Implementing Deep Learning tools and/or techniques in predicting football results – Literature Review

Introduction

The rise of Artificial Intelligence (AI) has been influential in revolutionising various fields and how organisations work, particularly using deep learning techniques. As defined by McCarthy (2007), AI is about “The science and engineering of making intelligent machines, especially intelligent computer programs" and deep learning mimics the workings of the human brain, utilising a "layered model structure, often in the form of neural networks, and the associated end-to-end learning algorithms" (Deng 2018). These models consist of multiple layers, where each layer processes input data and extracts increasingly complex features. This hierarchical learning enables deep learning models to handle large amounts of unstructured data, such as images, audio, and text, with impressive accuracy. As they process more data, they refine their parameters, improving accuracy and robustness without human intervention. This makes deep learning highly effective for complex tasks like image recognition, natural language processing, and even predicting the outcomes in sports such as football.

The aim of this literature review is to critically examine existing research on the application of deep learning techniques in predicting football results and its broader impact on the sporting industry. This review will explore the various deep learning algorithms employed, such as neural networks, convolutional and recurrent neural networks, and how these models are tailored to the unique challenges of sports prediction. As well as, the datasets involved, including player statistics, match outcomes, historical data, and other relevant factors that influence predictive accuracy. By evaluating these elements, the review seeks to highlight the current state of research, the effectiveness of deep learning in this domain, and potential areas for future exploration.

History of football prediction analysis

The origins of sports analysis can be traced back to England, where the introduction of the ‘Betting and Gaming Act of 1960’ marked a pivotal moment in the sports betting industry. This legislation created a legal framework for bookmakers to operate, which in turn generated a need for accurate and reliable sports data, especially in football, to ensure that fair odds were set for bettors (UK Parliament 2020). The collection and analysis of football data became crucial for bookmakers, as it enabled them to assess team performance, player statistics, and other relevant factors that could influence match outcomes. This laid the groundwork for more systematic and analytical approaches to sports prediction.

The advancement of sports analysis gained significant momentum with the advent of the internet, which revolutionised the way football statistics were collected, shared, and analysed. The online betting sport market is estimated to generate $45.18bn in 2024 (Statista 2024). The availability of real-time data and the increased accessibility to comprehensive datasets created a surge in demand for football statistics, both from fans and professional analysts. As a result, prediction models became more sophisticated, integrating various data points beyond just the basic metrics. Many of the most popular prediction models used by bookmakers and analysts during this time relied heavily on a team’s current form, using recent match outcomes, goal differences, and individual player performances to forecast future results. However, these

models remained relatively simplistic compared to modern approaches, as they focused on immediate past performances without fully exploring the potential of deeper variables.

In a more advanced prediction model, Prasetio (2016) applied logistic regression techniques, incorporating a range of key variables beyond just team form. In this model, several factors were identified as critical to predicting football match outcomes. For instance, ball possession was used to measure the control a team had over the game, which often correlates with success. Home advantage was another significant predictor, reflecting the historical trend of teams performing better in familiar environments. Additionally, the distance travelled by the away team was considered, as greater travel distances could lead to fatigue and lower performance levels. These variables provided a more nuanced and comprehensive approach to prediction, illustrating the growing complexity of sports analysis in recent years.

Deep learning techniques in Football prediction

Training a deep learning model involves several essential steps to ensure accurate performance. First, a suitable dataset must be gathered, which can include player statistics, match results, or team performance metrics. This data then undergoes preprocessing to clean and normalise it, preparing it for training. Preprocessing may involve handling missing data, scaling values, or encoding categories. Once the data is ready, the model architecture is defined by specifying the type and number of layers, neurons per layer, and key parameters like activation functions. With the architecture set, the training data is fed into the model, allowing it to learn patterns and relationships. After this, the model's performance is evaluated using a separate validation set, and hyperparameters such as learning rate and batch size are adjusted to enhance accuracy and generalisation.

Deep learning models are distinguished by their ability to autonomously process large amounts of data and refine internal parameters, improving prediction accuracy without human intervention. These models rely on two datasets: the training set, which teaches the model, and the validation set, which tests how well the model generalises unseen data. The training process continuously fine-tunes parameters to minimise prediction errors.

For sports prediction, three common deep learning architectures are typically employed: Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). ANNs are basic neural networks suited for capturing relationships between input features in a wide range of tasks. CNNs are well-suited for spatial data, such as images, though they are also adapted for structured data in sports. RNNs, including Long Short-Term Memory (LSTM) networks, excel at handling sequential data, making them ideal for tasks involving time-series data, like player movements or match progression. Each architecture offers specific advantages depending on the prediction task and data type.

Ogunseye & Balogun (2019) implemented an Artificial Neural Network (ANN) to predict the outcomes of matches for the English Premier League team, Manchester United, with match data spanning over nine seasons, from 2009 to 2018. The ANN used in their study was a feedforward network, which consisted of 6 input layers, 5 hidden layers, and 2 output layers. The input layers represented key features, such as team performance metrics, match conditions, and historical data, which were fed into the network to enable prediction.

To train the network, the sigmoid activation function was used to normalize output and predict probabilities, such as win, loss, or draw. A forward pass processed input data to generate predictions, and an error function was calculated based on the difference between predicted

and actual results. Backward propagation then adjusted the hidden node weights, using gradient descent to minimize errors. Through repeated forward and backward passes, the ANN refined its parameters, improving accuracy and effectively recognizing patterns in Manchester United's match data for better predictions. Although, Ogunseye & Balogun (2019) admitted that more data was needed to continue to improve the model and its accuracy.

As mentioned by Hsu (2021) a CNN classifier is highly effective in sports prediction due to its ability to automatically recognise and learn implicit patterns within complex datasets. In sports, where performance metrics, player movements, and match dynamics involve intricate spatial and temporal relationships, CNNs can capture these underlying patterns without the need for manual feature extraction. For instance, in predicting football match outcomes, CNNs can analyse player positioning, ball movement, and team formations over time, identifying hidden correlations that contribute to a team's success or failure. This makes CNNs particularly valuable in sports analytics, where recognizing subtle, non-obvious patterns is key to improving prediction accuracy.

RNNs are also a popular method for sports prediction. Similar to ANNs, RNNs use layers of interconnected neurons to process inputs and produce outputs, employing backpropagation, activation functions, and loss functions to update weights. However, RNNs have the distinct advantage of being able to process sequential data through feedback loops, while ANNs are limited to static, feedforward connections. Awadallah & Khandelwal (2020) developed a Long Short-Term Memory (LSTM) model, a specialized type of RNN, which can store information over extended periods and handle long sequences. This capability is particularly useful in sports prediction, where the order and context of data points are critical. For example, instead of considering only the total number of wins, draws, and losses, an LSTM can capture the recent pattern of those outcomes, which may provide more meaningful insights. As a result, their LSTM model achieved an accuracy of 58.2% when predicting a win, draw, or loss.

Tools and Frameworks for Deep Learning in Football Prediction

Horvat and Job (2019) emphasize the importance of balancing the size and length of training sets in sports prediction. In their study of NBA games, they found that too little data led to underfitting, where the model fails to capture complexities, while too much data without proper training caused overfitting, making the model perform poorly on new data. Striking the right balance is essential for reliable predictions and avoiding naive machine learning approaches.

For managing large datasets, libraries like Pandas (Python) and dplyr (R) are widely used for data filtering and cleaning, crucial in sports analytics. These tools simplify working with complex datasets, ensuring that clean, relevant data is analysed efficiently. Raschka & Mirjalili (2019) detail advanced libraries such as scikit-learn and TensorFlow. Scikit-learn is ideal for traditional machine learning tasks, while TensorFlow handles deep learning models like CNNs and RNNs. However, they caution that deep learning models require significant computational resources, which users must consider, to avoid performance bottlenecks.

Overall, balancing data size, choosing the right tools, and managing computational demands are key to accurate sports predictions.

Challenges and Limitations

As mentioned earlier, when dealing with large amounts of data, there is an increased potential for unreliable or noisy data to arise. This necessitates additional time and effort to clean and

ensure the quality of the data so that deep learning models can achieve the highest possible prediction accuracy. Without this step, the model's output could be skewed, leading to inaccurate results. Data preprocessing, including handling missing values, removing outliers, and ensuring the data is correctly formatted, is crucial to avoid compromising the performance of the model.

Moreover, the trade-off between 'underfitting' and 'overfitting' in neural networks presents another significant challenge. A model that is too simple may fail to capture the complexity of the data (underfitting), while a model that is too complex may perform well on training data but poorly on unseen data (overfitting). Striking a balance requires selecting the right learning rate, model complexity, training duration, and regularisation methods. Solutions like cross-validation, early stopping, and dropout can help find the optimal model configuration (Jabbar & Khan 2015).

Another critical challenge in developing deep learning models for football prediction is the potential misuse of these models in gambling. Based on the Problem Gambling Severity Index (PGSI) scores, 2.8% of adults in the UK were identified as being at risk for or involved in problem gambling, with 0.3% specifically classified as problem gamblers (NHS Digital 2022). Ethical considerations must prioritise the welfare of individuals, ensuring that people do not bet beyond their means or use predictions irresponsibly. Additionally, deep learning models cannot predict unpredictable events like injuries or weather conditions. Therefore, predictions should always come with disclaimers that they are probabilistic and not guaranteed, encouraging responsible use of the information.

Future Direction and Conclusion

The future of deep learning in football prediction will likely adopt a mixed-method approach. While ANNs provide a straightforward way to input data and make predictions, advancements in handling unstructured data, such as images, suggest that CNNs could enhance predictions by analysing formations or possession maps to identify patterns (Hsu 2021). Combining these methods could lead to improved accuracy in predictions.

In conclusion, we have explored the history and demand for football prediction models, many of which rely on data from the English Premier League. Models using ANNs, CNNs, and RNNs have shown promising results, but further improvements are possible, especially as data types and analytical tools continue to evolve. Ethical considerations must also remain a priority for data scientists and statisticians, as these models become more sophisticated which are likely to put the welfare of individuals at more risk.

References

McCarthy, J. (2007) What is artificial intelligence? Available at: https://www-formal.stanford.edu/jmc/whatisai.pdf (Accessed: 13th September 2024).).

Deng, L (2018) Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]. IEEE Signal Processing Magazine, 35(1), pp.180-177.

UK Parliament (2020) The gambling industry: Select Committee on the Social and Economic Impact of the Gambling Industry. Available at: https://publications.parliament.uk/pa/ld5801/ldselect/ldgamb/79/7905.htm (Accessed: 13th September 2024).

Statista (2024) Revenue in the Online Sports Betting segment is projected to reach US$64.82bn by 2029. Available at: https://www.statista.com/outlook/amo/online-gambling/online-sports-betting/worldwide (Accessed: 13 September 2024).

Prasetio, D (2016) August. Predicting football match results with logistic regression. In 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA) (pp. 1-5). IEEE.

Horvat, T. and Job, J. (2019) Importance of the training dataset length in basketball game outcome prediction by using naïve classification machine learning methods. Elektrotehniški vestnik-Journal of Electrical Engineering and Computer Science, sv, 86, pp.197-202.

Ogunseye, A.A., Balogun, O. and Global, F.S.P. (2019) Artificial neural network approach to football score prediction. Journal of Artificial Intelligence, 1(1).

Hsu, Y.C. (2021) Using convolutional neural network and candlestick representation to predict sports match outcomes. *Applied Sciences*, *11*(14), p.6594.

Awadallah, A.A. and Khandelwal, R. (2020) Football Match Prediction using Deep Learning (Recurrent Neural Network).

Raschka, S. and Mirjalili, V. (2019) Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt publishing ltd.

Jabbar, H. and Khan, R.Z. (2015) Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices, 70(10.3850), pp.978-981.

NHS Digital (2022) 'Health Survey for England 2021: Part 2 – Gambling' Available at: https://digital.nhs.uk/data-and-information/publications/statistical/health-survey-for-england/2021-part-2/gambling (Accessed: 13 September 2024).