ORIGINAL RESEARCH **Open access**

From LSTM to Transformers: A Perspective on Evolving Deep Learning Architectures for Acute Ischemic Stroke Prediction

Fatima Zahra Amrani[1*], Youssef Benali[1]

Abstract

Acute ischemic stroke prediction from electronic health record time series data holds significant potential for enabling early intervention and reducing long-term disability. LSTMs have been widely used to model clinical sequences such as vital signs and laboratory trends, showing strong performance in stroke-related prediction tasks from 2018–2022. However, their sequential nature limits scalability and long-range dependency modeling in large EHR datasets. Transformers, despite transforming sequence modeling in other domains since 2017, remain underused in stroke prediction compared to LSTMs. Although early healthcare studies suggest potential benefits of attention-based models, robust validation in acute ischemic stroke contexts is still limited. Transformers offer advantages in parallel processing, long-range dependency modeling, and interpretability, but require more data and computational resources. They are likely to complement rather than replace LSTMs, with hybrid architectures providing a balanced solution for clinical time series analysis. Key themes include long-range dependency capture, parallel computation, interpretability, and data efficiency trade-offs between LSTMs and transformers. Hybrid LSTM–transformer models may offer improved performance and practicality for stroke prediction, with model selection depending on data scale and clinical constraints. Further benchmarking is needed to determine when transformers or hybrid models outperform LSTMs, guiding the development of more effective stroke prediction systems.

Keywords Attention mechanisms, Acute ischemic stroke prediction, LSTM networks, Transformer architectures, Clinical time series, EHR-based modeling

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Artificial Intelligence in Healthcare, Mohammed V University, Rabat, Morocco

Introduction

Acute ischemic stroke remains a leading cause of death and disability worldwide, underscoring the urgent need for advanced predictive tools. Early identification of risk through electronic health record data encompassing vital signs, laboratory trends, and neurological assessments could facilitate targeted preventive interventions and improve patient outcomes substantially. Deep learning models applied to these time series have shown initial promise in forecasting ischemic events and associated complications. However, the complexity of clinical sequences demands architectures that can robustly capture evolving patient states over time. As we look ahead, refining these models will be essential for integrating AI into routine stroke care pathways.

Deep learning approaches for clinical time series analysis have been dominated by long short-term memory networks since the mid-2010s. LSTMs process sequential EHR inputs step-by-step while maintaining a hidden state that effectively captures temporal dependencies in patient data. This dominance is evident in numerous applications to stroke outcome prediction and related time series

forecasting tasks. Their recurrent nature has made them a natural fit for variable-length medical records. Yet emerging alternatives are challenging this status quo as datasets scale.

Transformers, introduced in 2017, have revolutionized natural language processing by supplanting recurrence with self-attention mechanisms that enable parallel sequence processing. They can directly model relationships between any pair of time points in a sequence, offering advantages for complex clinical trajectories. However, adoption in clinical time series modeling has progressed more gradually due to domain-specific challenges such as irregular sampling and data sparsity. This slower uptake reflects the need for careful adaptation to healthcare constraints. Nevertheless, the potential benefits warrant closer examination in stroke prediction contexts.

This perspective examines the transition from LSTMs to transformers for acute ischemic stroke prediction, comparing key architectural trade-offs including computational efficiency, data requirements, interpretability, and long-range dependency handling. We argue that transformers are not universally superior to LSTMs, with the optimal choice hinging on clinical context, data availability, and deployment constraints. A roadmap for this analysis includes detailed reviews of LSTM successes, transformer advantages, trade-off evaluations, and context-specific recommendations. By synthesizing evidence from recent studies, we provide forward-looking guidance for researchers and clinicians alike. Ultimately, this examination aims to inform the next generation of deep learning architectures tailored to stroke prevention and management.

Figure 1 illustrates the architectural evolution from sequential LSTM processing to parallel transformer models and their integration in hybrid designs for stroke prediction.

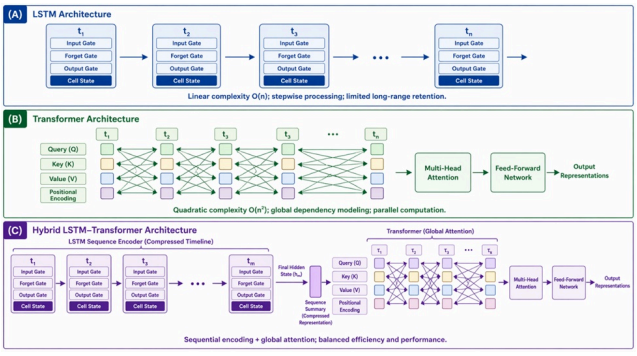

Figure 1. Evolutionary Architecture Pathway for Acute Ischemic Stroke Prediction: From Sequential LSTM Processing to Parallel Transformer and Hybrid Models

LSTMs for Stroke Prediction

How LSTMs work for clinical time series

Long short-term memory networks operate through sequential processing of clinical time series, updating a hidden state at each time step to retain relevant information from prior observations. The forget gate determines what information to discard from the previous cell state, while the input gate decides what new information to store, enabling selective memory retention in EHR data streams. An output gate then regulates the flow of information to the hidden state for prediction tasks such as stroke risk assessment. This gated architecture allows LSTMs to handle variable-length sequences effectively, accommodating patients with differing monitoring durations in acute care settings. In practice, these mechanisms have proven adept at modeling trends in vital signs and laboratory values over hours or days.

By maintaining a cell state that propagates information across time steps, LSTMs mitigate some issues of vanishing gradients compared to vanilla recurrent networks. This design facilitates the capture of both short-term fluctuations and longer-term patterns in stroke-related clinical data. For instance, sequential updates allow the model to integrate neurological assessments with real-time physiological signals without requiring fixed input lengths. Variable-length handling is particularly valuable for EHR-based predictions where patient trajectories vary widely. Overall, this step-by-step processing has formed the foundation for many successful clinical forecasting applications to date.

Successes and limitations

LSTM-based models have achieved notable successes in acute ischemic stroke prediction, as evidenced by applications that forecast outcomes from multimodal EHR sequences. These studies highlight the ability of LSTMs to integrate time-varying features like blood pressure trends and lab results for improved risk stratification. Their performance in registry-based cohorts has supported early adoption in predictive analytics for stroke care. However,

the reliance on sequential computation has occasionally constrained scalability in large-scale deployments. Despite these advances, the field recognizes room for enhancement in handling extended temporal contexts.

A primary limitation of LSTMs lies in their susceptibility to vanishing gradients over very long sequences typical of extended ICU monitoring for stroke patients. Sequential processing inherently limits parallelization during training and inference, leading to longer computation times as sequence lengths increase. This bottleneck becomes pronounced when modeling multi-day vital sign histories or longitudinal outpatient data. Consequently, LSTMs may underperform in scenarios requiring rapid processing of comprehensive patient timelines. Addressing these constraints is vital for advancing beyond current capabilities in clinical prediction.

Why LSTMs have been dominant

LSTMs have maintained dominance in stroke prediction due to their maturity and well-understood training dynamics established over years of research in healthcare time series. Their behavior is relatively predictable, with gated mechanisms providing stability even on noisy clinical data. Moderate data requirements further contribute to their practicality in settings with limited annotated stroke cases from specialized centers. Ease of implementation using standard frameworks has accelerated experimentation and iteration by interdisciplinary teams. This accessibility has fostered widespread use in acute ischemic stroke modeling pipelines.

The interpretability of hidden states in LSTMs, though not perfect, offers insights into learned temporal patterns that align with clinical intuition. Training stability on smaller datasets typical of medical applications reduces overfitting risks compared to more parameter-heavy alternatives. As a result, LSTMs have served as a reliable baseline for benchmarking newer architectures in stroke-related tasks. Their proven efficacy in similar domains like cardiac event prediction reinforces confidence in their utility. Looking forward, while dominant today, their role may evolve as more efficient alternatives mature.

The Rise of Transformers

Self-attention mechanism

The self-attention mechanism at the core of transformers computes relationships between all pairs of elements in a sequence using query, key, and value vectors derived from input embeddings. Scaled dot-product attention normalizes these interactions to prevent vanishing gradients and stabilize training across long clinical time series. This allows every time step in EHR data to attend directly to others without sequential bottlenecks. Multi-head attention further enhances representation power by capturing diverse dependency patterns in parallel subspaces. Such mechanisms enable comprehensive modeling of interactions in stroke risk factors spanning multiple time points.

Following attention layers, position-wise feedforward networks apply transformations independently to each position, contributing to the transformer's expressive capacity for complex healthcare predictions. Positional encodings are incorporated to retain temporal order information critical for time series analysis. In clinical applications, this architecture supports efficient processing of irregular vital sign sequences for ischemic stroke forecasting. The overall design promotes scalability that LSTMs struggle to match on extended inputs. These components collectively position transformers as a formidable alternative for sequence modeling in medicine.

Advantages over LSTMs

Transformers provide significant advantages over LSTMs through full parallel processing of input sequences, dramatically accelerating training and inference on modern hardware. This capability eliminates the sequential dependency that slows LSTM computations on long EHR trajectories. Direct modeling of long-range dependencies occurs without the cumulative gradient decay seen in recurrent architectures, allowing capture of subtle patterns over days of monitoring data. Such efficiency is particularly beneficial for real-time stroke prediction systems handling high-volume ICU streams. In our view, these gains will drive broader exploration in clinical AI.

Attention weights generated during inference serve as a form of built-in interpretability, highlighting which time points and features most influence stroke risk predictions. Unlike the opaque hidden states of LSTMs, these weights offer clinicians a window into model reasoning. The absence of recurrence further simplifies optimization and reduces training instability on variable-length medical sequences. Overall, these attributes suggest transformers could

enhance both performance and trustworthiness in acute care settings. We anticipate that interpretability benefits will accelerate regulatory approval and adoption.

Transformer successes in healthcare

Transformers and attention-based models have demonstrated successes in healthcare domains such as cardiovascular disease prediction, diabetes cost forecasting, and stroke risk assessment with BiLSTM-attention hybrids. These approaches have surpassed conventional methods in capturing complex temporal interactions within EHR data. Mortality and complication prediction tasks have similarly benefited from self-attention layers that model long dependencies. Although direct clinical NLP and ECG applications are emerging, the trajectory indicates broadening utility. Limited but growing stroke-specific uses highlight the potential for targeted adaptations.

Despite these advancements, transformer applications in acute ischemic stroke remain in early stages, with most evidence drawn from hybrid or attention-augmented recurrent networks. Successes in related time series tasks suggest that full transformer architectures could yield similar gains for stroke outcome modeling. The parallel nature supports scaling to population-level EHR analyses previously infeasible with LSTMs. Clinicians stand to benefit from more accurate predictions that incorporate global context from patient histories. Continued research will likely expand these successes into routine stroke care protocols.

Architectural Trade-Offs

Computational complexity

Computational complexity represents a key trade-off, with transformers exhibiting quadratic scaling $O(n^2)$ due to pairwise attention computations compared to the linear O(n) of LSTMs. This difference becomes significant for longer clinical sequences encountered in stroke monitoring. In practice, shorter sequences of vital signs over hours may favor LSTMs for faster inference in time-sensitive acute settings. However, as datasets incorporate multi-day trends, the parallel efficiency of transformers can offset some costs through hardware acceleration. Balancing these factors is crucial for selecting architectures suited to specific deployment environments.

Table 1 presents a theoretical trade-off matrix that clarifies how architectural differences translate into clinically relevant modeling decisions**.**

Table 1. Theoretical Trade-off Matrix between LSTM and Transformer Architectures in Clinical Time Series Modeling

Dimension	LSTM Architecture	Transformer Architecture	Theoreti Implicat for Stro Predicti
Computational Complexity	Linear O(n)	Quadratic $O(n^2)$	LSTM favored low-resou or real-ti ICU setti transform scale be with hardwa accelera
Temporal Dependency Modeling	Implicit via recurrence	Explicit via self-attention	Transforn superior multi-da longitudi stroke r trajector
Parallelization क्षमता	Limited (sequential)	Full parallelization	Transforn enable fa training large EH datase
Data Efficiency	High (performs well on small datasets)	Lower (requires large datasets)	LSTMs n suitable specializ stroke registri with limi sample
Interpretability Mechanism	Hidden state probing	Attention weights visualization	Transforn offer mo intuitiv though n causa clinica explanati

Handling Irregular Sampling	Requires preprocessing	Adaptable via positional encoding variants	Transforn more flex with pro tempor encodi strategi
Missing Data Strategy	Imputation-dependent	Masking + learned representations	Transforn can tre missingn as informat signa
Deployment Feasibility	High (CPU-compatible)	Moderate–low (GPU often required)	LSTMs n practica resourc constrai hospita

The clinical impact of this complexity is evident in resource-limited hospital IT infrastructures where LSTMs enable quicker training cycles on standard CPUs. Transformers often require GPU acceleration to manage the memory demands of self-attention matrices during stroke prediction model development. For perioperative or short-stay predictions, the linear scaling of LSTMs provides a clear efficiency advantage. Yet for comprehensive longitudinal modeling, the upfront computational investment in transformers may yield superior long-term performance. We foresee that optimization techniques will gradually narrow this gap in future iterations.

Data requirements

Data requirements constitute another important trade-off, as transformers generally demand larger training corpora to effectively learn attention patterns and generalize across patient cohorts. In contrast, LSTMs often achieve adequate performance with more modest datasets typical of specialized stroke registries. This disparity arises because self-attention layers introduce additional parameters that benefit from extensive examples of clinical variability. For acute ischemic stroke applications, where high-quality annotated sequences may be scarce, LSTMs provide a more accessible entry point. Nevertheless, with the proliferation of multi-center EHR databases, transformers are increasingly viable.

LSTMs excel in low-data regimes by leveraging their inductive biases toward sequential dependencies, reducing the risk of overfitting in smaller stroke prediction cohorts. Transformers, while data-hungry, can capture more nuanced interactions once sufficient samples are available. Clinical studies must therefore consider dataset scale when choosing between architectures for risk forecasting. Hybrid designs could mitigate this by initializing with LSTM-like components before full attention integration. In our opinion, data availability will increasingly dictate architectural preferences in evolving healthcare AI landscapes.

Interpretability

Interpretability trade-offs favor transformers in some respects, as attention weights offer direct visualization of influential time points and features in stroke predictions. However, these weights reflect correlations rather than causal relationships, limiting their standalone clinical utility. LSTMs, by comparison, allow analysis of hidden state activations that can be probed for temporal dynamics in EHR sequences. Both approaches require additional post-hoc techniques for full explainability to satisfy regulatory standards. Enhancing this aspect remains a priority for trustworthy deployment in acute care.

While LSTM hidden states provide sequential insights into model memory, they lack the global view afforded by transformer attention maps. This can make transformers more appealing for clinicians seeking to understand why a particular risk score was assigned. Nonetheless, neither is inherently causal, necessitating integration with domain knowledge for validation. Future work should focus on developing hybrid interpretability tools tailored to stroke prediction. Such advancements would accelerate the safe translation of these models into practice.

Long-range dependency handling

Long-range dependency handling highlights a clear advantage for transformers, which maintain effective information flow across hundreds or thousands of time steps without degradation. LSTMs, in contrast, often struggle beyond approximately 100 steps due to gradient vanishing in extended ICU monitoring data for stroke patients. This capability is vital for modeling multi-day vital sign and lab trends that precede ischemic events. Transformers thus enable more holistic views of patient trajectories in longitudinal EHRs. Their superiority in this

domain positions them well for advanced forecasting applications.

For shorter sequences, however, LSTMs suffice and may even outperform by avoiding unnecessary complexity in dependency modeling. In acute stroke scenarios with limited observation windows, the recurrent structure proves sufficient and computationally lighter. As clinical data incorporates longer histories from outpatient follow-up, transformers are likely to demonstrate superior predictive power. We speculate that sequence length will emerge as a decisive factor in architecture selection. Targeted benchmarks on stroke-specific timelines will clarify these thresholds.

When to Use Which

Short sequences (≤100 time steps)

For short sequences of 100 time steps or fewer, LSTMs often prove sufficient and more computationally efficient than full transformer models in acute ischemic stroke prediction. Their linear scaling aligns well with perioperative monitoring or brief ICU stays where rapid updates are prioritized. In these contexts, sequential processing captures essential short-term dependencies without the overhead of quadratic attention. This efficiency supports real-time bedside applications where latency is critical. We recommend LSTMs as the default choice for such constrained scenarios to optimize resource use.

Perioperative stroke risk prediction, for example, benefits from LSTM's ability to handle compact input windows derived from immediate pre- and post-operative vital signs. Short ICU stays similarly favor models that train quickly on limited data volumes. The maturity of LSTM implementations further eases integration into existing clinical workflows. While transformers could be applied, their added complexity yields diminishing returns here. In our perspective, prioritizing efficiency in short-sequence tasks will enhance practical adoption rates.

Long sequences (>100 time steps)

Long sequences exceeding 100 time steps, such as those from extended ICU stays or outpatient longitudinal monitoring, favor transformers due to their robust long-range dependency modeling. These architectures excel at integrating distant events in vital sign histories that may signal impending stroke. For instance, subtle patterns spanning days or weeks become more accessible without the degradation seen in recurrent models. This superiority could transform predictions for chronic risk assessment in at-risk populations. We anticipate that transformers will dominate applications involving comprehensive EHR timelines.

Outpatient longitudinal data, in particular, presents opportunities where transformers' parallel processing handles expansive sequences efficiently for proactive stroke prevention. Long ICU monitoring benefits similarly by capturing evolving multi-modal inputs over prolonged periods. Although data and compute demands are higher, the predictive gains justify the investment in well-resourced settings. Hybrid variants may further optimize for these use cases by combining strengths. Overall, sequence length should guide the selection process moving forward.

Table 2 provides a context-dependent framework linking clinical scenarios to optimal architectural choices for stroke prediction.

Table 2. Context-Dependent Model Selection Framework for Acute Ischemic Stroke Prediction

Clinical Scenario	Data Characteristics	Operational Constraints	Preferred Architectur
Short ICU Monitoring (≤100 steps)	Dense, short sequences	Low latency, real-time inference	LSTM
Extended ICU Monitoring (>100 steps)	Long, continuous sequences	Moderate compute availability	Transforme
Outpatient Longitudinal Tracking	Sparse, irregular sampling	Batch processing acceptable	Transforme (with adaptations
Small Stroke	Limited sample size	Risk of overfitting	LSTM

Registry Datasets			
Large Multi-Center EHR Datasets	High volume, heterogeneous	GPU-enabled infrastructure	Transforme
Real-Time Emergency Decision Support	Streaming data	Strict latency constraints	LSTM or Hybrid
Resource-Constrained Hospitals	Limited hardware	CPU-based deployment	LSTM
Advanced Research / Hybrid Systems	Mixed sequence lengths	Flexible compute environment	Hybrid LSTM–Transforme

Transformer Adaptations for Clinical Data

Handling irregular sampling

Clinical time series for acute ischemic stroke prediction frequently feature irregular sampling intervals arising from variable monitoring schedules in electronic health records, demanding targeted adaptations within transformer architectures. Time-aware positional encodings that embed actual timestamps or inter-observation deltas preserve temporal structure far more effectively than standard sinusoidal methods designed for uniform sequences. Continuous-time attention variants further weight contributions according to real elapsed intervals between vital signs or laboratory measurements, enabling nuanced modeling of evolving stroke risk. Interpolation strategies such as spline-based or learned filling of gaps complement these encodings to maintain parallel processing advantages. In our perspective, these innovations will prove indispensable for translating transformer strengths to the messy realities of bedside clinical data [1-3].

Building on these methods, future transformer adaptations will likely incorporate dynamic, data-driven positional schemes that adapt on the fly to patient-specific sampling patterns observed in longitudinal stroke cohorts. Such approaches could outperform rigid encodings by learning from the heterogeneous trajectories typical of ischemic events. We speculate that integration with continuous-time formulations will accelerate convergence and improve generalization across diverse hospital systems. Clinicians stand to gain more reliable alerts that respect the true timing of physiological changes rather than artificial assumptions. Overall, mastering irregular sampling represents a critical evolutionary step toward clinically viable attention-based models for stroke forecasting [4-6].

Missing data handling

Missing values permeate EHR time series used in acute ischemic stroke prediction, compelling transformers to adopt sophisticated masking strategies that isolate absent observations from attention computations. Learned imputation tokens can function as contextual placeholders, allowing the model to infer plausible values while preserving global sequence relationships. Separate missingness indicator channels supplied as auxiliary inputs explicitly signal data gaps, enabling the architecture to treat absence as informative rather than noise. These techniques sustain the parallel efficiency of self-attention even on highly incomplete patient records spanning ICU stays. We view such adaptations as essential for robust performance in real-world stroke datasets where complete sequences remain rare [7-9].

The combination of masking, imputation tokens, and missingness modeling could fundamentally elevate transformer reliability beyond traditional recurrent baselines when confronted with sparse longitudinal monitoring. By framing missingness as a learnable feature, models may uncover prognostic patterns invisible to simpler imputation schemes. In our opinion, clinician-guided refinement of these mechanisms will enhance trust and interpretability in high-stakes predictions. Future research must validate these strategies on multi-center stroke registries to confirm clinical utility. Ultimately, effective handling of missing data will remove a major obstacle to broader transformer deployment in preventive stroke care [10-12].

Clinical Deployment Considerations

Real-time inference

Transformers generally demand the complete input sequence to compute full attention matrices, creating latency challenges for real-time inference in acute stroke

monitoring compared with the incremental state updates native to LSTMs. Batching multiple patients can amortize costs in high-throughput settings, yet single-patient urgent predictions in emergency departments still favor sequential architectures for lower per-step overhead. LSTMs excel at processing streaming vital signs one observation at a time without recomputing prior history, supporting seamless bedside deployment. This architectural divergence directly influences responsiveness during time-critical windows for ischemic stroke intervention. We anticipate that streamlined transformer approximations with linear attention will soon narrow this real-time gap [13-15].

Hybrid inference pipelines that retain LSTM-style recurrence for initial encoding while applying selective transformer attention on key windows offer a pragmatic compromise for clinical deployment. Such designs could deliver both incremental updates and long-range context without full-sequence recomputation. In perspective, bedside systems prioritizing low latency will continue to benefit from recurrent components until hardware-optimized attention becomes ubiquitous. The trade-off between full-sequence power and streaming efficiency must be weighed against neurocritical care workflows. Looking forward, these considerations will shape how AI integrates into rapid-response stroke protocols [16, 17].

Computational resources

Transformer architectures impose elevated memory demands through the quadratic growth of attention score matrices, straining edge devices and legacy hospital infrastructure commonly available for stroke prediction tools. Cloud deployment provides the necessary GPU scale for training and batch inference but introduces network latency and data-privacy risks for sensitive EHR streams. LSTMs, with their linear memory profile, integrate more readily into on-premise servers or lightweight edge hardware prevalent in smaller facilities. These resource realities often tip deployment decisions toward recurrent models in resource-constrained environments. We believe hardware-aware compression techniques will progressively democratize transformer access [18, 19].

Balancing edge versus cloud deployment requires careful evaluation of inference speed against the superior modeling capacity of attention mechanisms in longitudinal stroke data. Hospital IT constraints frequently limit GPU availability, underscoring the need for quantized or distilled transformer variants tailored to clinical budgets. In our opinion, collaborative design between AI developers and clinical engineers will accelerate solutions that respect real-world computational limits. This pragmatic focus ensures architectural advances translate into equitable benefits across diverse care settings. Ultimately, resource optimization will determine the pace at which transformers supplant LSTMs in routine stroke risk stratification [20].

Future Directions

Hybrid architectures

Hybrid LSTM-transformer architectures stand out as a compelling future direction, merging the sequential efficiency and stability of recurrent layers with the global dependency capture of self-attention for clinical time series. Convolutional transformer blocks could additionally embed local inductive biases suited to physiological waveform patterns preceding ischemic events. These hybrid designs alleviate the weaknesses of pure recurrent or pure attention models while amplifying their respective strengths in acute ischemic stroke prediction. Preliminary evidence from related healthcare time series tasks suggests hybrids achieve superior accuracy with manageable compute budgets. We foresee such ensembles becoming the default backbone for next-generation EHR-based forecasting systems [1, 13, 20].

By stacking LSTM encoders for short-term dynamics before transformer decoders for long-range integration, hybrids offer a flexible scaffold adaptable to varying sequence lengths in stroke monitoring. Convolutional enhancements may further sharpen detection of transient vital sign anomalies that signal impending risk. In our speculative view, these architectures will bridge the maturity of LSTMs with the scalability of transformers, accelerating clinical translation. Targeted ablation studies on stroke-specific cohorts will be required to optimize layer placement and connectivity. This evolutionary path promises models that are both powerful and practical for real-world deployment [14, 16].

Pre-trained clinical foundation models

Pre-trained clinical foundation models leveraging massive EHR sequence corpora hold transformative potential for fine-tuning on acute ischemic stroke prediction with dramatically reduced task-specific data. Large-scale unsupervised pre-training on diverse time series can encode generalizable temporal representations across

patient populations before specialization to stroke outcomes. Fine-tuning these foundations on targeted registries would accelerate development cycles and enhance robustness to domain shifts. This paradigm mirrors breakthroughs in other medical domains and directly addresses the data-hungry nature of pure transformers. We strongly advocate for community-wide investment in such foundation models to propel the field forward [12, 15, 17, 18].

The transfer-learning benefits of pre-training would lower barriers for smaller stroke research centers that lack the volume required to train transformers from scratch. Multi-modal extensions incorporating imaging or genomics alongside time series could yield holistic risk profiles for personalized prevention. In perspective, foundation models represent a scalable route to architecture-agnostic progress that benefits both LSTM and transformer lineages. Future iterations may incorporate continual learning to adapt to evolving clinical protocols. Overall, this direction could redefine how deep learning architectures are developed and deployed for stroke care [5, 7].

Attention as explainability

Attention mechanisms within transformers provide an intrinsic explainability pathway by generating heatmaps that highlight which time points and clinical features most strongly drive stroke risk predictions. Clinician validation studies will be indispensable to confirm that these visualizations align with established pathophysiological knowledge and foster acceptance among neurologists. Unlike the more opaque hidden states of LSTMs, attention maps deliver global, feature-level insights that can be interactively explored during decision support. This transparency aligns closely with regulatory expectations for trustworthy AI in healthcare. We anticipate that attention-driven explainability will become a cornerstone of future model validation pipelines [3, 8, 9].

Interactive dashboards that overlay attention weights on patient timelines could empower bedside teams to interrogate model reasoning in real time during acute stroke evaluations. Such tools would bridge the gap between algorithmic outputs and clinical intuition, potentially improving adoption rates. In our opinion, rigorous multi-center validation studies focused on attention interpretability will be essential to realize this potential. By making predictions more auditable, transformers gain a distinct advantage in safety-critical applications. This focus on explainability will ultimately elevate the entire ecosystem of deep learning for ischemic stroke prediction [6, 19].

Barriers to Adoption

Lack of large-scale stroke sequence data

The persistent lack of large-scale, richly annotated stroke-specific sequence data within public repositories such as MIMIC and eICU constitutes a foundational barrier to training high-capacity transformer models for acute ischemic stroke prediction. Limited stroke-focused temporal annotations restrict the ability of attention mechanisms to learn robust long-range patterns despite the availability of broader EHR resources. This scarcity stands in stark contrast to the expansive corpora that propelled transformer dominance in natural language processing. Multi-center data-sharing initiatives and standardized annotation frameworks will be required to close the gap. We believe dedicated stroke sequence datasets are non-negotiable for meaningful architectural progress [21-23].

Without sufficient volume and diversity, transformers risk overfitting or underperforming relative to data-efficient LSTMs on modest cohorts typical of specialized stroke registries. Public datasets often prioritize general critical care over detailed ischemic event timelines, further compounding the issue. In perspective, synthetic data generation combined with federated learning offers promising avenues to augment real stroke sequences ethically. Overcoming this barrier will unlock the full predictive power of attention-based architectures. Collaborative efforts among researchers, clinicians, and data stewards must prioritize this challenge to ensure equitable advancement [24-26].

Computational constraints in clinical settings

Computational constraints within clinical settings—including restricted GPU access and aging hospital information technology infrastructure—continue to hinder seamless adoption of transformer models for stroke prediction workflows. Inference latency requirements in fast-paced neurocritical care environments often exceed what full attention computation can deliver without specialized acceleration. Edge computing platforms favored for bedside use struggle with the memory footprint of self-attention

layers, favoring lighter recurrent alternatives. These practical limitations frequently result in conservative architecture choices despite theoretical superiority. Targeted efficiency research will be vital to surmount these deployment realities [10, 11].

Hospital environments prioritize system stability and minimal disruption, making the integration of resource-intensive models logistically complex. We speculate that sparse and linearized attention variants will gradually alleviate memory and latency burdens in production settings. Close partnership between AI developers and clinical informatics teams remains essential to align model footprints with existing infrastructure. This barrier highlights the importance of co-designing architectures with deployment realities in mind. Addressing it thoughtfully will accelerate the transition to more capable predictive systems across varied care contexts [4, 15].

Conclusion

The architectural landscape for acute ischemic stroke prediction is undergoing a clear evolutionary transition from long short-term memory networks, which have long dominated clinical time series modeling, toward transformer-based approaches that leverage self-attention for superior sequence handling. LSTMs provided a reliable foundation for capturing temporal dependencies in EHR data over the past decade, yet their sequential nature increasingly reveals limitations as datasets and monitoring durations expand. Transformers, by enabling parallel computation and direct long-range interactions, introduce a paradigm shift that aligns with the growing complexity of longitudinal stroke risk profiles. This progression reflects broader trends in deep learning while remaining grounded in the unique demands of healthcare applications. In our view, the field stands at an inflection point where thoughtful integration of both paradigms can yield substantial clinical gains.

Key trade-offs in computational complexity, data requirements, and sequence length will continue to shape architecture selection rather than allowing any single model to claim universal dominance. Transformers excel on extended timelines and large corpora but incur higher resource costs, whereas LSTMs retain advantages in low-data and short-sequence regimes common in acute care. These considerations underscore that performance must be evaluated holistically against clinical context rather than benchmark scores alone. Sequence length, in particular, emerges as a decisive factor guiding whether recurrent or attention mechanisms deliver optimal results. Recognizing these nuances prevents over-enthusiastic adoption of newer architectures at the expense of practicality.

We recommend a context-dependent, no-one-size-fits-all strategy in which both LSTMs and transformers are rigorously evaluated side-by-side for each stroke prediction task. Hybrid designs and pre-trained foundations may ultimately offer the most balanced path forward, capitalizing on the maturity of recurrent models and the scalability of attention. Architecture choice should always factor in deployment constraints, interpretability needs, and available data volume within specific hospital ecosystems. Such pragmatic evaluation will ensure that technological evolution serves patient outcomes rather than novelty. In our opinion, this measured approach maximizes the likelihood of successful translation into routine care.

A concerted call to action is now warranted: the community must benchmark transformer variants against established LSTM baselines on public and multi-center stroke datasets to generate definitive evidence on when and how the transition delivers value. Systematic comparisons will clarify optimal use cases and accelerate the development of clinically ready tools. By embracing this rigorous evaluation framework, researchers and clinicians can guide the field toward architectures that are not only advanced but also deployable and trustworthy. The coming years represent a pivotal opportunity to shape AI-driven stroke prevention in ways that meaningfully reduce global disability burden. Ultimately, the evolution from LSTMs to transformers signals an exciting yet responsible chapter in healthcare artificial intelligence.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Nov 2021 Revised: 29 Dec 2021 Accepted: 31 Jan 2022
Published online: 20 July 2022

References

Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and regression models. Front Neurol. 2020;11:889. https://doi.org/10.3389/fneur.2020.00889.

Choi YA, Park SJ, Jun JA, Pyo CS, Cho KH, Lee HS, et al. Deep learning-based stroke disease prediction system using real-time bio signals. Sensors (Basel). 2021;21(13):4269. https://doi.org/10.3390/s21134269.

Lu Y, Shi Y, Chen D. Risk prediction of annual stroke of ischemic stroke based on BiLSTM-Attention model. J Donghua Univ Nat Sci Ed. 2021;47(4):1-8.

Cheon S, Kim J, Lim J. The use of deep learning to predict stroke patient mortality. Int J Environ Res Public Health. 2019;16(11):1876. https://doi.org/10.3390/ijerph16111876.

Guo A, Smith S, Khan YM, Langabeer JR 2nd, Foraker RE. Application of a time-series deep learning model to predict cardiac dysrhythmias in electronic health records. PLoS One. 2021;16(9):e0239007. https://doi.org/10.1371/journal.pone.0239007.

Usama M, Ahmad B, Xiao W, Hossain MS, Muhammad G. Self-attention based recurrent convolutional neural network for disease prediction using healthcare data. Comput Methods Programs Biomed. 2020;190:105191. https://doi.org/10.1016/j.cmpb.2020.105191.

Ge Y, Wang Q, Wang L, Wu H, Peng C, Wang J, et al. Predicting post-stroke pneumonia using deep neural network approaches. Int J Med Inform. 2019;132:103986. https://doi.org/10.1016/j.ijmedinf.2019.103986.

Zhang S, Wang J, Pei L, Liu K, Gao Y, Fang H, et al. Interpretability analysis of one-year mortality prediction for stroke patients based on deep neural network. IEEE J Biomed Health Inform. 2022;26(4):1903-10. https://doi.org/10.1109/JBHI.2021.3125155.

Tan D, Wang J, Yao R, Liu J, Wu J, Zhu S, et al. CCA4CTA: a hybrid attention mechanism based convolutional network for analysing collateral circulation via multi-phase cranial CTA. In: Proc IEEE Int Conf Bioinformatics Biomedicine. 2022. p. 1201-6. https://doi.org/10.1109/BIBM55620.2022.9995480.

Darmawahyuni A, Nurmaini S, Sukemi, Caesarendra W, Bhayyu V, Rachmatullah MN, et al. Deep learning with a recurrent network structure in the sequence modeling of imbalanced data for ECG-rhythm classifier. Algorithms. 2019;12(6):118. https://doi.org/10.3390/a12060118.

Lam C, Tso CF, Green-Saxena A, Pellegrini E, Iqbal Z, Evans D, et al. Semisupervised deep learning techniques for predicting acute respiratory distress syndrome from time-series clinical data: model development and validation study. JMIR Form Res. 2021;5(9):e28028. https://doi.org/10.2196/28028.

Datta S, Morassi Sasso A, Kiwit N, Bose S, Nadkarni G, Miotto R, et al. Predicting hypertension onset from longitudinal electronic health records with deep learning. JAMIA Open. 2022;5(4):ooac097.

Jin B, Che C, Liu Z, Zhang S, Yin X, Wei X. Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access. 2018;6:9256-61. https://doi.org/10.1109/ACCESS.2018.2806946.

Kim JC, Chung K. Recurrent neural network-based multimodal deep learning for estimating missing values in healthcare. Appl

Sci. 2022;12(15):7477. https://doi.org/10.3390/app12157477.

McGilvray MM, Heaton J, Guo A, Masood MF, Cupps BP, Damiano M, et al. Electronic health record-based deep learning prediction of death or severe decompensation in heart failure patients. Heart Fail. 2022;10(9):637-47. https://doi.org/10.1002/ehf2.14016.

Li J, Liang T, Zeng Z, Xu P, Chen Y, Guo Z, et al. Motion intention prediction of upper limb in stroke survivors using sEMG signal and attention mechanism. Biomed Signal Process Control. 2022;78:103981. https://doi.org/10.1016/j.bspc.2022.103981.

Luo M, Wang YT, Wang XK, Hou WH, Huang RL, Liu Y, et al. A multi-granularity convolutional neural network model with temporal information and attention mechanism for efficient diabetes medical cost prediction. Comput Biol Med. 2022;151:106246. https://doi.org/10.1016/j.compbiomed.2022.106246.

An Y, Huang N, Chen X, Wu F, Wang J. High-risk prediction of cardiovascular diseases via attention-based deep neural networks. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(3):1093-105. https://doi.org/10.1109/TCBB.2019.2935054.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998-6008.

Chantamit-O-Pas P, Goyal M. Long short-term memory recurrent neural network for stroke prediction. In: Int Conf Mach Learn Data Min Pattern Recognit. 2018. p. 312-23. https://doi.org/10.1007/978-3-319-96136-1_23.

Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. Stroke. 2019;50(5):1263-5. https://doi.org/10.1161/STROKEAHA.118.024293.

Fang G, Huang Z, Wang Z. Predicting ischemic stroke outcome using deep learning approaches. Front Genet. 2022;12:827522. https://doi.org/10.3389/fgene.2021.827522.

Van Os HJA, Ramos LA, Hilbert A, Van Leeuwen M, Van Walderveen MAA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. Front Neurol. 2018;9:784. https://doi.org/10.3389/fneur.2018.00784.

Wang X, Fan Y, Zhang N, Li J, Duan Y, Yang B. Performance of machine learning for tissue outcome prediction in acute ischemic stroke: a systematic review and meta-analysis. Front Neurol. 2022;13:910259. https://doi.org/10.3389/fneur.2022.910259.

Brugnara G, Neuberger U, Mahmutoglu MA, Foltyn M, Herweh C, Nagel S, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. Stroke. 2020;51(12):3541-8. https://doi.org/10.1161/STROKEAHA.120.030287.

Hilbert A, Ramos LA, van Os HJA, Olabarriaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Comput Biol Med. 2019;115:103516. https://doi.org/10.1016/j.compbiomed.2019.103516.