ENFOQUE UTE
REVISTA
Facultad Ciencias de la Ingeniería e Industrias

REVIEW PAPER

ENFOQUE UTE, VOL. 17, NO. 2, pp. 8-17, E-ISSN: 1390-6542
RECEIVED JULY 24, 2025; REVISED NOVEMBER 22, 2025; ACCEPTED FEBRUARY 2, 2026; PUBLISHED ONLINE (AHEAD OF PRINT) MARCH 6, 2026; PUBLICATION DATE APRIL 1, 2026.

Citation: A. B. Bestard Columbié, D. E. Cruz Morete, L. Soler Jay, and D. López Ramos, "MAMBA: SSM model as alternative to the transformers," *Enfoque UTE*, vol. 17, no. 2, pp. 8-17, Apr. 2026, doi: https://doi.org/10.29019/enfoqueute.1204

MAMBA: SSM model as alternative to the transformers

Ana Beatriz Bestard Columbié[1*], Daniela Elizabeth Cruz Morete[2], Lianet Soler Jay[3], and Dionis López Ramos[4]

***Abstract* — Mamba is a recent State Space Model (SSM) architecture to improve the computational and scalability limitations of transformer-based sequence models. In this review, we synthesize and compare Mamba's core design—interleaved SSM and feed-forward layers with hardware-aware memory management—to standard Transformers, highlighting its linear complexity and ability to process extremely long contexts. We analyze published benchmarks showing that Mamba outperforms or matches open-source baselines (e.g. Pythia, RWKV) of similar and even twice the size on zero-shot tasks, scales more efficiently on genomic sequences (processing >1 M tokens with only 74 K parameters), and supports variants such as Jamba (MoE extension), Falcon Mamba 7B, Mamba-2 (Structured SSM), and Mamba-4 with further speed or capacity gains. We discuss adaptations to vision (VIM) and dependency parsing (DepMamba), and emerging hybrids (e.g. Bamba, IBM Granite) that fuse SSM efficiency with Transformer accuracy. Finally, we interpret these findings in the context of real-world constraints—compute cost, energy, and tooling maturity—outlining where Mamba excels, and hybrid models may be preferable, and which areas require further optimization. Our conclusions suggest that Mamba and its derivatives offer a viable path toward more sustainable, scalable sequence modeling.**

Keywords: **language models; hybrid architectures; mamba; state space models; scalability.**

***Resumen* — Mamba es una arquitectura reciente de Modelo de Espacio de Estados (SSM) que supera las limitaciones computacionales y de escalabilidad de los modelos de secuencia basados en Transformadores. En esta revisión examinamos y comparamos el diseño clave de Mamba —que combina capas de modelo de espacio de estados (SSM) con capas feed-forward y uso optimizado de memoria— frente a los modelos transformadores estándar, destacando su complejidad lineal y capacidad para procesar contextos extremadamente largos. Analizamos marcos de referencia publicados que muestran que Mamba supera o iguala a referentes de código abierto (ejemplo, Pythia, RWKV) de tamaño similar e incluso del doble en tareas zero-shot, siendo más eficiente en secuencias genómicas (procesando más de 1 millón de tokens con solo 74 mil parámetros) y admite variantes como Jamba (extensión MoE), Falcon Mamba 7B y Mamba-2 (SSM estructurado) con mayores ganancias de velocidad o capacidad. Discutimos adaptaciones para visión (VIM) y análisis de dependencias (DepMamba), y nuevos híbridos (ejemplo, Bamba, IBM Granite) que combinan la eficiencia de los SSM con la precisión de los Transformers. Finalmente, interpretamos estos hallazgos en el contexto de restricciones del mundo real —costo computacional, energía y madurez de herramientas—, señalando dónde sobresale Mamba, dónde pueden preferirse modelos híbridos y qué áreas requieren mayor optimización. Nuestras conclusiones sugieren que Mamba y sus derivados ofrecen un camino viable hacia un modelado de secuencias más sostenible y escalable.**

Palabras Clave: **modelos de lenguaje; arquitecturas híbridas; mamba; modelos de espacio de estados; escalabilidad.**

I. INTRODUCTION

THE Transformer architecture, introduced in 2017, revolutionized sequence modeling by employing self-attention to capture token-wise dependencies [1]. Despite its success in natural language processing, vision, and other domains, Transformer models suffer from quadratic computational complexity, making them resource-intensive for long sequences and large contexts. This limitation drives the search for alternatives that balance efficiency, scalability, and modeling power.

Mamba emerges as a novel solution by replacing global attention with Selective State Space Models (SSMs), achieving linear complexity and hardware-aware memory usage [2]. Early evidence demonstrates its ability to rival or surpass Transformer baselines on zero-shot tasks and real-world benchmarks, while drastically reducing compute requirements.

Variants such as Jamba (Mixture of Experts) [3], Falcon Mamba 7B (long-context optimization) [4], and Mamba-2 (Structured SSM layers) [5] further explore this design space.

* ***Corresponding autor:*** ana.bestard@estudiantes.uo.edu.cu
1. Ana Beatriz Bestard Columbié is with Informatics Department, Faculty of Engineering in Telecommunications, Informatics and Biomedical at Universidad de Oriente Ave. Patricio Lumumba s/n, Santiago de Cuba, Cuba. E-mail: ana.bestard@estudiantes.uo.edu.cu). ORCID number https://orcid.org/0009-0006-5508-1057
2. Daniela Elizabeth Cruz Morete is with Informatics Department, Faculty of Engineering in Telecommunications, Informatics and Biomedical at Universidad de Oriente Ave. Patricio Lumumba s/n, Santiago de Cuba, Cuba. E-mail: daniela.cruz@estudiantes.uo.edu.cu). ORCID number https://orcid.org/0009-0006-1233-259X
3. Lianet Soler Jay is with Informatics Department, Faculty of Engineering in Telecommunications, Informatics and Biomedical at Universidad de Oriente Ave. Patricio Lumumba s/n, Santiago de Cuba, Cuba. E-mail: lianet.soler@estudiantes.uo.edu.cu). ORCID number https://orcid.org/0009-0002-4656-4752
4. Dionis López Ramos is with Informatics Department, Faculty of Engineering in Telecommunications, Informatics and Biomedical at Universidad de Oriente Ave. Patricio Lumumba s/n, Santiago de Cuba, Cuba. E-mail: dionis@uo.edu.cu). ORCID number https://orcid.org/0000-0001-9217-0696

DOI: https://doi.org/10.29019/enfoqueute.1204
Associate Editor: *Marcelo Mosquera*

This review aims to (1) summarize Mamba's architectural principles, (2) evaluate its performance against established models, (3) survey its emerging variants and applications, and (4) critically discuss trade-offs, implementation challenges, and future research directions. By contextualizing Mamba within the broader evolution of sequence models, we provide researchers and practitioners with a clear roadmap for adopting SSM-based approaches.

II. MATERIALS AND METHODS

This research applies a mixed qualitative–quantitative approach, to compare Mamba and Transformer architectures.

The overall procedure follows the algorithmic workflow illustrated in Figure 1.

Algorithm 1 Systematic Literature Review on MAMBA

Input: Research questions RQ
Output: Final report R

Step 1: Search
Define search string: "MAMBA" AND "state space model"
Search in databases: IEEE, ACM, arXiv, Scopus
Remove duplicate papers
$n_{initial} \leftarrow$ number of papers found

Step 2: Selection
Screen papers by title and abstract
Read full-text of relevant papers
Apply inclusion/exclusion criteria
$n_{final} \leftarrow$ number of selected papers

Step 3: Data Extraction
Extract: authors, year, method, results, datasets
Organize data in structured table
Identify patterns and trends

Step 4: Synthesis
Answer research questions using extracted data
Write final report R
return R

Fig. 1. Systematic Literature Review algorithm about papers related to Mamba concepts.

The key components and stages of this process are detailed below:

A. Systematic Literature Identification (Search): To ensure a thorough review of the existing literature, relevant academic databases were selected. The literature search was conducted using IEEE Xplore, ACM Digital Library, arXiv, Scopus, and Web of Science, covering both peer-reviewed publications and influential preprints related to Mamba and Transformer architectures.

B. Search Strategy and Temporal Scope (Search): Search strings were constructed using primary keywords and validated synonyms identified from a preliminary review of core Mamba publications to maximize retrieval sensitivity: ("MAMBA" OR "Mamba") AND ("state space model" OR "SSM" OR "selective state space") AND ("deep learning" OR "neural network" OR "sequence modeling"). The temporal scope spanned 2023–2026, corresponding to the introduction and evolution of the Mamba architecture. Few exceptions presented by the definition of concepts.

C. Study Screening and Selection (Selection): All retrieved citations were imported into a reference management tool for automated de-duplication, followed by a manual review to identify residual overlaps such as minor title variations, conference–journal duplicates, and discrepancies between preprint and final versions. Studies meeting any of the following exclusion criteria (EC) were also removed: non-peer-reviewed grey literature, papers under four pages, redundant publications, peripheral mentions of Mamba, non-English texts, inaccessible full-text versions, or non-primary research such as editorials and opinion pieces.

D. Full-Text Eligibility Assessment (Data Extraction): Publications that passed the title and abstract screening underwent full-text evaluation. Each study was assessed against the predefined inclusion and exclusion criteria. All exclusions made during this stage were accompanied by clear justification to ensure transparency and reproducibility.

E. Data Extraction and Technical Characterization (Data Extraction): For each eligible study, structured data extraction captured key attributes, including application domain, specific Mamba or Transformer architectural variants, reported evaluation benchmarks and datasets, quantitative performance metrics, Computational efficiency indicators (including FLOPs, memory usage, and inference latency), and viability of open-source implementations.

F. Comparative Analytical Framework (Synthesis): Following data extraction, the selected studies were analyzed according to five sequential analytical phases:
1) Transformers Limitations: Analyze Transformer challenges in long-sequence processing, focusing on resource demands and context loss.
2) Mamba Theoretical Analysis: Explore Mamba's architecture emphasizing its linear complexity and hardware optimization.
3) Evaluation Metrics: Establish clear quantitative metrics for objective performance comparison.
4) Experimental Comparison: Analyze and compare published benchmark results across diverse evaluation tasks.
5) Validation and Generalization: Assess applicability and provide deployment recommendations based on evidence synthesis.

III. RESULTS

The section presents different approaches and criteria evaluation of comparative features about SSM and Transformers.

A. Analysis of the Transformer architecture and its limitations in handling long data sequences.

1) Computational Complexity of Transformers: Transformer architectures use self-attention to compute relationships between all tokens in a sequence. This results in

quadratic computational complexity. As sequence length increases, computation grows rapidly—for example, 10 tokens require 100 operations, while 20 need 400. This high resource demand limits Transformers' efficiency in tasks involving long sequences [1].

2) Difficulty Capturing Long-Range Dependencies: While Transformers surpass Recurrent Neural Networks (RNNs) in handling long-range relationships, their attention mechanisms still struggle with extremely long sequences. This local focus impedes the capture of distant dependencies, which can affect the quality of predictions and text generation.
3) Inefficiency of Linear Approximations: Attempts to optimize Transformers via linear kernels often introduce new instabilities, such as unbounded gradients and attention dilution [6]. The former destabilizes training through improper scaling, while the latter spreads focus too thinly over long sequences, failing to prioritize semantically relevant tokens. Consequently, these variants struggle to match the performance-to-efficiency ratio required for complex long-range tasks.
4) Limited scalability: Transformers face significant scalability constraints as they grow in size. Their computational requirements make both training and deployment increasingly costly and time-intensive, limiting practical implementation. For instance, some large language models demand multimillion-dollar infrastructure, limiting accessibility and use where resources are scarce.

B. Mamba Architecture and Its Performance Results

The Mamba model represents a new generation of architectures for sequence processing, specifically designed to maximize computational efficiency and handle long sequences at low computational cost [7]. Unlike Transformers, which rely on self-attention mechanisms with quadratic complexity, Mamba is based on Selective State Space Models (SSMs), achieving linear complexity and enabling significantly greater scalability.

From a systems theory perspective, this reflects the fundamental distinction between Transformer's parameter-extensive Finite Impulse Response (FIR) processing—requiring explicit computation between all token pairs—versus Mamba's more general Infinite Impulse Response (IIR) formulation, which maintains a compact recurrent state for efficient long-range modeling [8].

Recent theoretical grounding via Rough Path Theory establishes that Mamba's selectivity mechanism achieves expressive equivalence to attention while maintaining linear scaling [9].

Its architecture (see Figure 2) consists of repeated Mamba blocks, each combining an SSM component with a feed-forward layer, where:

- The SSM layers maintain an internal state that is updated sequentially, capturing temporal dependencies.
- The feed-forward layers transform the captured representations, adding expressiveness to the model.

Within these blocks, nonlinearities, such as the SiLU activation (σ) and multiplicative gating (⊗), are applied to transform captured representations and enhance the model's expressiveness.

This sequential architecture enables efficient information processing, where each block refines the output of the previous one, ensuring a balance between computational efficiency and representational capacity.

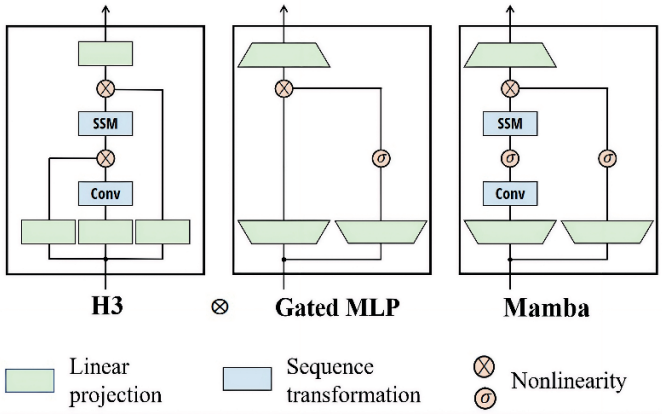

Fig. 2. The SSM architecture, modified from [2].

Additionally, Mamba makes intelligent use of GPU memory by combining SRAM (fast access) and HBM (high bandwidth) to optimize access to critical data and reduce bottlenecks [2].

C. Performance Evaluation of Mamba vs. Transformers

The model was evaluated on diverse zero-shot tasks against comparable open-source models, particularly Pythia and RWKV (see Table I). Results demonstrate that Mamba consistently outperforms same-size competitors and matches the performance of baseline models twice its size [2].

TABLE I
MAMBA ZERO-SHOT EVALUATIONS, TAKEN FROM [2]

Model	Token	Pile ↓	PIQA ↑	Arc-C ↑
Hybrid H3-130M	GPT2	—	64.2	24.2
Mamba-130M	NeoX	10.56	64.5	24.3
GPT-Neo 1.3B	GPT2	—	71.1	25.9
Pythia-1.4B	NeoX	7.51	71.0	28.5
RWKV-1.5B	NeoX	7.70	72.4	29.4
Mamba-1.4B	NeoX	6.80	74.2	32.8

Mamba shows superior scaling performance compared to baseline models such as HyenaDNA and Transformer++ when increasing model size and context length on the HG38 dataset. Its pre-training perplexity improves consistently with model size, allowing Mamba to match their performance while using 3 to 4 times fewer parameters [2] as shown in Figure 3.

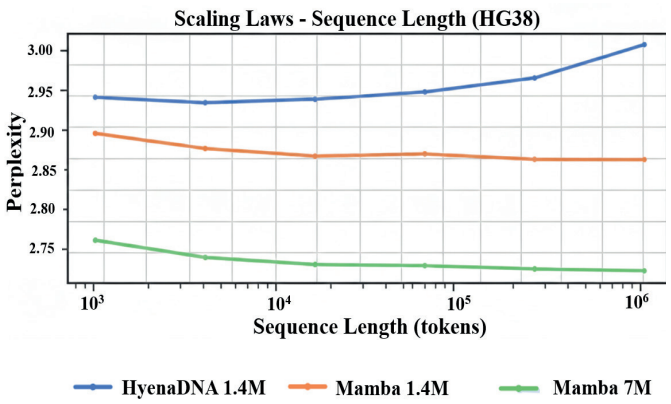

Fig. 3. DNA Scaling Laws, modified from [2].

With just 74k parameters, Mamba can process sequences of up to 1,048,576 tokens, while Transformers under the same conditions may encounter out-of-memory error. As illustrated in Table II, Mamba maintains perfect generalization accuracy (denoted by √) across sequence lengths that significantly exceed the original training length.

TABLE II
TEST ACCURACY AT SEQUENCE LENGTH, TAKEN FROM [2]

Model	Params	2^8	2^9	2^{10}
MHA-Abs	137K	100.0	58.6	26.6
MHA-xPos	137K	100.0	99.6	67.6
Hyena	69M	100.0	√	44.1
Mamba	74K	100.0	√	√

Systematic comparisons validate this length generalization across high-resolution domains achieving stable performance where Transformer architectures exhibit quadratic memory saturation and degradation beyond 64k context lengths [10].

Mamba's efficiency is particularly evident in long-context scenarios where Transformers typically encounter 'memory walls' or out-of-memory errors. Furthermore, empirical evaluations on the Long-Range Arena (LRA) and Pile benchmarks indicate that Mamba-based models effectively bridge the 'perplexity gap' that previously hindered pure State-Space Models (SSMs), matching the reasoning quality of state-of-the-art Transformers with significantly lower computational overhead [11]. This synergy of linear scalability and content-based reasoning establishes Mamba as a robust alternative for multi-modal and large-scale sequence modeling.

Mamba's advantages extend to code intelligence tasks, where systematic Pre-trained Language Models (PLM) evaluations demonstrate Mamba outperforms Transformer-based CodeGPT across code completion, generation, and clone detection tasks, pre-trained from scratch, evaluated with both full fine-tuning (FT) and parameter-efficient fine-tuning (PEFT), achieving superior efficacy and efficiency even at 7B scale with parameter-efficient fine-tuning [12].

Studies on the Mamba architecture show effectiveness in Information Retrieval (IR). After fine-tuning the model, it matches or outperforms Transformer models in both short-text (MS MARCO) and long-text (LoCoV0) benchmarks [13]. Beyond accuracy, it achieves significantly higher inference speeds and handles document lengths exceeding its original training window. These results confirm Mamba as a practical, high-speed alternative for large-scale document searching and semantic retrieval.

Even with minimal training, Mamba architectures distilled from pretrained Transformers achieve superior performance to all prior open-source non-Transformer models using only 3B tokens demonstrating Mamba's capacity to inherit Transformer capabilities while preserving linear scaling [14].

D. Research and Development on Mamba Variants

Recent advances in Mamba architecture have been driven by their linear scaling properties across diverse application domains, including language modeling, computer vision, medical imaging, and scientific computing. This section systematically categorizes these developments into five foundational areas, highlighting architectural innovations, empirical performance, and deployment advantages over Transformer baselines.

1) Foundational Evolutions and Large Language Models

This category encompasses core architectural refinements and scaling strategies for Mamba-based language models:

Mamba-2 and Structured State Space Duality (SSD): As an evolution of the original framework, Mamba-2 [5] addresses the hardware underutilization identified in its predecessor. Modern GPUs and TPUs are built for fast matrix operations, which the original Mamba underutilized. The new version introduces Structured State Space (SSD) layers, a more efficient computational method that integrates better with deep neural networks. It also generates key parameters in parallel with the input, making it significantly more compatible with large-scale training.

A notable real-world implementation is Codestral Mamba, a specialized commercial model for code generation. This application proves that Mamba-2 can handle massive programming contexts with constant-time inference.

In parallel, hybrid architectures have emerged, combining the strengths of SSMs and Transformers. Jamba, proposed by [3] is a compact architecture that integrates a Mixture-of-Experts (MoE) design. It handles contexts up to 256K tokens on a single 80 GB GPU and delivers three times the performance of models like Llama 2 13B and Mixtral 8x7B.

The Jamba family has since expanded to include instruction-tuned variants. Jamba 1.5, available in Large (94B parameters) and Mini (12B) versions, excels in conversational tasks and instruction following. Notably, the Large version employs ExpertsInt8 quantization to run on a single 8-GPU node with low latency and no quality degradation. Both models outperform similarly sized competitors, with inference up to 2.5× faster on long contexts. Jamba-1.5 Mini is the fastest for 10K token contexts.

Jamba Reasoning 3B, extends the family with an architecture for on-device deployment. At 3B parameters, it achieves 2–5× efficiency gains over competitors like DeepSeek and Llama while leading on instruction-following and reasoning benchmarks. Its lightweight footprint enables local use on iPhones, Androids, Macs, and PCs, supporting private, low-latency, offline-resilient applications without performance loss.

In contrast to these hybrid models, Falcon Mamba 7B represents a pure, large-scale implementation of the Mamba-1 architecture (without MLP layers). By incorporating RMS normalization, it ensures training stability and effectively overcomes the sequence-length limitations typical of earlier SSMs. Unlike traditional Transformers, it processes long contexts with constant generation speed and zero memory overhead growth.

Recent mechanistic studies reveal that its performance is driven by a "Gather-and-Aggregate" (G&A) mechanism localized within a small subset of its 64 layers. In this process, specialized heads extract and synthesize context—a strategy functionally similar to Transformer attention but adapted for continuous hidden states [4]. While these internal patterns are smoother than discrete attention, the model achieves high precision by utilizing multiple redundant Mamba channels.

As shown in Table III, Falcon Mamba 7B outperforms prominent models like Llama 3.1 8B and Mistral 7B, proving it to be a highly efficient and innovative alternative for complex, long-sequence tasks.

TABLE III
EVALUATION OF THE MAMBA AND TRANSFORMERS ARCHITECTURE, MODIFIED FROM [4]

Architecture	Model	IFEvall	BBH	MATH Lvl5	GPQA
Mamba (SSLM)	Falcon Mamba 7B	33.36	19.88	3.63	8.05
Transformer	Mistral 7B	22.66	24.04	2.64	5.59
Transformer	Llama 3.1 (8B)	12.7	25.29	4.61	6.15
Transformer	Falcon 2 (11b)	32.61	21.94	2.34	2.8

2) Computer Vision and Visual Understanding

Mamba variants adapt sequential SSM processing to 2D/3D visual data via innovative scanning strategies:

The application of Mamba architecture has recently expanded into computer vision. Unlike text, which is sequential, visual data requires holistic processing to extract meaningful context. To address this, researchers proposed VIM (Vision Mamba) [15], a vision-specific architecture that adapts Mamba blocks to compress positional and directional information through SSMs. Early results are promising: VIM not only retains Mamba's efficiency advantages but also achieves comparable performance to Vision Transformers (ViT) [16] while being three times smaller in size.

Beyond general vision, Vision Mamba (VIM) has shown significant potential in clinical settings. A recent study [17] fine-tuned the Vim architecture for Tuberculosis detection using chest X-rays, achieving a 94.32% accuracy. This application proved particularly effective, as VIM not only outperformed traditional CNNs but also reduced GPU memory usage by 80% compared to standard Transformer-based models.

However, despite its effective adaptation, Visual Mamba's sequential token access mechanism introduces new challenges, particularly a heightened sensitivity to standard quantization techniques. To overcome this limitation, the PTQ4VM (Post-Training Quantization for Visual Mamba) framework [18] introduces two novel strategies: Per-Token Static (PTS) quantization and Joint Learning of Smoothing Scale and Step Size (JLSS). As the first study on Visual Mamba quantization, this approach enables a 1.83× GPU speedup with negligible accuracy loss, allowing pretrained models to be optimized in under 15 minutes.

The rapid adoption of Mamba in computer vision is further corroborated by a comprehensive systematic review [19], which attributes this trend to Mamba's linear complexity against the quadratic self-attention costs of Transformers. This systematic review analyzes foundational Vision Mamba backbones and hybrid models enhanced with convolution, recurrence, and attention, covering applications across general vision (object detection, segmentation), medical imaging (2D/3D segmentation, classification), and remote sensing tasks. The survey identifies scanning techniques as critical for adapting Mamba's sequential processing to 2D visual data, positioning Mamba as a scalable alternative to Vision Transformers.

Extending these principles to dynamic visual data, VideoMamba [20] addresses the challenges of local redundancy and global dependencies in video understanding. By replacing the quadratic complexity of Video Transformers [21] with a linear-complexity operator, it enables high-resolution modeling for long sequences.

This architecture excels in four areas: it scales efficiently through self-distillation, detects subtle motions accurately, outperforms traditional models in long-term video analysis, and integrates seamlessly with multi-modal systems. These advancements position VideoMamba as a benchmark for efficient and comprehensive video analysis.

Finally, a novel architecture called MR-Stereos shown in [22], introduces Mamba Regularization to improve how robots and autonomous cars calculate depth. Unlike previous models that struggle with smooth surfaces, MR-Stereo uses Mamba to process 3D data as a sequence. This allows the system to understand the entire geometry of a scene with much less memory than a Transformer. It includes a Spatial Residual Convolution (SRC) module to ensure that fine details remain sharp, making it a highly efficient 'plug-and-play' solution for real-world navigation.

3) Clinical Informatics and Medical Imaging

The efficiency of Mamba architecture makes them well-suited for high-resolution medical data and clinical monitoring, where long sequences and computational constraints are critical. This has driven the development of specialized models across diverse healthcare applications.

In the field of genomics, for instance, the Caduceus model (also referred to as MambaDNA) [23] adapts the Mamba ar-

chitecture to handle the unique complexity of DNA sequences. Unlike standard versions, Caduceus introduces a bi-directional MambaDNA block, allowing it to process genetic information in both directions simultaneously.

It also features reverse-complement (RC) equivariance, which ensures the model recognizes DNA's double-stranded nature. These innovations allow Caduceus to process long-range sequences of over 131,000 tokens with high efficiency. In benchmarks, it notably outperformed models 10 times larger in tasks like Variant Effect Prediction, proving that specialized, pure Mamba architectures can surpass much larger Transformers in scientific precision.

Beyond genomics, Mamba's efficiency has been leveraged for image-based diagnostics. The DSA Mamba architecture [24] introduces a lightweight medical image classification architecture combining dynamic spatial attention with bidirectional state-space modeling. Its Dual Mamba Block extracts local and global features in parallel, while the Spatial-Channel Cross-Attention (SCCA) mechanism improves their integration. With superior accuracy and faster convergence than CNNs and Transformers, it proves highly efficient for real clinical deployment.

This diagnostic potential is further validated in dermatology. A recent benchmark study [25] evaluated Mamba against Vision Transformers (ViT) [16] on the HAM10000 dataset for skin cancer classification. The results demonstrated that Mamba achieves equivalent diagnostic accuracy while being significantly more efficient.

The core contribution of this study is showing that Mamba handles 'rare cases' (like Bowen's disease) better than traditional models. This validation suggests that Mamba is a production-ready alternative for early, non-invasive cancer detection in medical centers with limited computing resources.

For more complex segmentation tasks, Switch-UMamba [26], has been developed to overcome the limitations of CNNs and Transformers in segmenting complex medical structures. By combining a hybrid CNN–state space block with a dynamic scanning mechanism, it captures both local details and long-range dependencies in biomedical images. A key advantage is its use of a mixture-of-scans strategy, which adaptively selects the most suitable scanning patterns for each sample while maintaining low computational cost. Experimental results on multiple medical image segmentation benchmarks show that it consistently outperforms other Mamba-based and Transformer-based networks.

Mamba's applicability also extends to mental health. DepMamba, as shown in [27], is a Mamba-based architecture designed to handle long-range dependencies in real data with multimodal learning for mental-health assessment. It uses selective state-space modeling to support structured prediction tasks such as syntactic dependency parsing, achieving up to 3× faster processing and lower memory use than Transformer-based parsers. DepMamba also integrates audio and visual signals to better detect behavioral indicators of depression. Tests on real NLP and health datasets show competitive accuracy and superior inference efficiency, making DepMamba suitable for practical applications in clinical monitoring in language-based decision systems.

Similarly, Mamba's ability to model complex sequences has been successfully applied to autism diagnosis using functional magnetic resonance imaging (fMRI). By combining Transformer layers with Mamba modules, researchers have developed a hybrid approach that integrates brain connectivity data with temporal dependencies to study autism diagnosis method. This method uses the ABIDE dataset to extract features from functional brain networks, achieving a 70.5% classification accuracy, outperforming traditional baseline models in identifying neurological patterns [28].

In the field of drug discovery, MambaTransDTA [29] combines Mamba and Transformer architectures to predict drug-target affinity (DTA). Transformers excel at modeling local interactions, while Mamba is used to capture long-range dependencies in complex molecular sequences. This new model achieves a more comprehensive estimation of how drugs and targets interact, resulting in a significant reduction in prediction errors across major benchmarks like Davis and BindingDB. This hybrid approach proves to be a powerful tool for AI-driven medicine, offering higher accuracy and better generalization than previous models.

4) Spatio-Temporal Modeling and Time-Series Forecasting (TSF)

Mamba's linear complexity naturally suits long-horizon forecasting and multi-sensor spatiotemporal data fusion:

In time-series forecasting, TimeMachine [30] has emerged as a superior alternative to Transformers for long-term predictions. By using a quadruple-Mamba architecture, it captures complex patterns at multiple scales while maintaining a very small memory footprint, meaning it stays fast even as data grows. Although it outperforms most baselines, its results on specific datasets indicate that there is still room to improve how it aligns local and global information.

Complementing this line of research, S-Mamba [31] has been proposed as a highly efficient alternative for TSF. Unlike complex hybrid structures, S-Mamba tokenizes individual variates through a linear layer and employs a bidirectional Mamba layer to capture inter-variate correlations. By combining this with a Feed-Forward Network for temporal dependencies, the model achieves leading performance across thirteen public datasets while maintaining significantly lower computational overhead than Transformer-based models.

For traffic applications, ST-MambaSync [32] integrates both Mamba and Transformer technologies, improving accuracy and efficiency for real-time traffic management, road safety, and environmental impact reduction. Validated using real-world multi-sensor datasets, it shows a 0.70% reduction in Mean Absolute Error (MAE), 0.62% in Root Mean Square Error (RMSE), 64.86% faster inference, and 19.44% less training time compared to prior state-of-the-art models.

A similar approach has been applied to environmental monitoring. WOA-STMamba proposed by [33] predicts future pollution levels of fine particulate matter (PM2.5). Using convolutional encoding and Whale Optimization to adapt the Mamba architecture. Is validated with data from 36 monitoring stations in Beijing resulting in greater accuracy and robustness across

seasons and monitoring stations, with a 4.67–6.07% error reduction and a 0.57% increase in R^2. These results confirm that Mamba-based architectures effectively handle spatiotemporal data, providing fast and reliable insights for traffic and air-quality monitoring.

5) Edge Intelligence, Speech, and Scientific Computing

The applications in this domain target two extremes: resource-constrained deployment on edge devices and high-complexity physical modeling, both of which benefit from Mamba's linear scaling properties.

For autonomous navigation on edge devices, EdgeNavMamba [34] presents a novel framework utilizing an optimized Mamba-based object detector. It surpasses efficient models by integrating State Space Models (SSMs) for superior context-awareness and employs knowledge distillation, achieving a 31% parameter reduction. The framework demonstrates practical efficacy, achieving over 90% navigation success in MiniWorld and IsaacLab simulators. Its deployment on NVIDIA Jetson Orin Nano and Raspberry Pi 5 hardware validates a 73% reduction in energy per inference, confirming its viability for computationally critical applications in real-time robotics.

The push for on-device intelligence has led to the emergence of ultra-low-bit quantization methods for SSMs. A notable advancement is the application of the BitNet (1.58-bit) framework [35] to the Mamba architecture, creating specialized variants capable of running with ternary weights. Research demonstrates that by quantizing not only the linear layers but also the embedding and projection layers, Mamba-based models can achieve a 90% reduction in parameter bits with minimal degradation in perplexity [36]. This union between the linear scaling of SSMs and the extreme efficiency of BitNet quantization confirms Mamba's potential as a backbone for next-generation edge devices, where memory and energy constraints are critical.

In speech processing, multiple Mamba-based architectures have been evaluated against Transformer baselines. Models such as Mamba-TasNet [37], ConMamba [38], and VALL-M demonstrate comparable or superior performance to Transformer baselines (SepFormer, Conformer, VALL-E) across speech separation, recognition, and synthesis, achieving greater memory and speed efficiency for long-duration audio. However, Transformers remain preferable for short-duration tasks and scenarios requiring joint text-speech modeling, such as cross- or masked-attention mechanisms [39].

The development of Multi-Scale Dynamic Mamba (MSDM) [40] enhances dynamic facial expression recognition (DFER) in smart classrooms through efficient bidirectional Mamba blocks combined with multi-scale attention fusion. It captures global and local facial features via Multi-Scale Attention Fusion Module (MSAFM), models long-term expression dynamics with Dynamic Temporal Focus (DTF), and uses dual-resolution processing for robust feature extraction. Tested on seven real-world datasets including the new classroom-specific HM-Class dataset, MSDM achieves state-of-the-art performance with significantly fewer parameters and lower computational cost than Transformer-based models.

In remote sensing, the Learnable TransMamba Hybrid Network (LTMHN) [41] enhances remote sensing image super resolution by fusing Transformer attention with Mamba state-space models through learnable adapters. It preserves global context modeling while achieving linear computational complexity and includes an Adaptive Channel Enhancement Module (ACEM) that boosts channel activation by 60%. Tested on remote sensing datasets, LTMHN improves Peak Signal-to-Noise Ratio (PSNR) by 0.11 dB over MambaIR with only 68% of its computational cost.

For traffic flow prediction, Trans-Mamba improves traffic flow prediction through a cross-network architecture combining Transformer and Mamba blocks [42]. Its design features a learnable spatio-temporal embedding layer and stacked temporal-spatial blocks to capture dynamic long-range dependencies in traffic networks. Experiments on real-world datasets like PeMS08 show 2.59% reductions in Mean Absolute Error (MAE) and 4.14% in Root Mean Square Error (RMSE) compared to prior graph neural network methods.

Finally, in scientific computing, Mamba Neural Operator (MNO) [43] is a framework that applies Mamba to solve complex physics equations (PDEs). While traditional Transformers struggle with high-resolution simulations due to their heavy memory usage, MNO achieves a 90% error reduction by processing data with linear efficiency. However, a key limitation is that Mamba performs best on regular grids (like images), whereas Transformers are still more reliable for irregular or distorted shapes (geometries).

This development confirms that Mamba-based backbones, offer a more scalable path for high-resolution scientific modeling.

IV. DISCUSSIONS

Summary about the more encouraged and critical features about Mamba neural network architecture was developed in section.

A. Critical Perspectives on Mamba.

1) Hybrid Transformer–Mamba models often outperform pure Mamba: In many domains, hybrids: (e.g., Transformers for encoding, Mamba for decoding) that fuse Transformer encoders with Mamba decoders yield the best trade-offs—indicating neither architecture alone fully captures all sequence-modeling requirements [44].
2) Energy consumption and scalability limitations: Scaling Mamba models to billions of parameters increases energy consumption and training costs, prompting the development of efficient variants like Bi-Mamba using 1-bit representations to reduce compute demand [45].
3) Implementation complexity and tooling immaturity: Integrating Mamba's selective SSM kernels into production requires deep expertise in specialized numeric, and the ecosystem of optimized libraries remains less mature than the Transformer-focused tools (e.g., FlashAttention) [46]. In tasks such as document ranking, Mamba exhibits a lower training speed compared to optimized Transformers [47]. This suggests that Mamba's theoretical advantages are currently held back by the superior

hardware optimization of attention mechanisms, meaning its widespread use depends on the future maturity of its own computational tools.

4) Length Generalization and Training Dynamics: Although Mamba excels in long-context tasks, research indicates that recurrent architectures can degrade when processing sequences significantly longer than those used during training [48]. This issue is often linked to training dynamics rather than inherent architectural flaws. While traditional methods suggest using regularization or noise-based training to mitigate this, recent solutions like MambaExtend [49] have proven more effective. By optimizing the model's internal parameters without the need for retraining, this approach allows Mamba to handle million-token contexts with high precision. These advancements demonstrate that Mamba can effectively overcome the length extrapolation barriers that previously favored Transformers, ensuring scalability without altering its core structure.
5) Data Sparsity and Initialization Sensitivity: Despite its advantages, recent systematic reviews identify specific instabilities in short-sequence inputs or sparse data environments. Due to insufficient state initialization, Mamba-based architectures may face representation oscillation or underfitting in tasks with limited historical interactions, such as cold-start scenarios in recommendation systems [50]. This suggests that while Mamba excels at compressing long-range context, its performance is highly dependent on the density and duration of the initial input signal.
6) Copying Task: While early systematic reviews suggest that Mamba models may struggle with discrete 'copying' tasks and in-context learning compared to Transformers [51], recent theoretical analysis indicates this is not a structural failure but a scaling factor. Studies like [52] demonstrate that when the model's state size grows proportionally to the task complexity, Mamba matches Transformer performance in copying and reasoning operations while maintaining its linear efficiency, effectively refuting the notion of an inherent 'retrieval' weakness.
7) Explainability and Inner Representations: While Transformers are widely studied through their attention maps, Mamba's recurrent nature initially made it less interpretable. However, recent studies have introduced explainability tools by uncovering Mamba's hidden attention matrices [53]. This research demonstrates that Mamba's performance, fairness, and robustness can now be evaluated using metrics comparable to those of Transformers. This breakthrough addresses the 'black box' concern of State Space Models, providing the community with tools to peer into the model's decision-making process.
8) Advances in Model Transparency: A major concern for using Mamba in sensitive areas is its "black box" nature. To solve this, researchers adapted the Layer-wise Relevance Propagation (LRP) framework to the architecture [54]. This tool allows us to trace exactly how the model makes decisions by connecting outputs to specific input tokens. By making these internal processes visible, Mamba's behavior can now be audited for biases, reaching a level of transparency similar to Transformers.

B. Limitations and Future Works

The limitations of early State Space Models (SSMs), such as difficulties with content-dependent reasoning and inefficient hardware utilization during training, are being systematically addressed in the Mamba lineage. Mamba-1 established selective SSMs as a viable, efficient alternative to Transformers for long sequences. Building on this, Mamba-2 introduced the parallelizable Structured State-Space Duality (SSD) framework, which significantly accelerated training and inference. Initial studies confirm Mamba-2 matches or surpasses Mamba-1's performance, excelling in complex tasks due to its larger state capacity, while being faster to train.

This evolution continues with Mamba-4, which adopts an inference-first design to enhance state-tracking and model quality without a memory cost increase. The efficacy of this architectural progression is evidenced by its adoption in next-generation hybrids. Models like Bamba and IBM's Granite series use Mamba-2 as a foundation, achieving transformer-like performance with greater efficiency—Bamba reportedly enables inference twice as fast as traditional Transformers.

Mamba's efficiency democratizes long-context processing, challenges like energy consumption and tooling remain. This trajectory confirms that the future of efficient AI likely lies not in pure SSMs or Transformers, but in their synergistic hybrids, which offer a balanced path for practical deployment.

V. CONCLUSIONS

The results demonstrate that Mamba outperforms conventional architectures such as Transformers and Receptivance Weighted Key Value (RWKV) in terms of computational efficiency, moving from quadratic to linear complexity, and long-context processing (exceeding 1 million tokens with as few as 74K parameters), achieving the objective of developing a scalable and sustainable model.

The emergence of variants such as Jamba, which utilizes a Mixture of Experts (MoE) design, and Mamba-2, which introduces Structured State Space Models (SSM), and mamba-4 confirms the architecture's adaptability across specialized domains like genomics and edge computing. While limitations persist regarding extreme scalability beyond 10 billion parameters and current hardware dependency, these findings suggest that the future of efficient artificial intelligence lies in hybrid approaches.

Future research should: 1) optimize implementations for heterogeneous hardware, 2) explore Mamba-Transformer hybrids for complex reasoning tasks, and 3) develop standardized Selective State Space Models (SSM) benchmarks. These findings support Mamba's adoption in resource-constrained applications (genomics, edge computing), while contexts requiring absolute precision like enterprise Large Language Models (LLMs) may benefit from hybrid approaches.

ACKNOWLEDGMENT

The Faculty of Engineering in Telecommunications, Informatics and Biomedical, at Universidad de Oriente supported in events the presentation of the main results in this research.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ARTIFICIAL INTELLIGENCE STATEMENT

The authors declare that generative artificial intelligence tools were used for the following purposes: searching and identifying academic sources and relevant bibliography to support the research. The tool(s) used include: Perplexity AI. The authors take full responsibility for the content of the manuscript.

REFERENCES

[1] A. Vaswani *et al.*, "Attention is All you Need", in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces", in *Proceedings of the First Conference on Language Modeling*, Philadelphia, Pennsylvania, USA., Oct. 2024. [Online]. Available: https://openreview.net/forum?id=tEYskw1VY2

[3] O. Lieber, B. Lenz, and G. Cohen, "Jamba: A Hybrid Transformer-Mamba Language Model", presented at the 13th International Conference on Learning Representations (ICLR), 2025.

[4] A. Bick, E. P. Xing, and A. Gu, "Understanding the Skill Gap in Recurrent Language Models: The Role of the Gather-and-Aggregate Mechanism", in *Proceedings of Machine Learning Research*, ML Research Press, 2025, pp. 4324–4344. [Online]. Available: https://nchr.elsevierpure.com/en/publications/understanding-the-skill-gap-in-recurrent-language-models-the-role/

[5] T. Dao and A. Gu, "Transformers are SSMs: generalized models and efficient algorithms through structured state space duality", in *Proceedings of the 41st International Conference on Machine Learning*, in ICML'24, vol. 235. Vienna, Austria: JMLR.org, Jul. 2024, pp. 10041–10071.

[6] Z. Qin *et al.*, "The Devil in Linear Transformer", in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7025–7041. https://doi.org/10.18653/v1/2022.emnlp-main.473.

[7] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces", Oct. 2023, [Online]. Available: https://openreview.net/forum?id=AL1fq05o7H

[8] T. Hrycej, B. Bermeitinger, and S. Handschuh, "Integrating the Attention Mechanism Into State Space Models", in *2025 IEEE Swiss Conference on Data Science (SDS)*, Jun. 2025, pp. 170–173. https://doi.org/10.1109/SDS66131.2025.00033.

[9] N. Muca Cirone, A. Orvieto, B. Walker, C. Salvi, and T. Lyons, "Theoretical Foundations of Deep Selective State-Space Models", *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 127226–127272, Dec. 2024, https://doi.org/10.52202/079017-4041.

[10] E. Yanar *et al.*, "A Comparative Analysis of the Mamba, Transformer, and CNN Architectures for Multi-Label Chest X-Ray Anomaly Detection in the NIH ChestX-Ray14 Dataset", *Diagnostics*, vol. 15, no. 17, p. 2215, Jan. 2025, https://doi.org/10.3390/diagnostics15172215.

[11] B. N. Patro and V. S. Agneeswaran, "Mamba-360: Survey of state space models as transformer alternative for long sequence modelling: Methods, Applications, and Challenges", *Eng. Appl. Artif. Intell.*, vol. 159, p. 111279, Nov. 2025, https://doi.org/10.1016/j.engappai.2025.111279.

[12] S. Liu, J. Keung, Z. Yang, Z. Mao, and Y. Sun, "Can Mamba Be Better? An Experimental Evaluation of Mamba in Code Intelligence", in *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Nov. 2025, pp. 1856–1868. https://doi.org/10.1109/ASE63991.2025.00155.

[13] H. Zhang, C. Chen, L. Mei, Q. Liu, and J. Mao, "Mamba Retriever: Utilizing Mamba for Effective and Efficient Dense Retrieval", in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, in CIKM 24. New York, NY, USA: Association for Computing Machinery, Oct. 2024, pp. 4268–4272. https://doi.org/10.1145/3627673.3679959.

[14] A. Bick, K. Y. Li, E. P. Xing, J. Z. Kolter, and A. Gu, "Transformers to SSMs: Distilling Quadratic Knowledge to Subquadratic Models", *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 31788–31812, Dec. 2024, https://doi.org/10.52202/079017-0999.

[15] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: efficient visual representation learning with bidirectional state space model", in *Proceedings of the 41st International Conference on Machine Learning*, in ICML'24, vol. 235. Vienna, Austria: JMLR.org, Jul. 2024, pp. 62429–62442.

[16] A. Berroukham, K. Housni, and M. Lahraichi, "Vision Transformers: A Review of Architecture, Applications, and Future Directions", in *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, Dec. 2023, pp. 205–210. https://doi.org/10.1109/CiSt56084.2023.10410015.

[17] Y. Hedhoud, T. Mekhaznia, and M. Amroune, "Vision Mamba for efficient Tuberculosis Detection based on Chest X-Rays: A comparative study with CNN and Vision transformers", in *2025 7th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, Apr. 2025, pp. 1–6. https://doi.org/10.1109/PAIS66004.2025.11126051.

[18] Y. Cho, C. Lee, S. Kim, and E. Park, "PTQ4VM: Post-Training Quantization for Visual Mamba", in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Feb. 2025, pp. 1176–1185. https://doi.org/10.1109/WACV61041.2025.00122.

[19] H. Zhang *et al.*, "A Survey on Visual Mamba", *Appl. Sci.*, vol. 14, no. 13, p. 5683, Jan. 2024, https://doi.org/10.3390/app14135683.

[20] K. Li *et al.*, "VideoMamba: State Space Model for Efficient Video Understanding", in *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXVI*, Berlin, Heidelberg: Springer-Verlag, Oct. 2024, pp. 237–255. https://doi.org/10.1007/978-3-031-73347-5_14.

[21] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video Transformers: A Survey", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12922–12943, Nov. 2023, https://doi.org/10.1109/TPAMI.2023.3243465.

[22] S. Suo, J. Liu, and H. Miao, "Mamba-Stereo: Mamba Regularization for Stereo matching", *Pattern Recognit.*, vol. 170, p. 112120, Feb. 2026, https://doi.org/10.1016/j.patcog.2025.112120.

[23] Y. Schiff, C.-H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov, "Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling", *Proc. Mach. Learn. Res.*, vol. 235, pp. 43632–43648, Jul. 2024.

[24] Z. Wang *et al.*, "DSA Mamba: A Model for Advanced Medical Image Classification", *Expert Syst. Appl.*, p. 130064, Oct. 2025, https://doi.org/10.1016/j.eswa.2025.130064.

[25] V. Nikitin and V. Danilov, "Transformer vs. Mamba as Skin Cancer Classifier: Preliminary ResultS", *KPI Sci. News*, vol. 137, no. 1–4, Dec. 2024, https://doi.org/10.20535/kpisn.2024.1-4.301028.

[26] Z. Zhang, Q. Ma, T. Zhang, J. Chen, H. Zheng, and W. Gao, "Switch-UMamba: Dynamic scanning vision Mamba UNet for medical image segmentation", *Med. Image Anal.*, vol. 107, p. 103792, Jan. 2026, https://doi.org/10.1016/j.media.2025.103792.

[27] J. Ye, J. Zhang, and H. Shan, "Depmamba: Progressive fusion mamba for multimodal depression detection", presented at the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.

[28] L. Zhao and Y. Zhang, "Research on Autism Diagnosis Method Based on Transformer and Mamba", in *2025 6th International Conference on Machine Learning and Computer Application (ICMLCA)*, Oct. 2025, pp. 1190–1193. https://doi.org/10.1109/ICMLCA66850.2025.11336595.

[29] X. Lou, J. Cai, Q. Liu, and S. W. I. Siu, "MambaTransDTA: A Hybrid Mamba-Transformer Architecture for Accurate Drug-Target Binding Affinity Prediction", *J. Chem. Inf. Model.*, vol. 66, no. 1, pp. 259–270, Jan. 2026, https://doi.org/10.1021/acs.jcim.5c02361.

[30] M. A. Ahamed and Q. Cheng, "TimeMachine: A Time Series is Worth 4 Mambas for Long-Term Forecasting", in *ECAI 2024*, IOS Press, 2024, pp. 1688–1695. https://doi.org/10.3233/FAIA240677.

[31] Z. Wang *et al.*, "Is Mamba effective for time series forecasting?", *Neurocomput.*, vol. 619, no. C, Feb. 2025, https://doi.org/10.1016/j.neucom.2024.129178.

[32] Z. Shao, Z. Wang, X. Yao, M. G. H. Bell, and J. Gao, "ST-MambaSync: Complement the power of Mamba and Transformer fusion for less computational cost in spatial–temporal traffic forecasting", *Inf. Fusion*, vol. 117, p. 102872, May 2025, https://doi.org/10.1016/j.inffus.2024.102872.

[33] C. Zhang, M. Cheng, X. Li, X. Wang, B. Zhao, and X. Zhu, "WOA-STMamba: A spatiotemporal Mamba model enhanced by Whale Optimization for short-term PM2.5 concentration prediction", *J. Environ. Chem. Eng.*, vol. 13, no. 5, p. 118366, Oct. 2025, https://doi.org/10.1016/j.jece.2025.118366.

[34] R. Aalishah, M. Navardi, and T. Mohsenin, "EdgeNavMamba: Mamba-Optimized Object Detection for Energy-Efficient Edge Devices", presented at the 2025 IEEE 11th International Conference on Edge Computing and Scalable Cloud (EdgeCom), IEEE Computer Society, Nov. 2025, pp. 62–67. https://doi.org/10.1109/EdgeCom66327.2025.00017.

[35] H. Wang *et al.*, "BitNet: 1-bit Pre-training for Large Language Models", *J. Mach. Learn. Res.*, vol. 26, no. 125, pp. 1–29, 2025.

[36] Z. Yu, T. Kojima, Y. Matsuo, and Y. Iwasawa, "Slender-Mamba: Fully Quantized Mamba in 1.58 Bits From Head to Toe", in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds., Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 4715–4724.

[37] X. Jiang, C. Han, and N. Mesgarani, "Dual-path Mamba: Short and Long-term Bidirectional Selective Structured State Space Models for Speech Separation", in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10888514.

[38] H. Hou, X. Gong, and Y. Qian, "ConMamba: A Convolution-Augmented Mamba Encoder Model for Efficient End-to-End ASR Systems", in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, Nov. 2024, pp. 711–715. https://doi.org/10.1109/ISCSLP63861.2024.10800511.

[39] X. Jiang, Y. A. Li, A. Nicolas Florea, C. Han, and N. Mesgarani, "Speech Slytherin: Examining the Performance and Efficiency of Mamba for Speech Separation, Recognition, and Synthesis", in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889391.

[40] Y. Liang *et al.*, "MSDM: A Lightweight Multi-Scale Dynamic Mamba for Dynamic Facial Expression Recognition in Smart Classrooms", *IEEE Trans. Affect. Comput.*, no. 01, pp. 1–17, Dec. 2025, https://doi.org/10.1109/TAFFC.2025.3646217.

[41] M. Zhang, J. Li, H. Jing, K. Wu, and Q. Rong, "LTMHN: Learnable TransMamba Hybrid Network for Remote Sensing Image Superresolution", *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 18, pp. 23548–23562, 2025, https://doi.org/10.1109/JSTARS.2025.3607029.

[42] Q. Liu *et al.*, "Trans-Mamba: The Cross-network of Transformer and Mamba for Traffic Flow Prediction", in *2024 IEEE Smart World Congress (SWC)*, Dec. 2024, pp. 219–226. https://doi.org/10.1109/SWC62898.2024.00064.

[43] C.-W. Cheng, J. Huang, Y. Zhang, G. Yang, C.-B. Schönlieb, and A. I. Aviles-Rivero, "Mamba neural operator: Who wins? transformers vs. state-space models for PDEs", *J. Comput. Phys.*, vol. 548, p. 114567, Mar. 2026, https://doi.org/10.1016/j.jcp.2025.114567.

[44] X. Zhu, Q. Ruan, S. Qian, and M. Zhang, "A hybrid model based on transformer and Mamba for enhanced sequence modeling", *Sci. Rep.*, vol. 15, no. 1, p. 11428, Apr. 2025, https://doi.org/10.1038/s41598-025-87574-8.

[45] S. Tang, L. Ma, H. Li, M. Sun, and Z. Shen, "Bi-Mamba: Towards Accurate 1-Bit State Space Models", *Transactions on Machine Learning Research*, 2025.

[46] S. Das, R. Sen, and S. Devendiran, "Mamba Models a possible replacement for Transformers?", *Python Sci. Conf.*, pp. 332–344, Jun. 2024, https://doi.org/10.25080/XHDR4700.

[47] Z. Xu, J. Yan, A. Gupta, and V. Srikumar, "State Space Models are Strong Text Rerankers", in *Proceedings of the 10th Workshop on Representation Learning for NLP (RepL4NLP-2025)*, V. Adlakha, A. Chronopoulou, X. L. Li, B. P. Majumder, F. Shi, and G. Vernikos, Eds., Albuquerque, NM: Association for Computational Linguistics, May 2025, pp. 152–169. https://doi.org/10.18653/v1/2025.repl4nlp-1.12.

[48] A. Terzic, M. Hersche, G. Camposampiero, T. Hofmann, A. Sebastian, and A. Rahimi, "On the Expressiveness and Length Generalization of Selective State Space Models on Regular Languages", *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 19, pp. 20876–20884, Apr. 2025, https://doi.org/10.1609/aaai.v39i19.34301.

[49] S. Azizi, S. Kundu, M. E. Sadeghi, and M. Pedram, "MambaExtend: A Training-Free Approach to Improve Long Context Extension of Mamba", presented at the 13th International Conference on Learning Representations, 2025. [Online]. Available: https://openreview.net/forum?id=LgzRo1RpLS

[50] Q. Miao, L. Jia, K. Xie, K. Fu, and Z. Yang, "A comprehensive survey and taxonomy of mamba: Applications, Challenges, and Future Directions", *Inf. Fusion*, vol. 130, p. 104094, Jun. 2026, https://doi.org/10.1016/j.inffus.2025.104094.

[51] A. Salam, R. Mahmud, T. Islam, S. Mukta, and S. Shatabda, "A Comprehensive Survey on Mamba: Architectures, Challenges, and Opportunities", *Computer*, vol. 58, no. 8, pp. 64–76, Aug. 2025, https://doi.org/10.1109/MC.2025.3571322.

[52] R. Ren, Z. Li, and Y. Liu, "Exploring the Limitations of Mamba in COPY and CoT Reasoning", in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds., Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 12539–12563. https://doi.org/10.18653/v1/2025.emnlp-main.634.

[53] A. A. Ali, I. Zimerman, and L. Wolf, "The Hidden Attention of Mamba Models", in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1516–1534. https://doi.org/10.18653/v1/2025.acl-long.76.

[54] F. Rezaei Jafari, G. Montavon, K.-R. Müller, and O. Eberle, "MambaLRP: Explaining Selective State Space Sequence Models", *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 118540–118570, Dec. 2024, https://doi.org/10.52202/079017-3764.