Adapting deep learning models between regional markets

Tonkin, Isaac; Gepp, Adrian; Harris, Geoff; Vanstone, Bruce

Neural Computing and Applications

DOI:
https://doi.org/10.1007/s00521-022-07805-1

Published: 01/01/2023

Publisher's PDF, also known as Version of record

Neural Computing and Applications
https://doi.org/10.1007/s00521-022-07805-1

ORIGINAL ARTICLE

Adapting deep learning models between regional markets

Isaac Tonkin[1] · Adrian Gepp[1] · Geoff Harris[1] · Bruce Vanstone[1,2]

Received: 17 October 2021 / Accepted: 6 September 2022

Abstract
This paper extends a series of deep learning models developed on US equity data to the Australian market. The model architectures are retrained, without structural modification, and tested on Australian data comparable with the original US data. Relative to the original US-based results, the retrained models are statistically less accurate at predicting next day returns. The models were also modified in the standard train/validate manner on the Australian data, and these models yielded significantly better predictive results on the holdout data. It was determined that the best-performing models were a CNN and LSTM, attaining highly significant Z-scores of 6.154 and 8.789, respectively. Due to the relative structural similarity across all models, the improvement is ascribed to regional influences within the respective training data sets. Such unique regional differences are consistent with views in the literature stating that deep learning models in computational finance that are developed and trained on a single market will always contain market-specific bias. Given this finding, future research into the development of deep learning models trained on global markets is recommended.

Keywords Deep learning · Machine learning · Candlesticks · Technical analysis

1 Introduction

This paper investigates the application of deep learning models between two regionally distinct financial markets. The aim is to determine if such financial models can replicate their performance on regionally distinct, though comparable, markets. Specifically, this study examines this problem using a range of deep learning techniques including fully connected, convolutional, and recurrent networks. Evaluation of the performance of neural network architectures across various regional markets is of interest due to the different microstructure factors present within each market. These factors impact the ability of models to generalise across markets.

Motivation for such research is also driven by the ongoing need of professionals to use the latest tools when investing in financial markets. As a result of the industry-wide push to identify investment opportunities more accurately, there is a continuing drive towards the adoption of predictive algorithms such as neural networks, which have been shown to obtain significant results in several fields, including computer vision, natural language processing, health analytics, engineering, and game-playing [1–9]. The attraction to such investment pursuits has driven the value of the world's stock markets to the point that it is well-known that these markets comprise a substantial quantity of global wealth. A recent estimate places the value of the global equities market at more than USD $110 trillion [10].

The past two decades have also seen enormous growth in the predictive power of neural networks as well as the development of several new neural network classes such as generative adversarial networks and transformers [11, 12]. A symbiotic increase in computational power has fuelled the growth and widespread adoption of deep neural networks. This increasing use of neural networks is

Adrian Gepp, Geoff Harris and Bruce Vanstone have equally contributed to this work.

✉ Adrian Gepp
adgepp@bond.edu.au

Isaac Tonkin
itonkin@bond.edu.au

Geoff Harris
gharris@bond.edu.au

Bruce Vanstone
b.vanstone@bangor.ac.uk

[1] Centre for Data Analytics, Bond Business School, Bond University, Gold Coast, QLD 4226, Australia

[2] Bangor Business School, Bangor University, Bangor, Gwynedd LL57 2DG, Wales

Published online: 27 September 2022

predominantly due to their ability to act as automated pattern recognition machines that can be trained on real-world data without an explicit theoretical basis as to the complete inner workings of these models. As pattern recognition machines, neural networks have been used previously to identify patterns in the financial markets [13–24].

Two of the most widely used neural network classes in the financial modelling literature are convolutional neural networks (CNN) and recurrent neural networks (RNN). Convolutional neural networks are typically used for computer vision tasks, such as image recognition or video classification. However, due to their ability to extract increasingly abstract and generalised features, they have achieved substantial successes in a broad range of fields, particularly in finance [25]. A common input to convolutional neural networks in finance is raw numerical data. As an example, Gudelek et al. [26] use a 2-D CNN to predict the following day's stock price for exchange-traded funds (ETFs). They use trend and momentum indicators as inputs to their CNN and are able to obtain a significant and positive return from their backtesting procedure. Hoseinzade and Haratizadeh [27] extend the finance and deep learning literature through the use of a 3-D CNN. They use a set of 82 variables including economic factors, technical indicators and index-specific factors as inputs to this 3D-CNN. Beyond the standard application of convolutional neural networks to raw numerical data, there is a small, but growing body of the literature that utilises visual financial inputs as an alternative style of input data. From the small number of studies, it appears that the use of visual inputs (e.g. price or candlestick charts) is able to produce significant results [28–30]. In addition, research is ongoing to improve our ability to understand what features of the input data (or features extracted within hidden layers) are having the most important effect on the learning and prediction process. Traditional methods such as Hinton diagrams [31, 32] are still in use in the financial literature [17], but there is an increasing body of work that seeks to improve upon the traditional approaches to gain insight into what are often deemed to be black boxes [33, 34].

Recurrent neural networks make allowance for temporal factors and are therefore often used for modelling time-series data, particularly in finance [17, 25]. There are many varieties of recurrent neural networks; however, the most commonly used are the long-short term memory (LSTM) [35] and the gated recurrent unit (GRU) [36]. One particular study of note [17] utilised recurrent neural networks (as well as a suite of deep and machine learning models) to predict the next day return for stocks in the S &P500 index. They were able to achieve significant positive returns using relatively small RNNs. Nelson et al. [37] obtained similarly promising results, but they also provide a suite of technical indicators as well as price history as additional inputs to the LSTM. Matsumoto and Makimoto [38] investigated the performance of various machine and deep learning models in equity investing and found that LSTM models outperformed a range of other candidate models on the S &P500. Similar results were also obtained by Fischer and Krauss [19]. It is also becoming more common for a variety of neural networks to be combined to create a hybrid network, such as the LSTM-CNN employed by Kim and Kim [39]. A similar study by Liu et al. [40] also used an LSTM-CNN to perform strategy analysis, as well as to improve stock selection and timing, and found that their hybrid neural network was able to outperform two benchmarks: the respective index and the classic momentum strategy.

Overall, as demonstrated above, there are a variety of ways in which financial data can be input to machine learning models. Candlesticks are one such representation of the standard stock time-series [41]. It has been argued for some time that both finance academics and practitioners would benefit from better understanding the predictive information contained in candlestick charts (see, e.g. [42, 43]). To this end, research has begun that examines the ability of deep neural networks to extract such patterns [44, 45]. Indeed, a systematic review of the contemporary literature investigating the development and application of machine learning to the equities markets has been recently completed by [46] that documented the different machine learning categories that dominate the literature. They make three main conclusions in their review. The first is that there needs to be an increased emphasis on the generalisability of results from machine learning studies. As a result, they suggest that models and approaches should be evaluated across several distinct markets in the future research. The second conclusion notes that the use of machine learning techniques for financial modelling work (regardless of whether it is a black box or not) needs to have due consideration for the financial theory in terms of the inputs to the model, the algorithms utilised, and the subsequent performance analysis. They also conclude that artificial neural networks are best suited for regression-style problems in this area, while support vector machines are better suited for classification tasks.

In keeping with the direction of the literature, the work of Ghoshal and Roberts [17], which was published in an earlier volume of this journal, developed and compared several networks trained on 22 years of US equity data. They found that optimised neural networks outperform standard technical and other shallow learning methods. By examining the weight-space visualisation (Hinton diagrams) of their CNN, they provide a visual interpretation of what the network has recognised as significant candlestick sequences. Their validated best model is statistically significantly better than random choice at predicting the

direction of the next day's returns. However, and in keeping with the first conclusion of Strader et al. [46], they did not attempt to develop or apply their model to other markets. Such extensions are commonplace in the financial literature. Works such as [47, 48] investigate the efficacy of methodologies developed in one market that are then applied to other distinct regional markets. As such, the motivation of this current work is to extend the work of Ghoshal and Roberts [17] to the Australian equities market since in doing so, it will address the key research gap that exists as their approach has not yet been applied to other markets or over various market cycles.

The contributions of the current paper are twofold. First, it provides a continuation study of Ghoshal and Roberts' [17] work on Australian data as supported by the call for future research by Strader et al. [46]. We address their comment on the need to assess results over various market cycles. Secondly, the universal workflow of machine learning [25] is used to independently develop models that best fit the Australian data. Together these contributions address the benchmarking and application of neural network models developed on one market to similar, but geographically separated markets. As such, the key objective of this work is to determine the performance differential of deep learning architectures between regional markets. An additional objective of this work is to begin to address the lack of generalisability of results that has been identified as a key issue in the financial literature [46].

The remainder of this work proceeds as follows: First, the methodology is presented in Sect. 2, including emphasis on the data, deep learning techniques and the training methodology. The results are then discussed in Sect. 3, where it is shown that the findings are statistically significant. Finally, the work concludes in Sect. 4 and future research directions are discussed.

2 Methodology overview

This section provides an overview of the methodological details involved in this study. Neural networks have dominated in popularity over the past decade as the go-to modelling methodology for pattern recognition applications [25]. The application of these pattern recognition models to financial data is a natural step in the application of deep learning models, and their use within finance research has grown substantially over the past decade [46]. The use of deep neural network architectures that follow in this paper necessarily follows Ghoshal and Roberts [17] given that this is a comparative study. A description of the data used herein, and the methods used to ensure a valid comparison with the previous work are described. Details specific to the construction of balanced datasets are provided to ensure conformance and reproducibility with the methodology adopted in Ghoshal and Roberts [17].

2.1 Candlestick data

In keeping with this work being a continuation study, candlestick data from the Australian stock market were collected for training, validation, and test data. Ghoshal and Roberts [17] justify the use of candlestick data as they note that it is widely believed by technical analysts to be a leading indicator of future price movements. The raw data were collected for each company from the publicly available Yahoo Finance website. This included the daily Open, High, Low, Close, Adjusted Close price and Volume data. In their study, Ghoshal and Roberts [17] selected the US S &P500 as their market. For consistency, comparable stocks from the Australian ASX50 were collected. This approach ensured that, just as with the S &P500, these 50 Australian stocks have a significant influence on the local market. All data were adjusted in the usual manner on a per-day basis over the entire period by applying the ratio of the adjusted close and the close price on each day to that day's candlestick values.

For consistency with the original research, this study involved a binary classification for the dependent variable. The dependent variable was either a next-day upward move in the closing price or a next-day downward move in the closing price. In Table 1, the standard data split, broken up by classification, is shown. In this study, great care was taken to ensure that the stocks selected would have sufficient liquidity to allow for realistic results to be obtained. As the Australian market is significantly less liquid than the US market, the top set of stocks with sufficient liquidity is much more limited. Ghoshal and Roberts [17] were able to use a selection of 500 US stocks, where this study was limited to just the top 50 Australian stocks. As such, the judicious manner in which the Australian stocks were selected did not impact upon the continuation of the original study due to the comparable liquidity of the two sets of equities. In addition, the resultant model architectures that were trialled were selected with care to ensure there were sufficient training vectors for the network to be appropriately trained. In order to determine the final set of model hyperparameters, the standard training and validation procedure was applied with the optimization goal of maximising the attained accuracy. Embedded within this optimization was also consideration of the number of trainable parameters used by the model. In the event that two models attained similar performance, but one had substantially fewer trainable parameters than the other, the smaller model would be selected. The entire training and validation process was conducted consistent with standard practice to avoid biasing final out-of-sample results.

Table 1 Summary of final data

Data split	Date range	Number of 'UP'	Number of 'DOWN'	Total number of samples
Training	01/01/1999–31/12/2011	76,436	76,436	152,872
Validation	01/01/2012–31/12/2015	25,071	23,777	48,848
Testing	01/01/2016–31/12/2019	25,411	24,108	49,519

2.2 Sequencing and balancing the datasets

As is well-known with binary classification tasks, and as commented upon in Ghoshal and Roberts [17], convergence of the model parameters during training was crucially dependent upon having balanced input datasets. This is a well-discussed, well-known issue in the deep learning literature when training binary classification systems. As can be seen in Table 1, this issue is addressed here in the standard manner of ensuring the training data is equally balanced between the binary values of the daily returns, assessed close-to-close. In balancing the training data, there are three outcomes in the raw data (assuming the stock continues to trade): the price either increases, remains the same, or decreases. The approach that Ghoshal and Roberts [17] adopted was to introduce noise generated from a Gaussian distribution (with a mean of zero and standard deviation of 0.001) to jitter those data sets with zero-return days into one of either the 'UP' or 'DOWN' classes. This approach satisfactorily reduces the classification task to a binary one for the purposes of training the model. Furthermore, to ensure the entire training dataset is balanced, a threshold is calculated that would place 50% of the jittered returns above that value and 50% beneath. To reduce bias, this dataset-specific threshold was calculated using only the training set and then applied to the training, validation and testing datasets.

In keeping with the approach taken by Ghoshal and Robert [17], the input datasets were then batched into tensors of shape (20, 4). These tensors create a 20-day historical window for each input vector consisting of the Open, High, Low and Close prices of each day. This window overlaps the temporally ordered input training vectors when developing one input training vector per day. It is argued [17] that the rationale for including this historical data window is that it provides a context within which the candlestick pattern of each day could be examined by the neural network. To ensure a fair comparison during the current continuation study, this historical window was kept at 20 days for the Australian data as well.

Finally, each (20, 4) tensor generated (following Ghoshal and Roberts [17]) was individually normalised to ensure the visual appearance of the normalised candlestick is identical to the unscaled candlestick. The (20, 4) input data tensors were stacked into one of either the training, validation or testing datasets, based on date, and consistent with the approach of Ghoshal and Roberts [17]. The resultant tensors had dimensionality (# samples in data split, 20, 4).

2.3 Training methodology and implementation

Ghoshal and Roberts [17] did not explicitly learn any of the standard candlestick patterns. Their approach was to allow the deep learning models to extract potential candlestick patterns as opposed to explicitly learning set patterns based on established candlestick pattern theory. A significant difference between the two approaches is the predictive ability that emerges. Whilst learning theoretical patterns yields apparently superior predictive power, unsupervised learning may be better suited where the nuances contained within the training data (such as geographical location) can be inferred by the network itself.

In justification of their adopted approach, Ghoshal and Roberts [17] used a variety of classic statistical models in addition to machine learning algorithms. The deep learning models produced the most significant results of all their models with Z-scores up to 36.546. That particular result was achieved using a CNN with a single convolutional layer and a filter length equivalent to one trading day. Given a 1-day kernel considers each candlestick individually, and the other 2- or 3-day kernels consider the 2- or 3-day candlestick patterns, this suggests that it is preferable to allow the CNN to identify the combination of individual candlesticks itself, rather than explicitly instructing it to consider more than one candlestick at a time. It is well-known that inputs that carry no information are simply ignored by a network, and thus, in the context of this study, the model determines during training the appropriate number of candlesticks for its representations. The best-performing deep learning model architectures developed by Ghoshal and Roberts [17] were chosen for comparison with the current study. In addition, Ghoshal and Roberts' [17] finding that deep learning methods outperform classical machine learning methods for this task further motivates this work's emphasis on deep learning techniques.

Specifically, this work makes use of fully connected, recurrent, and convolutional layers. The multi-layer

perceptron (MLP) is a classic artificial neural network and is known as a fully connected or densely connected network [25]. Each neuron in each layer is connected to each neuron in the immediately preceding and succeeding layer (where applicable). It consists of at least three layers: an input layer, any number of hidden layers and an output layer. Dense layers are found in a number of other network architectures, typically as classifiers. Although they are also commonly used independently, meaning that the dense layers complete both the feature extraction and classification.

Recurrent neural networks (RNNs) retain information from the current input using internal structures. This retained information is replaced with each subsequent input [25]. As a result of this information progression, they are commonly used for natural language processing or financial time-series modelling [19, 49]. Dense layers are used in recurrent neural networks to classify the outputs from the recurrent layers. There are many types of RNN, although the RNN architectures used in this study are the long short-term memory (LSTM) [35] and the gated recurrent unit (GRU) [36]. Both models were designed to address the vanishing gradient problem [25]. In addition, they share many similarities, but differ primarily because of the gates used. As a result, the GRU also has fewer trainable parameters.

Convolutional neural networks utilise convolutional layers to complete feature extraction. Convolutional layers are spatially invariant, meaning that features learnt in one area of the input may be applied to other areas of that input. This is a significant improvement from the MLP, since they are spatially variant and as a result require additional parameters to learn the same features. Dense layers are employed in CNNs to classify the features that have been extracted by the convolutional layers. CNNs also learn pattern hierarchies, meaning that local features are extracted first and then combined to create more generalised global features [25]. The primary application of CNNs is to computer vision tasks; however, they can also be used on data such as financial time-series. For additional information on these networks, as well as the mathematical formulations, the reader is referred to the work of Goodfellow et al. [50].

In order to compare the results of the models with those of Ghoshal and Roberts [17], the same set of metrics are adopted. Specifically, the metrics used are accuracy, precision, recall, F-score, area under the receiver operating characteristic curve (AUC), Z-score and P value. The standard formulation for accuracy is adopted, which represents the proportion of predictions which are correct. This is calculated using the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Precision and recall can also be calculated using these counts. Precision represents the number of true positives relative to the total number of predicted positives, while recall represents the number of true positives relative to the total number of actual positives. The F-score is a combined representation of precision and recall. Following the methodology and notation of Ghoshal and Roberts [17], we calculate the AUC using the popular scikit-learn Python package and then use this result to obtain the test statistic, U, of the Mann–Whitney–Wilcoxon test as per Mason and Graham [51]. Here, the number of positive and negative samples in the holdout set are represented by n_P and n_N, respectively. The equivalence is then used to obtain the Z-score. We note that the accuracy and Z-score metrics are not directly derived from one another, and however, both metrics do reflect similar attributes of model performance. As such, each provides a different perspective on the overall results. We adopt the standard formulae (in keeping with Ghoshal and Roberts [17]) to define the following measures:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$U = AUC \times n_P \times n_N$$

$$Z = \frac{U - \mu_U}{\sigma_U}$$

$$\mu_U = \frac{n_P \times n_N}{2}$$

$$\sigma_U = \sqrt{\frac{n_P \times n_N \times (n_P + n_N + 1)}{12}}$$

Finance-specific metrics such as profit and compound annual growth rate (CAGR) are also included for comparison purposes. As per [17], the profit is formulated as a multiple of the starting balance, while CAGR is a representation of the annual rate of return.

$$CAGR = \left(\frac{\text{Balance}_{\text{End}}}{\text{Balance}_{\text{Start}}}\right)^{1/t} - 1$$

The models and data manipulation were implemented in Python 3.7.6 using TensorFlow 2.2.0. The typical training time for each model on an NVIDIA GeForce RTX 2080Ti averaged between five and ten minutes each with early stopping enabled. Automated routines were developed specific to this study to ensure a fair and efficient coverage of the parameter phase-space for both models. Training the US-validated model on the Australian data was a straightforward procedure as no model validation was

required. The training and validation data splits were used to ensure the best possible alternative model was developed for the Australian market. The comparison of the results of these models is now detailed.

3 Discussion of results

Tables 2, 3 and 4 present the structure of the MLP, CNN and RNN models trained and validated on the Australian regional data. Interestingly, like the US-validated models, the Australian-validated models share the same general structure, in terms of the number of layers and the number of neurons within those layers.

Utilising the Z-scores as a measure of the efficacy of each model [17], the continuation study did not return the same level of significance compared to the original financial study overall but did perform better on a few of the metrics, as shown in Table 5. The results for US data were obtained from [17], and those reported for Australian data were completed by this study. These initial results were inconclusive as to the efficacy of the model replication. It was proposed to extend the study to include a model chosen by validation of several original models trained on the Australian data. This extension therefore investigates the effect of market-specific microstructure factors on the selection of the final model architecture.

Upon training these models, every originally developed and Australian data validated model generated more favourable results (for most metrics) than those produced by the American data validated (albeit retrained on the same Australian data) model. From the perspective of deep learning model construction, these better results were achieved with significantly fewer trainable parameters, as shown in Table 6. Given these current results could not statistically discriminate between the predictive capabilities of the best-performing models, the LSTM and 1-Day CNN, the standard practice of choosing the architecture with the fewer weights was used to nominate the best model. This standard practice is the application of Occam's razor for deep learning models [52]. In this case, given the CNN has utilised 15 times as many weights, the LSTM is thus chosen to be the superior model. In addition, the LSTM is designed for tasks with temporal elements, whereas the CNN is not. This acts as additional support for this selection of the LSTM model. It may also help to explain why comparable results were obtained by the LSTM, but with far fewer trainable parameters.

It is interesting to note that Ghoshal and Roberts' [17] final models were also substantially larger than those validated in this study (see Table 6). Consequently, it is proposed that, because the Australian market is smaller and less liquid than the US market, fewer significant features can be extracted. This is in keeping with network capacity being proportionate to the size and liquidity of the market of interest. Utilising larger networks validated on larger markets results in masking out the key small regional market features due to the excess capacity of the networks. That is, the excess capacity masks out the other factors that clearly had a larger relative influence in the Australian market. As a result of this, a model developed on multiple regional markets would necessarily need a significantly deeper structure to extract the individual regional market factors.

The practical applications of this work appear in an examination of a simple trading strategy over the holdout test period using the Australian-validated LSTM. Following the methodology adopted by Ghoshal and Roberts [17], the strategy takes positions in those stocks for which the predicted probability on the day exceeds the centile threshold. We determine this centile threshold using the training set and each position is an equally weighted proportion of the portfolio value. The same range of transaction costs used by Ghoshal and Roberts [17] is implemented here, and the cumulative profit and CAGR results are presented in Table 7. There are some striking differences in the results shown in this table, and however, given the differences between the markets (such as regulation and the number of market participants), this variation is to be expected and within reasonable bounds. In addition, Ghoshal and Roberts [17] reported breakeven at a transaction cost of 0.35%, whilst the Australian model breaks even at 0.26%.

Table 2 The validated MLP architecture for the Australian market

Layer #	Layer type	Neurons	Dropout	Activation function	Output shape
1	Input	–	–	–	(20,4)
2	Flatten	–	–	–	(80)
3	Dense	32	–	ReLU	(32)
4	Dropout	–	0.3	–	(32)
5	Dense	32	–	ReLU	(32)
6	Dropout	–	0.3	–	(32)
7	Dense	2	–	Softmax	(2)

Table 3 The validated CNN architecture for the Australian market

Layer #	Layer type	Neurons	Dropout	Activation function	Kernel size	Output shape
1	Input	–	–	–	–	(20,4,1)
2	Conv2D	8	–	ReLU	(1,4)	(20,4,8)
3	Flatten	–	–	–	–	(640)
4	Dense	16	–	ReLU	–	(16)
5	Dropout	–	0.3	–	–	(16)
6	Dense	8	–	ReLU	–	(8)
7	Dropout	–	0.3	–	–	(8)
8	Dense	2	–	Softmax	–	(2)

Table 4 The validated LSTM/GRU architecture for the Australian market

Layer #	Layer type	Neurons	Dropout	Activation function	Output shape
1	Input	–	–	–	(20,4)
2	LSTM/GRU	8	–	Tanh	(8)
3	Dense	16	–	ReLU	(16)
4	Dropout	–	0.3	–	(16)
5	Dense	8	–	ReLU	(8)
6	Dropout	–	0.3	–	(8)
7	Dense	2	–	Softmax	(2)

Table 5 Summary of Ghoshal and Roberts' model results

Model	Precision	Recall	F-Score	AUC	*P* Value	Z-Score	Test accuracy (%)
MLP (US)	0.497	0.496	0.496	0.511	<0.0001	23.766	50.60
MLP (AU)	0.515	0.945	0.666	0.503	0.1246	1.152	51.46
1-Day CNN (US)	0.509	0.512	0.51	0.518	<0.0001	36.546	51.3
1-Day CNN (AU)	0.519	0.741	0.611	0.509	0.0003	3.405	51.50
2-Day CNN (US)	0.510	0.510	0.510	0.515	<0.0001	31.291	51.20
2-Day CNN (AU)	0.524	0.596	0.558	0.513	2.87E-07	5.000	51.52
3-Day CNN (US)	0.508	0.510	0.509	0.515	<0.0001	31.423	51.20
3-Day CNN (AU)	0.525	0.550	0.537	0.513	4.45E-07	4.915	51.37
LSTM (US)	0.506	0.510	0.508	0.510	<0.0001	19.616	50.80
LSTM (AU)	0.523	0.658	0.583	0.513	1.43E-07	5.132	51.71
GRU (US)	0.503	0.508	0.506	0.512	<0.0001	24.880	50.90
GRU (AU)	0.529	0.434	0.477	0.514	6.69E-08	5.274	51.16

Table 6 Summary of model results on Australian data

Model	Precision	Recall	F-score	AUC	*P* value	Z-Score	Test accuracy (%)	Model parameters
G &R MLP	0.515	0.945	0.666	0.503	0.1246	1.152	51.46	9,474
AU MLP	**0.531**	**0.525**	**0.528**	**0.518**	**1.73E-12**	**6.958**	**51.82**	**3,714**
G &R 1-Day CNN	0.519	0.741	0.611	0.509	0.0003	3.405	51.50	45,354
AU 1-Day CNN	**0.524**	**0.756**	**0.619**	**0.516**	**3.77E-10**	**6.154**	**52.23**	**10,450**
G &R LSTM	0.523	0.658	0.583	0.513	1.43E-07	5.132	51.71	5,282
AU LSTM	**0.535**	**0.542**	**0.538**	**0.523**	**0**	**8.789**	**52.33**	**714**
G &R GRU	0.529	0.434	0.477	0.514	6.69E-08	5.274	51.16	5,202
AU GRU	**0.535**	**0.504**	**0.519**	**0.521**	**1.11E-16**	**8.165**	**52.07**	**634**

Bold indicates new results produced in this study

Table 7 Comparison of cumulative profits (as a multiple of the starting balance) and compound annual growth rates (%)

Transaction costs	G &R US results		Australian Results	
	Profit	CAGR (%)	Profit	CAGR (%)
Frictionless	48.2	42.50	8.47	70.58
0.1% per transaction	34.1	38.20	3.73	38.94
0.25% per transaction	13	27.13	1.08	2.05

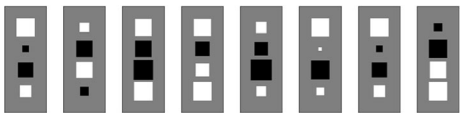

Fig. 1 Weight-space visualisation of convolutional layer (Hinton Diagrams) for the current study

Fig. 2 Hinton diagrams for Ghoshal and Roberts' study

The weight-space visualisations shown in Fig. 1 and 2 are known as Hinton diagrams. White and black squares indicate positive and negative values, respectively, while the size of the square indicates the magnitude of the value. Ghoshal and Roberts [17] experimented with a variety of filter sizes, 1, 2 and 3 day, and generated the associated Hinton diagrams. Figure 1 presents the Hinton diagrams from the 1-day CNN of the current study, which is notably different to the diagrams in Ghoshal and Roberts [17]) (see Fig. 2), and consequently demonstrates the difference in features extracted for each market. While an exact interpretation is known to contain a subjective element, there can be no doubt about the markedly different patterns produced. This represents further evidence that the models themselves have extracted inherently different features.

4 Conclusion and future research

This paper has extended an important recent US study on Australian data. Upon retraining the original US-validated architectures on Australian data, the results underperformed their earlier performance, suggesting that the US models could not exploit the regional specifics of an Australian market. In comparison, the newly validated Australian models significantly outperformed these original architectures. These results are attributable to the difference in microstructure factors across markets, which impact upon the selection of the final network architecture. Specifically, we find that the Australian-validated LSTM and CNN obtain the most significant results, although the LSTM achieves slightly superior results with 15 times fewer parameters. Given the outperformance of the Australian-validated models over those validated in the US, we propose that regional-specific models are required, that is, the model architectures need to be optimised for each market of interest. As such, a machine learning expert is still required to develop the network architecture as models developed on other markets cannot be effectively applied to a new market. This has implications for both practitioners and researchers in computational finance.

In addition, this study is part of an effort to overcome the lack of generalisability of results, which was an issue identified in a published survey and analysis of the relevant financial modelling literature. This is a notable contribution since many existing studies do not consider the effect that market-specific factors have upon model performance, that is, the transferability of results between markets. A simple trading strategy is developed, which produced above-market returns on the holdout test data. It is suggested that further work be completed to investigate the effect of slippage and other real-world considerations such as the validity of assessing returns close-to-close.

Future research should be conducted on additional regional markets to confirm the findings of this work. Additionally, an open research question is whether a model can be developed and trained across several regional markets with comparable accuracy to the models already developed. Such a model would undoubtedly require a very deep neural network to enable it to infer regional factors. Necessarily this would require additional independent variables, over and above the candlestick data, to be included in the training data for any proposed global model.

Acknowledgements The authors would like to thank the attendees of the Bond University Business School Research Seminar series for their valuable feedback when this work was presented in late 2020.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

1. Silver D et al (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484–489. https://doi.org/10.1038/nature16961
2. Buetti-Dinh A et al (2019) Deep neural networks outperform human expert's capacity in characterizing bioleaching bacterial biofilm composition. Biotechnol Rep. https://doi.org/10.1016/j.btre.2019.e00321
3. Litjens G et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005
4. Kumar A et al (2022) Generative adversarial network (GAN) and enhanced root mean square error (ERMSE): deep learning for stock price movement prediction. Multimed Tools Appl 81(3):3995–4013. https://doi.org/10.1007/s11042-021-11670-w
5. Wang X, Gong C, Khishe M, Mohammadi M, Rashid T (2021) Pulmonary diffuse airspace opacities diagnosis from chest x-ray images using deep convolutional neural networks fine-tuned by whale optimizer. Wireless Pers Commun. https://doi.org/10.1007/s11277-021-09410-2
6. Shrestha K et al (2021) A novel solution of an elastic net regularisation for dementia knowledge discovery using deep learning. J Exp Theor Artif Intell. https://doi.org/10.1080/0952813X.2021.1970237
7. Cha Y-J, Choi W, Büyüköztürk O (2017) Deep learning-based crack damage detection using convolutional neural networks. Comput Aided Civ Infrastruct Eng 32(5):361–378. https://doi.org/10.1111/mice.12263
8. Spencer BF Jr, Hoskere V, Narazaki Y (2019) Advances in computer vision-based civil infrastructure inspection and monitoring. Engineering 5(2):199–222. https://doi.org/10.1016/j.eng.2018.11.030
9. Szegedy C, et al (2014) Going deeper with convolutions. https://doi.org/10.48550/arxiv.1409.4842
10. Bloomberg LP (2021) Bloomberg world exchange market capitalization. Bloomberg Database
11. Vaswani A, et al (2017) Attention is all you need. https://doi.org/10.48550/arxiv.1706.03762
12. Goodfellow I et al (2014) Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K (eds) Advances in neural information processing systems, vol 27, pp 2672–2680
13. Vanstone B, Finnie G (2006) Combining technical analysis and neural networks in the Australian stockmarket. del Pobil AP (ed.), Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Soft Computing, ASC 2006, 125–130
14. Li H, Ng WWY, Lee JWT, Sun B, Yeung DS (2008) Quantitative study on candlestick pattern for Shenzhen Stock Market. IEEE (ed.), 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, 54–59
15. Vanstone B, Finnie G, Hahn T (2010) Stockmarket trading using fundamental variables and neural networks. Aust J Intell Inf Process Syst 11(1):41–47
16. Gabrielsson P, Johansson U (2015) High-frequency equity index futures trading using recurrent reinforcement learning with candlesticks. IEEE (ed.), 2015 IEEE Symposium Series on Computational Intelligence, pp 734–741
17. Ghoshal S, Roberts S (2020) Thresholded ConvNet ensembles: neural networks for technical forecasting. Neural Comput Appl 32:15249–15262
18. Krauss C, Do X, Huck N (2017) Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500. Eur J Oper Res 259(2):689–702. https://doi.org/10.1016/j.ejor.2016.10.031
19. Fischer T, Krauss C (2018) Deep learning with long short-term memory networks for financial market predictions. Eur J Oper Res 270(2):654–669. https://doi.org/10.1016/j.ejor.2017.11.054
20. Mishev K, Gjorgjevikj A, Vodenska I, Chitkushev L, Trajanov D (2020) Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE Access 8:131662–131682. https://doi.org/10.1109/ACCESS.2020.3009626
21. D'Amato V, Levantesi S, Piscopo G (2022) Deep learning in predicting cryptocurrency volatility. Phys A. https://doi.org/10.1016/j.physa.2022.127158
22. Li Y, Fu K, Zhao Y, Yang C (2022) How to make machine select stocks like fund managers? use scoring and screening model. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2022.116629
23. Obaid K, Pukthuanthong K (2022) A picture is worth a thousand words: measuring investor sentiment by combining machine learning and photos from news. J Financ Econ 144(1):273–297. https://doi.org/10.1016/j.jfineco.2021.06.002
24. Li Y, Pan Y (2022) A novel ensemble deep learning model for stock prediction based on stock prices and news. Int J Data Sci Anal 13(2):139–149. https://doi.org/10.1007/s41060-021-00279-9
25. Chollet F (2017) Deep learning with Python. Manning Publications, Shelter Island
26. Gudelek M, Boluk S, Ozbayoglu A (2018) A deep learning based stock trading model with 2-d cnn trend detection, Vol. 2018-January, 1–8
27. Hoseinzade E, Haratizadeh S (2019) Cnnpred: Cnn-based stock market prediction using a diverse set of variables. Expert Syst Appl 129:273–285. https://doi.org/10.1016/j.eswa.2019.03.029
28. Jearanaitanakij K, Passaya B (2019) Predicting short trend of stocks by using convolutional neural network and candlestick patterns. Proceedings of 2019 4th International conference on information technology: encompassing intelligent technology and innovation towards the new era of human life, InCIT 2019, pp 159–162. https://doi.org/10.1109/INCIT.2019.8912115
29. Birogul S, Temur G, Kose U (2020) Yolo object recognition algorithm and buy-sell decision model over 2d candlestick charts. IEEE Access 8:91894–91915. https://doi.org/10.1109/ACCESS.2020.2994282
30. Chen J-H, Tsai Y-C (2020) Encoding candlesticks as images for pattern classification using convolutional neural networks. Financ Innov. https://doi.org/10.1186/s40854-020-00187-0
31. Hinton G, Shallice T (1991) Lesioning an attractor network: investigations of acquired dyslexia. Psychol Rev 98(1):74–95. https://doi.org/10.1037/0033-295X.98.1.74
32. Bremner F, Gotts S, Denham D (1994) Hinton diagrams: viewing connection strengths in neural networks. Behav Res Methods Instrum Comput 26(2):215–218. https://doi.org/10.3758/BF03204624
33. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: 34th International conference on machine learning, ICML 2017, vol 7, pp 4844–4866
34. de Sá C (2019) Variance-based feature importance in neural networks. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in

bioinformatics) 11828 LNAI:306–315. https://doi.org/10.1007/978-3-030-33778-0_24
35. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
36. Cho K, van Merrienboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder–decoder approaches. https://doi.org/10.48550/arxiv.1409.1259
37. Nelson D, Pereira A, De Oliveira R (2017) Stock market's price movement prediction with LSTM neural networks. In: Proceedings of the international joint conference on neural networks, pp 1419–1426. https://doi.org/10.1109/IJCNN.2017.7966019
38. Matsumoto K, Makimoto N (2020) Time series prediction with lstm networks and its application to equity investment. Adv Stud Financ Technol Cryptocurr Mark. https://doi.org/10.1007/978-981-15-4498-9_4
39. Kim T, Kim H (2019) Forecasting stock prices with a feature fusion lstm-cnn model using different representations of the same data. PLoS ONE. https://doi.org/10.1371/journal.pone.0212320
40. Liu S, Zhang C, Ma J (2017) Cnn-lstm neural network model for quantitative strategy analysis in stock markets. Lect Notes Comput Sci 10635 LNCS:198–206. https://doi.org/10.1007/978-3-319-70096-0_21
41. Nison S (1991) Japanese candlestick charting techniques: a contemporary guide to the ancient investment techniques of the far east. Institute of Finance, New York. ISBN: 0139316507
42. Caginalp G, Laurent H (1998) The predictive power of price patterns. Appl Math Financ 5(3–4):181–205
43. Chen S, Bao S, Zhou Y (2016) The predictive power of Japanese candlestick charting in Chinese stock market. Phys A 457:148–165
44. Jasemi M, Kimiagari A, Memariani A (2011) A modern neural network model to do stock market timing on the basis of the ancient investment technique of Japanese candlestick. Exp Syst Appl. pp 3884–3890
45. Hu G, et al (2018) Deep Stock Representation Learning: From Candlestick Charts to Investment Decisions. IEEE (ed.), ICASSP, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, pp 2706–2710
46. Strader T, Rozycki J, Root T, Huang Y-HJ (2017) Machine learning stock market prediction studies: review and research directions. J Int Technol Inf Manag 28(4):63–83
47. Vanstone B, Hahn T (2017) Australian momentum: performance, capacity and the GFC effect. Account Financ 57(1):261–287. https://doi.org/10.1111/acfi.12140
48. Vanstone B, Gepp A, Harris G (2018) The effect of sentiment on stock price prediction. Lect Notes Comput Sci 10868 LNAI:551–559. https://doi.org/10.1007/978-3-319-92058-0_53
49. Liu G, Guo J (2019) Bidirectional lstm with attention mechanism and convolutional layer for text classification. Neurocomputing 337:325–338. https://doi.org/10.1016/j.neucom.2019.01.078
50. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. The MIT Press, Cambridge
51. Mason S, Graham N (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. Q J R Meteorol Soc 128(584 PART B):2145–2166. https://doi.org/10.1256/003590002320603584
52. Rasmussen C, Ghahramani Z (2001) Occam's Razor. In: Leen T, Dietterich T, Tresp V (eds) Advances in neural information processing systems, vol 13. MIT Press, Cambridge, MA, USA, pp 294–300