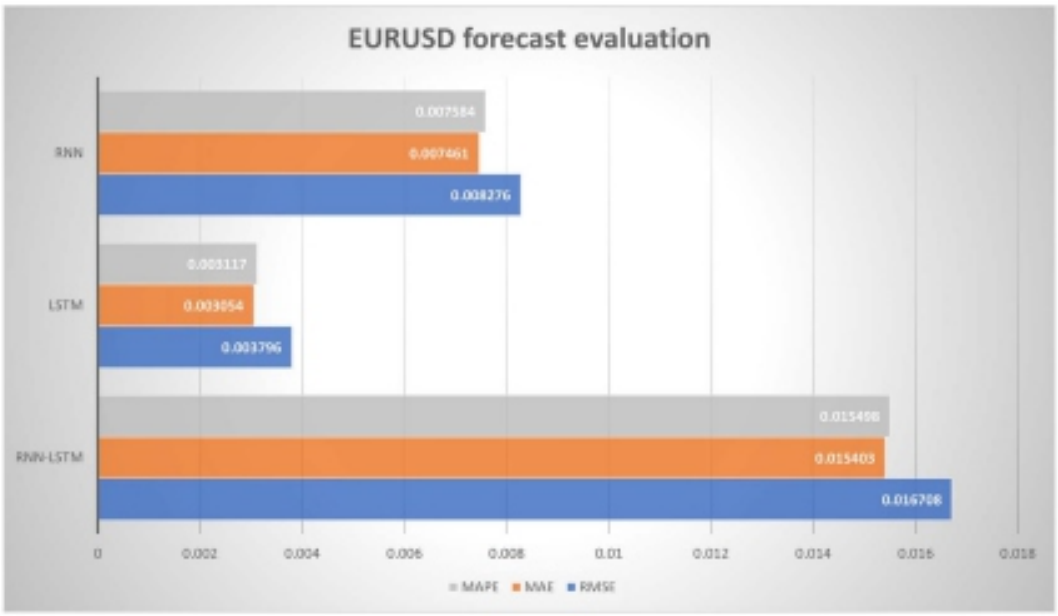


(b)

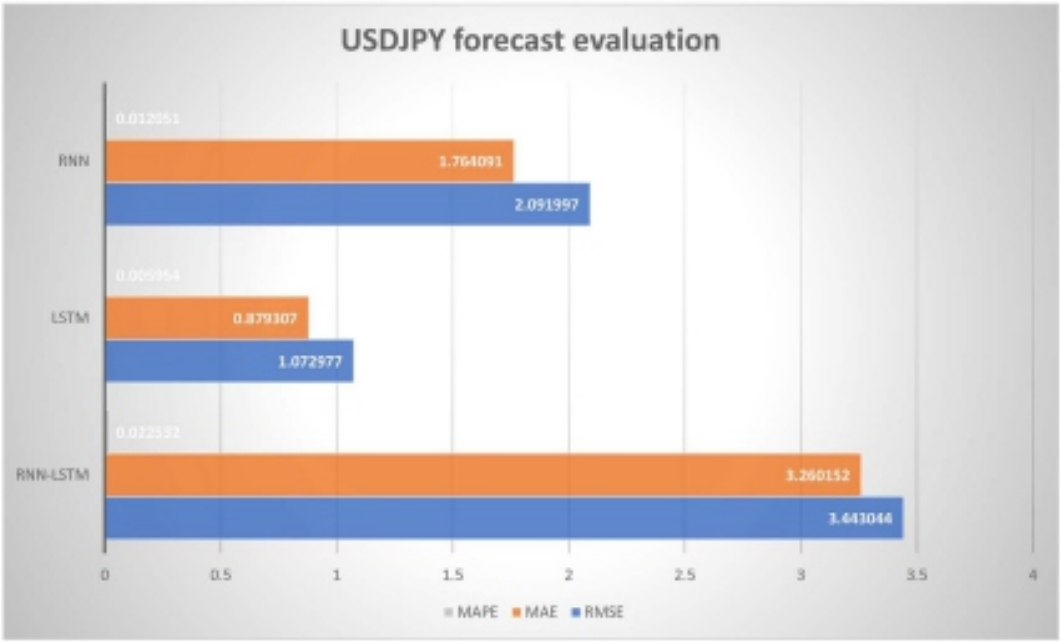


(c)

Figure 7. Evaluation metrics of GBP/USD, USD/ZAR, and AUD/NZD for all models.



(a)



(b)

Figure 8. Evaluation metrics of EUR/USD and USD/JPY for all models.

This indicates that our hybrid RNN-LSTM model is more computationally efficient and requires less time to train compared to other stock and Forex predic tion models. Additionally, the use of transfer learning and the fine-tuning of hyperparameters also helped to improve the performance of the model. These results demonstrate the potential of our proposed hybrid RNN-LSTM model for real-time stock and Forex prediction.

Overall, our suggested hybrid RNN-LSTM base model outperforms RNN and LSTM base models when forecasting the currency pairings in the target domain, as evidenced by the prediction of three of the

forecasted currency pairs more accurately by our proposed RNN-LSTM base model, and the AUDNZD was forecasted more precisely than other currency pairs examined. Figure 9 displays the out-of-sample forecast for the AUDNZD.

6. Conclusions

The results of our study indicate that the hybrid RNN-LSTM base model pro vides better accuracy compared to the RNN and LSTM base models when forecasting the target domain currency pairings. Our proposed model requires a lower number of epochs to converge and has a shorter average computation time of 4.29 minutes, making it a more efficient option for Forex traders. Also, using a smaller set of closely related currency pairs in the source domain improved the model's performance. This study highlights the effectiveness of using transfer learning and hybrid RNN-LSTM models for financial forecasting, particularly in the Forex market.

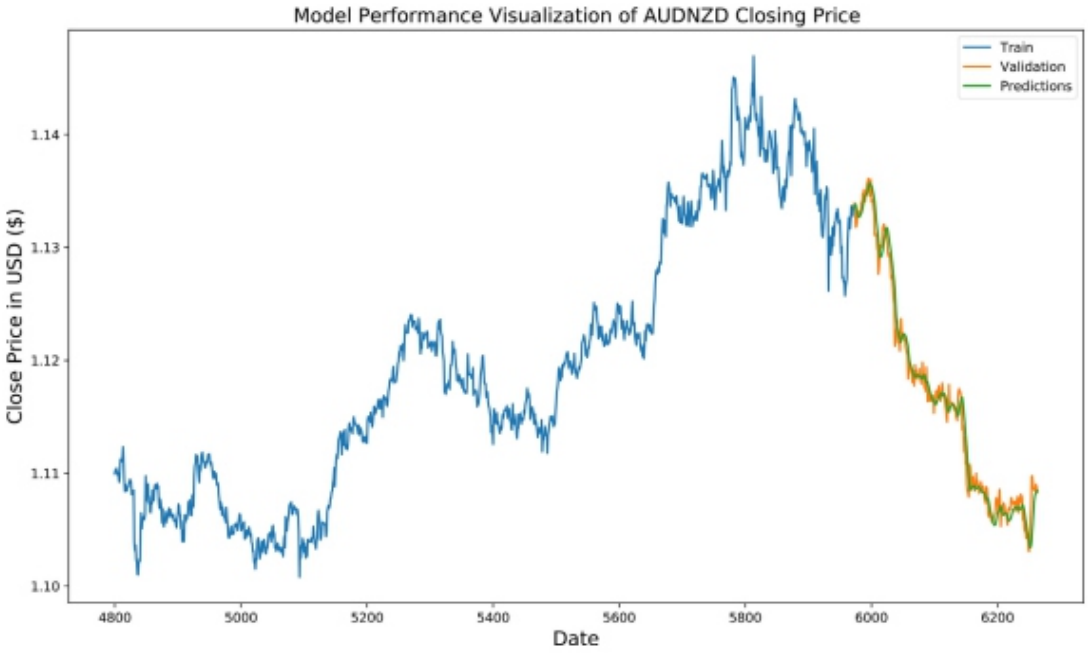


Figure 9. Out-sample forecast of AUDNZD.

Additionally, fine-tuning the hyperparameters and exploring different optimization algorithms and loss functions can also be considered in future stu dies to further improve the performance of the proposed model. Additionally, incorporating other factors such as economic news and sentiment analysis, can also be considered as they play a crucial role in the Forex market and can help to increase the accuracy of the predictions. Overall, our proposed hybrid RNN-LSTM base model has shown promising results in predicting the Forex market, but there is still room for improvement and further exploration in the field. The web dep loyment for this study for real time dataset can be accessed through the follow ing linkhttps://mrfaru-fx-forecasting-fx-forecasting-app-tprth8.streamlit.app/.

Acknowledgment

My supervisors, Prof. Anthony Waititu and Dr. Lawrence Nderu, deserve my gratitude for guiding me through this project and allowing me to give it my all. Without them, I knew that this trip would be difficult for me. I appreciate your continued help, family and friends. Most importantly, I want to thank my lovely mother Jamila Habale and my devoted and loving wife Nuru Mohamed, who never ceases to

inspire me.

Conflicts of Interest

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

References

[1] Cirillo, S., Lloyd, S. and Nordin, P. (2014) Evolving Intraday Foreign Exchange Trading Strategies Utilizing Multiple Instruments Price Series. https://arxiv.org/abs/1411.2153v1

[2] Anastasakis, L. and Mort, N. (2009) Exchange Rate Forecasting Using a Combined Parametric and Nonparametric Self-Organising Modelling Approach. Expert Systems with Applications, 36, 12001-12011. https://doi.org/10.1016/j.eswa.2009.03.057

[3] Islam, M.S., Hossain, E., Rahman, A., Hossain, M.S. and Andersson, K. (2020) A Review on Recent Advancements in FOREX Currency Prediction. Algorithms, 13,186. https://doi.org/10.3390/a13080186

[4] Tay, F.E. and Cao, L. (2001) Application of Support Vector Machines in Financial Time Series Forecasting. Omega, 29, 309-317.https://doi.org/10.1016/S0305-0483(01)00026-3

[5] Zhang, N., Lin, A. and Shang, P. (2017) Multidimensional k-Nearest Neighbor Model Based on EEMD for Financial Time Series Forecasting. Physica A: Statistical Mechanics and Its Applications, 477, 161-173.https://doi.org/10.1016/j.physa.2017.02.072

[6] Chen, H., Xiao, K., Sun, J. and Wu, S. (2017) A Double-Layer Neural Network Framework for High-Frequency Forecasting. ACM Transactions on Management Information Systems, 7, 4. https://doi.org/10.1145/3021380

[7] Henrique, B.M., Sobreiro, V.A. and Kimura, H. (2019) Literature Review: Machine Learning Techniques Applied to Financial Market Prediction. Expert Systems with Applications, 124, 226-251. https://doi.org/10.1016/j.eswa.2019.01.012

[8] Balabanov, T., Zankinski, I. and Dobrinkova, N. (2012) Time Series Prediction by Artificial Neural Networks and Differential Evolution in Distributed Environment. Large-Scale Scientific Computing 8th International Conference, LSSC 2011, Sozo-pol, 6-10 June 2011, 198-205. https://doi.org/10.1007/978-3-642-29843-1_22https://link.springer.com/chapter/10.1007/978-3-642-29843-1_22

[9] Rehman, M., Khan, G.M. and Mahmud, S.A. (2014) Foreign Currency Exchange Rates Prediction Using CGP and Recurrent Neural Network. IERI Procedia, 10, 239-244.https://doi.org/10.1016/j.ieri.2014.09.083

[10] Gu, Q., Dai, Q., Yu, H. and Ye, R. (2021) Integrating Multi-Source Transfer Learning, Active Learning and Metric Learning Paradigms for Time Series Prediction. Applied Soft Computing, 109, Article ID: 107583.https://doi.org/10.1016/j.asoc.2021.107583

[11] Nguyen, T.T. and Yoon, S. (2019) A Novel Approach to Short-Term Stock Price Movement Prediction Using Transfer Learning. Applied Sciences (Switzerland), 9,22. https://doi.org/10.3390/app9224745

[12] Cem Kadilar, D., Şdmşek, M., Gör Çağdaş Hakan Aladağ, A., Üniversitesi, H.,Fakültesi, F. and Bölümü, D. (2009) Forecasting the Exchange Rate Series with Ann:The Case of Turkey. Istanbul University Econometrics and Statistics e-Journal, 9,17-29. https://dergipark.org.tr/en/pub/iuekois/issue/8991/112073

[13] Hochreiter, S. (1998) The Vanishing Gradient Problem during Learning Recurrent Neural Nets and Problem Solutions. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6, 107-116.https://doi.org/10.1142/S0218488598000094

[14] Ni, L., Li, Y., Wang, X., Zhang, J., Yu, J. and Qi, C. (2019) Forecasting of Forex Time Series Data

Based on Deep Learning. Procedia Computer Science, 147, 647-652.https://doi.org/10.1016/j.procs.2019.01.189

[15] Yıldırım, D.C., Toroslu, I.H. and Fiore, U. (2021) Forecasting Directional Move ment of Forex Data Using LSTM with Technical and Macroeconomic Indicators. Financial Innovation, 7, 1-36. https://doi.org/10.1186/s40854-020-00220-2

[16] Hossain, M.A., Karim, R., Thulasiram, R., Bruce, N.D. and Wang, Y. (2019) Hybrid Deep Learning Model for Stock Price Prediction. Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018, Bengaluru, 18-21 November 2018, 1837-1844. https://doi.org/10.1109/SSCI.2018.8628641

[17] Wu, D., Wang, X. and Wu, S. (2022) Jointly Modeling Transfer Learning of Indus trial Chain Information and Deep Learning for Stock Prediction. Expert Systems with Applications, 191, Article ID: 116257.https://doi.org/10.1016/j.eswa.2021.116257

[18] Azevedo, A. and Santos, M.F. (2008) KDD, SEMMA and CRISP-DM: A Parallel Overview. IADIS European Conference Data Mining, Amsterdam, 24-26 July 2008,182-185.

[19] Rathore, V.S., Worring, M., Mishra, D.K., Joshi, A. and Maheshwari, S. (2018) Emerging Trends in Expert Applications and Security: Proceedings of ICETEAS 2018. Springer, Berlin, 485-495. https://doi.org/10.1007/978-981-13-2285-3

[20] Huynh, H.D., Dang, L.M. and Duong, D. (2017) A New Model for Stock Price Movements Prediction Using Deep Neural Network. Proceedings of the 8th Inter national Symposium on Information and Communication Technology, Nha Trang,7-8 December 2017, 57-62. https://doi.org/10.1145/3155133.3155202

[21] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. Neural Com putation, 9, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735[22] Farahani, A., Pourshojae, B., Rasheed, K. and Arabnia, H.R. (2020) A Concise Re view of Transfer Learning. Proceedings 2020 International Conference on Comput ational Science and Computational Intelligence, CSCI 2020, Las Vegas, 16-18 De cember 2020, 344-351. https://doi.org/10.1109/CSCI51800.2020.00065

[23] Ubert de Almeida, B., Ferreira Neves, R. and Horta, N. (2018) Combining Support Vector Machine with Genetic Algorithms to Optimize Investments in Forex Mar kets with High Leverage. Applied Soft Computing Journal, 64, 596-613.https://doi.org/10.1016/j.asoc.2017.12.047.

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).
2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);
3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);
4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005