Review

A Comprehensive Literature Review on the Use of Restricted Boltzmann Machines and Deep Belief Networks for Human Action Recognition

Majid Joudaki *

Posted Date: 1 April 2025

doi: 10.20944/preprints202502.1119.v2

Keywords: Restricted Boltzmann Machines; deep belief networks; human action recognition; spatial–temporal feature extraction; unsupervised learning; hybrid architectures; video analysis; transfer learning; benchmark datasets; deep learning

 doi:10.20944/preprints202502.1119.v2

Review

A Comprehensive Literature Review on the Use of Restricted Boltzmann Machines and Deep Belief Networks for Human Action Recognition

Majid Joudaki

Department of Computer Engineering, Faculty of Engineering, Ayatollah Boroujerdi University, Boroujerd, Iran; m.joudaki@gmail.com

Abstract: This literature review provides a comprehensive synthesis of research on the use of Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs) for human action recognition (HAR) from 2012 to the present. The review begins by introducing the theoretical foundations of RBMs and DBNs, detailing their architectures, training algorithms (notably contrastive divergence), and various extensions—including convolutional and recurrent adaptations—that have been developed to better capture the spatial–temporal dynamics inherent in video data. Key contributions in the field are systematically analyzed, with emphasis on hybrid models that integrate RBM/DBN pretraining with modern deep learning techniques to enhance feature extraction and improve recognition accuracy. The review also examines the major benchmark datasets used in HAR research (such as KTH, HMDB51, UCF101, NTU RGB+D, and Kinetics), discussing preprocessing strategies, evaluation metrics, and the challenges associated with overfitting, computational complexity, and model interpretability. In addition, recent trends such as the incorporation of attention mechanisms, self-supervised learning, and multi-modal data fusion are explored. By highlighting both the historical significance and the evolving advancements of RBM/DBN methodologies, this review provides insights into the current state of HAR research and outlines promising directions for future investigation, including the integration of generative pretraining with emerging architectures for robust and efficient real-time action recognition.

Keywords: Restricted Boltzmann Machines; deep belief networks; human action recognition; spatial–temporal feature extraction; unsupervised learning; hybrid architectures; video analysis; transfer learning; benchmark datasets; deep learning

1. Introduction

Human Action Recognition (HAR) is the process of automatically identifying and classifying human activities from visual data such as video sequences or image streams. HAR has gained increasing importance in many applications—including surveillance, human–computer interaction, video indexing, and sports analytics—due to its potential to enhance security, improve user interfaces, and provide automated analysis in multimedia systems (Aggarwal & Ryoo, 2011; Weinland et al., 2011).

In recent years, deep learning methods have revolutionized HAR by enabling systems to learn hierarchical representations directly from raw data (Imani et al., 2025). Among the wide range of deep learning approaches, Restricted Boltzmann Machines (RBMs) and their extension to Deep Belief Networks (DBNs) have been employed for unsupervised feature learning and pretraining for subsequent classification tasks (Hinton & Salakhutdinov, 2006; Hinton, 2009). Although convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have become more dominant in many HAR applications, RBMs and DBNs continue to offer advantages—particularly in settings where unlabeled data are abundant and a generative model of the input is desired (Lee et al., 2009; Sutskever et al., 2009).

 doi:10.20944/preprints202502.1119.v2

The time frame of this review spans from 2012 to the present. During this period, significant advances have been made in adapting RBMs and DBNs for HAR, including improvements in training algorithms (such as contrastive divergence and its variants), modifications to model architectures (e.g., convolutional and recurrent extensions), and integration with other deep models to better capture spatio-temporal dynamics. This review systematically examines these contributions, the datasets used, challenges encountered, and the trends and opportunities for future research in the field.

In the sections that follow, we first introduce the key theoretical concepts behind RBMs and DBNs and then move on to review the literature on their applications in HAR. We provide a detailed synthesis of methodological advances, describe common datasets and benchmarks, and discuss limitations of these models when applied to HAR. Finally, we conclude by highlighting promising future research directions.

2. Background on RBMs and DBNs

This section reviews the theoretical underpinnings of Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs), which have played a seminal role in early deep learning research and continue to influence unsupervised and generative modeling approaches for HAR.

2.1. Restricted Boltzmann Machines

2.1.1. Architecture and Functioning

An RBM is a generative stochastic neural network that models a probability distribution over its inputs. It is composed of two layers: a visible layer $\mathbf{v} \in \mathbb{R}^d$(representing the observed data) and a hidden layer $\mathbf{h} \in \{0,1\}^p$ (representing latent features). Unlike standard Boltzmann machines, RBMs have no intra-layer connections, resulting in a bipartite structure where every visible unit is connected to every hidden unit through symmetric weights $\boldsymbol{W} \in \mathbb{R}^{d\times p}$ (Hinton, 2002; Fischer & Igel, 2012).

The energy function of an RBM is defined as (1):

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{b}^{\mathrm{T}}\mathbf{v} - \mathbf{c}^{\mathrm{T}}\mathbf{h} - \mathbf{v}^{\mathrm{T}}\boldsymbol{W}\mathbf{h}, \tag{1}$$

where **b** and **c** are the bias vectors for the visible and hidden units, respectively. The joint probability distribution over the visible and hidden units is given by the Boltzmann distribution as (2):

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\mathbf{exp}\big(-E(\mathbf{v},\mathbf{h})\big), \tag{2}$$

with partition function $Z = \sum_{\mathbf{v},\mathbf{h}} \exp(-E(\mathbf{v},\mathbf{h}))$.

A key advantage of the RBM as shown in (3) is that its conditional distributions factorize:

$$P\big(h_j = 1|\mathbf{v}\big) = \sigma\left(c_j + \sum_{i=1}^{d} w_{ij}v_i\right). P(v_i = 1|\mathbf{h}) = \sigma\left(b_j + \sum_{j=1}^{p} w_{ij}h_j\right). \tag{3}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic sigmoid function (Hinton, 2002). This conditional independence enables efficient Gibbs sampling for both training and inference (Tieleman, 2008).

2.1.2. Training: Contrastive Divergence

Training RBMs involves maximizing the likelihood of the observed data. However, exact maximum likelihood estimation is intractable because computing the partition function Z is exponentially complex. Hinton (2002) introduced the contrastive divergence (CD) algorithm, which approximates the gradient of the log-likelihood by running a short Gibbs chain (often a single step, CD-1). The gradient update for a weight w_{ij} is given by (4):

$$\Delta w_{ij} \propto \langle v_i h_j\rangle_{data} - \langle v_i h_j\rangle_{model}, \tag{4}$$

 doi:10.20944/preprints202502.1119.v2

where $\langle\cdot\rangle_{data}$ is the expectation computed over the data distribution and $\langle\cdot\rangle_{model}$ is the expectation with respect to the model's distribution (Hinton, 2002; Fischer & Igel, 2012). This approximation has been shown to work well in practice, even though it introduces a bias in the gradient estimate.

2.1.2. Extensions and Variants

Over the past decade, researchers have proposed numerous extensions to the standard RBM to better suit various data modalities and application domains. For example, Gaussian–binary RBMs have been developed for modeling real-valued data by replacing the binary visible units with Gaussian units (Hinton & Salakhutdinov, 2006). Variants such as conditional RBMs (CRBMs) incorporate temporal dependencies for sequential data, making them particularly relevant for HAR tasks where motion dynamics are critical (Taylor et al., 2007; Sutskever et al., 2009). In addition, convolutional Restricted Boltzmann Machines (Conv-RBMs) have emerged as a powerful extension of the traditional RBM. Conv-RBMs address the challenges associated with high-dimensional inputs, such as images and video frames, by leveraging the principle of convolutional filtering. In these models, instead of fully connecting each visible unit to every hidden unit, local receptive fields are defined and weights are shared across spatial locations. This architecture significantly reduces the number of parameters, preserves spatial invariance, and enables the model to effectively learn hierarchical representations of local features. A seminal work by Lee et al. (2009) demonstrated that convolutional deep belief networks—built upon stacked Conv-RBMs—can learn robust feature hierarchies from image data and achieve scalable unsupervised learning. This approach has since been widely adopted in various applications, including human action recognition, where capturing local spatial patterns is crucial for accurate classification.

2.2. Deep Belief Networks

2.2.1. Stacking RBMs

A Deep Belief Network (DBN) is constructed by stacking multiple RBMs in a layerwise fashion. The first RBM is trained on the raw input data, and then the hidden activations of this RBM are used as the input to the next RBM. This greedy, layerwise pretraining process helps initialize the network weights in a region of parameter space that is close to a good solution, which is then fine-tuned using supervised learning or further unsupervised learning (Hinton et al., 2006; Bengio et al., 2007).

2.2.2. Fine-Tuning and Feature Extraction

After pretraining the stack of RBMs, the DBN can be "unrolled" into a feed-forward neural network that is further fine-tuned by backpropagation to improve classification performance. In the context of HAR, the DBN can be used to extract high-level spatio-temporal features from video data. These features are subsequently fed into a classifier (often a softmax layer) to perform the final action recognition task (Hinton & Salakhutdinov, 2006; Zhang et al., 2014).

2.2.3. Advantages and Limitations

The unsupervised pretraining phase of DBNs helps to alleviate issues such as vanishing gradients and can improve generalization when labeled data are scarce. However, challenges remain in the fine-tuning process and in scaling the models to very large datasets. Despite these challenges, DBNs have been successfully applied in various domains, including HAR (Uddin & Kim, 2017; Abdellaoui & Douik, 2020).

3. Key Research Contributions (2012–Present)

This section provides a systematic review of influential works that have applied RBMs and DBNs to human action recognition since 2012. The focus is on methods that have contributed novel

 doi:10.20944/preprints202502.1119.v2

architectures, training strategies, or hybrid models aimed at improving recognition performance on challenging datasets.

3.1. Early Applications and Baseline Studies

In the early part of the decade, several studies extended the ideas of unsupervised pretraining via RBMs and DBNs to the HAR domain. Although some pioneering work (e.g., Taylor et al., 2007) predates 2012, these foundational methods set the stage for later developments.

3.1.1 Initial Demonstrations

Abdellaoui and Douik (2020) presented a novel HAR system that uses a DBN composed of a series of RBMs to extract spatio-temporal features from video sequences. Their approach uses DBN pretraining for feature reconstruction and classification, demonstrating promising recognition accuracy on benchmark datasets such as KTH and UIUC. The system also underscored the potential of generative models to provide additional insights by allowing data reconstruction and synthesis.

Another early study compared RBMs with standard matrix factorization techniques (e.g., Independent Component Analysis) for estimating intrinsic networks (INs) from video data. These studies highlighted that RBMs could capture nonlinear relationships between pixels (or voxels, in the case of fMRI) more effectively than linear methods, paving the way for their use in HAR tasks (Hinton & Salakhutdinov, 2006; Sutskever et al., 2009).

3.2. Hybrid and Convolutional Approaches

As deep learning research advanced, researchers began to explore hybrid models that integrate RBMs and DBNs with other architectures. These methods aimed to combine the generative power of RBMs/DBNs with the spatial feature extraction capabilities of convolutional neural networks (CNNs) or the temporal modeling strength of recurrent neural networks (RNNs).

3.2.1. Convolutional RBMs and DBNs

Convolutional Restricted Boltzmann Machines (ConvRBMs) extend the standard RBM by incorporating weight sharing and local connectivity, making them more suited to image data where spatial locality is important (Lee et al., 2009; Krizhevsky & Hinton, 2010). In the context of HAR, convolutional DBNs have been used to extract hierarchical features from video frames. For instance, Zhang et al. (2014) proposed a real-time HAR system that uses a modified DBN architecture with convolutional layers. Their approach significantly improved recognition accuracy on real-time video streams, demonstrating the utility of convolutional RBMs/DBNs in handling the spatial complexity inherent in human motion data. Recent research has also focused on developing hybrid architectures that address the dual challenge of capturing both spatial and temporal dynamics in video-based HAR.

3.2.2. Recurrent Extensions

Temporal dynamics are a critical aspect of HAR. To better model sequential data, several studies have integrated recurrent mechanisms with RBMs. The recurrent temporal RBM (RTRBM) (Sutskever et al., 2009) introduces connections that capture temporal dependencies, enabling the model to learn features that evolve over time. Uddin and Kim (2017) further extended these ideas by proposing a robust DBN-based approach that combines recurrent connections with deep belief pretraining to improve the robustness and generalization of HAR systems. These models have been shown to effectively capture the temporal coherence of human actions, improving performance on datasets where motion dynamics are subtle and complex.

3.3. Transfer Learning and Fine-Tuning Strategies

One of the key challenges in HAR is the limited availability of labeled training data relative to the complexity of human actions. Several works have investigated how unsupervised pretraining

 doi:10.20944/preprints202502.1119.v2

with RBMs/DBNs can serve as an effective form of transfer learning, initializing deep networks with good feature representations that are later fine-tuned on specific HAR tasks.

3.3.1. Pretraining Benefits

Hinton and Salakhutdinov (2006) originally demonstrated that unsupervised pretraining with RBMs/DBNs can help initialize weights in a way that facilitates subsequent supervised learning. This concept has been adopted by HAR researchers to address overfitting and to achieve higher recognition accuracy, particularly when labeled data are scarce (Bengio et al., 2007; Zhang et al., 2014). By learning robust feature representations from unlabeled video data, these models can generalize better when fine-tuned on smaller, task-specific datasets.

3.3.2. Fine-Tuning Techniques

Recent studies have explored different fine-tuning strategies to further optimize HAR performance. For instance, some researchers have experimented with freezing early layers of a pretrained DBN and only updating the top layers during supervised training, while others have opted for joint training of all layers using backpropagation (Rodenburg, 2021; jpj, 2021 on Cross Validated). The consensus emerging from these studies is that fine-tuning using the pretrained weights—rather than replacing them with randomly initialized parameters—can significantly reduce training time and improve classification accuracy.

3.4. Comparative Studies and Performance Evaluations

Several comparative studies have been conducted to benchmark RBM/DBN-based methods against other deep learning architectures (e.g., CNNs, RNNs) as well as traditional machine learning approaches. These studies have evaluated models on a variety of metrics, including recognition accuracy, computational efficiency, and robustness to noise.

3.4.1. Performance on Standard Datasets

For example, Ali and Wang (2014) compared a DBN-based HAR system with several state-of-the-art methods on the KTH and HMDB51 datasets. Their results indicated that the DBN approach achieved competitive recognition accuracy, particularly when combined with appropriate feature preprocessing and data augmentation techniques. Similarly, Uddin and Kim (2017) reported that their robust DBN approach outperformed traditional methods on the NTU RGB+D dataset, demonstrating the viability of deep generative pretraining for complex HAR tasks.

3.4.2. Advantages Over Linear Methods

A recurring theme in the literature is that RBMs and DBNs can capture nonlinear, high-dimensional patterns in the data that are often missed by linear factorization methods such as Principal Component Analysis (PCA) or Independent Component Analysis (ICA). Studies have shown that the latent features extracted by RBMs/DBNs are more invariant to variations in viewpoint and lighting, leading to improved recognition in real-world scenarios (Hinton & Salakhutdinov, 2006; Abdellaoui & Douik, 2020).

3.5. Summary of Key Contributions

In summary, the period from 2012 to the present has seen the following major contributions regarding RBMs and DBNs in HAR:

- Early demonstrations of DBN-based HAR systems that leverage unsupervised pretraining for effective feature extraction (Abdellaoui & Douik, 2020).
- Hybrid architectures, including convolutional and recurrent extensions, which address the spatial and temporal complexities of human actions (Zhang et al., 2014; Sutskever et al., 2009).

 doi:10.20944/preprints202502.1119.v2

- Transfer learning strategies that demonstrate the benefits of RBM/DBN pretraining for initializing deep networks when labeled data are limited (Hinton & Salakhutdinov, 2006; Uddin & Kim, 2017).
- Comparative studies that benchmark RBM/DBN-based approaches against both traditional linear methods and other deep learning models, illustrating their potential for robust and invariant feature learning (Ali & Wang, 2014).

4. Datasets and Benchmarks

An essential component of evaluating HAR systems is the selection of appropriate datasets and benchmarks. In this section, we review the most commonly used datasets in the literature and discuss preprocessing techniques, feature extraction methods, and performance metrics used in RBM/DBN-based HAR studies.

4.1. Commonly Used Datasets

The following datasets have been frequently used to evaluate HAR methods based on RBMs and DBNs:

4.1.1. KTH Action Dataset

The KTH dataset is one of the earliest and most widely used benchmarks in HAR. It consists of six classes of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed by 25 subjects in four different scenarios (indoor, outdoor, etc.) (Schuldt et al., 2004). Although relatively small, KTH provides a controlled environment for testing basic HAR algorithms.

4.1.2. HMDB51

HMDB51 is a larger and more challenging dataset containing 51 action classes with over 7,000 video clips collected from movies, YouTube, and other sources. The dataset includes significant variations in viewpoint, illumination, and background clutter, making it a rigorous testbed for advanced HAR systems (Kuehne et al., 2011).

4.1.3. UCF101

UCF101 is one of the largest publicly available action recognition datasets, containing 101 action classes and more than 13,000 video clips. The dataset covers a wide range of activities in unconstrained environments, providing a realistic benchmark for evaluating the generalization capability of HAR models (Soomro et al., 2012).

4.1.4. NTU RGB+D

The NTU RGB+D dataset is particularly noteworthy for its inclusion of multiple modalities (RGB video, depth maps, and skeleton data). With over 56,000 video samples across 60 action classes, this dataset has become a gold standard for evaluating HAR systems that integrate multi-modal data (Shahroudy et al., 2016).

4.1.5. Kinetics

The Kinetics dataset, introduced by Kay et al. (2017), contains hundreds of thousands of video clips covering several hundred action classes. Although more commonly used with CNN-based architectures, Kinetics has also been employed in comparative studies involving RBM/DBN-based approaches to assess scalability and performance in large-scale settings.

4.2. Preprocessing and Feature Extraction

Preprocessing steps in HAR often include the following:

 doi:10.20944/preprints202502.1119.v2

- Frame Extraction and Normalization: Video frames are typically extracted at a fixed frame rate, and pixel values are normalized to a common range.
- Optical Flow Computation: For methods that require motion information, optical flow may be computed to capture temporal dynamics.
- Dimensionality Reduction: Although RBMs/DBNs can operate on high-dimensional data, some studies incorporate PCA or other dimensionality reduction techniques to reduce computational complexity.
- Data Augmentation: Techniques such as random cropping, horizontal flipping, and rotation are commonly used to increase the diversity of training data and improve generalization.

RBM/DBN-based HAR systems typically use these preprocessing steps to provide consistent inputs. In many studies, the RBM or DBN itself is used for unsupervised feature extraction, learning latent representations that capture both spatial appearance and temporal dynamics (Abdellaoui & Douik, 2020; Uddin & Kim, 2017).

4.3. Performance Metrics and Evaluation Criteria

The performance of HAR systems is typically measured using:

- Recognition Accuracy: The percentage of correctly classified action instances.
- Precision, Recall, and F1-Score: Metrics that account for class imbalances, particularly important in datasets with many classes (e.g., HMDB51, UCF101).
- Confusion Matrices: Used to visualize misclassifications and identify which action classes are often confused.
- Computational Efficiency: Training time and inference speed are critical when deploying HAR systems in real-time applications.
- Robustness: Sensitivity to variations in illumination, viewpoint, occlusion, and noise is also evaluated, particularly in datasets like NTU RGB+D and Kinetics.

4.3.1. Comparative Table of Datasets

Table 1, summarizes key characteristics of commonly used HAR datasets:

Table 1. commonly used HAR datasets and their characteristics.

Dataset	No. of Classes	No. of Clips/Samples	Modality	Key Characteristics	Reference
KTH	6	~600	RGB Video	Controlled environment, limited background variations	Schuldt et al., 2004
HMDB51	51	~7,000	RGB Video	Real-world videos, significant variations in viewpoint and clutter	Kuehne et al., 2011
UCF101	101	>13,000	RGB Video	Unconstrained settings, diverse action categories	Soomro et al., 2012
NTU RGB+D	60	>56,000	RGB, Depth, Skeleton	Multi-modal data, large-scale, multiple viewpoints	Shahroudy et al., 2016
Kinetics	400+	300,000+	RGB Video	Large-scale, diverse set of actions, realistic conditions	Kay et al., 2017

5. Challenges and Limitations

While RBMs and DBNs have provided significant insights into unsupervised feature learning for HAR, several challenges and limitations must be addressed.

5.1. Overfitting and Generalization

One of the primary challenges in HAR is overfitting, particularly when training deep generative models on limited labeled data. Although unsupervised pretraining helps to alleviate overfitting by initializing the network in a favorable region of the parameter space, fine-tuning can still lead to overfitting on small datasets (Hinton & Salakhutdinov, 2006). Researchers have employed techniques

 doi:10.20944/preprints202502.1119.v2

such as dropout (Srivastava et al., 2014) and weight decay (Hinton, 2002) to mitigate these issues, yet the balance between model capacity and generalization remains delicate. imbalanced actions in action recognition datasets. In predictive analytics applications, both customer churn prediction and human action recognition (HAR) face challenges related to class imbalance in their respective datasets. Recent studies, such as Xie et al. (2023), explored the effectiveness of advanced machine learning algorithms like XGBoost and LightGBM, alongside upsampling techniques (e.g., SMOTE and ADASYN), to address such challenges. These insights can be applied to HAR, where imbalanced action occurrence in video datasets may benefit from similar approaches to upsampling or balancing during the preprocessing phase. Notably, the competitive performance of LightGBM, achieving excellent F1-scores and ROC AUC, offers promising avenues for exploring novel hybrid modeling strategies within HAR systems.

5.2. Computational Complexity

Training RBMs and DBNs—especially when extended to handle video data with spatial and temporal dimensions—can be computationally expensive. The use of contrastive divergence approximations helps to alleviate some of the computational burden, but training on large-scale datasets (e.g., UCF101, Kinetics) still demands significant computational resources and careful hyperparameter tuning (Tieleman, 2008). The need for specialized hardware, such as GPUs or TPUs, is common in this research area.

5.3. Convergence and Stability

The training process for RBMs using contrastive divergence is known to be sensitive to the choice of learning rate, momentum, and the number of Gibbs sampling steps. Convergence can be slow, and the models are prone to instabilities such as mode collapse or divergence if the hyperparameters are not carefully tuned (Fischer & Igel, 2012). Additionally, the unsupervised pretraining phase may not always produce representations that are optimal for the downstream HAR task, necessitating careful fine-tuning strategies (Rodenburg, 2021).

5.4. Comparison with Other Deep Learning Approaches

While RBMs and DBNs have demonstrated success in HAR, they have been largely eclipsed by more recent deep architectures, such as CNNs, RNNs, and Transformers, which are particularly effective in learning spatial and temporal representations. CNNs excel at capturing spatial hierarchies in image data, while RNNs (and their variants such as LSTMs and GRUs) are well suited for modeling temporal dependencies in video streams (Simonyan & Zisserman, 2014; Donahue et al., 2015). In many cases, the performance gap between DBN-based HAR systems and these newer architectures has narrowed or even reversed, especially when large-scale annotated datasets are available.

5.5. Dataset Bias and Domain Adaptation

HAR datasets often exhibit biases related to specific camera angles, backgrounds, or environmental conditions. Such biases can affect the performance of deep models and limit their generalization to real-world scenarios. Although RBMs and DBNs have the advantage of unsupervised pretraining—which can help in learning more generalizable features—domain adaptation remains an ongoing challenge, particularly when transferring models trained on one dataset to another (Wang et al., 2018).

5.6. Interpretability

While one of the appealing aspects of RBMs and DBNs is their ability to learn latent representations, the interpretability of these learned features can be limited. In HAR, it is often desirable not only to classify actions accurately but also to understand which features or patterns contribute to the decision. Although some studies have attempted to visualize the spatial and

 doi:10.20944/preprints202502.1119.v2

temporal filters learned by DBNs (e.g., by mapping hidden weights back to the input space), the black-box nature of these models remains a challenge compared to more interpretable methods.

5.7. Handling Data Imbalance: Insights from Customer Churn Prediction

While challenges such as overfitting and convergence have been extensively discussed, an additional critical issue in both HAR and other domains is class imbalance. In customer churn prediction studies, researchers have effectively addressed imbalance by employing oversampling techniques such as SMOTE and ADASYN [Imani et al., 2025; Imani, Mehdi, et al., 2024]. Similar to the imbalance observed in telecom customer datasets—where churn events occur less frequently compared to non-churn events—HAR datasets can also suffer from underrepresentation of certain action classes.

- Proposed Strategies:
    a. Application of Oversampling Techniques: The successful application of SMOTE, ADASYN, and even novel sampling methods like GNUS in customer churn prediction suggests that similar approaches could enhance HAR performance. By artificially balancing the dataset, the model's ability to generalize to infrequent but critical actions may be improved.
    b. Integration with Hyperparameter Optimization: In addition to oversampling, advanced hyperparameter tuning strategies have been demonstrated to improve model robustness in churn prediction [Imani & Arabnia, 2023]. A combined approach where sampling techniques are jointly optimized with model parameters may reduce training bias and enhance overall classification accuracy in HAR systems.

Integrating these strategies into HAR pipelines could provide a promising avenue for mitigating the adverse effects of data imbalance, potentially leading to more reliable recognition performance across diverse action categories.

6.6. Cross-Domain Insights: From Customer Churn to Action Recognition

Recent advances in machine learning for customer churn prediction offer valuable insights that can be cross-applied to HAR. Studies in the churn domain have focused on refining ensemble methods and optimizing hyperparameter settings alongside combined data sampling techniques [Imani, Joudaki, Beikmohamadi, & Arabnia, 2025; Imani, Mehdi, et al., 2024]. These methodologies address challenges such as class imbalance and overfitting—challenges that are similarly prevalent in HAR research.

- Key Observations and Recommendations:
    a. Ensemble Methods and Sampling: The comprehensive analyses comparing Random Forest, XGBoost, and advanced boosting techniques under varying imbalance levels demonstrate that ensemble learning, when combined with targeted oversampling, can substantially improve prediction performance. Translating these findings, HAR systems might benefit from experimenting with ensemble strategies that incorporate oversampling techniques to better capture the nuances of rarely occurring actions.
    b. Hyperparameter Tuning: The emphasis on systematic hyperparameter optimization in churn prediction, where models are fine-tuned to achieve stability and robustness, is directly applicable to HAR models, particularly in the context of training RBMs/DBNs. Optimizing learning rates, momentum parameters, and sampling strategies in tandem could yield more stable convergence and improved accuracy.
    c. Cross-Domain Methodological Synthesis: By integrating these cross-domain insights, future HAR studies can adopt a more holistic approach to model training. A systematic exploration of the interplay between data sampling and hyperparameter optimization—grounded in the methodologies proven in customer churn research—could lead to breakthroughs in addressing the limitations of current HAR systems.

This cross-domain perspective not only enriches the methodological toolkit available for HAR but also fosters a more unified understanding of common challenges in predictive analytics across different fields.

 doi:10.20944/preprints202502.1119.v2

6. Recent Advances and Trends

Recent work in HAR has explored novel ways to improve upon classical RBM/DBN architectures. The following subsections detail some of the major trends and emerging areas of research.

6.1. Hybrid Models: Integrating RBMs/DBNs with CNNs and RNNs

6.1.1. Convolutional Extensions

One promising direction has been the integration of convolutional architectures with RBMs and DBNs. Convolutional RBMs (ConvRBMs) exploit local connectivity and weight sharing to model spatial invariances, which are critical for video-based HAR (Lee et al., 2009; Krizhevsky & Hinton, 2010). These models have been used as building blocks for deep networks that learn hierarchical feature representations from raw pixel data. For example, Zhang et al. (2014) demonstrated that a convolutional DBN could extract robust spatio-temporal features that lead to high recognition accuracy on real-time video streams. For instance, Joudaki et al (2025) proposed a novel efficient hybrid technique that combines a two-dimensional convolutional restricted Boltzmann machine (2D Conv-RBM) with a long short-term memory (LSTM) network. In their work, the 2D Conv-RBM is utilized to efficiently extract local spatial features—such as edges, textures, and motion patterns—from individual video frames. These features are then sequentially processed by an LSTM, which effectively models temporal dependencies across frames.

6.1.2. Recurrent and Temporal Models

To address the inherently sequential nature of human actions, several researchers have proposed recurrent extensions of RBMs. The recurrent temporal RBM (RTRBM) (Sutskever et al., 2009) introduces time-dependent connections that enable the model to capture temporal dependencies in sequential data. Subsequent work has combined recurrent architectures with DBN pretraining to produce models that are both robust and capable of modeling complex temporal patterns in HAR data (Uddin & Kim, 2017). In other paper, a new recurrent method based on DBNs is proposed. In the proposed method, the ability to process and interpret two-dimensional video frames and understand the concept of time through recursive implementation is added to DBNs(Joudaki & Ebrahimpour, 2024).

6.2. Transfer Learning and Self-Supervised Learning

6.2.1. Transfer Learning via Unsupervised Pretraining

Unsupervised pretraining remains one of the most compelling advantages of RBMs and DBNs. In scenarios where labeled data are limited, pretraining on a large corpus of unlabeled video data allows the model to learn useful representations that can be fine-tuned for a specific HAR task. This transfer learning approach has been shown to reduce overfitting and improve generalization performance (Hinton & Salakhutdinov, 2006; Bengio et al., 2007).

6.2.2. Self-Supervised Learning

More recently, the field has seen a surge of interest in self-supervised learning methods that create proxy tasks to learn representations from unlabeled data. Although these methods are most commonly associated with CNNs and Transformers, the principles can be applied to generative models such as DBNs. For instance, by designing auxiliary tasks (e.g., temporal order verification or future frame prediction), RBM/DBN-based systems can be encouraged to learn more robust and temporally coherent features for HAR (Misra et al., 2016).

 doi:10.20944/preprints202502.1119.v2

6.3. Hardware Accelerations and Scalable Training

The increasing availability of high-performance computing resources, such as GPUs and TPUs, has greatly facilitated the training of deep generative models. Advances in parallel and distributed computing have enabled the training of larger and deeper RBMs/DBNs on large-scale HAR datasets (Tieleman, 2008). In addition, optimization techniques such as mini-batch training and adaptive learning rate methods (e.g., Adam, RMSProp) have been integrated with RBM/DBN training pipelines, further reducing training time and improving convergence (Kingma & Ba, 2014).

6.4. Interpretability and Visualization Techniques

Interpreting the latent representations learned by RBMs and DBNs has become an area of active research. Recent studies have proposed methods to visualize the "receptive fields" of hidden units by mapping the learned weights back to the input space. These visualizations not only provide insights into what the model has learned about human motion but also help in diagnosing model failures and guiding architectural improvements (Erhan et al., 2009; Zeiler & Fergus, 2014).

6.5. Comparative Evaluations and Benchmarking

There is an increasing trend toward rigorous comparative studies that benchmark RBM/DBN-based HAR systems against contemporary approaches. Such studies often evaluate models on multiple datasets under varying conditions to assess robustness, scalability, and computational efficiency. Recent work has emphasized the need for standardized evaluation protocols, as well as the importance of publicly available code and reproducible experiments (Uddin & Kim, 2017; Abdellaoui & Douik, 2020).

7. Future Directions

Based on the literature reviewed, several promising avenues for future research emerge in the context of RBMs and DBNs for HAR.

7.1. Enhancing Model Architectures

7.1.1. Hybrid and Multi-Modal Models

Future research could focus on further integrating RBMs/DBNs with other deep learning paradigms. For instance, hybrid models that combine the unsupervised generative power of RBMs/DBNs with the spatial invariance of CNNs or the temporal modeling capacity of RNNs may yield significant improvements in HAR accuracy. Additionally, exploring architectures that effectively fuse data from multiple modalities (e.g., RGB, depth, skeleton) could improve performance in real-world scenarios (Shahroudy et al., 2016).

7.1.2. Attention Mechanisms

Incorporating attention mechanisms into DBN architectures could allow the models to focus on the most relevant spatial regions or temporal segments of a video. Such an approach may enhance the interpretability and performance of HAR systems, especially in complex scenes with multiple overlapping actions (Vaswani et al., 2017).

7.2. Improved Training Strategies

7.2.1. Advanced Optimization Techniques

While contrastive divergence has been a workhorse for RBM training, newer optimization techniques and sampling methods (such as persistent contrastive divergence, parallel tempering, or score matching) could be further explored to improve convergence speed and stability. Combining

 doi:10.20944/preprints202502.1119.v2

these methods with adaptive learning rate algorithms may yield further improvements (Tieleman & Hinton, 2009).

7.2.2. Self-Supervised and Semi-Supervised Learning

As mentioned earlier, self-supervised learning has shown great promise in other domains and could be adapted to RBM/DBN-based HAR systems. Designing pretext tasks that exploit the temporal structure of video data (for example, predicting the order of frames or reconstructing missing segments) may help the model learn richer representations without extensive manual labeling (Misra et al., 2016). In a similar vein, semi-supervised learning techniques that leverage both labeled and unlabeled data could address the common problem of data scarcity in HAR.

7.3. Scalability and Real-Time Deployment

The practical deployment of HAR systems in real-world applications (such as surveillance or interactive systems) demands models that are both accurate and computationally efficient. Future research should focus on scaling RBM/DBN-based systems to larger datasets while optimizing their inference speed. Techniques such as model compression, pruning, and quantization could be adapted to RBM/DBN architectures to reduce their computational footprint without sacrificing performance (Han et al., 2016).

7.4. Interpretability and Explainability

Improving the interpretability of RBM/DBN-based HAR systems is another key research direction. Future work could explore methods to better visualize the learned features and understand the decision-making process of the network. Techniques such as saliency mapping, layer-wise relevance propagation, and concept activation vectors, which have been applied successfully to CNNs, might be extended to generative models like DBNs (Zeiler & Fergus, 2014; Montavon et al., 2018).

7.5. Robustness to Environmental Variations

HAR systems must operate reliably under diverse environmental conditions—varying lighting, occlusions, and background clutter. Future research should investigate how RBMs and DBNs can be made more robust to these factors. Approaches may include data augmentation strategies, domain adaptation techniques, and the incorporation of adversarial training methods to make the models less sensitive to noise and variation (Goodfellow et al., 2014).

7.6. Fusion with Other Modalities and Sensor Data

As wearable devices and sensor networks become more prevalent, there is growing interest in multi-modal HAR. Future studies could explore how RBM/DBN architectures can be extended to fuse information from visual, inertial, and physiological sensors to create more comprehensive models of human activity. Multimodal DBNs that jointly learn representations from these diverse data sources could lead to more accurate and robust HAR systems (Zhang et al., 2017).

8. Conclusions

This review has provided an in-depth analysis of the role of Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs) in human action recognition (HAR) from 2012 to the present. We began by outlining the fundamental concepts behind RBMs and DBNs, including their architectures, training methodologies (with an emphasis on contrastive divergence), and extensions such as convolutional and recurrent variants. Key research contributions were examined, highlighting early demonstrations of DBN-based HAR systems, the development of hybrid architectures that integrate convolutional and recurrent components, and the use of transfer learning

 doi:10.20944/preprints202502.1119.v2

strategies to address the challenges posed by limited labeled data. Comparative studies were discussed, showing that RBM/DBN approaches can achieve competitive recognition accuracy on standard benchmarks such as KTH, HMDB51, UCF101, NTU RGB+D, and Kinetics.

Despite these advances, several challenges remain. Overfitting, computational complexity, convergence instability, and limited interpretability are persistent issues that have spurred further research into optimization techniques and architectural modifications. Moreover, as deep learning research has increasingly favored CNNs, RNNs, and Transformer-based models for HAR, RBMs and DBNs now occupy a more specialized niche—particularly in scenarios where unsupervised learning or generative modeling is critical. Recent trends indicate that hybrid models integrating RBM/DBN components with modern deep learning architectures, self-supervised pretraining strategies, and attention mechanisms hold significant promise. Advances in hardware acceleration and scalable training methods further bolster the case for revisiting and refining these early generative models for HAR. Future research directions include enhancing model architectures for multi-modal fusion, improving training stability with advanced optimization techniques, and developing methods for better interpretability and robustness to environmental variations.

In conclusion, while RBMs and DBNs have been partly overshadowed by newer architectures in many HAR applications, they continue to offer unique advantages—particularly in the context of unsupervised feature learning and generative modeling. Their ability to learn rich, hierarchical representations from unlabeled data remains a valuable asset in scenarios where labeled data are limited or where interpretability of the learned features is desired. As research continues to evolve, integrating the strengths of RBM/DBN approaches with contemporary deep learning methods could lead to more robust, efficient, and interpretable HAR systems.

References

1. Abdellaoui, M. & Douik, A., 2020. Human Action Recognition in Video Sequences Using Deep Belief Networks. IIETA Journal, pp.37–44. doi:10.18280/ts.370105.
2. Aggarwal, J.K. & Ryoo, M.S., 2011. Human activity analysis: A review. ACM Computing Surveys, 43(3), pp.16:1–16:43. doi:10.1145/1922649.1922653
3. Ali, K.H. & Wang, T., 2014. Learning features for action recognition and identity with deep belief networks. In: Proceedings of the International Conference on Audio, Language and Image Processing, Shanghai, China, pp.129–132.
4. Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H., 2007. Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems. pp.153–160.
5. Donahue, J. et al., 2015. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2625–2634.
6. Erhan, D. et al., 2009. Visualizing higher-layer features of a deep network. University of Montreal Technical Report.
7. Fischer, A. & Igel, C., 2012. An introduction to restricted Boltzmann machines. In: Pattern Recognition, 47(1), pp.25–39.
8. Goodfellow, I., Bengio, Y. & Courville, A., 2014. Deep Learning. MIT Press.
9. Hinton, G.E., 2002. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8), pp.1771–1800.
10. Hinton, G.E. & Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786), pp.504–507. doi:10.1126/science.1127647.
11. Hinton, G.E., 2009. Deep belief networks. Scholarpedia, 4(5), pp.5947.
12. Imani, M., Ghaderpour, Z., Joudaki, M., & Beikmohammadi, A. (2024, April). The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction. In 2024 10th International Conference on Web Research (ICWR) (pp. 202-209). IEEE. , doi: 10.1109/ICWR61162.2024.10533320.

 doi:10.20944/preprints202502.1119.v2

13. Jayaraman, D. & Grauman, K., 2016. Slow and steady: Information propagation in video recognition. In: Proceedings of the European Conference on Computer Vision (ECCV).
14. Joudaki, M. Ebrahimpour Komleh, H. Introducing a New Architecture of Deep Belief Networks for Action Recognition in Videos. Journal of Machine Vision and Image Processing 2024, 11(1), 43-58.
15. Joudaki, Majid, Mehdi Imani, and Hamid R. Arabnia. "A New Efficient Hybrid Technique for Human Action Recognition Using 2D Conv-RBM and LSTM with Optimized Frame Selection." Technologies 13.2 (2025): 53.
16. Kay, W. et al., 2017. The Kinetics Human Action Video Dataset. arXiv preprint arXiv:1705.06950.
17. Kingma, D.P. & Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
18. Krizhevsky, A. & Hinton, G., 2010. Convolutional deep belief networks on CIFAR-10. Unpublished Manuscript.
19. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. & Serre, T., 2011. HMDB: A large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV), pp.2556–2563.
20. Lee, H., Grosse, R., Ranganath, R. & Ng, A.Y., 2009. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML), pp.609–616.
21. Misra, I., Zitnick, C.L. & Hebert, M., 2016. Shuffle and learn: Unsupervised learning using temporal order verification. In: European Conference on Computer Vision (ECCV), pp.527–544.
22. Montavon, G., Samek, W. & Müller, K.R., 2018. Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, pp.1–15.
23. Rodenburg, F., 2021. [Discussion on deep belief networks and transfer learning]. Cross Validated, May 1. Available at: https://stats.stackexchange.com/ [Accessed 1 February 2025].
24. Schuldt, C., Laptev, I. & Caputo, B., 2004. Recognizing human actions: a local SVM approach. In: 17th International Conference on Pattern Recognition, pp.32–36.
25. Shahroudy, A. et al., 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.1010–1019.
26. Simonyan, K. & Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp.568–576.
27. Soomro, S., Zamir, A.R. & Shah, M., 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
28. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), pp.1929–1958.
29. Sutskever, I., Hinton, G.E. & Taylor, G.W., 2009. The recurrent temporal restricted Boltzmann machine. In: Advances in Neural Information Processing Systems, pp.1601–1608.
30. Tieleman, T., 2008. Training restricted Boltzmann machines using approximations to the likelihood gradient. In: Proceedings of the 25th International Conference on Machine Learning (ICML), pp.1064–1071.
31. Uddin, M. & Kim, J., 2017. A robust approach for human activity recognition using 3-D body joint motion features with deep belief network. KSII Transactions on Internet & Information Systems, 11(2), pp.1118–1133. https://doi.org/10.3837/tiis.2017.02.028
32. Vaswani, A. et al., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, pp.5998–6008.
33. Wang, H., Nie, W. & Huang, G.B., 2018. Deep metric learning with robust attention for person re-identification. IEEE Transactions on Multimedia, 20(5), pp.1097–1109.
34. Weinland, D., Ronfard, R. & Boyer, E., 2011. A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding, 115(2), pp.224–241.
35. Welling, M., Rosen-Zvi, M. & Hinton, G.E., 2005. Exponential family harmoniums with an application to information retrieval. In: Advances in Neural Information Processing Systems, pp.1481–1488.
36. Zeiler, M.D. & Fergus, R., 2014. Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV), pp.818–833.

 doi:10.20944/preprints202502.1119.v2

37. Zhang, H., Zhou, F., Zhang, W., Yuan, X. & Chen, Z., 2014. Real-time action recognition based on a modified deep belief network model. In: IEEE International Conference on Information and Automation (ICInfA), pp.225–228.
38. Zhang, S., Zhao, X. & Li, Z., 2017. Multimodal human action recognition via fusing convolutional and recurrent features. IEEE Transactions on Cybernetics, 47(10), pp.3007–3020.
39. Imani, M., Joudaki, M., Beikmohamadi, A., & Arabnia, H. R. (2025). Customer Churn Prediction: A Review of Recent Advances, Trends, and Challenges in Conventional Machine Learning and Deep Learning. Preprints. https://doi.org/10.20944/preprints202503.1969.v1
40. Imani, Mehdi, Ali Beikmohammadi, and Hamid Reza Arabnia. "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels." Technologies 13.3 (2025): 88.
41. Imani, Mehdi, and Hamid Reza Arabnia. "Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis." Technologies 11.6 (2023): 167.