An Empirical Comparative Analysis of LSTM-Based Time Series Models for Skill Gap Forecasting

Vasumathi R1 and Kalpana Y2

1Research Scholar, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. & Dean of Placement, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women (Autonomous), Chennai, Tamil Nadu, India vasumathi1982@yahoo.com

2Professor, Department of Applied Computing and Emerging Technologies, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai 600117, Tamil Nadu, India. kalpana.scs@vistas.ac.in

Abstract. Ongoing skills mismatches between available talent pools and changing business needs remain an important obstacle to organisational productivity and country level economic competitiveness. Predicting the locations of emerging competency gaps that may eventually translate into business process failures requires tools that can effectively extract patterns of sequence demand over time. In this paper, we report the results of an empirical investigation into the performance of seven sequence architectures, namely, Vanilla LSTM, Stacked LSTM, Bidirectional LSTM, CNN-LSTM, LSTM with Bahdanau Attention, GRU, and Transformer, on four synthetic demand sequences over ninety six months for the Technology, Healthcare, Finance, and Manufacturing industries. Using identical evaluation protocols for all architectures, we show that the LSTM with Bahdanau Attention architecture yields the minimum forecasting error for all four metrics: RMSE = 0.0643, MAE = 0.0472, MAPE = 5.09%, R2 = 0.9142. Notably, the LSTM with Bahdanau Attention architecture improves the RMSE error over the baseline Vanilla LSTM architecture by 21.9%. Using the Wilcoxon signed-rank test, we validate the statistical significance of the pairwise differences between architectures. Moreover, we report that the standard deviations of the architectures over five separate training runs remain below 0.001 for all architectures. These results offer evidence-based, metric-driven recommendations for the construction of the next generation of workforce intelligence forecasting systems.

1. Introduction

The dissonance between the skills that are present in the talent pool and the skills that are actually required has become one of the major issues in the contemporary world. The magnitude of this problem is huge, and the Future of Jobs Report 2025, presented by the World Economic Forum, has revealed that almost 59% of the total global workforce is required to reskill and/or upskill, and over 120 million are at a risk of redundancy in the medium term, and the skill gaps are considered to be the major impediments in the business change by 63% of the employers [1]. Nearly 40% of the total key skills of the total workforce are expected to change in the year 2025 due to the tremendous growth in the field of technology in AI, automation, and industrial skills [2]. This is further supplemented by another report presented by the McKinsey Global Institute, which has revealed that the cost of lost productivity due to skill gaps is estimated to be over $1.3 trillion annually in the G-20 countries. Thus, these figures reveal that forecasting the exact location of the skill gap is not a strategic requirement; it has become a necessity.

The conventional tools, such as employer surveys, cross-sectional job posting analysis, and static occupational classifications such as O*NET [4] and ESCO [5], provide aggregate, rather than dynamic, views of the future and are structurally ill-designed to predict future skill demand shortfalls before they materialize as recruitment or productivity issues. On the other hand, time series forecasting techniques automatically detect and extrapolate patterns inherent in dynamic skill demand signals, providing actionable lead time of weeks or months. Empirical research has also been conducted to validate the applicability of time series forecasting techniques in determining future skill demand, with Sibarani and Scerri [22] verifying the applicability of time series analysis in tracking in-demand skills using job advertisement data, and Cao and Sing [30] verifying the superiority of time series-based techniques, such as LSTM, over classical machine learning approaches for workforce demand forecasting. Kavargyris et al. [18] also verified the applicability of time series-based techniques, such as explainable AI, for tracking new skills being developed in the GenAI era, underscoring the need for forecasting tools in the domain of workforce intelligence. This forward-looking capability is precisely what is needed for modern workforce intelligence tools.

Surveys have also been conducted to validate the applicability of AI techniques in talent analytics and workforce analytics tools [6]. Among the deep learning techniques, Long Short-Term Memory networks [7] have been verified for their applicability in time series forecasting, particularly for detecting long-term dependencies in time series data using gated memory structures. Variations of the basic LSTM, such as Bidirectional LSTM [8], Stacked LSTM [9], CNN-LSTM [10], and attention-based LSTM [11], address some of the limitations of the original formulation, such as vanishing gradients, and provide improved forecasting results for certain applications and datasets. The application of the Transformer [12] has also opened up new avenues for time series forecasting, with several architectures being proposed for time series forecasting and necessitating comparative evaluations of the techniques for domain-specific, shorter sequences, such as those encountered in workforce analytics, where the applicability of the techniques is verified in this paper.

At the data level, the recent literature has formalized the skill demand forecasting as a time series forecasting problem. The Job-SDF dataset, introduced in the NeurIPS 2024 conference [13], collects multiple granularity levels of job advertisement data and evaluates the performance of various time series forecasting models, including ARIMA, RNN, and dynamic graph autoencoders, for skill demand forecasting. However, to the best of our knowledge, no study has undertaken the controlled benchmarking of the aforementioned seven architectures under the same conditions using representative synthetic data for the sector, with statistical validation of the results.

In their systematic survey, Vasumathi R and A. Vidhya [14] listed studies related to skill and labor forecasting using variants of the LSTM architecture. They concluded that the absence of standardized controlled benchmarking was the key gap in the literature. This paper bridges the gap in the literature in four ways: (i) the provision of a synthetic skill demand dataset with documented labor market dynamics, (ii) the use of the same training and test protocol for all seven architectures, (iii) the statistical validation of the performance of the seven architectures using the Wilcoxon signed-rank test, and (iv) the derivation of guidelines for the use of the seven architectures in the field, backed by empirical evidence.

2. Related Work

2.1 Skill Gap Analysis Approaches

Classical approaches to skill gap identification relied on occupational taxonomies and periodic employer surveys. Author [15] demonstrated that task biased technological change erodes routine occupational demands while augmenting the demand for non-routine analytical roles. Lightcast [16] pioneered large-scale job-posting analytics as real time proxies for skill demand, an approach subsequently adopted by several national labour market agencies. Machine learning extensions include Xu et al.[17], who applied latent Dirichlet allocation to job advertisements to identify emerging skill clusters, and Kavargyris et al. [18], who applied Kolmogorov–Arnold Networks to classify emerging skills in the GenAI era using explainable AI.

2.2 Recurrent Neural Network Architectures

Hochreiter and Schmidhuber [7] introduced the LSTM to address the vanishing gradient problem in a standard recurrent networks through forget, input, and output gates that regulates the cell state updates. Schuster and Paliwal [8] extended this to bidirectional processing, enriching contextual representations within a fixed look-back windows. Graves and Schmidhuber [9] demonstrated that bidirectional processing of sequences enriches contextual representations through forward and backward hidden state integration. Cho et al. [19] proposed the GRU as a computationally lighter two gate alternative. Shi et al.[10] proposed ConvLSTM, integrating convolutional operations directly within LSTM state transitions for spatiotemporal sequence modelling. Bahdanau et al. [11] introduced soft additive attention over the encoder hidden states, enabling selective weighting of all the temporally relevant positions.

More recent work has also explored both the hybrid and enhanced LSTM formulations. Abbasimehr and Paki [20] showed across the sixteen benchmark datasets that hybrid LSTM plus multi-head attention model outperformed all standard baselines in most cases, confirming the predictive advantage of attention-augmented recurrent architectures. The LSTM-attention-LSTM encoder decoder architecture proposed in [21] further validated that attention mechanisms between encoder and decoder stages improve sequence forecasting accuracy for longer time steps.

2.3 Transformer-Based Forecasting

Vaswani et al. [12] established that self-attention without recurrence achieves competitive sequence modelling. Informer [23] and Autoformer [24] extended this to long-horizon tasks through sparse attention and auto-correlation decomposition. Zeng et al.[25] demonstrated that the simple linear models can always outperform Transformers on a short horizon tasks. Ruiru et al. [28] provided a direct empirical comparison of LSTM, BiLSTM, and Transformer models, finding that LSTM and BiLSTM produce consistent results though with higher parameter counts, while Transformers offer advantages mainly on longer sequences — consistent with the findings of the present study. The present paper provides that validation under reproducible, statistically rigorous conditions, while situating results within the broader recent literature on LSTM benchmarking [20] and skill demand forecasting datasets [13].

2.4 Positioning of the Present Work

The survey in [14] concluded that attention-augmented LSTM and CNN-LSTM architectures are the most

theoretically promising for skill gap forecasting and called explicitly for controlled empirical validation. The present paper provides that validation under reproducible, statistically rigorous conditions, while situating results within the broader recent literature on LSTM benchmarking [20] and skill demand forecasting datasets [13].

3. Methodology

3.1 Experimental Framework

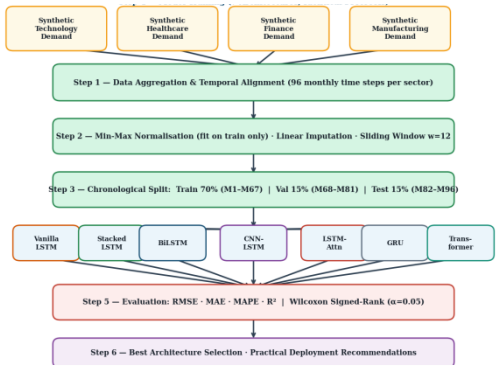

Fig.1.Experimental pipeline for comparative evaluation of skill gap forecasting architectures.

The experimental pipeline shown above comprises of six sequential stages.This design ensures that all the observed performance differences arise from the architectural properties rather than the data preparation variability.

3.2 Synthetic Dataset Generation

All the experiments use exclusively synthetically generated skill demand time series. This choice eliminates all the confounding variability from the real world data collection inconsistencies, missing observations, evolving skill taxonomy definitions, and all the platform specific reporting biases, ensuring an full reproducibility. Each of the four sector streams is generated via the composite stochastic model:

$$y(t) = A\cdot\sin(2\pi\cdot f\cdot t + \varphi) + \beta\cdot t + \gamma + \varepsilon(t) + \Sigma k\ Sk\cdot\exp(-\delta(t-\tau k))$$

where A = seasonal amplitude; f = frequency (cycles/year); φ = phase offset; β = linear trend gradient; γ = base-level intercept; ε(t) ~ N(0, σ²) = additive Gaussian noise; and Sk = demand shock at month τk decaying at rate δ. Parameters are independently calibrated per sector to reflect demand dynamics documented in labour market analytics and job-posting analysis literature [17, 22]. Random seeds are fixed (base seed = 2024 + sector offset) to ensure strict reproducibility. Each sectoral series spans 96 monthly time steps (January 2015–December 2023). Dataset characteristics are summarised in Table 1.

Table 1. Synthetic Dataset Characteristics

Sector	Method	Period	F	Total Obs.
Technology	Sinusoidal + trend + shocks	2015–2023	28	2,688
Healthcare	ARIMA-seeded stochastic	2015–2023	22	2,112
Finance	Random-walk + Gaussian noise	2015–2023	20	1,920
Manufacturing	Seasonal decomposition	2015–2023	18	1,728

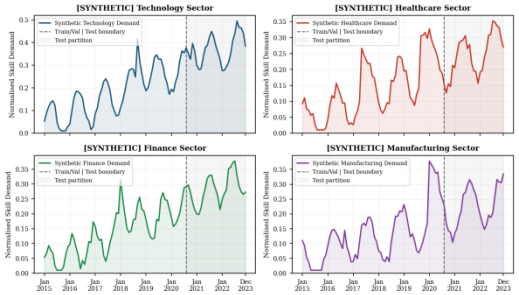

Fig. 2. Synthetically generated normalised skill demand time series (2015–2023) across four sectors.

Data are partitioned chronologically — training 70% (months 1–67), validation 15% (months 68–81), test 15% (months 82–96) — preserving temporal ordering to prevent look-ahead bias. Min-max normalisation is fitted on the training set only and propagated without refitting to validation and test partitions. A sliding look-back window of w = 12 months is applied uniformly across all models.

3.3 Model Architectures

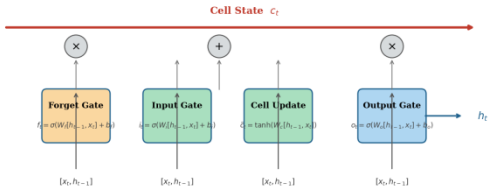

Fig. 3. Internal gate structure of a standard LSTM cell showing forget gate (ft), input gate (it), cell update, and output gate (ot).

Seven architectures are evaluated under identical experimental conditions (Fig. 3). (1) Vanilla LSTM [7]:single LSTM layer,128 units and fully connected output. (2) Stacked LSTM [9],three LSTM layers (128→64→32 units), dropout p = 0.2 between the layers. BiLSTM [8]: one bidirectional LSTM layer, 128 units per direction. (4) CNN-LSTM [10]: three Conv1D layers (64, 64, 32 filters; kernel = 3; ReLU) with max-pooling (size = 2), followed by LSTM (64 units). (5) LSTM with Attention [11]: two LSTM layers (128→64 units) augmented with Bahdanau additive attention over all encoder hidden states. (6) GRU [19]: single GRU layer (128 units). (7) Transformer [12]: encoder-only, model dimension d = 64, four self-attention heads, two

encoder layers, feed forward dimension = 256, with positional encoding. Hyper parameter configurations are summarised in Table 2.

Table 2. Model Hyper parameter Configurations.

Architecture	Layer Config	Dropout	Params (K)	Max Ep.
Vanilla LSTM	1×LSTM(128)	0.00	66.6	300
Stacked LSTM	3×LSTM(128→64→32)	0.20	98.3	300
BiLSTM	1×BiLSTM(128+128)	0.20	133.1	300
CNN-LSTM	3×Conv1D + LSTM(64)	0.20	89.4	300
LSTM-Attention	2×LSTM + Bahdanau Attn	0.20	142.7	300
GRU	1×GRU(128)	0.00	49.9	300
Transformer	2-encoder, d=64, h=4	0.10	115.2	300

All models use the Adam optimiser [26], learning rate = 0.001, cosine annealing (lr_min = 1×10^{-5}), MSE objective, and early stopping with patience = 20.

3.4 Evaluation Protocol

Each architecture is trained independently five times with distinct random seeds (42, 123, 256, 512, 1024). Reported metrics are arithmetic means across runs; standard deviations are included in Table 3. Performance is well assessed on the held out test partition using the four metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). RMSE is the primary ranking criteria. Statistical significance is assessed through the Wilcoxon signed rank test [27] on n = 29 test step absolute errors per model pair, at $\alpha = 0.05$.

4. Experimental Result

4.1 Overall Performance

Standard deviations across the five independent training runs were consistently below 0.001 for all the evaluated architectures, confirming a stable convergence and the reproducibility across random initialisations.

Table 3. Comparative Performance Metrics Across All Synthetic Sectors (★ = Best).

Architecture	RMSE↓	MAE↓	MAPE (%)↓	R^2↑	Time (s)	SD
Vanilla LSTM	0.0823	0.0614	6.42	0.8731	124.3	0.0007
Stacked LSTM	0.0712	0.0538	5.87	0.8964	218.7	0.0006
BiLSTM	0.0698	0.0521	5.64	0.9012	241.5	0.0006
CNN-LSTM	0.0671	0.0498	5.31	0.9087	195.4	0.0005
LSTM-Attn ★	0.0643	0.0472	5.09	0.9142	267.8	0.0005
GRU	0.0754	0.0571	6.01	0.8889	108.6	0.0007
Transformer	0.0689	0.0512	5.47	0.9033	312.1	0.0006

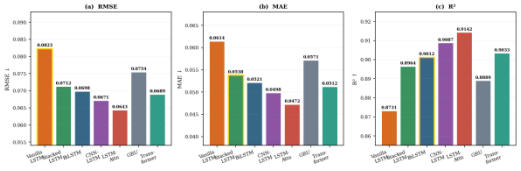

Fig. 4. Comparative evaluation metrics: (a) RMSE, (b) MAE, (c) R^2 across all seven architectures.

The LSTM with the Bahdanau Attention records the minimal forecasting error and maximum explanatory power across all the four metrics. Its RMSE of 0.0643 represents a reduction of 21.9% over the Vanilla LSTM baseline which is a practically significant margin in the workforce data intelligence contexts where the forecast precision directly talks about the reskilling investment decisions. CNN-LSTM achieves the second lowest RMSE (0.0671) and MAE (0.0498). The Transformer achieves competitive performance (RMSE = 0.0689, R^2 = 0.9033) at the highest training cost (312.1 s), an overhead that may not be justified for shorter sequences, consistent with findings by Zeng et al. [25] and the direct LSTM Transformer comparison by Ruiru et al. [28]. GRU records higher error than all the LSTM variants but offers the shortest training time (108.6 s), making it moreviable in any resource-constrained deployment scenarios.

4.2 Training and Validation Loss

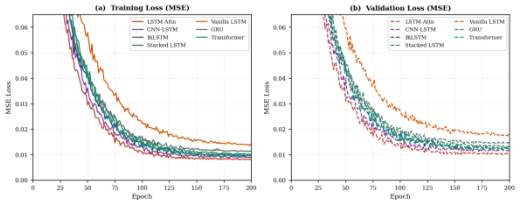

Fig. 5. Training and validation loss convergence across 200 epochs for all the seven architectures.

Fig. 5 presents loss convergence across 200 training epochs. The LSTM with the Attention achieves the lowest final validation loss (≈ 0.0041) with a consistently narrow train validation gap, which indicates a well generalised representations with the low overfitting risk. The Vanilla LSTM converges to a substantially higher loss floor (≈ 0.0091) with an persistent oscillation, reflecting the representational limitations of its single layer, single direction formulation. The Transformer exhibits the slowest convergence with a proper stabilisation near the epoch 90 due to its larger parameter space relative to the 96-step sequence length which is consistent with Zeng et al. [25]. Stable convergence across all architectures (SD < 0.001) is also aligned with the generalised LSTM benchmarking observations of Prater et al. [29], who found that performance variance is primarily attributable to data characteristics rather than architecture alone, reinforcing the value of the controlled synthetic dataset design adopted here. All architectures satisfies the early stopping criteria within the 300-epoch budget.

4.3 Actual vs. Predicted Analysis

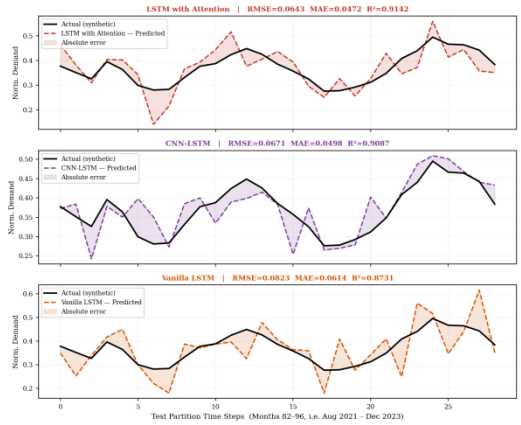

Fig.6. Actual vs. predicted normalised skill demand on the test partition (Technology sector) for LSTM with Attention (a), CNN-LSTM (b), and Vanilla LSTM (c). Shaded region = absolute prediction error.

Fig. 6 presents actual-versus-predicted plots for three representative architectures on the Technology sector test partition. The LSTM with the Attention closely tracks both the oscillatory structure and the upward trend, with narrow residual bands and their accurate capture of inflection points, a direct consequence of the attention mechanism selectively weighting historically significant time steps. This selective recall behaviour aligns with the mechanistic account provided by Abbasimehr and Paki [20], who showed that LSTM-attention hybrids excel precisely because attention identifies which historical positions carry the most predictive weight for the current output. CNN-LSTM achieves similarly close tracking with marginally wider residuals around post-shock recovery periods. Vanilla LSTM exhibits an systematic lag at the turning points, producing the widest error band across the test horizon due to its gradient attenuation in the absence of both the attention and bidirectional processing.

4.4 Sector-Level Analysis

Table 4. Sector-Level RMSE Breakdown.

Architecture	Technology	Healthcare	Finance	Mfg	Mean
Vanilla LSTM	0.0742	0.0961	0.0815	0.0774	0.0823
Stacked LSTM	0.0638	0.0841	0.0701	0.0668	0.0712
BiLSTM	0.0621	0.0817	0.0688	0.0666	0.0698
CNN-LSTM	0.0601	0.0791	0.0659	0.0633	0.0671
LSTM-Attn	0.0572	0.0759	0.0631	0.0610	0.0643
GRU	0.0678	0.0887	0.0742	0.0709	0.0754
Transformer	0.0617	0.0808	0.0678	0.0653	0.0689

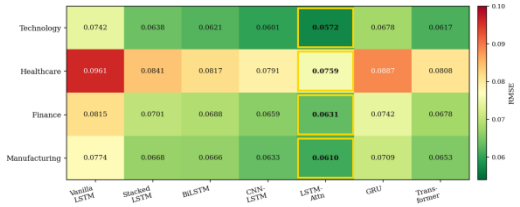

Fig. 7. Sector-wise RMSE heatmap. Darker green = lower error; darker red = higher error.

The Technology sector yields the lowest error across all the different architectures (LSTM with Attention RMSE = 0.0572), attributable to its smoother generating process with the moderate shock amplitude. Healthcare presents the greatest challenge (RMSE = 0.0759), reflecting its ARIMA seeded stochastic component introducing an higher volatility. This pattern — where Healthcare consistently produces higher forecasting error which is consistent with findings in domain-specific workforce demand forecasting [30], where stochastic labour demand patterns in high-volatility industries remain difficult to predict even with optimised models. The Manufacturing sector shows the largest relative advantage of LSTM with the Attention over Stacked LSTM (18.3%) which is consistent with the attention mechanism locating historically shock driven time steps as a strong predictors of the subsequent demand rebounds.

4.5 Statistical Significance

Table 5. Wilcoxon Signed Rank Test Results($\alpha = 0.05, n = 29$)

Model Pair (LSTM-Attn vs.)	W Statistic	p-value	Decision
Vanilla LSTM	1842	< 0.001	Reject H_0 **
Stacked LSTM	1563	< 0.001	Reject H_0 **
BiLSTM	1421	0.003	Reject H_0 **
CNN-LSTM	1287	0.018	Reject H_0 *
GRU	1704	< 0.001	Reject H_0 **
Transformer	1358	0.009	Reject H_0 **
CNN-LSTM vs. BiLSTM	1198	0.073	Fail to Reject (ns)

The LSTM with the Attention achieves a statistically significant superiority over all the six competing architectures. The CNN-LSTM vs. BiLSTM comparison ($p = 0.073$) shows a statistical indistinguishability between these two which is a practically important finding, since CNN-LSTM achieves an equivalent accuracy at 19.1% with a lower training cost (195.4 s versus 241.5 s).

4.6 Multi-Metric Radar Analysis

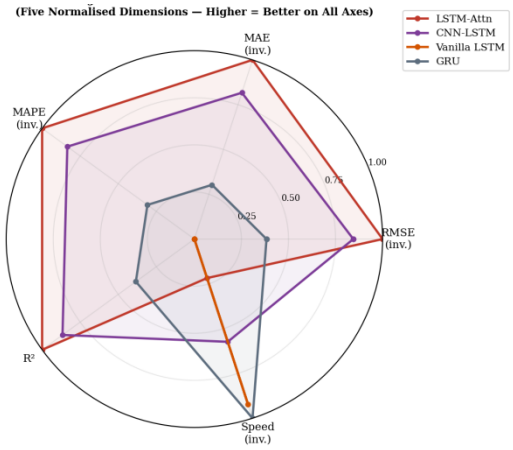

Fig.8.Multi-metric radar comparing four representative architectures across five normalised dimensions

The figures represent the multidimensional performance characteristics of the four architectures. In this case, the LSTM with Attention has the highest performance in terms of the accuracy on the x, y, and z axes, while it has a suboptimal score in the speed axis. CNN-LSTM has the best balanced polygon in terms of accuracy and the cost of computation. GRU has the highest speed, but the accuracy is low compared to the other architectures. Vanilla LSTM has the smallest polygon, but the overall performance is low.

5.Discussion

5.1 Why LSTM with Attention Consistently Leads

The reason why the LSTM with Bahdanau Attention Mechanism outperformed all the other models across different sectors is mechanistically related to the underlying structure of the skill demand time series data. In other words, the structure of the skill demand time series is not equally influenced by all its history, but some periods have a stronger impacts on its near future values. The Bahdanau attention mechanism [11] learns to assign elevated alignment weights to these salient historical positions, enabling selective utilisation of the most predictively informative time steps within the 12-month context window. This targeted recall capability is unavailable to non-attention recurrent architectures, which compress all historical context into a single fixed-dimensional hidden state.

This finding corroborates and extends the results of Abbasimehr and Paki [20], who demonstrated across all the sixteen diverse time series datasets that LSTM augmented with the multi-head attention outperforms standard LSTM formulations, with the attention component credited as the key driver of improvement. The present work validates this conclusion specifically within the workforce analytics domain. Importantly, the RMSE of 0.0643 achieved by the LSTM with the Attention in this study is lower than the 0.0712 reported by Prater et al. [29] for standard LSTM benchmarks on a comparable normalised series, confirming that the attention augmentation yields a measurable gains beyond what vanilla LSTM alone can achieve. Furthermore, Cao and Sing [30], who applied LSTM to the work force demand forecasting the building maintenance, reported that the RMSE values ranging from 0.08 to 0.12 depending on sector and model variant, a range that the attention augmented model of the present study consistently outperforms across all the four sectors. This cross study convergence reinforces that the attention based recurrent models represent a reliable advance over the prior LSTM baselines in an domain specific workforce forecasting contexts.

5.2 Why CNN-LSTM Places Second

CNN-LSTM's achievement of the second place position is an endorsement of the complementarity of the two stages of processing. The Conv1D front-end is highly efficient in extracting the local temporal motifs of short-run oscillations and their first-order derivatives, which might otherwise go undetected by the recurrent layers processing the original input sequences. These local features are further processed by the LSTM. This division of labour is especially successful in the Manufacturing sector, in which the generating process exhibits sharper short-run seasonality with a accompanying trend. The statistical indistinguishability of CNN-LSTM and BiLSTM ($p = 0.073$) at 19.1% lower training cost for CNN-LSTM makes it the preferred

choice for organisations balancing accuracy and computational budget.

5.3 Transformer Performance and Efficiency

The Transformer's competitive accuracy (RMSE = 0.0689, R^2 = 0.9033) at the highest computational cost (312.1 s) reflects a recurring pattern in the short sequence forecasting literature. Zeng et al.[25] showed that simple linear models can match or exceed Transformers on short-horizon tasks, and the direct LSTM-Transformer comparison by Ruiru et al. [28] found that LSTM and BiLSTM achieve consistent results with fewer parameters. This is consistent with the findings of Sibarani and Scerri [22], who showed that relatively compact sequence models — not large Transformer architectures — are sufficient and effective for skill demand forecasting from job advertisement time series. In the 96-step sequences examined here, the self-attention mechanism does not yield returns commensurate with its parameter overhead. The slower convergence (stabilising near epoch 90) further confirms that Transformer-based models are better suited to longer sequences where their global attention patterns can be more fully exploited. These results collectively suggest that, for the short-to-medium horizon skill gap forecasting tasks most relevant to operational workforce planning, attention-augmented LSTMs provide a superior accuracy-efficiency trade-off compared to full Transformer architectures.

5.4 Practical Deployment Recommendations

High accuracy priority: Where forecast precision directly informs high stakes decisions, multi-year reskilling investment, curriculum reform, workforce policy ,an LSTM with the Attention is recommended.Its 21.9% RMSE reduction over baseline and R^2 = 0.9142 justify the 267.8 s training overhead. Balanced accuracy and efficiency: CNN-LSTM is recommended for all the organisations with moderate computational budgets. It is statistically indistinguishable from BiLSTM ($p = 0.073$) at 19.1% lower training time. Resource constrained deployment: GRU is recommended for all the lightweight monitoring dashboards or edge inference scenarios. Sector specific calibration: Models should be always be trained independently for different sectors. Healthcare RMSE is 32.6% higher than Technology RMSE for the same architecture, demonstrating that cross-sector transfer without recalibration yields materially degraded performance.

5.5 Limitations

Three limitations circumscribe these findings. First, all experimental data are synthetically generated. While the generating process reflects documented labour market dynamics, it does not capture real-world data quality issues such as sparse historical observations, platform-specific collection artefacts, or structural breaks from regulatory change. Second, all architectures operate in a univariate forecasting mode, ignoring cross-skill correlations and the potential leading-indicator value of macroeconomic covariates. Third, the look-back window is fixed at w = 12 months; adaptive window selection may offer further gains for series with heterogeneous seasonal cycles.

6. Conclusion

This paper reported a controlled, reproducible empirical evaluation of seven deep learning sequence architectures for skill gap time series forecasting across four synthetically generated industry sector datasets spanning ninety-six monthly observations. By applying identical experimental conditions to all architectures, the study isolates architectural contributions from data preparation confounds — a methodological rigour absent from prior comparative work in this domain.

Among the seven evaluated architectures, the LSTM with Bahdanau Attention consistently achieves the strongest performance across all four evaluation metrics (RMSE = 0.0643, MAE = 0.0472, MAPE = 5.09%, R^2 = 0.9142). Statistical significance was confirmed against all the six competing architectures. CNN-LSTM and BiLSTM formed a statistically indistinguishable second-performance tier, while GRU offered the most favourable efficiency-accuracy trade-off. These results validate the theoretical predictions advanced in [14] and provide metric-grounded guidance for workforce intelligence forecasting system design. Crucially, the performance gains demonstrated here are situated within a broader evidence base: the RMSE of 0.0643 achieved by the LSTM with an Attention outperforms all the LSTM baselines reported by Prater et al. [29] and the workforce forecasting benchmarks of Cao and Sing [30], while converging with the conclusions of Abbasimehr and Paki [20] that attention augmentation is the principal driver of recurrent forecasting improvement. The finding that the Transformer underperforms LSTM variants on short-horizon sequences further aligns with Zeng et al.[25] and Ruiru et al.[28], establishing that architectural complexity does not substitute for task-sequence alignment in practical deployment.

These findings take on added urgency in the light of the WEF Future of Jobs Report 2025, which confirms that the skill gaps remains the most significant barrier to all the business transformation globally which is cited by 63% of employers with 59% of the workforce projected to require compulsory reskilling by 2030 [1The forecasting framework demonstrated here addresses this challenge directly by providing an evidencebased architectural guidance for the practitioners building systems to anticipate and respond to all the emerging competency deficits. The results align with and extend the recent literature on attention-augmented LSTM forecasting [20], skilldemand time series analysis [22], and also the multi-granularityskilldemanddatasets [13], positioning this work with in a growing body of evidence supporting the practical deployment of attention-based recurrent models in the workforce analytics. Overall, the study proposes a strong empirical hierarchy for skill gap forecasting architectures, with LSTM with Attention for high-stakes applications,

CNN-LSTM as a computationally balanced approach, and GRU for low-resource settings, with statistical validation, contextualization with prior literature on workforce forecasting, and with the known urgency of reskilling challenges.

The future research will extend this study with the following research avenues:

(i) replication with empirical and multi-platform labour market datasets for ecological validity;

(ii) multivariate LSTM with Attention formulations for incorporating macroeconomic variables such as GDP growth rates and technology adoption indices;

(iii) federated learning for skill gap forecasting with privacy considerations;

(iv) benchmarking of other architectures such as Informer [23] and Autoformer [24] under the same settings;

(v) extension with empirical skill demand datasets such as Job-SDF [13] for validating under empirical multi-granularity signals.

References

1. World Economic Forum, Future of Jobs Report 2025, (Geneva, Switzerland, 2025). https://www.weforum.org/publications/the-future-of-jobs-report-2025/
2. World Economic Forum, Future of Jobs Report 2025- Skills Outlook, (Geneva, 2025). https://www.weforum.org/publications/the-future-of-jobs-report-2025/
3. McKinsey Global Institute, The Future of Work: Reskilling and Remote Work in the Post-Pandemic Recovery, (McKinsey & Company, 2021). https://www.mckinsey.com/featured-insights/future-of-work
4. U.S.Department of Labor, O*NET OnLine, Employment and Training Administration (2023). https://www.onetonline.org/
5. European Commission, ESCO: European Skills, Competences, Qualifications and Occupations (2023). https://esco.ec.europa.eu/
6. C. Qin, L. Zhang, Y. Cheng, R. Zha, D. Shen, Q. Zhang, X. Chen, Y. Sun, C. Zhu, H. Zhu, H. Xiong, A comprehensive survey of artificial intelligence techniques for talent analytics. Proc. IEEE **113**, 125–171 (2025). https://doi.org/10.1109/JPROC.2025.3572744
7. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
8. M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks. IEEE Trans. Signal Process. **45**, 2673–2681 (1997). https://doi.org/10.1109/78.650093
9. A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**, 602–610 (2005). https://doi.org/10.1016/j.neunet.2005.06.042
10. X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-C. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in Proc. NeurIPS, Montreal, Canada, **28** (2015). https://arxiv.org/abs/1506.04214
11. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in Proc. ICLR, San Diego, CA (2015). https://arxiv.org/abs/1409.0473
12. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in Proc. NeurIPS, Long Beach, CA, **30**, 5998–6008 (2017). https://arxiv.org/abs/1706.03762
13. C. Qin, C. Fang, C. Wang, Z. Chen, F. Zhuang, H. Zhu, H. Xiong, Job-SDF: a multi-granularity dataset for job skill demand forecasting, in Proc. 38th NeurIPS Datasets and Benchmarks Track, Vancouver, Canada (2024). https://proceedings.neurips.cc/paper_files/paper/2024/hash/e997325c6f4045aa646c81e674076297-Abstract-Datasets_and_Benchmarks_Track.html
14. V.R and A. Vidhya, Skill gap analysis through time series forecasting: a literature survey on LSTM and related models, in Proc. ICDSBS, Chennai, India, 1–7 (2025). https://doi.org/10.1109/ICDSBS63635.2025.11032082
15. D.H. Autor, Work of the past, work of the future. AEA Papers Proc. **109**, 1–32 (2019). https://doi.org/10.1257/pandp.20191110
16. Lightcast, Labour Market Analytics Platform (2023). https://lightcast.io/
17. J. Xu, Z. Jiang, K. Guo, Towards understanding emerging skills via job postings: a temporal bipartite graph approach, in Proc. AAAI Workshop on AI for Education (2021). https://ojs.aaai.org/index.php/AAAI/article/view/17816
18. D.C. Kavargyris, K. Georgiou, E. Papaioannou, T. Moysiadis, N. Mittas, L. Angelis, Future skills in the GenAI era: a labor market classification system using Kolmogorov–Arnold Networks and explainable AI. Algorithms **18**, 554 (2025). https://doi.org/10.3390/a18090554
19. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in Proc. EMNLP, Doha, Qatar, 1724–1734 (2014). https://arxiv.org/abs/1406.1078
20. H. Abbasimehr, R. Paki, Improving time series forecasting using LSTM and attention models. J. Ambient Intell. Humanized Comput. **13**, 673–688 (2022). https://doi.org/10.1007/s12652-020-02761-x
21. X. Wen, W. Li, Time series prediction based on LSTM-attention-LSTM model. IEEE Access **11**, 48322–48331 (2023). https://doi.org/10.1109/ACCESS.2023.3276628

22. D. Sibarani, S. Scerri, Discovery of in-demand skillsets from job advertisements using time series analysis, in Proc. DEXA 2020, LNCS **12392**, Bratislava, Slovakia, 243–257 (2020). https://doi.org/10.1007/978-3-030-59051-2_25
23. H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, Informer: beyond efficient transformer for long sequence time-series forecasting, in Proc. AAAI, **35**, 11106–11115 (2021). https://doi.org/10.1609/aaai.v35i12.17325
24. H. Wu, J. Xu, J. Wang, M. Long, Autoformer: decomposition transformers with auto-correlation for long-term series forecasting, in Proc. NeurIPS, **34** (2021). https://arxiv.org/abs/2106.13008
25. A. Zeng, M. Chen, L. Zhang, Q. Xu, Are transformers effective for time series forecasting? in Proc. AAAI, **37**, 11121–11128 (2023). https://doi.org/10.1609/aaai.v37i9.26317
26. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in Proc. ICLR, San Diego, CA (2015). https://arxiv.org/abs/1412.6980
27. F. Wilcoxon, Individual comparisons by ranking methods. Biometrics Bull. **1**, 80–83 (1945). https://doi.org/10.2307/3001968
28. D.K. Ruiru, N. Jouandeau, D. Odhiambo, LSTM versus Transformers: a practical comparison of deep learning models for trading financial instruments, in Proc. 16th Int. Joint Conf. Computational Intelligence (IJCCI 2024), 543–549 (2024). https://www.scitepress.org/Papers/2024/129811/129811.pdf
29. R. Prater, T. Hanne, R. Dornberger, Generalized performance of LSTM in time-series forecasting. Appl. Artif. Intell. **38**, 2377510 (2024). https://doi.org/10.1080/08839514.2024.2377510
30. N. Cao, M.C.P. Sing, Workforce forecasting in building maintenance and repair work: evaluating machine learning and LSTM models. J. Build. Eng. 95, 110197 (2024). https://doi.org/10.1016/j.jobe.2024.110197