Proceedings of CONF-SEML 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA31282

A Review of the Applications and Frontier Progress of Deep Learning in Human Behavior Prediction: A Critical Examination

Chengan Tao

Minzu University of China, Beijing, China
3051485253@qq.com

Abstract. Human behavior correct modeling is very essential for next gen intelligent system, but it still has a lot of practical difficulties to be addressed even though it can deal with high dimensional data better than traditional Machine Learning (ML). Multimodal data fusion, loss of long-term temporal information, lack of interpretability. These are the problems. This paper looks into how deep learning is used in three main parts: feeling figuring out, social interaction patterns, and everyday actions watching. These findings reveal a distinct ket-set: structures like Long Short-Term Memories (LSTMs) and Convolutional Neural Netwroks (CNNs), they kasper tasks simple things really good, their ability add the exact randomness catching real societies is harder though. As well as that, Graph Neural Networks (GNNs) can model relationships but have computational scalability issues. Finally, this study recommended using Explainable Artificial Intelligence (XAI) along with privacy preserving method to solve the problem of "black box" and fill the gap between practical performance and reliable application of XAI.

Keywords: Deep learning, human behavior prediction, emotion recognition, social impact, interpretability

1. Introduction

1.1. Research background and challenges

Human behavior prediction is no longer a theory, rather a essential demand for Human-Computer Interaction (HCI). It's far reaching and it's applied over all the intelligent Healthcare Diagnostic Assistant and Pedestrian Trajectory Prediction for Autonomous Vehicle. But putting lab findings into use still encounters lots of difficulties. Although transformations of strengths of neural architectures, such as Transformer and Graph Neural Networks (GNNs) might exist, many problems still exist. These methods often fall short of achieving user-friendly results in practical applications.

First, data heterogeneity is a big problem. Real-world behavioral data is rarely clean, it is usually made up of unstructured text as well as some structured sensor data and noisier [1]. Second, it does not have much robustness to the environment. Autonomous driving will have movement as well, pedestrians will also be very unpredictable. Existing static models find it hard to keep up with such

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA31282

immediate changes in intent [2]. Finally, there is no prediction result as well. Current study is mostly about the behavior right now, it does not take into consideration any changes in habits over time [3].

1.2. Paradigm shift and existing problems

When dealing with big and high dimensional data, traditional machine learning algorithms hit a wall regarding their own performanc Deep learning is pushing this paradigm transformation forward and uses the automatic feature extraction of Recurrent Neural Network (RNN) and graph neural network (GNN) to find new unseen and hidden patterns.

Although there's lots of work that summarize these algorithms [4], but most of them have two big shortcomings:

Model-centric vs. Problem-centric: Most evaluations just list architectures, and they only compare CNNs with RNNs, ignoring interdisciplinary analysis of common difficulties like the similarity between data scarcity in Healthcare and Transportation.

Lack of critical thinking: There is no discussion on the "trust bottleneck." Although the accuracy is getting better, the "black box" nature of deep learning causes huge ethical problems such as the lack of interpretability and privacy [5].

1.3. Research objectives

To fill in these knowledge gaps, it does a thorough review of three major places in the deep learning world: feeling feelings, talking to others, and predicting daily behaviors. Instead of just looking at performance stats, we zero in on three strategic zones: (1) solid multimodal data recoiling methods, (2) the comparative strengths of the main build for managing spatiotemporal connections, and (3) new frontiers within explainable artificial intelligence (XAI). The goal of this review is to provide a path to developing highly precise prediction machinery which is necessarily reliable.

2. Scope and methods

2.1. Literature search strategy

To do a good review, this review searched several large scientific literature databases, including IEEE Xplore and ACM Digital Library, Web of Science and PubMed. Search time span was set as 2015-2025, only latest development was considered. Keywords like "deep learning," "human behavior prediction," "affective computing," and "explainable artificial intelligence" were included. In order to guarantee the quality of the researched articles, the research paid extra attention to articles from notable journals and top conferences.

2.2. Key research questions

This review focuses on two basic research issues (RQs):

RQ1: Heterogeneous Data Fusion Mechanisms. Given that people normally do things with more than just one kind, how can machine-made smart stuff use more than just some numbers with letters around them? This paper studies how to integrate unstructured data.

RQ2: Reshaping Interpretability (XAI). Deep learning has its own type of opacity (the 'black box' property of deep learning), how can interpretable AI (XAI) be applied for behaviour prediction? Techniques such as attention method can help make someone more certain about

DOI: 10.54254/2755-2721/2026.BA31282

important calls, especially for doctors, those controlling themselves, and people running their own show.

3. Theoretical foundations: data refinement and architecture

In order to build an accurate model, you need more than just great algorithms. Poeple have to fundamentally rethink their data representation. But deep learning is different than traditional analyses: here success is tied closely not to the quality of raw input data and feature density.

3.1. Heterogeneous data get and improve

Human behavioral data is complex and comes from all sorts of sources. To process a physical signal from the body is very different from a semantic text on a page.

Structured vs. Unstructured Data: As for data like structured data hearts, rate, GPS, coordinates can also be sent to direct network. However, just as they are also under the weather. On another hand, unstructured opinions like social media text is especially tough because it's very context sensitive [5]. For example, whether it is sarcasm, it requires deep semantic understanding and many people use transformers.

High-dimenisonal timeseries denosing: Data sparcity is still a bottleneck. Just filling in those missing sensor data always give us very horrible biases. So, most recent research uses GANs and autoencoders that can reconstruct missing time steps and reconstructs the inherent time steps [6]. More over end to end feature extraction by use of Convolutional Neural network (CNN) is also found successful in filtering environmental noice.

Action Embedding: The model has to understand the sequence so we need to map discrete actions into a continuous vector space. Action embedding with some of the ideas from Word2Vec can find potential relationships like the correlation between buy milk and buy bread, giving a math basis for sequence prediction [7,8].

3.2. Core architectures comparison

Architecture is definitely important. Table 1 comparison of three main architectures.

Table 1. Deep learning architecture comparative analysis in behavior prediction

Architecture	Core Mechanism	Advantages	Limitations	Typical Application Scenarios
RNNs / LSTMs	Gating Mechanisms & Temporal Memory	It's really good at grabbing short term temporal docks and it can deal with varying sized sequences too [9].	Long sequence has problem of gradient vanishing; Serial computation restricts training [10].	Action sequence completion, short-term trajectory prediction.
CNNs	Local Convolution & Pooling	Efficient extraction of local spatiotemporal features (e.g., gait patterns); high training speed [11].	Limited receptive field restricts global context understanding; lacks explicit modeling of temporal logic [12].	Pose recognition, video-based behavior classification.
GNNs	Message Passing & Graph Aggregation	Capable of processing non-Euclidean data; effective for modeling social interactions [13].	Highly sensitive to graph structure quality; computationally prohibitive for large-scale dynamic graphs [14].	Social network propagation, crowd flow analysis.

DOI: 10.54254/2755-2721/2026.BA31282

Analysis: As indicated, single architectures often fail to cope with complex scenarios. Therefore, the field is shifting toward Hybrid Architectures, combining CNNs for spatial feature extraction with LSTMs for temporal modeling.

4. Empirical applications of main areas

This part looks at whether deep learning models can do a good job on three things: feelings, socializing, and normal day-to-day things.

4.1. Emotion & affective state prediction

Affective computing is to understand the emotion from behavior. There's a clear shift going on from single modal to many modal fusion.

In text-based sentiment analysis, the revolution has been brought by transformers such as BERT. By using the Self-Attention mechanism, they can pick up on distant context much better than the first CNNs, making it way better at spotting tricky stuff like sarcasm [15].

But in the real world, poeple all express emotion in a variety of ways. Multimodal Fusion is a widely used method, however the current approach generally uses "Late Fusion," which can ignore Temporal Alignment. For instance, it's a before an expression of something really angry vocal cue. Simple weighted fusion can't pick up this kind of causal delay, resulting in late predictions during real-time chat [16].

4.2. Social interaction and group dynamics

Social behavior prediction extends to modeling influence within network.

As for information propagation, predicting whether someone will retweet or not is basically predicting an information cascade. Like DeepCas uses random walks for graph embedding. It does work, but it has trouble when it's ever-changing.

Group decision model must pay attention to predict the polarization trend. Existing GNNs can do well with static graphs, but they'll often fall apart really quickly when it comes to things like sudden stuff like rumors popping up. Thus, it could be that it is better at predicting a steady state distribution than predicting changes in non-linearLayout [17].

4.3. Daily activities and intent recognition

In smart environment, it is not just simple recognition now. Focus is on early predicting the Intent.

Smart home human activity recognition (HAR) has made progress. While CNN-LSTM hybrid model can get a very high accurate level on standardized actions, but it's difficult on interleaved activities like cook with telephone.

Autonomous driving, Pedestrian crossing intent prediction is very important: From research we can see that research using only visual skeleton extraction is not robust when there is occlusion. Use scene semantics in multi-stream network can improve accuracy but increase computation latency. It is a big trouble for safety-related systems every single second [18].

5. Challenges and ethical constraints as well as future trends

Performance can be raised but they will have a hard time to get promoted due to no promotion and privacy issues.

DOI: 10.54254/2755-2721/2026.BA31282

5.1. The interpretability crisis: breaking the "black box"

In the high stakes area of health care it is not acceptable that deep learning should be opaque.

Current XAI methods like SHAP explain after a decision is made. But now some people say it might not be a good match that these simplification can really show what the AI is thinking, so people will have wrong ideas about whether they can trust the computer [19]. The future is building ahead-of-time interpretable models. For example, some sparsity constraints could be imposed on the attention so that model will focus only important parts of the image, then doctors can look on what AI think about image and check whether its right or not.

5.2. Data privacy and federated learning

The behavioral data includes sensitive information, such as the path of the location and the health status.

Traditional centralized training requires data to be aggregated which will create a single point of failure. Federated Learning (FL) provides a solution through local model training. But FL also has to deal with Non-IID data in behavior prediction, because users' usage habits vary greatly and it is very hard for a global model to converge. Balancing personalization and global privacy, is still a main field of study [20].

5.3. Emerging trends: generative agents

The rise of generative AI brings about new things. Now researchers are using large language models-LLMs to create generative agents. It's possible to create high-fidelity social experiences inside a virtual world with LLMs that have both a memory as well as a character. This can also be seen as a way to help fix the problem not having a lot of information from social science studies [21].

6. Conclusion

This paper gives a systematic and critical analysis of deep learning in human behavior prediction. By adding up all the evidence about emotion recognition, social interactions and daily activities, this review present what is currently on the market: Architectures like RNNs, CNNs, and GNNs have their own pros, but they also run into major problems with multimodal asynchronous fusion, long-term temporal ties, and changing graph evolution, regardless of user.

The results show that it is no longer enough to merely pursue predictive accuracy. In order to be robust, flexible and easy to understand, it has become more important for the success of the model. Future research needs to move towards interpretable architectures rather than explanations after the fact, as well as optimizing architectures like federated learning which preserve privacy. Only after people reconcile technical advancement with ethical consideration, deep learning will be able to bring about real trustable intelligent systems.

References

[1] Xu, J., Arpan, L.M. and Chen, C. (2015) The moderating role of individual differences in responses to benefit and temporal framing of messages promoting residential energy saving. Journal of Environmental Psychology, 44, 95-108.

[2] Liu, L. and Wang, J. (2013) Case study on the implications of the medical library embedded librarian service. Res Libr Sci, 17, 95–98.

DOI: 10.54254/2755-2721/2026.BA31282

[3] Sajana, T. and Narasingarao, M.R. (2021) Human Behavior Prediction and Analysis Using Machine Learning-A Review. Turkish Journal of Computer and Mathematics Education, 12(5), 870–876.
[4] Alzubaidi, L., Zhang, J., Humaidi, A.J., et al. (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 8, 53.
[5] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2021) Deep learning-based text classification: A comprehensive review. ACM Computing Surveys (CSUR), 54(3), 1-40.
[6] Fang, C. and Wang, C. (2020) Time series imputation with generative adversarial networks. Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI).
[7] Casale, P., Pujol, O. and Radeva, P. (2012) Personalization and user verification in wearable systems using biometric walking patterns. Personal and Ubiquitous Computing, 16(5), 563-580.
[8] Wang, S., He, X., Cao, Y., Liu, M. and Chua, T.S. (2019) Neural Graph Collaborative Filtering. Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 165-174.
[9] Lindemann, B., Müller, T. and Jazdi, N. (2021) A survey on long short-term memory networks for time series prediction. Procedia CIRP, 99, 650-655.
[10] Al-Ghamdi, M. (2023) LSTM inefficiency in long-term dependencies regression problems. Semarak Ilmu, 5(1), 12-20.
[11] Turan, M. (2025) Comparison of CNN and LSTM networks on human intention prediction in physical human-robot interactions. Scholars' Mine-Masters Theses, Missouri S&T.
[12] Lundervold, A.S. and Lundervold, A. (2019) An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik, 29(2), 102-127.
[13] Chen, H. and Li, X. (2024) NAH-GNN: A graph-based framework for multi-behavior and high-hop interaction recommendation. PLOS One, 19(5), e0321419.
[14] Wirth, F., et al. (2024) Geometric Graph Neural Network Modeling of Human Interactions in Crowded Environments. arXiv preprint arXiv: 2410.17409.
[15] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171-4186.
[16] Poria, S., Cambria, E., Bajpai, R. and Hussain, A. (2017) A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125.
[17] Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A. and Vandergheynst, P. (2017) Geometric deep learning: going beyond euclidean data. IEEE Signal Processing Magazine, 34(4), 18-42.
[18] Rasouli, A. and Tsotsos, J.K. (2019) Autonomous vehicles that interact with pedestrians: A survey of theory and practice. IEEE Transactions on Intelligent Transportation Systems, 21(3), 900-918.
[19] Rudin, C. (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.
[20] Li, T., Sahu, A.K., Talwalkar, A. and Smith, V. (2020) Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.
[21] Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P. and Bernstein, M.S. (2023) Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1-22.